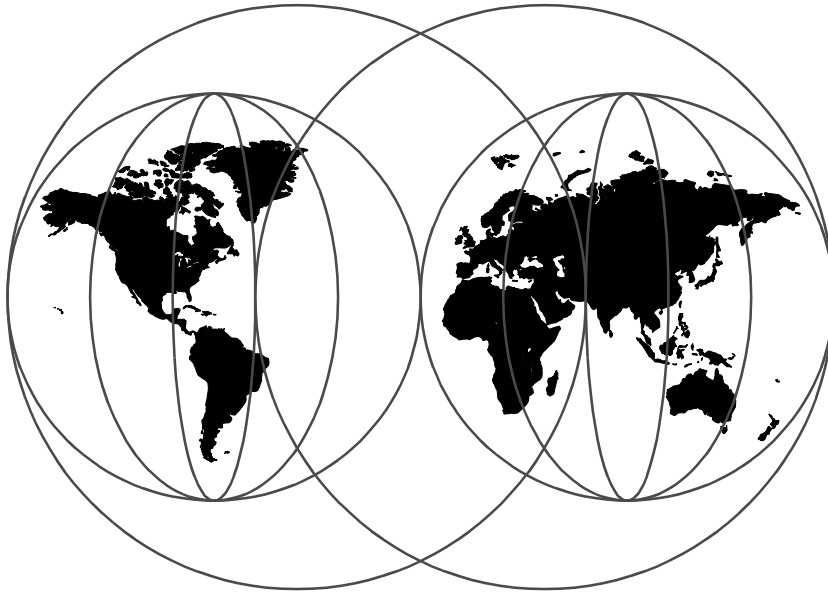


Intelligent Miner for Data: Enhance Your Business Intelligence

Joerg Reinschmidt, Helena Gottschalk, Hosung Kim, Damiaan Zwietering



International Technical Support Organization

<http://www.redbooks.ibm.com>

SG24-5422-00



International Technical Support Organization

**Intelligent Miner for Data:
Enhance Your Business Intelligence**

June 1999

Take Note!

Before using this information and the product it supports, be sure to read the general information in Appendix C, "Special Notices" on page 189.

First Edition (June 1999)

This edition applies to Version 2.1.3 of the Intelligent Miner for Data for use with the Windows NT Version 4, AIX Version 4.3.1, Sun Solaris Version 2.7, OS/400 V4R3, and OS/390 V1R3 Operating Systems.

Comments may be addressed to:

IBM Corporation, International Technical Support Organization
Dept. QXXE Building 80-E2
650 Harry Road
San Jose, California 95120-6099

When you send information to IBM, you grant IBM a non-exclusive right to use or distribute the information in any way it believes appropriate without incurring any obligation to you.

© Copyright International Business Machines Corporation 1999. All rights reserved

Note to U.S Government Users – Documentation related to restricted rights – Use, duplication or disclosure is subject to restrictions set forth in GSA ADP Schedule Contract with IBM Corp.

Contents

Figures	ix
Tables	xi
Preface	xiii
The Team That Wrote This Redbook	xiii
Comments Welcome	xv
<hr/>	
Part 1. Introduction to Business Intelligence and Data Mining	1
Chapter 1. Business Intelligence and Data Mining	3
1.1 Business Intelligence	3
1.1.1 The Evolution of Leverage	3
1.1.2 From Data to Decisions	5
1.1.3 The Data Warehousing Concept	6
1.2 OLAP and Data Mining	8
1.2.1 Hypothesis Verification or Information Discovery	9
1.2.2 Data Mining Input	9
1.2.3 Data Mining Output	10
1.3 A Definition of Data Mining	10
1.4 Data Mining Applications and Operations	11
1.4.1 Data Mining Applications	13
1.4.2 Data Mining Operations	13
Chapter 2. Getting Started with Data Mining	15
2.1 Business Drivers	15
2.2 Technology Enablers	16
2.3 Organizational Factors	17
2.3.1 The Organizational Culture	17
2.3.2 The Business Environment	17
2.3.3 The People	17
2.3.4 The I/T Architecture	19
2.3.5 The Data	19
2.3.6 The Data Mining Tools	19
2.3.7 The Project	20
2.4 The Data Mining Process	20
2.4.1 Business Analysis	21
2.4.2 Data Analysis	23
2.4.3 Data Gathering	24
2.4.4 Data Preparation	25
2.4.5 Data Mining	27

2.4.6	Result Interpretation	27
2.4.7	Business Application	28
2.4.8	Business Feedback	29
2.5	The IBM BI Methodology	29
2.5.1	Data Discovery	30
2.5.2	Tasks and Deliverables	30
2.5.3	Roles	31
2.6	Summary and Outlook	32
Chapter 3. Data Considerations for Data Mining		33
3.1	Data Sources	33
3.1.1	Data in Operational Systems	34
3.1.2	Data in a Data Warehouse	35
3.1.3	Data Replication	39
3.2	Data Transformation	41
3.3	Where to Put the Result Data	43
3.3.1	For Analysts	43
3.3.2	For Decision Makers	44
3.3.3	For Applications	44
3.3.4	For Resources	44
3.4	Technical Considerations	44
3.4.1	Security	45
3.4.2	Performance	46
3.4.3	Maintenance	48
Chapter 4. Introduction to IBM Intelligent Miner for Data		51
4.1	Overview of the Intelligent Miner	51
4.2	Working with Databases	52
4.3	The User Interface	53
4.4	Data Preparation Functions	54
4.5	Statistical Functions	56
4.6	Mining Functions	56
4.6.1	Associations	57
4.6.2	Sequential Patterns	58
4.6.3	Clustering	58
4.6.4	Classification	59
4.6.5	Prediction	60
4.6.6	Similar Time Sequences	60
4.7	Creating and Visualizing the Results	60
4.8	Creating Data Mining Sequences	61
Chapter 5. Implementing IM in the ITSO Environment		63
5.1	The ITSO BI Environment	63
5.1.1	The Environment	63

5.1.2	Networking Configuration	66
5.1.3	The Data	66
5.1.4	Table Relationships	68
5.2	Implementing Data Mining Techniques	69
5.2.1	Associations	69
5.2.2	Sequential Patterns	71
5.2.3	Clustering	72
5.2.4	Classification	74
5.2.5	Prediction	76
5.2.6	Similar Time Sequences	78
Part 2. Installation and Configuration of Intelligent Miner for Data		81
Chapter 6. Implementation on Windows NT		83
6.1	Prerequisites	83
6.1.1	Hardware Requirements	83
6.1.2	Software Prerequisites	85
6.1.3	Networking Requirements	89
6.2	Product Installation	90
6.3	Verifying the Installation	96
6.4	Running the Server	97
Chapter 7. Implementation on AIX		99
7.1	Prerequisites	99
7.1.1	Hardware Requirements	100
7.1.2	Software Prerequisites	101
7.1.3	Networking Requirements	105
7.2	Product Installation	108
7.3	Installation Verification	115
7.4	Running the Server	116
Chapter 8. Implementation on Sun Solaris		119
8.1	Prerequisites	119
8.1.1	Hardware Requirements	119
8.1.2	Software Prerequisites	120
8.1.3	Networking Requirements	122
8.2	Product Installation	124
8.3	Installation Verification	127
8.4	Running the Server	128
Chapter 9. Implementation on OS/400		129
9.1	Prerequisites	129
9.1.1	Hardware Requirements	129

9.1.2 Software Prerequisites	129
9.1.3 Networking Requirements.	131
9.1.4 Relational Databases	132
9.2 Product Installation.	134
9.3 Installation Verification and Starting the Server	135
Chapter 10. Implementation on OS/390.	137
10.1 Installation Planning	138
10.1.1 IBM Service Information	138
10.1.2 Installation Planning Values	139
10.2 Understanding the Prerequisites.	140
10.2.1 Hardware Requirements	140
10.2.2 Software Prerequisites	141
10.2.3 Networking Requirements.	142
10.2.4 Verifying the TCP/IP Configuration	142
10.2.5 Networking Between the IM Server and Client	147
10.3 Product Installation and Customization.	148
10.3.1 Installation Procedure.	148
10.3.2 Installation Customization.	154
10.4 Running the Server	172
10.4.1 Starting the Server	172
10.4.2 DB2 Considerations	173
Appendix A. Using DB2 V 2.1 or DataJoiner on AIX.	175
A.1 Running with DB2 for AIX	175
A.2 Working with Data Joiner (Version 2.1.1)	176
Appendix B. Intelligent Miner for Data Installation Sample JCL	179
B.1 IDMRECEV	179
B.2 IDMALLOC	179
B.3 IDMDDDDEF	181
B.4 IDMHFS	182
B.5 IDMAPPCCK	182
B.6 IDMAPPLY	183
B.7 IDMDB2	183
B.8 IDMDEMO.	184
B.9 IDMVERFY	184
B.10 IDMACCCK.	185
B.11 IDMACCEP.	185
B.12 IDMSECUR.	185
B.13 IDMSTART	186
B.14 IDMCFLD	187

Appendix C. Special Notices	189
Appendix D. Related Publications	193
D.1 International Technical Support Organization Publications	193
D.2 Redbooks on CD-ROMs	193
D.3 Other Publications	194
How to Get ITSO Redbooks	195
IBM Redbook Fax Order Form	196
Glossary	197
List of Abbreviations	203
Index	205
ITSO Redbook Evaluation	211

Figures

1. Evolution from Queries to Data Mining	4
2. Decisions, Information, and Data	5
3. Application Focus and Subject Focus	7
4. A Business Intelligence Environment.	8
5. Applications, Operations and Techniques	12
6. Roles in Data Mining	18
7. The Data Mining Process.	20
8. Percentages of Time Spent in Each Process Step	21
9. Business Analysis	22
10. Data Analysis	23
11. Data Gathering.	24
12. Result Interpretation.	28
13. Data Discovery Activities and Tasks	30
14. Deliverables and Work Products	31
15. 3-Tier Data Structure	33
16. Customer Transaction Table	36
17. Product and Channel Tables	36
18. Channel Usage Example Table	37
19. Transposition Example	38
20. IBM's Open and Comprehensive Data Replication Solution	39
21. Data Warehouse Table Pattern and Data Mart Pattern.	42
22. Update Maintenance Flow	49
23. Intelligent Miner Block Diagram	52
24. Intelligent Miner Main Window	54
25. Sample Clustering Visualization	61
26. Sequence Settings Window.	62
27. The ITSO BI Environment	64
28. Table Relationships	68
29. Data Transformation for an Associations Model	69
30. Data Transformation for a Sequential Patterns Model.	71
31. Data Transformation for a Clustering Model	73
32. Clustering Result	74
33. Data Transformation for a Classification Table	75
34. Classification Tree	76
35. Data Transformation for a Prediction Model	77
36. Prediction Result	77
37. Data Transformation for a Similar Time Sequences Model.	78
38. Similar Time Sequences Result.	79
39. Intelligent Miner for Data on Windows NT	83
40. Windows NT System Configuration	84

41. Windows NT version	87
42. Windows NT Command Line Processor for DB2.	87
43. Setting the Path Variable	88
44. Windows NT TCP/IP Properties.	89
45. Message on Available Colors	91
46. IM for Windows NT, Welcome Screen.	91
47. Selected Components	92
48. User Selection	93
49. Multi-User Setup	93
50. Directory for Miningbases	94
51. Select Start Menu Folder	95
52. Start Menu Folder on Desktop.	95
53. Windows NT Services Panel	96
54. Intelligent Miner for Data on AIX	99
55. List Installed Software on AIX	103
56. Software Selection on AIX.	103
57. Display Installed JAVA Software on AIX	104
58. Display Installed DB2 Software on AIX	105
59. AIX TCP/IP SMIT Panel.	106
60. Defined Network Interface on AIX	106
61. AIX TCP/IP Minimum Configuration.	107
62. AIX Hosts File	108
63. Create Group SMIT Panel on AIX	109
64. Create Group on AIX	109
65. Create User SMIT Panel on AIX	110
66. Create User on AIX	111
67. Change User Password on AIX	111
68. Log in as a IM User	112
69. Software Installation Smit Panel on AIX	113
70. Installation Device Selection on AIX	114
71. Append IM Command Directory.	115
72. Intelligent Miner for Data on SUN Solaris	119
73. Installed Software list on Sun Solaris.	121
74. Detail Software Information on Sun Solaris	122
75. Sun Solaris Workstation Information	123
76. Admintool on Sun Solaris.	123
77. Host Information on SUN Solaris.	124
78. Add Group on Sun Solaris.	124
79. Add User on Sun Solaris	125
80. Intelligent Miner for Data on OS/400	129
81. Intelligent Miner for Data on OS/390	137

Tables

1. Data Mining Application Areas	13
2. Sample Mining History	50
3. Sales Information	66
4. Article Information	67
5. Article Information	68
6. Windows NT Server Storage Requirements	85
7. Windows Client Storage Requirements	85
8. Windows NT Server Software Prerequisites	86
9. Windows Client Software Prerequisites	86
10. AIX Server Hardware Requirements	100
11. AIX Client Hardware Requirements	100
12. AIX Server Software Prerequisites	101
13. AIX Client Software Prerequisites	101
14. Sun Solaris Server Hardware Requirements	120
15. Sun Solaris Server Software Prerequisites	120
16. OS/400 Required PTFs	131
17. OS/390 Publications Useful During Installation	137
18. OS/390 PSP Upgrade and Subset ID	138
19. OS/390 Component IDs	139
20. OS/390 Installation Planning Table	139
21. OS/390 Driving System Software Prerequisites	140
22. OS/390 Total DASD Space Required	141
23. OS/390 Target System Software Prerequisites	141
24. OS/390 Sample Job List	148
25. OS/390 SMP/E Options Subentry Values	149
26. Target Libraries	180
27. Distribution Libraries	180

Preface

Competitive business pressures and a desire to leverage existing information technology investments have led many firms to explore the benefits of data mining technology. This technology is designed to help businesses discover hidden patterns in their data—patterns that can help them understand the purchasing behavior of their key customers, detect likely credit card or insurance claim fraud, predict probable changes in financial markets, and so on. Using Intelligent Miner, you can increasingly leverage the data warehouse and more quickly derive business value from that investment.

This redbook will help you design and implement a solution to enhance an existing Business Intelligence environment with the functionality of data mining. It will also help you install, tailor, and configure the Intelligent Miner for Data product on all available platforms.

First, we introduce the concepts of Business Intelligence, Data Warehousing, Online Analytical Processing (OLAP), and data mining. We discuss the benefit of data mining and the necessary tasks that must be performed when planning for an implementation of this technology, including such topics as data sources, data transformation, data placement, and security.

Next, we introduce the Intelligent Miner for Data product, and describe the environment used at the International Technical Support Organization, including some data mining techniques using this data warehouse.

Finally, we provide instructions for installing the product on all available platforms, covering Windows NT, AIX, Sun Solaris, OS/400, and OS/390.

The Team That Wrote This Redbook

This redbook was produced by a team of specialists from around the world working at the International Technical Support Organization San Jose Center.

Joerg Reinschmidt is an Information Mining and Knowledge Management Specialist at the International Technical Support Organization, San Jose Center. He writes extensively and teaches IBM classes worldwide on Information Mining, Knowledge Management, DB2 Universal Database, and Internet access to legacy data. Before joining the ITSO in 1998, Joerg worked in the IBM Solution Partnership Center (SPC) in Germany as a DB2 Specialist, supporting independent software vendors (ISVs) to port their applications to use IBM data management products.

Helena Gottschalk is a Data Mining Specialist in the IBM e-business and Multi-Industry Solutions and Services Unit in Sao Paolo, Brazil. She holds a master's degree in electrical engineering from the Federal University of Rio de Janeiro (UFRJ).

Hosung Kim is a CRM Solution Specialist for the IBM Banking Industry in Korea. He has implemented database marketing solutions for 4 years. He has several years of experience in UNIX systems, especially IBM Parallel System SP2, and 4 years of experience in databases (IBM DB2 and Oracle). He has worked as a data analyst and system administrator.

Damiaan Zwietering is an I/T Specialist for IBM Global Services in The Netherlands. He holds a bachelor's degree in computer science from the Enschede Polytechnic. He graduated from Utrecht University with a master's thesis in cognitive artificial intelligence. He has worked for several years on customer projects in the Business Intelligence field, in presales, technical consulting, architecture, modeling, and implementation.

Thanks to the following people for their invaluable contributions to this project:

Maria Sueli Almeida
International Technical Support Organization, San Jose Center

Ute Baumbach
Software Solutions Development, IBM Germany

Hyunhee Choi
Software Development Institute, IBM Korea

Herman van Dellen
IBM Global Services, The Netherlands

Claudia Gardner
IBM Santa Teresa

Vasilis Karras
International Technical Support Organization, Poughkeepsie Center

Timo Kussmaul
Software Solutions Development, IBM Germany

Stanley Kwong
IBM Santa Teresa

Jarek Misczyk
International Technical Support Organization, Rochester Center

Bob Perlite
IBM Santa Teresa

Friedemann Schwenkreis
Software Solutions Development, IBM Germany

Dr. Jaap Verhees
ELC Information Services B.V., The Netherlands

Comments Welcome

Your comments are important to us!

We want our redbooks to be as helpful as possible. Please send us your comments about this or other redbooks in one of the following ways:

- Fax the evaluation form found in "ITSO Redbook Evaluation" on page 211 to the fax number shown on the form.
- Use the on-line evaluation form found at <http://www.redbooks.ibm.com>
- Send your comments in an Internet note to redbook@us.ibm.com

Part 1. Introduction to Business Intelligence and Data Mining

Chapter 1. Business Intelligence and Data Mining

Businesses do not run in isolation. Therefore, running a business successfully does not just depend on how you run *your* business, but how you run it *in comparison to others*. The key to making a difference may well be the use of the data stored in the systems that you use for your daily business.

We will also see that the main result of leveraging that data will be the ability to predict facts about your business environment. This enables you to perform *proactively* instead of just *reactively*. Data on historical behavior can provide valuable knowledge about the future. It can help you execute optimized business operations, grow your market share, increase your customer share, and build customer loyalty through clearly focused services.

So, how do we extract this knowledge from our data? The key to this is *data mining*, which has become one of the most popular techniques for building intelligent decision support (IDS) systems, applying tools and methods used in statistical mathematics and machine learning.

In this chapter, we investigate the basic principles of the use of data mining in a Business Intelligence (BI) environment to gain knowledge about your business. We review the steps that you take to derive knowledge from data, and the data mining techniques that you can use for productive and robust data mining operations.

1.1 Business Intelligence

This redbook is about implementing IBM's data mining tool, the Intelligent Miner for Data, into an existing Business Intelligence (BI) environment. The following section describes the concept of BI and how it evolved over the past decades. Our definitions of terms may not be the ultimate ones, but they are the frames of reference used in this book.

1.1.1 The Evolution of Leverage

You probably store large amounts of data about your day-to-day business processes. This data is a rich source of information about your business, its processes, and its customers. Increasing competitive pressures might be one of the most important drivers for seriously trying to use the information contained in all that data, and using it to gain the knowledge needed to stay in business. If the economic situation is favorable, you probably do not have to worry about this potential. But once the competition moves ahead, or the environment deteriorates, it might be too late.

On the basis of information, you can make strategic decisions and keep yourself ahead of the competition. However, the road from enterprise data to information, and finally to the anticipation of business events, is not an easy one.

Figure 1 shows the evolution that has taken place over the last few years. It starts by being able to execute queries against operational data, resulting in reports or charts. The next logical step is to analyze the resulting data with traditional statistics or OLAP tools, looking for trends or trying to verify a hypothesis. You can also try to model the relationships in your data to find out the behavior of your business under varying circumstances. Once you have done this, you can use such a model to alert you when special circumstances require your attention. In the last step, we can speak of Business Intelligence (BI) where knowledge about your business is used to drive the decisions you make.

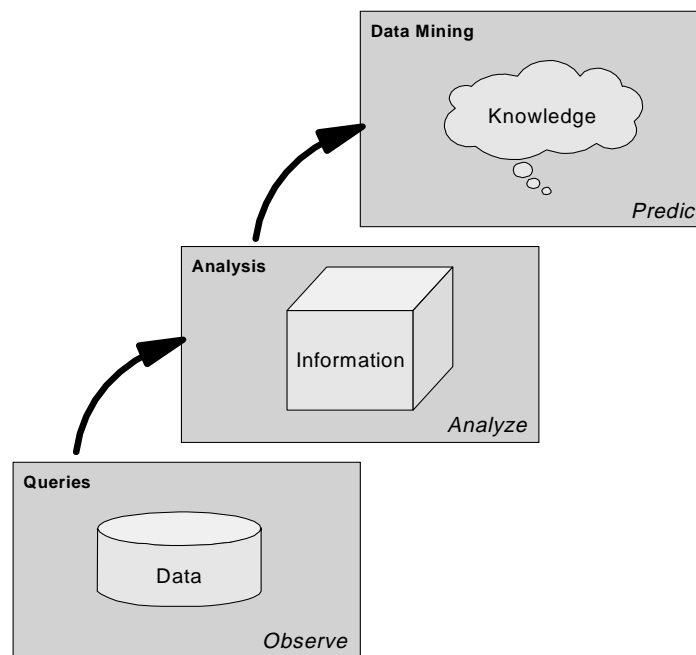


Figure 1. Evolution from Queries to Data Mining

The first steps shown are based upon questions, or knowledge, that you input yourself and validate against the available data. The third step in Figure 1 is data mining, where tools generate knowledge based on the original data itself. This new knowledge can be used to model your business without

depending on any assumptions that do not originate from your corporate data. You can go beyond the self-fulfilling prophecy of modeling your business to discovering what you know.

Data mining is not just hype, but is the logical next step in leveraging what might be your company's most strategic asset: its data. Taking this step does not mean that you must have taken the previous steps, but considering these steps will assist you in the preparation. Also, data mining will not render the existing systems obsolete, but will enhance them by enriching the data they use.

1.1.2 From Data to Decisions

Business Intelligence is all about leveraging the assets in your business to gain profit from the available data, whether it is scattered around in disparate systems, or integrated in a central repository. It provides a path to gain the knowledge needed to make well informed decisions about your business, as shown in Figure 2.

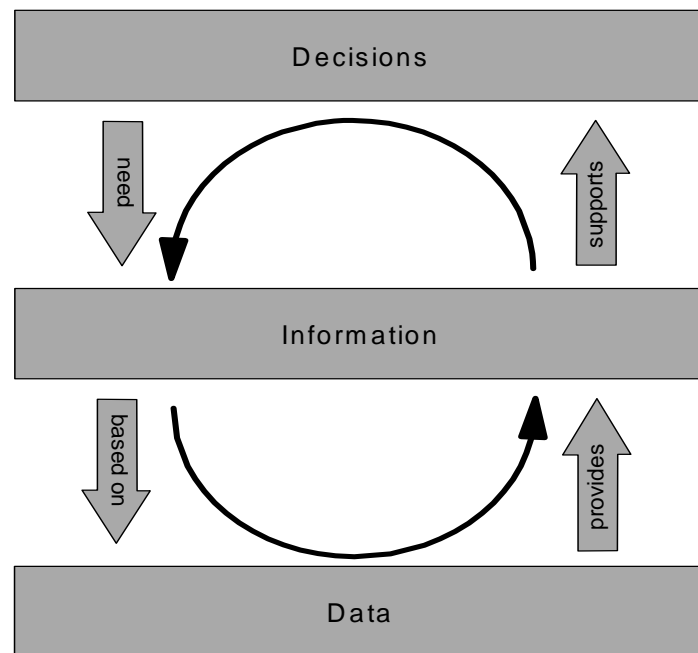


Figure 2. Decisions, Information, and Data

You can interpret this figure as a cycle. Making decisions requires information that is based on data. Data provides the information to support decisions, and so on.

Data in itself provides no judgment or interpretation and no basis of action. The context and use of data turns it into information. Connecting pieces of available information leads to knowledge that can then support decisions.

In the ideal situation, the outcome of those decisions is fed back into the BI environment, completing the cycle shown in Figure 2 on page 5. This enables a learning organization in which decisions can be made based upon real knowledge instead of just gut feeling. Actually, the optimal environment enables cross-fertilization between data warehousing and data mining. The data warehouse enables access to a wealth of integrated data that can be mined. Data mining delivers results that are integrated back into the data warehouse, and with that become an integrated part of your organizational knowledge. From this knowledge interesting new areas for mining can be found.

In fact, much of the development work performed today is directed at integrating data mining into your BI environment. It may even end up in your database engine, extending the query capabilities with something like “select the 1000 most likely candidates for buying product X from my customers”. The actual data mining tools will become an integral part of the BI environment and can merge into vertical applications that directly support your line of business.

1.1.3 The Data Warehousing Concept

In the previous section we already mentioned the data warehouse. You may have guessed that there is much data manipulation going on in a BI environment, and, even more, if you add data mining. That is where the need for a data warehouse comes in.

Do you need a data warehouse for data mining? Not necessarily, but it will help you a lot. Most of the preparation work for data mining, that we will see when we cover the data mining process, will already be done when a complete BI environment is in place. While the potential of data mining can be the selling point for a BI environment in your organization, the risk of gathering your data just for mining is, that the effort will be lost after this single shot.

Actually, several reasons exist for building a data warehouse. In Figure 3, you can see how it integrates data from several source systems into a corporate view of your data.

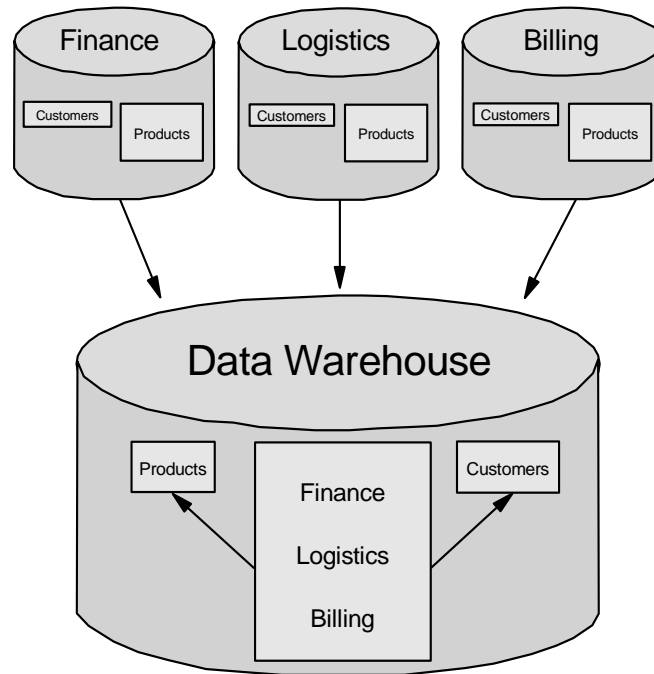


Figure 3. Application Focus and Subject Focus

The figure shows that the focus shifts from operational applications to subjects that are important for your business, such as products and customers.

Besides the integration of data, you might need a separate system for reasons related to the way you use your data by various reasons:

- Queries on operational systems would execute against a data model that was not designed for this, running on a system that has another purpose.
- Queries will compete for resources with the transactional processes running on the operational system which can cause unacceptable delays in real time processing.
- Data that is constantly changing makes it difficult to compare analyses.
- Information must be correlated across independent application systems to clarify all the relationships.

- Operational data is tailored for transaction speed rather than human understanding.

Ad hoc access from a large number of users can also raise security concerns. The delivery of business information from all this data is then gated by the capacity of the IT department in your organization.

The solution to many of these problems, although not a simple one, is building a BI environment with a central repository for all your company data, containing an integrated history of all your operational systems enriched with any other sources that might be of interest. Figure 4 provides an overview of the components of a typical BI environment.

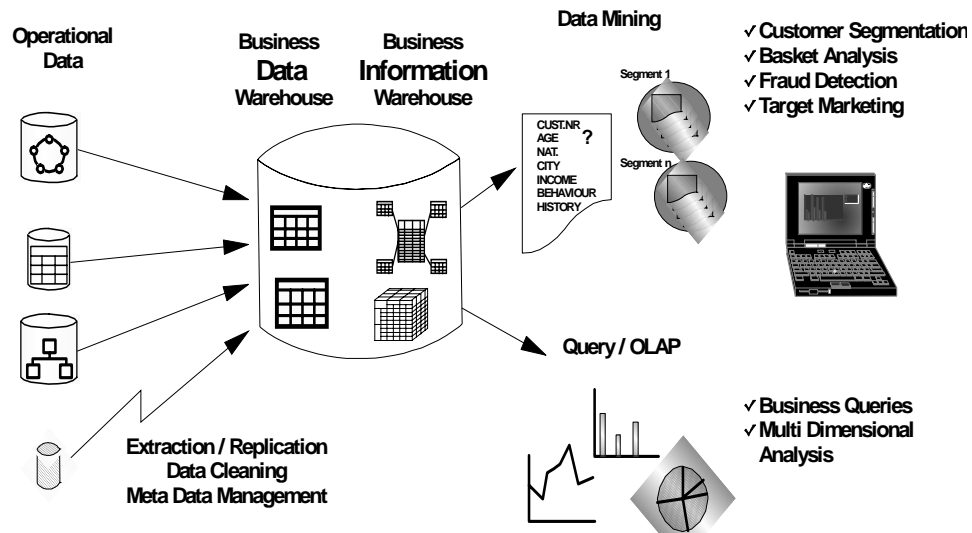


Figure 4. A Business Intelligence Environment

1.2 OLAP and Data Mining

What exactly does data mining offer to enhance an existing BI environment? Online Analytical Processing (OLAP) tools already provide an interactive environment where a user can analyze business data at “thinking speed” instead of having to wait at least a day for the results of their query. OLAP lives by the fact that the result of each query will immediately generate new questions, which must be processed before you forget what you were looking for.

1.2.1 Hypothesis Verification or Information Discovery

Some people already call the kind of analysis described above "data mining". That is the point where we will go one step further. With OLAP, you will only find information that you looked for in the first place. We call this *verification-driven* analysis. Data mining systems will go out and find new information all by themselves, without human interference or input, so the analysis is *discovery-driven*. Data mining systems employ several techniques to determine the key relationships and trends in the data. Data mining tools can look at numerous relationships at the same time, highlighting those that are dominant or exceptional. In this way, you are able to gain new business knowledge from your existing data.

1.2.2 Data Mining Input

Data mining can effectively deal with inconsistencies in your data. Even if your sources are clean, integrated, and validated, they may contain data about the real world that is simply not true. This *noise* can, for example, be caused by errors in user input or just plain mistakes of customers filling in questionnaires. If it does not occur too often, data mining tools are able to ignore the noise and still find the overall patterns that exist in your data.

On the other hand, your data may contain patterns that are only true for a small subset of your data. This effect may be statistically uninteresting, but it also could be just this subset of your customers that are the most interesting because of their behavior. Data mining algorithms can find such *localities* (local effects) so that they will not be lost in generalization.

The effects mentioned in the paragraphs above, noise and localities, are exactly opposite. Depending on the approach, the mining tool could see noise as an interesting effect or discard localities as noise. There are no hard and fast rules about this, but in general these issues can be solved with feedback from the business application of your data mining models.

Data mining will look at your data from different angles at the same time. This prevents it from discarding attributes that do not seem to be relevant on their own. It will find *interdependencies* between attributes that enable the extraction of all relevant information from your data, even when hidden in the combination of several attributes.

1.2.3 Data Mining Output

The output of data mining can provide you with more flexibility. For example, if you have a budget to mail information to 1000 people about a new product, queries or OLAP analysis directly on your data will never be able to select exactly that number of people from your database. By enhancing your data with an attribute that you can use in your query or OLAP analysis, data mining enables you to find the 1000 people *most likely* to respond. This example also shows that data mining is not *replacing* OLAP, but *enhancing* it.

1.3 A Definition of Data Mining

We can state the definition of data mining in a more formal way:

The *process* of extracting *valid*, *useful*, *unknown*, and *comprehensible* information from data and using it to make business decisions.

Let us look at the highlighted words in this definition more closely:

Process	Data mining is not a tool in a box that you simply buy and run against your BI environment, and that will automatically start generating interesting business insights. In fact, we will devote a whole section of this chapter to describe the process.
Valid	The information that is extracted should be correct and statistically significant to support well-founded decisions. Validity means correctness, but also completeness. You want not just the right customers from your database, but all of them. This requires that both the original data and the process of data mining are valid.
Useful	The data mining process may deliver results that are correct and significant, but this knowledge must be of use to the business. For example, if the results tell you to diversify your marketing actions into an inordinate amount of channels, you may not be able to act upon this knowledge. Also, the results must enable you to act before one of your competitors does so.
Unknown	Data mining is intended to generate new information. If the process only delivers trivial results, the business drivers for data mining will disappear. This is the property that distinguishes between verification and discovery.

Comprehensible The results of the data mining process must be explainable in business terms. If not, they might just be a statistical model that you use for rating your customers, as an example. The model itself should at least provide insight into the way that customers are rated and the factors that influence this rating. Being able to provide this insight is required by privacy law in some countries.

The definition above shows you the minimum requirements for what we call data mining. You can use it to evaluate whether data mining adds extra value to your environment.

1.4 Data Mining Applications and Operations

At this point you may be wondering about what data mining means for your business. In this section, we present some of the practical applications of data mining, where each application uses one or more data mining operations. We first provide an overview of the types of operations you will encounter, and the kind of information that each provides. Figure 5 shows some examples of applications, operations, and techniques used in data mining, with some of the relationships between them.

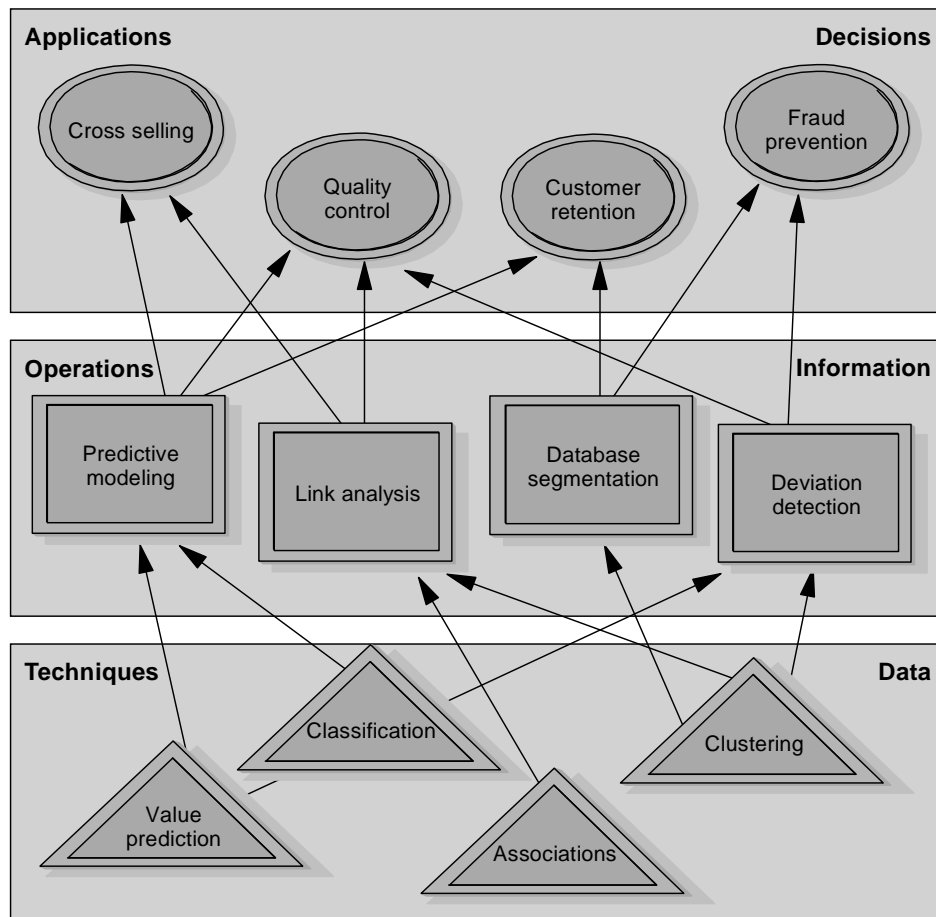


Figure 5. Applications, Operations and Techniques

This figure does not pretend to be complete, but gives an idea of the three levels that you encounter. *Applications* are seen at the business level, where decisions are made. *Operations* are handled by a data mining expert at the information level. Then one or several data mining tools can be used to provide the *techniques* to operate on the data, as shown at the bottom of Figure 5.

1.4.1 Data Mining Applications

The actual application of data mining in your environment depends partly on your business, and partly on your imagination and that of the mining expert. Table 1 provides an overview of data mining applications that have been used so far. We have distributed a number of examples over three main categories.

Table 1. Data Mining Application Areas

Market Management	Risk Management	Process Management
Target Marketing	Forecasting	Inventory Optimization
Relationship Management	Customer Retention	Quality Control
Channel Management	Churn or Attrition Analysis	Demand Forecasting
Market Basket Optimization	Underwriting	Business Scorecards
Cross Selling	Competitive Analysis	
Market Segmentation	Healthcare Fraud	
Web Usage Analysis		

1.4.2 Data Mining Operations

As you have seen in Figure 5 on page 12, the applications are supported by data mining operations. The main categories of these operations are listed below, with a short example of their applications.

- Predictive modeling** Predicting the value of an attribute by using examples. Example applications: assigning risk categories to new customers, or predicting the likelihood for customers to respond to a mailing.
- Database segmentation** Using attributes to find groupings of records where the records in each group have similar attributes, yet the difference between groups is clear. Example applications: use this to group your customers based on their behavior, or as a preparatory step for predictive modeling.
- Link analysis** Finding links between records within transactions or over time. Example application: analyzing which products sell together to optimize your store layout or inventory. You could also use this type of operation to analyze questionnaires or series of medical treatments.

Deviation detection

Finding records, or series of records, in your database that contain values you would not expect. Example application: use this to identify fraudulent behavior patterns or control the quality of your production process.

We leave the discussion of the techniques underlying these operations for the following chapters. You will find more details there about how they can be implemented, what kind of data they need as input, and how the generated output may look.

Chapter 2. Getting Started with Data Mining

So you have this data warehouse or BI environment, with sources, targets, query facilities and, maybe, even OLAP. You have read about data mining, and now you ask yourself whether this technology can be used in your own business.

2.1 Business Drivers

First of all, consider this: If your current environment suits your business needs, that is fine—you may not need more. But perhaps you are wondering if your investments should not return more than they do right now. Or you might wonder about customers that switch to other companies, and why they do so. That might be the point where data mining comes in, because of a business need, rather than as a technical possibility. That is the only way that implementing data mining into your current BI environment will succeed.

Examples of business drivers toward data mining are:

- **Saturated markets:** It is hard to find the right customer for your product, or the right product for your customer.
- **Blurred industry boundaries:** Mergers, takeovers and diversification create sudden changes in the market.
- **Lack of clear differentiation:** It is not clear what kind of customers you or your competitors serve or would like to serve.
- **Fast growing alternate channels:** The internet is drawing much attention from certain customers.
- **Compressed time to market:** Shortened cycle times are needed to keep following, or even predicting, your customers' behavior.

The common theme shared by these drivers is the lack of information about the behavior of the market. At the same time much data is available, but there is no clear way of extracting the information needed.

Example questions that data mining may help you to answer are:

- Who are my most profitable customers and what are their purchasing habits?
- How do I optimize my store layout for profit?
- What kind of products return the highest profits, either directly or indirectly (because they sell together with other products)?

- How do I optimize inventory while maximizing profitable sales and what are the important factors (season, store location, etc.)?
- Which of my new customers are in my most effective market segments?
- How do I make my business more attractive to customers, retain them, and maximize lifetime value?
- Who commits fraud and how do I recognize it?
- How do I predict and prevent failure in my production process?

What these questions have in common is that none contain any assumptions. Instead of testing whether the time of year is an important factor in your marketing efforts (verification), data mining can tell you which factors are important (discovery).

2.2 Technology Enablers

This focus on the business end of the situation does not mean that you might not realize right now what data mining can do for you. It was not possible, until recently, to answer the kinds of questions mentioned in the previous section. There are three important technological developments that enable the business application of data mining:

1. Research on the application of machine learning techniques used to tackle practical business questions, enabling answers that are not just scientifically interesting, but that relate to the real world.
2. The development in hardware and software technology, building systems that can sift through enormous amounts of data in real time. This means that they can make predictions about behavior before it actually happens, enabling you to make the right decision in time.
3. Open connectivity between databases and tools allows easy access to any data from any software package.

In other words, these developments enable data mining because we have enough power available, we know how to apply it, and the data is accessible. It also means that it is now feasible for anybody—with some background in statistics—to apply data mining on large amounts of data to solve real business questions.

2.3 Organizational Factors

Now that you know why to begin using data mining, it is time to pay some attention on where to start. In other words, what organizational environment is needed? There are several prerequisites to successfully start and complete a data mining project. We cover those questions in the following sections.

2.3.1 The Organizational Culture

First of all, the culture in your organization must support the flow of data and information that is needed. It must also accommodate the flow of results from the data mining effort. This means you need an open, communicative culture, where people actively cooperate in the exchange of information. This is especially needed in the interaction between business departments and technical departments. People must be willing to accept new information and, based on this, change the way they work.

If people protect their data and are not willing to share both the needed and resulting information, your organization will need an internal or external consulting effort to change this. It will not be an easy task, but it is essential to the success of the data mining effort.

2.3.2 The Business Environment

We already paid attention to the fact that the business must be leading your data mining activities. Executives at a high level in your organization must be willing to sponsor the project. There must be a clear understanding of the business issue you want to tackle, with a clear statement of the objectives to be able to fix the scope and expectations.

Knowledge about the general business environment and the metrics used to measure its various aspects must be readily available. Keep in mind that implementing a tool in your *technical* environment always means integrating the use of the tool into your *business* environment.

2.3.3 The People

The people factor is all about availability of the right specialists at the right time. The people that are involved in data mining activities fall into three general roles, as shown in Figure 6. The roles can be filled by fewer people, but normally that is not the case.

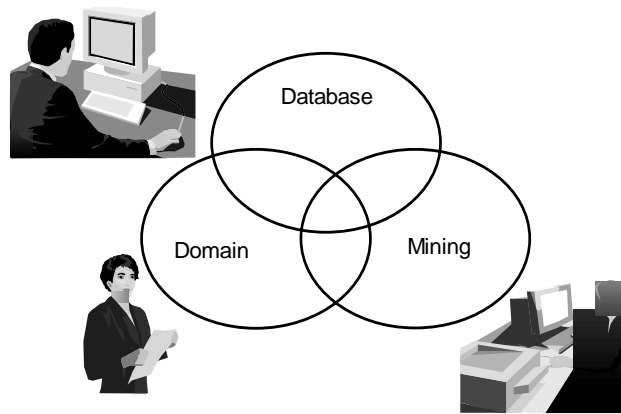


Figure 6. Roles in Data Mining

Following is a description of these roles:

Domain experts	People that know the business environment, the processes, the customers, and the competitors. They are typically people in higher business management functions.
Database administrators	People that know where and how the company's data is stored, how to access it, and how to relate it to other data stores.
Mining specialists	People with a background in data analysis who have at least basic statistical knowledge. They are able to apply data mining techniques and interpret the results in a technical way. They must be able to establish relationships with the domain experts for business guidance on their results, and to the database administrators for access to the data required for their activities.

Generally, you will find the first two roles already present in your company. The third role might require an external advisor the first time your company goes through the data mining process. People in your organization that might fill this role after the knowledge transfer has taken place include, for example, your marketing analysts.

One of the main difficulties in getting the right people, either inside or outside of your organization, is the variety of domains that have to be combined in data mining. You will need liaisons between business, analysis, and technical environments to sustain the continuous bidirectional flow of information.

2.3.4 The I/T Architecture

While your organizational culture will support the information flow, the I/T architecture must support the data flow. You need fast, scalable, and open access to the available data and the flexibility to extract and update subsets of that data in the environment that you will use for data mining. An environment that supports BI and the easy creation of data marts from a central warehouse is a good example of such an environment.

Besides accessing and transmitting data, your I/T architecture must also allow for either the processing capacity needed for data mining, or the easy addition of extra capacity. This could mean adding an extra system dedicated to mining, or running it in addition to the existing processes on one of your systems. In 3.4.2, “Performance” on page 46 we will describe a more detailed picture of the factors influencing processing performance.

2.3.5 The Data

Of course, the data must be available. The amount of raw data usually is not the problem, but the amount of clean, usable, relevant, and integrated data may be less than you think. One of the first steps in the data mining process, as we describe in 2.4, “The Data Mining Process” on page 20, is analyzing your data with this in mind. However, you should have a good idea of what is or is not available, before you start thinking about data mining.

There is no fixed rule about the amount of data needed to start mining. As a rule of thumb, several thousand records, and ten or more attributes, are a good starting point. These numbers further depend on the data mining technique you will employ.

2.3.6 The Data Mining Tools

The tool, or tools, that you use for data mining must be able to support data access, preprocessing, mining, visualization, storage, and maintenance of the results. This can be supported from a single package, or might need several tools. In any case, tight integration between the tools is essential.

You must also pay attention to the scalability of the tools that you plan to use. The well-known credo “start small, but think big” is essential in this case. You will always want to add extra data, explore more history, or achieve results faster. The processing time should not grow much more than linearly with the amount of data, either in the amount of attributes or the amount of records.

2.3.7 The Project

Finally, your data mining efforts must be managed as you would manage any other complex project. That entails a clear understanding of the measures of success in order to be able to gauge progress. There must be a commitment of the skills and resources needed to sustain the project, and it must be managed following a proven methodology.

You will find more on running data mining projects in 2.5, “The IBM BI Methodology” on page 29.

2.4 The Data Mining Process

We already mentioned the important fact that data mining is not just a tool, but a process that should be integrated into the business that it is trying to support. In this section we go into more detail about how this process will look when applied to a BI environment. You will get an idea of what is involved in the application of data mining in practice.

Figure 7 may give the impression that this is a linear process. In fact, results from each step may lead to the conclusion that more information is needed from previous steps, which will then be repeated. These iterations ensure that the final outcome is tailored exactly to your business. We will pay attention to the reasons for looping back at each step we consider.

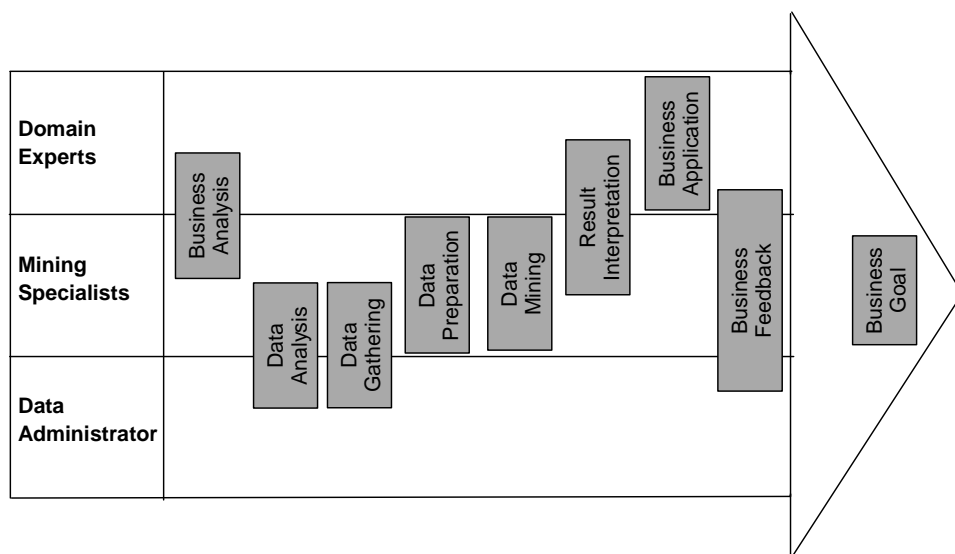


Figure 7. The Data Mining Process

The levels in Figure 7 represent the roles of people involved in the process. You can see that each step involves one or more of these roles, which also correspond to the levels of decisions, information, and data. Iteration cycles between two steps occur frequently when levels are switched, for example between business analysis and data analysis, or between data mining and result interpretation.

To give a general idea of the distribution of the effort across the steps in the data mining process, we list them in Figure 8, together with the typical percentage of time spent in each step. In this case, we have treated the process as linear, but each iteration would mean extra time spent in the previous steps.

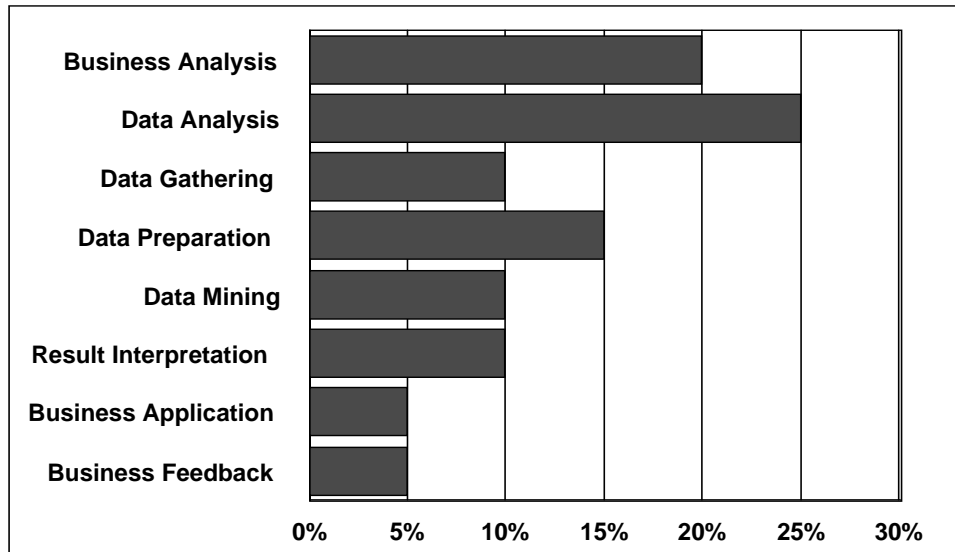


Figure 8. Percentages of Time Spent in Each Process Step

In the following section you find detailed descriptions of the steps listed in Figure 8.

2.4.1 Business Analysis

Ideally, all activities in your company are in some way related to the corporate mission statement through strategy and business objectives. Data mining enables you to handle your objectives at a much higher level than previously. For example, a term like “customer satisfaction” might not mean anything to you in a measurable way.

Data mining can help you clarify the factors that are important in such a case. In this way, it can help you drive your business towards high level goals by concentrating your objectives on the factors you can directly influence at a more practical level.

This means that the goals you set for your data mining activities will always relate to your business at a strategic level. Examples are:

- Increasing customer satisfaction
- Decreasing fraud
- Optimizing inventory

From these goals you try to set clear deliverables and requirements. In the case of increased customer satisfaction, you might specify this as: "gaining insight in the factors that influence customer satisfaction, in a way that enables us to influence those factors", which indirectly will enable you to increase customer satisfaction.

Notice that at the first iteration you may not have to be more specific than this. If the specification of your business requirements is not clear enough, later stages will loop back to this stage for more specific requirements.

Business analysis involves the domain expert and the mining specialist. The former concentrates on specifying the business requirements, while the latter guards the feasibility of those requirements from a data mining point of view, and specifies the mining operations needed to satisfy them, as shown in Figure 9.

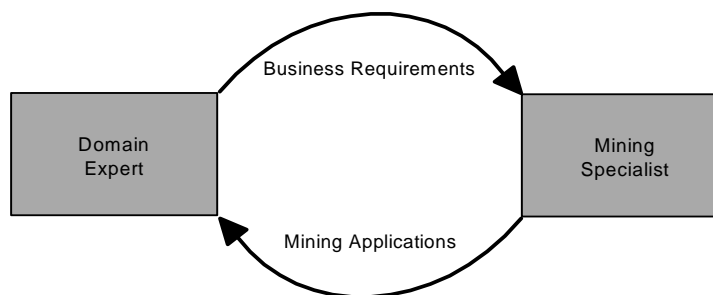


Figure 9. Business Analysis

2.4.2 Data Analysis

The next step is to find out how the business requirements from the first step are represented in the data. This connection is made through the data mining operations specified in the previous step, because the mining specialist knows what kind of data is needed to run these operations.

To investigate the quality of that data using statistical measures, it might be necessary to clean the data, fill in missing values, or integrate data from several systems. If your starting point is a fully operational data warehouse environment, this step may be relatively easy. Be prepared to spend much energy at this point if that is not the case. Data analysis can make good use of an existing BI environment for running queries, generating reports and charts, or OLAP for interactive analysis.

Data analysis will provide an initial insight into the data transformations necessary in later steps, such as cleansing and integration. It might also point out that acquisition of external information is necessary, for example demographic data on customers, that are not required for running the day to day business processes.

If data analysis does not supply the necessary information, it might become a mining process in itself by employing data mining to specify the data necessary for reaching the business goal. Data visualization can be a useful technique in this step, because the human eye tends to note unusual effects in graphical representations.

The roles involved in this step are those of the mining specialist, who will perform most of the tasks, and the database administrator, who will support the activities by providing access to the available data, as shown in Figure 10.

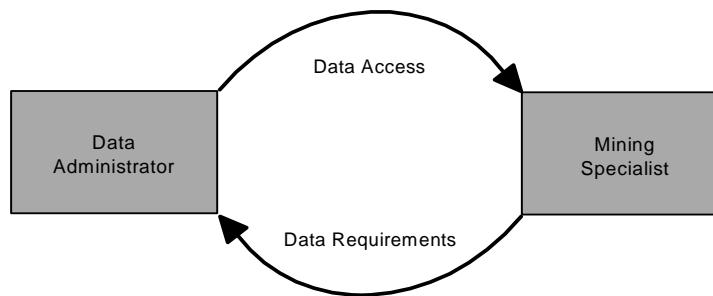


Figure 10. Data Analysis

Looping back to the business analysis step will be necessary if you find that the business requirements cannot be expressed in terms of the data that is available, or if it is not clear how to run the data mining operations using that data. You would then have to adapt your business requirements to the possibilities, or find additional means of satisfying them.

2.4.3 Data Gathering

Data gathering is the step of building a “data mart” especially for data mining. This data mart can be virtual, providing a view on your data warehouse data, or a copy of selected data from your data warehouse. During the data gathering process the data is cleansed and integrated with data from other sources according to the data analysis that took place. Sampling can also take place in this step to reduce the amount of data or to correct for skewed distributions. For example, suppose we try to predict fraud when the actual percentage of fraudulent cases is only 2% of the population. If we then select all fraudulent cases and randomly select the same amount of cases from the rest of the population, this could result in a statistically invalid model, because only 4% of the total amount of data is being used.

A data mart, in the sense of a specialized subset of data from the data warehouse for a specific purpose, is normally not a usable source for data mining. The data is aggregated to a certain level and for a specific goal, whereas data mining needs as much depth and breadth of the data as is available in the original data.

Data gathering involves the mining specialist and database administrator roles. It uses the output from the data analysis step to specify which data is needed, and collects the data by specifying the queries to execute either at the next step or right away for building a separate data repository. (See Figure 11.)

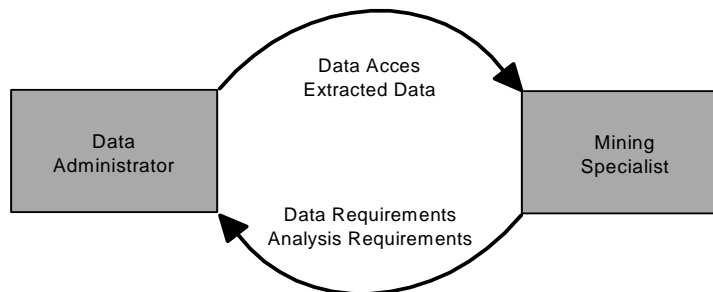


Figure 11. Data Gathering

Looping back to the data analysis step is necessary if the requirements and research from data gathering are not sufficient to build a satisfactory repository of data for mining. Data analysis might use statistical measures that cannot correctly predict the actual data that is gathered.

2.4.4 Data Preparation

When the required data is available for mining, it usually needs some preparatory steps before the actual mining can take place. The need for these preparations is, for the most part, assessed in the data analysis step.

2.4.4.1 Data Quality

The data may show some values, called *outliers*, that are far outside the normal range of what is expected. These values can be treated in several ways. If they are still realistic, a logarithmic conversion of the data will convert the data back to a smaller range. Otherwise, it may be necessary to remove either the records containing such values, or the attribute within all the records.

A more frequent problem is that of *missing values*. Again, values can be missing for some of the records, or one of the attributes may have a high amount of missing values. In the first case, you will probably not be able to use those records, in the second case you would probably drop the attribute.

Another approach to handling missing values is *imputation*. You try to guess the missing value by one of several techniques to prevent discarding records or attributes that might also contain valuable information. Some imputation techniques, increasing in sophistication, are:

1. Fill in a random value taken from the other records.
2. Take the mode, median or average of the attribute from the other records.
3. Make a statistical model of the distribution of the value in the other records and randomly choose a value according to that distribution.
4. Try to predict the missing value with statistical or mining techniques from the values found in similar records.

The last technique requires more work, and also contains the risk of results that reinforce themselves, because the data we use is expected to be correct in order to build our model in the end. This may introduce a bias that defeats the important aspect of data mining; that is, information is generated from data only, without any assumptions.

2.4.4.2 Data Manipulation

Some techniques require *normalization* of the data, where the values found for a certain attribute are converted to have a distribution approaching that of the standard normal distribution with an average of 0 and a standard deviation of 1.

Sometimes the available attributes show a high degree of *intercorrelation*, meaning that the same information is present in several attributes. To prevent this information from dominating too much, we can use several techniques for *dimension reduction*. These techniques try to reduce the amount of attributes to the minimum that still contains the original information. Sometimes they are also useful to speed up the data mining process because of the reduced number of attributes.

One of the data mining techniques sometimes used in the data preparation step is *clustering*. It is described in more detail in 4.6.3, “Clustering” on page 58. The reason for using clustering is that it splits up the data into more or less homogeneous groups. When these groups are very different, it might be better not to try to handle them in one model, but to build models for each separate group. When we apply clustering in this step, it might become a small scale data mining process in itself.

The last important action in data preparation is to split the records into a *training set* and a *test set*. This ensures that, in the end, we have data available that was not used to build the model. This data is used to validate the model. It prevents *overfitting*, which means that the model is completely fitted to the training set, and is not general enough to handle records outside that set of data.

Sometimes *cross-validation* is used. In that case, we also treat the sets the other way around, building a model on the test set and testing it on the training set. This can even lead to *n-fold cross-validation*, using multiple sets, and therefore multiple models that are tested on all other sets.

The activities during data preparation are performed by the mining specialist. They follow up the business analysis and use the data available from the data gathering to prepare for data mining as specified in the data analysis.

Data preparation will expose shortcomings in the previous phase, which then may have to be repeated, sometimes including additional data analysis or even business analysis.

2.4.5 Data Mining

The data mining step consists of actually running the data mining techniques against the data. This might involve running several techniques beside each other, for comparing the results, or after each other, where the output of one technique is used as the input for the next.

We use the training and test sets defined in data preparation to validate the models that are built. Visualization also plays an important role in model validation. The results are summarized using various graphics techniques, which provides an easy way to estimate the model quality, especially when both data sets are combined in one display. Visualization generally provides a starting point for more detailed analysis of the results using statistical measures.

The activities are performed by the mining specialist, following the business requirements that are translated into data requirements, using the prepared data from the previous step.

Data mining might not provide adequate results. In that case we might have to reiterate several of the preceding steps. In the worst case, the conclusion is that the available data does not enable building an adequate model.

2.4.6 Result Interpretation

The interpretation of the results relies heavily on the visualized output from the data mining step. The same graphics that were used to assess model quality are now used to explain the results in business terms.

This step might involve a presentation of results to the business users, as this is the first tangible output of the process they have been waiting for. It can be the point where the project sponsor decides whether or not to continue the support.

Result interpretation requires close cooperation between the data mining specialist and the domain expert. Together they must translate technical results into their meaning for the business and assess the validity of conclusions drawn from these results, as shown in Figure 12.

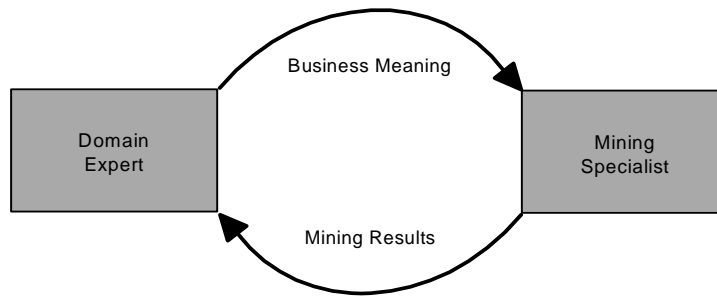


Figure 12. Result Interpretation

This step will usually form a cycle together with the previous step. Iteration is necessary when results are valid in technical terms, but do not mean anything interesting in business terms. An example would be when results try to model customer behavior, but actually model company behavior. That might be useful, but it is for the domain expert to decide.

2.4.7 Business Application

After all parties have agreed on the meaning of the results for the business, you need to make decisions on how to use the results. This decision will not only be a logical outcome of the results, it must also enable the verification of that outcome in the business environment.

One example is the prediction of a high profit segment within your total customer universe. You must decide how to approach these customers and, at the same time, how to measure whether they are indeed as profitable as the model predicts.

The decision about business applications is supported by the domain expert, who relies on the information from the mining result interpretation. In a small company the domain expert might be the decision maker, but normally it will be someone in an advisory role.

Looping back to result interpretation will be necessary if the domain is not sure whether the decision to be taken is fully backed by the data mining results, or if there is any misunderstanding between him and the actual decision maker.

2.4.8 Business Feedback

In this step, results from the business environment are fed back into the BI environment to be analyzed together with the output of the data mining model. For example, suppose you compare the predicted response to a mailing campaign with the actual response. This could also trigger a new data mining process when you try to understand the factors that caused you to predict incorrectly regarding some of your customers. The results of the data mining process can be used to justify further business objectives.

You will have to be careful when you start treating your customers or optimizing your processes using models that are based on past behavior. If you rely only on these models, you will never get any information outside the model you applied. So, when you are using data mining, traditional market research or statistics must still be used to constantly verify the feasibility of your models against neutral data and market trends. This is especially true for the first time that you apply the model. You will start treating your customers differently, so past behavior is no longer valid in the current context. Be prepared to update your models frequently in the first iterations of the data mining process.

As you can see in Figure 7 on page 20, this last step integrates everything from the previous steps over all three levels of decisions, information, and data. It enables the learning organization by assimilating the knowledge as one of the main assets of the organization itself.

Business feedback involves all three roles in the data mining process. The domain expert will be able to gauge the results. The mining specialist will assess the validity of the models, and prepare the data for feeding back into the data warehouse, supported by the data administrator.

Looping back previous steps will be necessary if the actual performance is below what was expected there. It could mean that the expectations were unrealistic, the model was invalid (forcing a loop back to previous steps), or that market behavior has changed since the creation of the model.

2.5 The IBM BI Methodology

IBM offers the full range of support from business consultancy to product implementation through standard offerings or tailored projects by IBM Global Services. The experience built during the last decades has culminated into a reference framework for BI projects called the *IBM BI Methodology*. This methodology encompasses a reference data warehouse architecture that provides the necessary framework for designing your BI environment.

2.5.1 Data Discovery

Part of this methodology is the offering for *decision optimization*, which covers data mining in its *data discovery* function. It emphasizes not just the data mining process itself, but also the knowledge transfer that enables your organization to continue the path it will take. Because the offering fits in the overall BI methodology, your data mining efforts will always fit into your BI environment.

The use of data mining on an ongoing basis is an explicit part of the methodology, which is designed to start the process as described in 2.4, “The Data Mining Process” on page 20, and to keep it running from that point on. Each activity consists of several tasks which may be further divided into the task steps as shown in Figure 13.

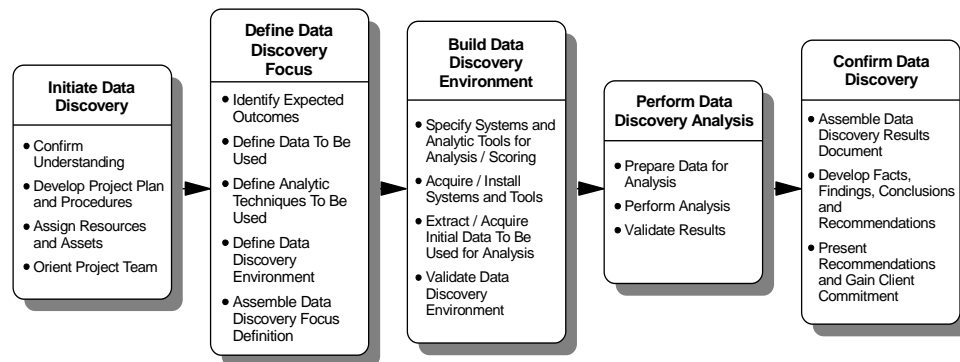


Figure 13. Data Discovery Activities and Tasks

2.5.2 Tasks and Deliverables

The methodology describes the following for each task:

1. Purpose
2. Description
3. Technique: how to fulfill the purpose, for example, using a facilitated session or an interview
4. Inputs: what is needed to start this task
5. Outputs: the deliverables
6. Roles: what kind of people is involved
7. Estimated duration

The outputs of each activity allow you to review current material as the project progresses, rather than reviewing everything at the completion of the project. Figure 14 shows the deliverables and an example of the durations to expect.

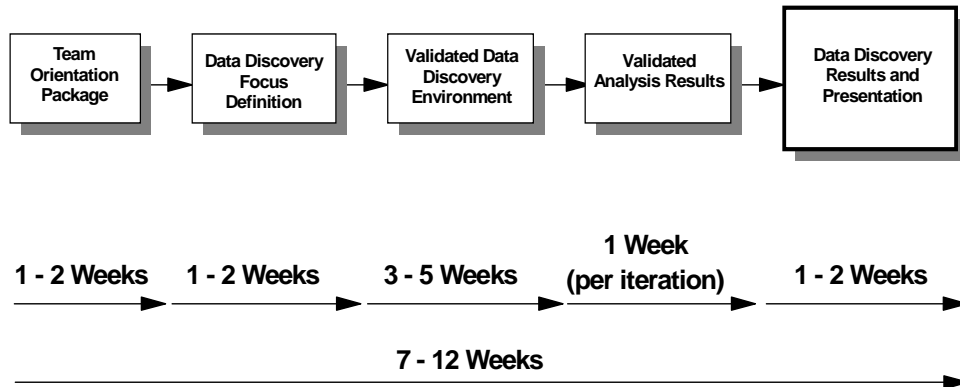


Figure 14. Deliverables and Work Products

2.5.3 Roles

The roles people play in the data discovery offering encompass the same roles mentioned in 2.3.3, “The People” on page 17, but they include additional I/T and project management roles. After all, the offering tries not just to run the data mining process, but also to embed it in your organization by knowledge transfer and architectural design. The following roles are needed to implement the offering:

- Project Manager
- IT Architect
- Solution Architect
- Data Architect
- Data Modeler
- Data Mining Expert
- Business Analyst

Each of these roles can be filled by either a specialist from your own organization or an IBM specialist.

2.6 Summary and Outlook

These first two chapters have shown what data mining can mean for your business and how it can enhance your existing BI environment. In the following chapters we will consider more technical aspects of an implementation, introduce the IBM Intelligent Miner for Data, and describe how to implement it into an existing environment. For this we will use an example environment and show how to prepare its data for mining.

Chapter 3. Data Considerations for Data Mining

The most time-consuming procedure in the entire mining process usually is preparing the input data for data analysis. This means that if you want to make a good model to satisfy your business goal, you must pay attention to the flow of data. In this chapter we talk about data sources, data format, and data transformation for mining. We also pay attention to the location of the result data of the mining process.

3.1 Data Sources

There are two major approaches in accessing data for data mining. Data can come from a data warehouse, if you have already built one for the end-users to get easy access to data. The other approach is to access the operational data directly and extract the data in the appropriate format of aggregation for data mining. Additionally, you sometimes might need to purchase customer data, like demographic data, from external sources to enhance your existing data.

The preferred approach is to directly access a data warehouse that has a data delivery system that can quickly produce a specific format of the data mart. This approach focuses on the data preparation process for mining and the objective is to make the process as dynamic and reusable as possible. To the extent possible, the system is a dynamic function that takes the data preparation specifications as an input parameter and generates the data for mining without requiring any type of recoding or recompiling.

As shown in Figure 15, we can define a 3-tier data structure in the BI environment.

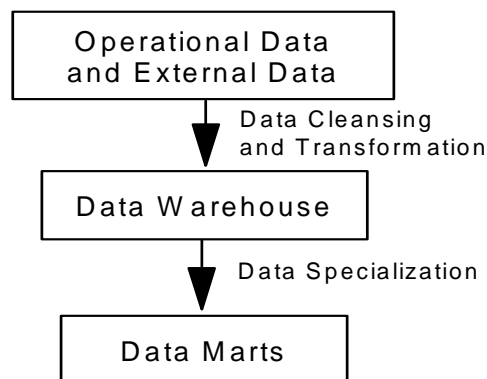


Figure 15. 3-Tier Data Structure

Data from any of these tiers can be on any platform or the same platform, depending on volume, security, and performance. We will describe the volume, security, and performance considerations at the end of this chapter.

3.1.1 Data in Operational Systems

Operational systems are used to manage the day-to-day business activities. As such, they are critical to the ongoing vitality of the enterprise. They usually do not allow for historical data analysis, yet this is a major concern for business analysts. Historical data assets have been sitting idle in a variety of inaccessible storage media.

For example, if a business analyst wants to find out who will be a possible defaulter, he might ask a data administrator to gather 3 years of transaction trends and demographic data for all customers. The data administrator will then need to extract not only 3 years of customer transaction data from accounting systems, but also demographic data from other storage media.

Finally, if you create two files for customers (transaction data and demographic data), you need to merge them based on the customer number. The merged file can be the input of data mining. In this procedure there must be a good communication between the business analyst and the data administrator.

Normally the data in operational systems is not easy to access by the business analyst because the location and meaning of data are diverse across the domains. This situation is an obstacle to the business analysts. Insufficient pre-analysis of data can cause the distortion of initial business requirements.

Operational systems often perform at the limit of the hardware and software with which they are implemented. They are often a key part of customer service, a major factor in the success of an individual retail enterprise in the very competitive retail industry. Therefore, the ongoing operation of the operational systems is of highest priority. The reason for using copy-based informational systems with respect to protecting existing operational applications, fall primarily into the areas of data accountability and application performance.

3.1.2 Data in a Data Warehouse

A data warehouse is a set of databases consisting of cleansed, reconciled, and enhanced data, integrated into logical business subject areas for the purpose of improving decision making. Although a data warehouse is not essential for data mining, it provides the basis for the powerful data analysis techniques available today, such as data mining and multidimensional analysis, as well as the more traditional query and reporting tools. The data warehouse is designed to be a neutral holding area for informational data and is intended to be the sole source of quality company data for decision making. The data warehouse is different from the operational data in that it is subject-oriented, integrated, time-variant, and exposed.

Subject-oriented	Data warehouses are designed to satisfy the needs of business users, not for day-to-day operational applications. Not all the information in the operational systems is useful for a data warehouse.
Integrated	Data in a data warehouse is clean and consistent across the domains, and is stored in a form business users can understand.
Time-variant	Unlike operational systems, which contain only detailed current data, data warehouses can supply both historical and summarized information.
Exposed	Unlike operational systems, business users can access the different domain data in data warehouse for decision making.

Once you have your enterprise-wide data warehouse operational, it means you have integrated and flexible data across the domains. Although there may be a need to create new tables or change the related tables in a data warehouse because business requirements are changing continuously, it is more flexible to deal with relational tables than flat files.

For the following example, assume that you have a table with 3 years of customer banking transactions in the data warehouse, as shown in Figure 16. Figure 17 shows a product and communication media table that also is located in the data warehouse. (Both of these tables are usually generated on a day-to-day basis from your operational data using data replication tools).

Cust_Tran						
Cust_No	TR_Date	TR_Time	Product_No	Channel_No	Amount	...
.
123	02/12/1998	12:12:31	99353	01	150	.
234	02/12/1998	12:13:25	99323	03	84	.
123	02/12/1998	18:29:30	99237	02	740	.
325	02/13/1998	10:10:25	99237	02	200	.
325	02/13/1998	11:21:12	99845	01	320	.
211	02/13/1998	15:40:21	99353	03	32	.
.
.

Figure 16. Customer Transaction Table

Product			Channel	
Product_No	Product_Group	Product_Name	Channel_No	Channel_Name
99353	001	Saving Account	01	ATM
99323	001	Fixed Account	02	Internet
99237	003	Installment	03	Phone
99845	002	Special Deposit	.	.
.

Figure 17. Product and Channel Tables

Now let us also assume that your business requirement is “What is the proper communication media for a campaign targeting the new customers of the last month?” If a business analyst requires the usage frequency of channels for each customer, how will you do it? If this data resides in an operational system or storage media as a file format, you will need to create a program to extract the data you need.

To know the business requirements, you might need a summary table of channels for each of your customers (see Figure 18).

There are several ways to build a Channel_Usage table. You can use a transformation tool for transposing or your own program. Intelligent Miner (IM) provides various processing functions based on a Structured Query Language (SQL) interface.

Channel_Usage

Cust_No	ATM	Internet	Phone	...
.	.	.	.	
123	21	80	2	
211	45	10	20	
234	86	0	1	
325	4	12	95	
.	.	.	.	
.	.	.	.	

Figure 18. Channel Usage Example Table

Following are the SQL command steps to build a Channel_Usage table (Figure 18) out of the source tables (Figure 16 and Figure 17) in DB2.

1. Create Channel_Usage and temporary table:

```
CREATE TABLE Channel_Usage
(Cust_No          INTEGER NOT NULL,
 ATM              SMALLINT NOT NULL WITH DEFAULT,
 Internet         SMALLINT NOT NULL WITH DEFAULT,
 Phone            SMALLINT NOT NULL WITH DEFAULT,
 PRIMARY KEY (Cust_No));
```

```
CREATE TABLE TMP_Channel_Usage
(Cust_No          INTEGER NOT NULL,
 Channel_No       SMALLINT,
 Usage            SMALLINT,
 PRIMARY KEY (Cust_No, Channel_No));
```

2. Extract each channel usage frequency and insert it into the TMP_Channel_Usage table:

```
INSERT INTO TMP_Channel_Usage
SELECT Cust_No, Channel_No, COUNT(Channel_No)
FROM Cust_Tran GROUP BY Cust_No, Channel_No;
```

3. Insert Cust_No into the Channel_Usage table:

```
INSERT INTO Channel_Usage(Cust_No)
SELECT DISTINCT(Cust_No)
FROM TMP_Channel_Usage;
```

4. Update Channel_Usage with each channel usage frequency:

```
UPDATE Channel_Usage T1 SET (ATM)
SELECT Usage FROM TMP_Channel_Usage T2
WHERE T2.Channel_No=1 and T1.Cust_No=T2.Cust_No;
```

Repeat this command for as many channels as you have. Figure 19 shows the table values and formats for these steps.

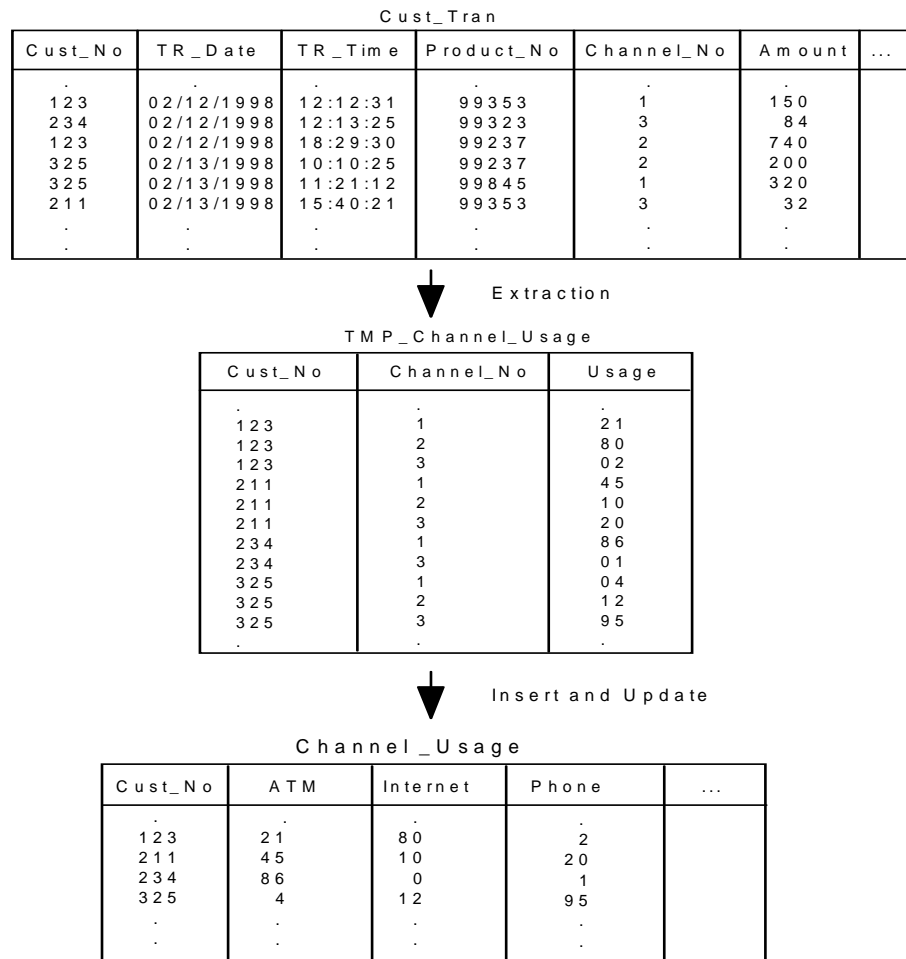


Figure 19. Transposition Example

For keeping historical data, you need to capture the point-in-time picture of information and reconcile the data for mining. Staging of data and data replication techniques are necessary to perform the reconciliation required for data mining without impacting the operational environment.

3.1.3 Data Replication

An organization's data configuration may be composed of a variety of data stores (for example, relational and hierarchical databases, and flat files) on a variety of dispersed systems. Creating and maintaining the reconciled and derived data can be a complex task. It requires understanding of the source and target data formats. It also requires a methodology for subsetting, cleansing, transforming, and transferring the data using copy tools. The role of data replication is to assist administrators in performing these tasks.

In most cases, business requirements, performance, and security issues separate the data warehouse from the operational data. This results in another concern, namely that the data warehouse needs its data to be consistent with the operational data at a point in time.

The data replication architecture of IBM will make it easier to manage the delivery process, and reduce the cost of both tools development and execution. Figure 20 shows an IBM product based data replication solution.

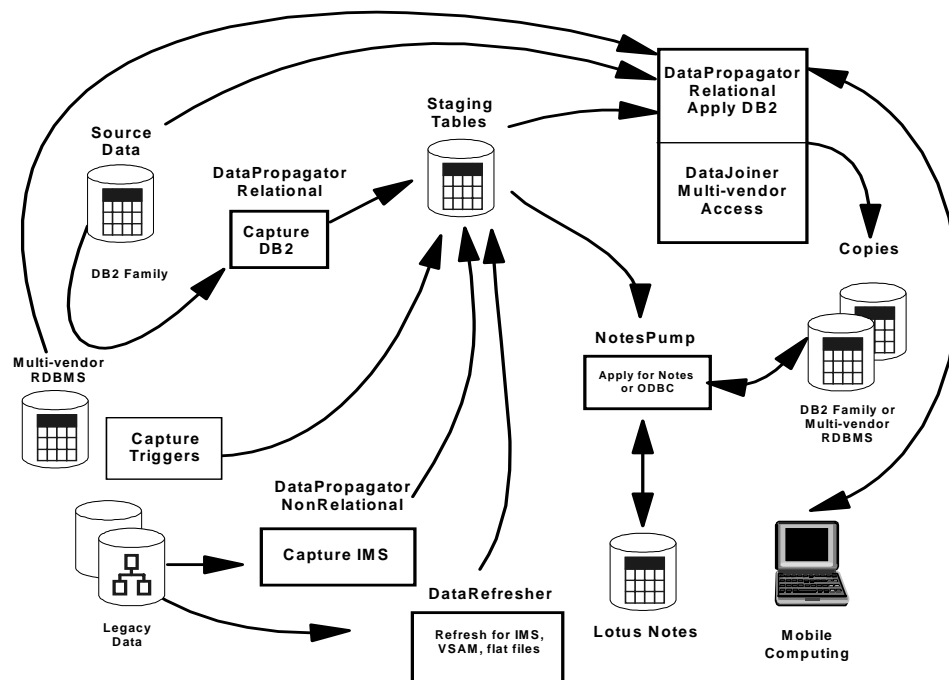


Figure 20. IBM's Open and Comprehensive Data Replication Solution

DataPropagator Relational (DPropR) and DataPropagator Non-Relational (DPropNR) fulfill the functional requirements of the data replication architecture. IBM has many kinds of data integration tools such as Visual Warehouse, DataJoiner, and ETI-Extract.

We have focused on a particular data replication environment in this section. This set of products is part of the data delivery technologies that include the following:

- Data access protocol
- Capturing
- Applying
- Refresh propagation
- Archiving

IBM has published distributed relational database architecture (DRDA) as an open interface for remote access to relational databases. The architecture supports both remote unit of work (RUW) and distributed unit of work (DUW).

The capture component uses an interface to the database log files or journals to detect and save changes (update propagation) to the registered tables. The database management system writes records to the log as a normal recovery function. The record format is slightly different for tables identified for data propagation. The DataPropagator Relational capture program reads these log records and stores them in a table called a staging table.

The apply components takes the changes to the target tables and applies them to the corresponding base tables. The source used by the apply components varies, depending on the environment (in particular, whether capture is running in the base table's database system) and on the type of propagation request. The apply components generally operate on a time interval basis; you specify how often the apply should run for a particular subscription.

A variety of tools implement refresh propagation. DataRefresher performs refresh propagation on a variety of sources, including IMS, VSAM, and DB2 for MVS. Refresh tools are often extended by specialized code available as user exits. The Data Facility SORT (DFSORT), an IBM MVS-based program, is a tool that illustrates this feature. It is fast, economical, and performs sorting and manipulation of data.

Archive systems are a distinct class of data copying and data delivery because they specifically manage historical data. DB2 Data Archive Retrieval Manager/MVS (DARM) supports this type of data delivery. Archive products will become more important as historical data and the trend analysis against it become more popular.

3.2 Data Transformation

There are two different stages of data transformation in a 3-tiered data structure.

Stage 1: While building the data warehouse you will need transformation to integrate the data across different domains. For example, an insurance company has separate application domains for car, health, home, and life policies. It is desirable to have a single, consolidated view of all the relationships that a business has with a given customer. You can perform data aggregation, data cleansing, and data mapping in this stage.

Stage 2: When you prepare the data for data mining, you will denormalize the tables, change the values (data cleansing), make a mapping, or add new columns to the tables. This step needs basic statistical concepts, mining functions, and business rules.

This differentiation is important because sometimes the term *data transformation* causes confusion between business analysts and data warehouse designers. Normally, the data cleansing in stage 1 is done to consolidate the type and format of the data that comes from various sources, and to check the invalid data or missing data when integrating. The data cleansing in stage 2 is different in that it is related to data mining techniques and statistical problems for analysis. For example, sometimes you need to add the missing values with mean values, or to exclude out of range values to reduce the noise.

Figure 21 shows a suitable data warehouse design pattern and data mart pattern for data mining.

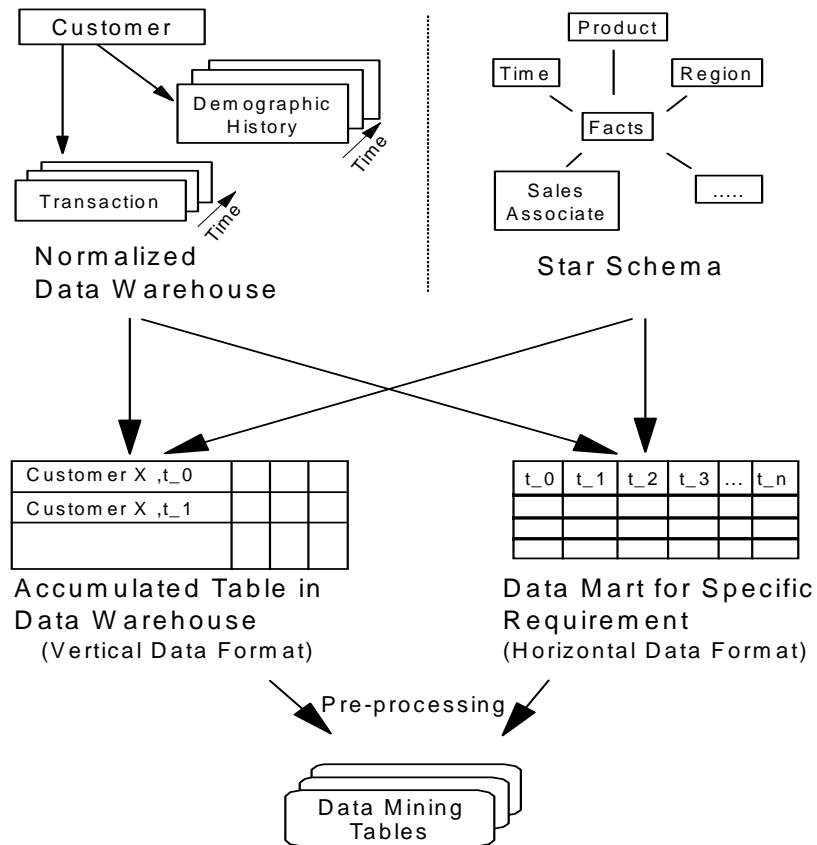


Figure 21. Data Warehouse Table Pattern and Data Mart Pattern

Because there are various requirements of business users to keep their changing environment, data warehouse designers normally make an effort to satisfy the business requirements and current IT environment. They try to optimize the purpose, usability, performance, and maintenance, but sometimes it does not support everybody's needs. So, you need to create a table (Data Mart) for the specific requirements from existing tables.

As shown in Figure 21, we can use tables that are prepared for data mining directly, or tables which need preprocessing before running data mining.

3.3 Where to Put the Result Data

For the entire data mining process, we consider two kinds of results. The first kind is the model that is generated by the data mining process. The second kind is the result of applying the model to the data that is available. Most data mining tools can create output data from the model application as database tables or flat files. You can use the output data to create a report, to support decisions, or keep it for further analysis by domain users. The model itself is usually shown as a visualization.

When you share the output data among domain users, be sure to provide business rules and meta data across all domains. For example, although you can see link analysis results only as a visualization, you can design a table to store the output data in your data warehouse. Metadata is used to document the data descriptions so they are well understood by the business users.

The location of the result data is related to your business components, performance, security, and maintenance policies. There are many factors influencing the business components that sustain your company. For example, end-users, applications, decision makers, and IT resources are important factors in your company. So, when you think about the placement of result data, you should ask yourself the following questions:

1. Which business domain will use these results?
2. Who will make decisions based on these results?
3. Can the existing application access these results easily?
4. Do we have enough resources (hardware)?
5. Does it create performance problems for other applications?
6. Does it violate my security policy (access authority)?
7. What is the cost to maintain it?

The first four factors consider business and I/T environment issues. The others will be handled in more detail in 3.4, "Technical Considerations" on page 44.

3.3.1 For Analysts

The actual people that will analyze the result data should have direct access to the results, either in database tables or as result files they can access with a visualizer. These tables or files could be located in the central data warehouse, but a better location would be in a specific data mart for analysis.

This enables the analysts to perform their tasks without any impact on other users or operational systems.

3.3.2 For Decision Makers

The results that are available to decision makers are usually highly aggregated and visual. These results, prepared by analysts, would most likely be located in the system the analysts use. However, in a highly evolved data warehouse architecture there could be a separate Executive Information System (EIS), which would then be the desirable location for the results.

3.3.3 For Applications

For some kinds of applications, the results must be available online. One example would be a call center that needs the profitability score for a customer who is calling. Another example in which the model must be online would be an insurance company that wants to assess the risk of a new customer while on the phone with the customer.

The best location for result data would be in the operational system itself, added to the tables that contain current information. The model would be embedded in the application, or easily accessed from that application using an application programming interface (API) in the data mining tool.

3.3.4 For Resources

The resources that we consider in this section are storage space and processing capacity. We will look at each factor in turn.

Generally it will not cause storage problems to store the additional mining results along with your operational data (either in your data warehouse or in your operational system), because it consists of only one or two small fields.

For online scoring we might need a substantial amount of processing capacity. This capacity is, generally, best located on a separate mining system instead of the operational system. It might be taken from the data warehousing system, although this is normally optimized for I/O performance. If so, it is not a good choice for providing processing performance, because this will strongly impact the system's tuning.

3.4 Technical Considerations

This section describes some of the technical considerations for security, performance, and maintenance of the data mining component in your BI environment.

3.4.1 Security

Building the data warehouses and data marts, along with security and availability requirements, have placed data remotely from the operational transaction activities. Copying data allows for data placement closer to knowledge workers responsible for making decisions. Ownership of data implies identifying an individual who is responsible for the quality and currency of the informational object. It is advisable to have a copy of the data in an isolated informational environment where the access security can be handled independent of existing operational system policies. This includes granting access to specific sets of data by userids or by a group identifier.

3.4.1.1 Accessing Mining Input Data

As described in 3.1, “Data Sources” on page 33, there can be three possible locations in which the mining input data can reside:

- Mining input data that is located in operational systems should be extracted and accumulated for data mining. You can also use either flat files or database tables (in a separate data mart) as source for mining. In case of flat files, only the data administrator will have access authority to the operational system’s data and the mining specialist will have the authority to access the flat file.
- When you consider mining input data in a data warehouse, access security appears to be in direct conflict with the very philosophy of the data warehouse—it is created to store integrated corporate data to support decisions. Applying security measures keeps users from accessing data and defeats its original purpose. On the other hand, the data warehouse contains a large amount of strategic data that is vital to your business and must, therefore, be protected. If you have a strict security policy in your data warehouse, you can create a view on the related tables for the mining input data. The data mining specialists can then be granted “read” authority to that information.
- If you provide the mining input data as a data marts, it is easier to determine the security level. Because you already have a specific reason to make the data marts, only those people that are related to data mining will have access authority.

3.4.1.2 Accessing Mining Result Data

The value of the mining result data is so important to your business strategy that it needs a high security level for general business users. But authorized mining specialists must be able to read, create, and update the mining result data regardless of the location.

Once you have determined the location where the mining result data will be kept in the data warehouse, you should be careful about the users that need access to this result data. You can select the columns that are useful to the end user business applications and then create a corresponding view on the mining result data. Again, this will be easier if the result data is stored in specific data marts that allow access to only a specific group.

3.4.1.3 Access Auditing

Auditing system and data access provides a reporting of who is accessing the data, and a way of identifying access for illegal purposes. However, damage done to your business by illegal access must be evaluated thoroughly. Many businesses believe that their employees are reliable and trustworthy and therefore access is not a problem. The auditing of access then provides only evidence of wrongdoing for which action can be taken.

A security architecture for the corporation must include the data warehouse and data marts, at the same time encouraging use of the data warehouse for supporting decisions.

3.4.2 Performance

There are many factors that influence performance. We will detail several of those factors in the following sections.

3.4.2.1 Location of Input Data

We recommend that the location of mining input data be local to the mining server to prevent the mining process from becoming I/O bound. If this is not feasible, the data can be stored in a lookaside buffer or local cache on the mining server, so that only the first access will be I/O bound. Each subsequent data access then occurs locally on the server. This requires sufficient local storage (up to the amount of the original data) and will increase performance for mining techniques that perform multiple passes over the input data.

3.4.2.2 Input Data Volume and Variables

There is no rule stating that the quality of mining results depends on the amount of input data available, although a minimum amount is required. More important than data quantity is data quality. For optimal cost and performance, some experts advocate a sampling strategy that extracts a reliable, statistically representative sample from the full detail data. Mining a representative sampling instead of the whole volume drastically reduces the processing time required to get crucial business information, in addition to

minimizing the resources such as the required disk space. The result, of course, is a less costly process that yields faster results of equal accuracy.

On the other hand, details in the data such as local effects can be lost due to sampling. It is now possible to execute mining techniques against amounts of data in the terabyte range on platforms up to multiple-node mainframes. Sampling can be used as a preparatory step to find the main effects, while the actual mining run uses the maximum amount of data available.

Reducing the amount of attributes in your input data not only reduces the distortion of results but can also increase the performance of data mining. There are three major techniques used for reduction:

Combination	Creating new variables by combining existing variables with some kind of formula, for example, by summing or multiplying them.
Factor analysis	A statistical technique that tries to eliminate variables that have no relation to the information we want to extract.
Principal components	Conversion of the available attributes to a certain minimum number of attributes that can still represent the original information.

Again, you may lose details due to the reduction, so we want to stress the importance of good data analysis as a preparatory step.

3.4.2.3 Mining Techniques

Performance also depends on the mining technique that you use, although this will normally not be the deciding factor. Depending on many factors, a mining run on all available data may take several hours. You should try to estimate the running time by first mining on a sampling of data.

Most mining techniques will scale linearly with the amount of data. However, be aware of factors such as local effects in your data, which will increase the model's complexity, or certain limits that are reached, such as the amount of data that fits in local memory. In such cases the actual run time may be much higher than you expect. A good rule of thumb is to calculate the amount of data in two or more consecutive runs with different sample sizes, and then add about 25% extra time, to stay on the safe side.

3.4.2.4 Parallelism

We can discern two different kinds of parallelism associated with data mining: within the server or within the mining technique.

Server Parallelism

Parallelism within the server means availability of multiple processors that can run any of the tasks executed on the server. This kind of parallelism is not explicitly used. The operating system distributes tasks over processors, meaning that if you execute several mining runs in parallel, or a database engine together with the mining engine, you will automatically gain from the parallel configuration.

Mining Parallelism

Some mining techniques lend themselves to parallel execution on multiple servers or multiple subsystems within the same server. This may drastically reduce the run time, sometimes linearly with the amount of processors. However, planning, preparation, and tuning of the overall architecture will require a considerable amount of time.

You should consider exploiting this feature only if you have large amounts of data available, or if the mining process itself is time constrained. For example, you might have to evaluate a large amount of customers each day, for which a batch window of only one hour is available. You will gain the most from the application of parallelism with repeatable applications because the considerable effort of tuning can then be reused.

3.4.2.5 Database Design

Data mining techniques will generally execute table scans on a single large table, so the main bottleneck is I/O throughput. Internal database parallelism can help to deliver the data to the mining technique as fast as possible. You should try to avoid joining tables during the mining run, unless one of the tables is very small.

Generally a database should be well indexed in order to achieve optimum performance. Some of the mining techniques will benefit from indexing, for example, Associations Discovery and Sequential Patterns. Both will automatically create an "order by" clause within the SQL statement, and therefore an (ascending) index on the appropriate columns will help to speed up the run.

3.4.3 Maintenance

The maintenance factors associated with data mining depend on your business cycle and business area. In the following sections we will identify four areas where maintenance is an important factor. The need for updating parts of your mining process follows the flow that is shown in Figure 22.

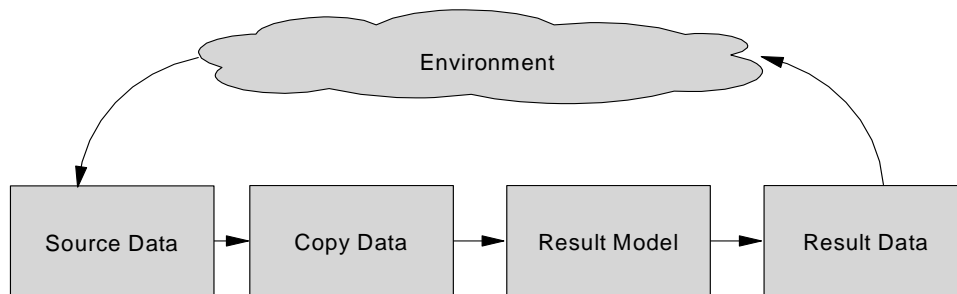


Figure 22. Update Maintenance Flow

This figure represents the need to update each step of the process when something changes in the previous step.

There are no special considerations for maintaining your data mining source data, other than those considerations applicable to any other database or data warehouse. Because you will make decisions based on information derived from this data, it is very important that your data represents the actual status in the real world.

If you use a specific data mart for data mining, it generally needs to be updated before you build a new model on the data. You could create an automated process by using one of the copy management techniques mentioned in 3.1.3, “Data Replication” on page 39.

The validity of a model on which you built on your data is dependent on the volatility of your business environment. If the environment is changing rapidly, you would have to regenerate the model frequently. If you have your maintenance chain set up correctly, you can validate your model on the changed data by regularly applying it to incoming data and comparing the output with the actual status of the data.

Result data is generated by applying your current model on your current data. Therefore, the additional data, such as scores or segments that you generate, will be invalidated by changes in other attributes in the records. Again, any update in the source data will eventually invalidate the model and also force regenerating the result data using an updated model.

You can administrate the process by maintaining a database table with attributes as shown in Table 2. A table such as this should become part of the metadata in your data warehouse as an audit log for updates in your mining application.

Table 2. Sample Mining History

Attribute	Description
Model_No	Sequence number
Model_Group	Purpose
Model_Status	Current status (for example: test, production, valid, invalid)
Create_Date	Creation date
Update_Date	Update date
Model_Creator	Creator or owner
Related_Dept	Department of model users
In_Data_Type	Flat file or database table
In_Data_Name	Name of input data
In_Data_Loc	Location of input data
Result_Type	Flat file or database table
Result_Name	Name of result data
Result_Location	Location of result data

Chapter 4. Introduction to IBM Intelligent Miner for Data

The Intelligent Miner (IM) was developed to help the data mining user to identify and extract unknown highly-valued business data from conventional data. The user of this data mining toolkit, as explained in Chapter 1, "Business Intelligence and Data Mining" on page 3, must be familiar with statistics, mathematics, and machine learning, and must understand the business requirements.

The objective of the IM is to expedite the information discovery process while maintaining the quality of the extracted information. The IM offerings are intended for use by data analysts and business technologists in areas such as marketing, finance, product management, and customer relationship management.

In this chapter we describe the functionalities of the product and the steps of the data mining process when using the product.

4.1 Overview of the Intelligent Miner

IM is based on a client-server architecture. The server can run on OS/390, OS/400, AIX, Sun/Solaris, or WindowsNT, and the client can be installed on either of AIX, OS/2, WindowsNT, or Windows95.

It has the ability to handle large quantities of data, shelter users from the inner workings of the underlying mining technology, present results in an easy to understand fashion, and provide programming interfaces. Increasing numbers of mining applications that deploy mining results are being developed by customers, IBM, and IBM partners.

Through an intuitive graphical user interface (GUI) you can visually design data mining operations. You can choose tools and customize them to meet your requirements. The available tools cover the whole spectrum of data mining functions. In addition, IM selects data, explores it, transforms it, and visually interprets the results for productive and efficient knowledge discovery.

The IM block diagram in Figure 23 shows the role of each professional: The data analyst handles the development, and the business analyst handles the application work.

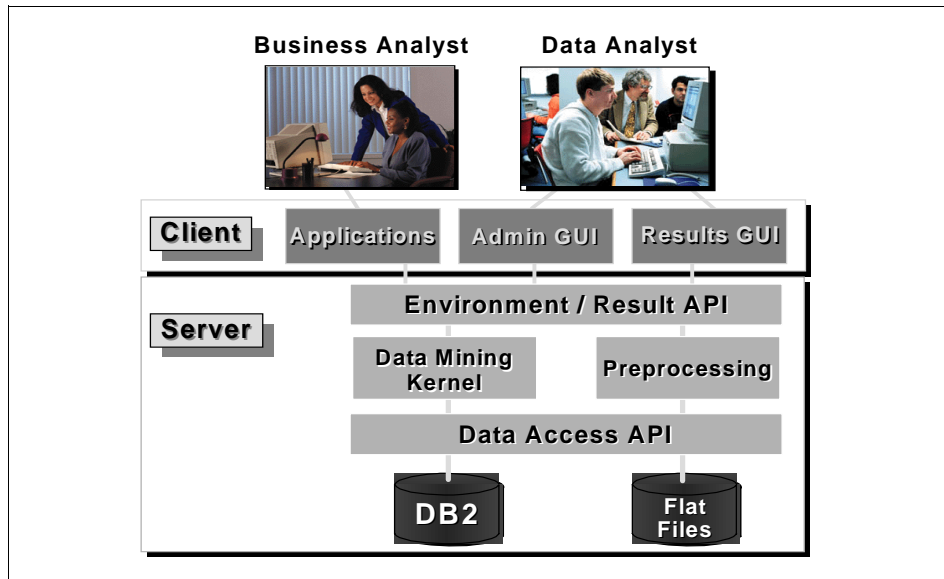


Figure 23. Intelligent Miner Block Diagram

The server runs the mining and processing functions, and stores the historical data and the mining results. The client manipulates the data with the visualization tools, and can be used to visually build a data mining operation, run it on the server, and have the results returned for visualization and further analysis. In addition, the IM application programming interface (API) provides C++ classes and methods as well as C structures and functions for application programmers.

4.2 Working with Databases

The input data in IM can be either flat files or database tables. If you want to access DB2 tables, you only need the authorization to access that database and the permission to query the appropriate tables. If you want to access tables in Oracle, Sybase, Informix, and/or SAS, or any other open database connectivity (ODBC) compliant datasource, you will need to install IBM's DataJoiner as a middleware product. For more information on how to access 'non-IBM' datasources for data mining, refer to *'Mining Relational and Nonrelational Data with IBM Intelligent Miner for Data, Using Oracle, SPSS, and SAS as Sample Data Source, SG24-5278.*

To access database management systems through the IM server, you do not need to install any database software on the client system. By default, DB2 will be accessed from the Intelligent Miner server, unless the environment variable `IDM_CLI_USED` is set in the client.

If, for example, you use the IM AIX client, and the `IDM_CLI_USED` is set in the client, then listing of table schemas and tables on the Intelligent Miner client GUI is done through DB2 CLI calls. To be able to do this, the DB2 Client Application Enabler (CAE) must be installed on the IM client system. For the Korn Shell on AIX, the command to set this environment variable would be `'export IDM_CLI_USED'`.

4.3 The User Interface

The IM user interface is simple and intuitive, and provides consistency across all operations. The interface's state-of-the-art GUI facilities include online help, task guides, and a graphical representation of the mining operations and its functions.

The main window of the IM GUI is divided into three areas (see Figure 24):

- **Mining base container:** A tree view of object folders containing knowledge discovery tools.
- **Contents container:** An area for customized objects.
- **Workarea:** An area where customized objects from the contents container can be imported and assigned to a single folder (Mining base).

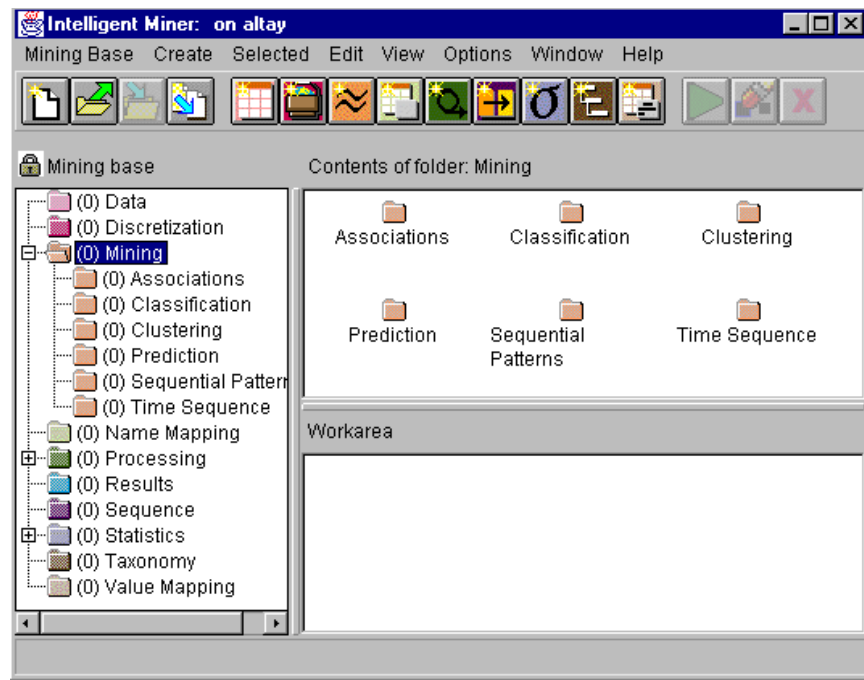


Figure 24. Intelligent Miner Main Window

4.4 Data Preparation Functions

Once the desired input data has been selected, it is usually necessary to perform certain transformations on the data. IM provides a wide range of data preparation functions that help to quickly transform the data to be analyzed. The data preparation functions of IM are:

- **Aggregate values** - to aggregate values of existing fields, for example, monthly salary to annual salary;
- **Calculate values** - to create new fields with the result of a calculation of existing fields;
- **Clean up data sources** - to delete database tables or views from input or output data;
- **Convert to lower or upper case** - to convert one or more fields in the input data;
- **Copy records to file** - to copy records from a database table or view to a flat file (you can also sort by the field you specify);

- **Discard records with missing values** - to remove input data records containing a missing (NULL) value in any of the fields you specify;
- **Discretize into quantiles** - to assign input data records to the number of quantiles you specify;
- **Discretize using ranges** - to assign input data records by splitting the value range of a continuous field into intervals, and then mapping each interval to a discrete value;
- **Encode missing values** - to encode missing values in the input data by specifying one or more fields to search for missing values;
- **Encode nonvalid values** - to encode values not found in the first input data if they do not match valid values from the second input data;
- **Filter fields** - to filter the input data fields to get only the fields you specify;
- **Filter records** - to filter the input data records to get only the records you specify;
- **Filter records using a value set** - to compare fields values in a first input data with values in a value set specified for a second input data;
- **Get random sample** - to reduce an input data to a smaller sample by specifying the size of the sample as a percentage of the input data;
- **Group records** - to summarize groups of records into a single record that contains aggregated values of the group;
- **Join data sources** - to join two databases tables or views based on one or more pairs of join fields from the input data;
- **Map values** - to map values found in the first input data to values found in the second input data;
- **Pivot fields to records** - to split each record of the input data into multiple records;
- **Run SQL statements** - to submit SQL statements.

Data preparation functions are performed through the GUI, reducing the time and complexity of data mining operations. You can transform variables, impute missing values, and create new fields through the touch of a button. This automation of the most typical data preparation tasks is aimed at improving your productivity by eliminating the need for programming specialized routines.

4.5 Statistical Functions

After transforming the data, we analyze it with statistical functions. Statistical functions facilitate the analysis and the preparation of data, as well as providing forecasting capabilities. For example, you can apply statistical functions like regression to understand hidden relationships in the data, or use factor analysis to reduce the number of input variables. Statistical functions included are:

- **Factor analysis** - discovers the relationships among many variables in terms of a few underlying, but unobservable, quantities called factors
- **Linear regression** - used to determine the best linear relationship between the dependent variable and one or more independent variables
- **Principal component analysis** - used to rotate a coordinate system so that the axes better match the data distribution. The data can be now described with fewer dimensions (axes) than before
- **Univariate curve fitting** - finds a mathematical function that closely describes the distribution of your data
- **Univariate and bivariate statistics** - descriptive statistics, especially means, variances, medians, quantiles, and so on

4.6 Mining Functions

All mining functions can be customized using two levels of expertise. Users who are not experts can accept the defaults and suppress advanced settings. However, experienced users who want to fine-tune their application have the ability to customize all settings according to their requirements.

You can also define the mode in which your data mining model will be performed. Possible modes are:

- | | |
|-------------------------|--|
| Training mode | In training mode, a mining function builds a model based on the selected input data. |
| Test mode | In test mode, a mining function uses new data with known results to verify that the model created in training mode produces adequate results. |
| Application mode | In application mode, a mining function uses a model created in training mode to predict the specified fields for every record in the new input data. |

You can also use data mining functions to analyze or prepare the data for a further mining run.

Based on IBM research, validated through real-world applications, IM has incorporated a number of data mining algorithms as the critical suite to address a wide range of business problems.

The algorithms are categorized as follows:

- Associations
- Sequential patterns
- Clustering
- Classification
- Prediction
- Similar time sequence

The following sections describe these algorithms in more detail.

4.6.1 Associations

The association algorithm, developed at the IBM Almaden Research Center in San Jose, CA, for example, compares lists of customers purchases with each other. The algorithm looks for patterns such as whether, when you buy paint, you also buy paint brushes. More specifically, it assigns probabilities; for example, if you buy paint, there is a 20% chance that you will buy a paint brush. This is a counting algorithm. In this example, it would locate all transactions that contained paint and count the occurrences of paint brushes.

The advantage of this approach is that it compares all possible associations. It also finds multiple associations, for example, if you buy paint and paint brushes, there is a 40% chance you will also buy paint thinner. When the algorithm runs, it creates hundreds or thousands of such rules. The user can select a subset of rules that have either higher confidence levels (a high likelihood of B given A) or support levels (the percent of transactions in the database that follow the rule). It is up to the user to read the rules and decide if the rules are:

- **Chance correlations** (for example, paint and hair rollers were on sale the same day and therefore were correlated by chance)
- **Known correlations** (for example, the paint and paint brush correlation is something that would have been known)
- **Unknown but trivial correlations** (for example, red gloss paint and red non-gloss paint correlation may be something unknown, and is unimportant to know)
- **Unknown and important correlations** (for example, paint and basketballs; this may be something previously unknown and very useful in both organization of advertising, and product placement within the store)

Association discovery is used in, for example, market basket analysis, item placement planning, promotional sales planning, and so forth.

4.6.2 Sequential Patterns

The rule generation method is a variation of the association technique. However, instead of looking at 10,000 purchases, the algorithm looks at 10,000 sets of purchases. These sets are lists of purchases from a sequence of shopping trips by a single customer. For example one set of lists might be the Jones family purchases: canteens and tent stakes in January, knapsacks and video tapes in February, and a sleeping bag in March. Here the sequential association algorithm looks at all sets of lists and returns rules such as: if canteens are purchased from the January catalog, there is a 30% chance that sleeping bags will be bought from the March catalog.

Sequential pattern detection can be used to discover associations over time. This is especially useful for direct marketers to design special targeted advertising supplements, such as credit card statement inserts, etc.

4.6.3 Clustering

Clustering is used to segment a database into subsets, the clusters, with the members of each cluster having similar properties. You can perform clustering by using either the demographic or the neural network algorithm, depending on the type of the input data set.

For example, each customer type has a characteristic profile of what they like to buy, what they dislike, and to what they are indifferent. Records of purchases can be tagged with a person's name to discover what their purchase patterns are, which reflects their likes and dislikes. When we start the clustering, we have no preconceived notions, that is when we look for patterns, it is a discovery process. We find whatever patterns exist in the data. The algorithm to do this type of discovery is the neural network. Looking at purchase histories is similar to looking through a list of people's names and what they purchased. Purchases are looked at over a period of 6 months to 2 years, depending on the type of merchandise. These purchase histories are drawn from credit card transaction data, preferred customer data, frequent shopper club data, lay-away plan data, or any other purchase data with a customer name associated with it. The methodology and technology to do this has been developed at IBM and is called *neural segmentation*. Typically, we would identify 10-20 different types of patterns. These segments are the *first output* of the neural analysis.

Once a segment is identified, the customer names can be identified. This examination of the demographics of each segment is the *second output* of the analysis.

With clustering, we also look at the “product linkage”. For instance, there might be a group of people who buy men's suits, men's ties, women's high-fashion shoes, and chocolate. On the other hand, they do not buy baby clothes, housewares, and greeting cards, and are middle-of-the-road in sports equipment. This tells us that we might be able to attract more customers for a men's suit sale, if we also have chocolate on sale. Or, better yet, if we give away a box of chocolates with the purchase of a suit, we might sell more suits. This information is the *third output* of the analysis.

In summary, we divide a population into segments by looking at how they behave (which, in our example, means what they bought). We then look at the segments and see who they are by looking at demographics. Then we examine product linkages. The result of these three steps is that, for example, a retailer gains a better understanding of the many types of customers that buy in the store. The retailer can then tailor a marketing strategy that appeals to each of these types of customers, offering products they like through advertising media that will reach them. This approach hits a midpoint between mass marketing and individual marketing. It is much more effective than mass marketing, yet much less expensive than individual marketing.

Clustering is well used in the area of cross-marketing, cross-selling, customizing marketing plans for different customer types, deciding on media approach, understanding shopping goals, and so forth.

4.6.4 Classification

A major database marketing activity is to plan a promotional strategy that takes advantage of the information known about the customers. Each customer is assigned to a class representing his likelihood to respond to a promotion.

Classification is the process of automatically creating a model of classes from a set of records. The induced model consists of patterns, essentially generalizations over the records, that are useful for distinguishing the classes. Once a model is induced, it can be used to automatically predict the class of other unclassified records. You can use tree induction (modified CART regression tree) or neural networks algorithms (back propagation) to compute the classes. Like neural networks, trees develop arbitrary accuracy

and use validation data sets to avoid spurious detail. Unlike neural networks, trees are easy to understand and modify.

4.6.5 Prediction

As in classification, the goal is to build a data model as a generalization of the records. However, the difference is that the target is not a class membership but a continuous value, or ranking. With prediction, you can also use the neural networks algorithm and use the radial basis function (RBF) algorithm.

4.6.6 Similar Time Sequences

The purpose of this process is to discover all occurrences of similar subsequences in a database of time sequences. Given a database of time sequences, the goal is to find sequences similar to a given one, or find all occurrences of similar sequences. The powerful alternatives afforded by multiple methods are enhanced by the fact that several of the methods are supported by more than one mining technique.

Multiple techniques are often used in combination to address a specific business problem.

4.7 Creating and Visualizing the Results

Information that has been created using statistical or mining functions can be saved for further analysis in the form of result objects. Figure 25 illustrates a result generated by the IM clustering function.

Result objects can be used in several ways:

- To visualize or access the results of a mining or statistical function
- To determine what resulting information you want to write to an output data object
- To be used as input data, when running a mining function in test mode to validate the predictive model representation by the result
- To be used as input data, when running a mining function in application mode to apply the model to new data
- To discuss the results with business users

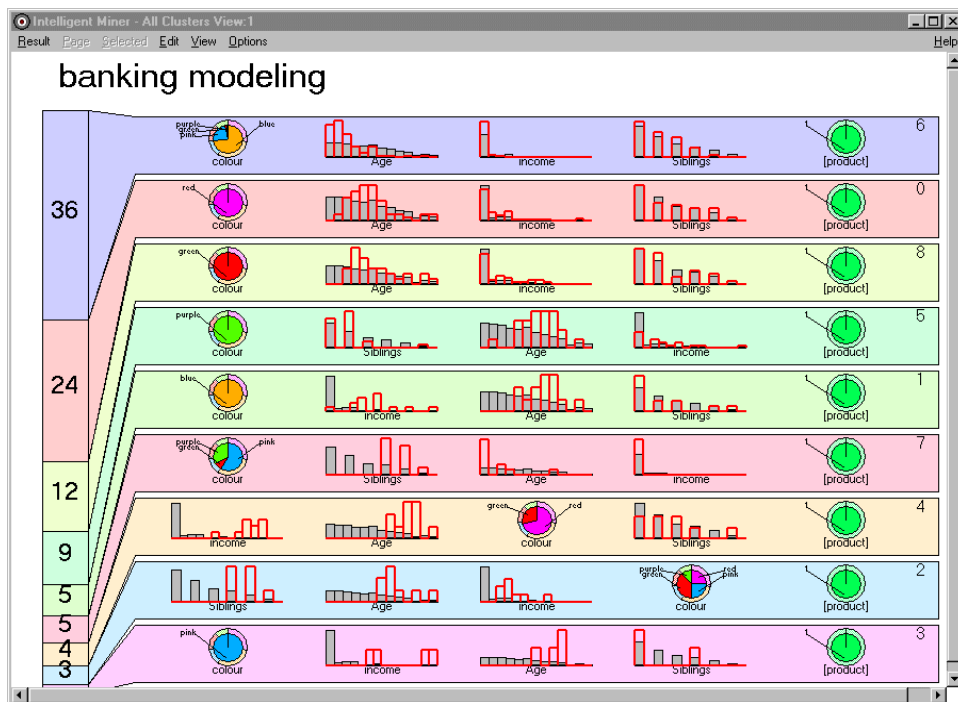


Figure 25. Sample Clustering Visualization

4.8 Creating Data Mining Sequences

IM provides a means to construct data mining operations as a sequential series of related data preparation, statistical, and mining functions. IM will execute the objects in the order in which they have been specified. A sequence can also contain other sequences. You can construct standard sequences that can be reused in similar data mining operations forming part of a more complex sequence.

In IM, sequence setting objects are created through the sequence task guide. You can select objects from the mining database and place them in the sequence work area setting in the order in which they are to run. In addition, you can specify whether the sequence should continue if any of the setting objects fails. Figure 26 illustrates a sequence settings window.

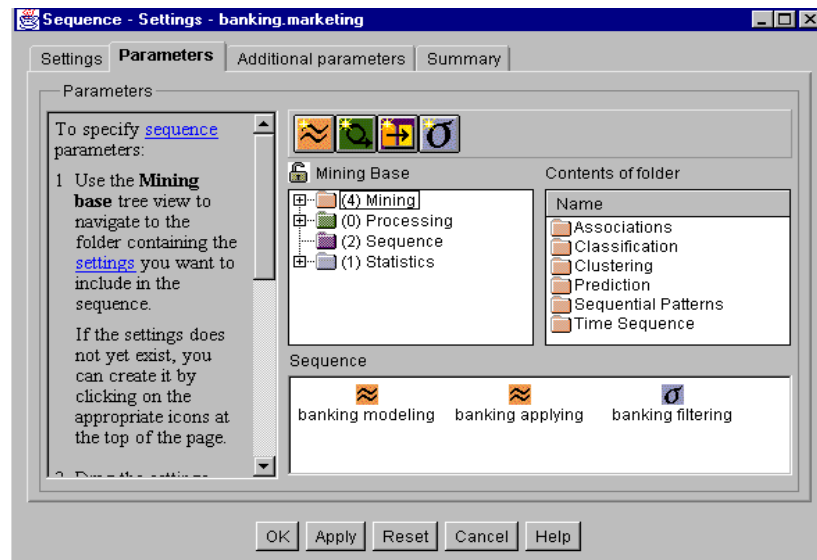


Figure 26. Sequence Settings Window

Chapter 5. Implementing IM in the ITSO Environment

In this chapter we describe the BI environment at the ITSO, San Jose center. Then we consider each technique provided by Intelligent Miner, and the way the source data from our environment was prepared to implement the different data mining techniques.

5.1 The ITSO BI Environment

In this section, we describe the network and BI systems setup of the ITSO, San Jose center, where we ran the project. We also describe the software and hardware that were used.

5.1.1 The Environment

The ITSO BI environment includes OS/390, VM and VSE host server systems, and AIX and Windows NT servers and workstations. We also included an AS/400 and Sun/Solaris server to cover the whole range of platforms on which Intelligent Miner (IM) might be installed.

The OS/390 and VSE systems are running as guest systems on different VM systems located at the ITSO, Poughkeepsie center. On the workstation side, we used two AIX systems and two Windows NT Server systems, one of them installed on a Netfinity server. The AS/400 is located at the ITSO, Rochester center, and we used a Sun/Solaris server running at the IBM Santa Teresa Laboratories.

Figure 27 shows an overview of the systems used in the ITSO BI Center.

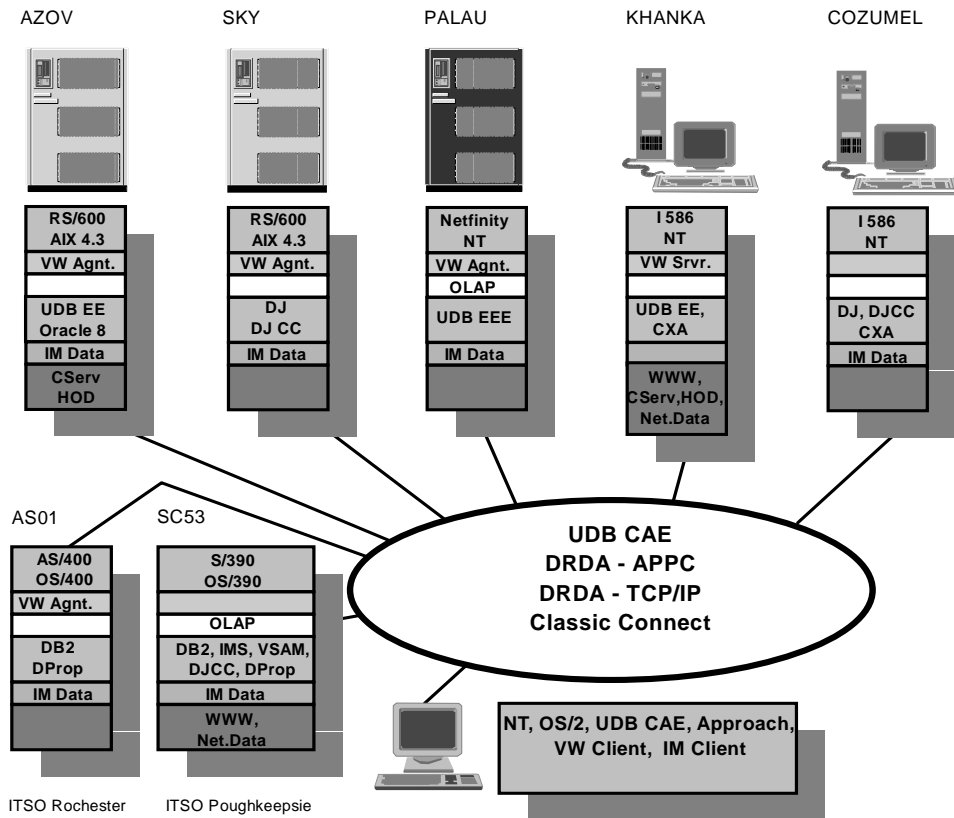


Figure 27. The ITSO BI Environment

The systems are named by their TCP/IP hostname (workstations) or their VTAM SSCP-name (hosts).

5.1.1.1 AIX Data Warehouse, Communication, and OLAP Server

AZOV is an AIX 4.3 system customized as a Data Warehouse and OLAP server. IBM Communication Server is installed for APPC communication, and DB2 UDB Enterprise Edition provides database server functionality. Intelligent Miner is installed here for direct access to the data warehouse database. This system also hosts an Oracle 8.0.4 database server.

5.1.1.2 AIX DataJoiner Classic Connect Gateway

SKY is the second AIX system and is used for non-IBM database access through DataJoiner. DataJoiner Classic Connect is installed for access to the nonrelational data on OS/390. IM is installed to access non-IBM databases

directly through the DataJoiner databases. SKY provides no local databases, but acts as a gateway to other systems through DataJoiner and Classic Connect.

5.1.1.3 WindowsNT Data Warehouse Server and OLAP Server

PALAU is a Windows NT Server installed on Netfinity. It is customized as a data warehouse and OLAP server supplied with a DB2 UDB Extended Enterprise Edition for multiprocessor functionality. It also runs Intelligent Miner for direct access to the warehouse database on this Windows NT system.

5.1.1.4 WindowsNT Visual Warehouse Server and Gateway

KHANKA is the Visual Warehouse server that manages all data flows in this environment. The IBM Communication Server is applied for APPC communication to the hosts and the Cross Access ODBC driver provides access to the nonrelational data on the VSE system. DB2 UDB Enterprise Edition provides database server functionality. Web server, Net.Data, and Host On Demand complete the server functions of this system.

5.1.1.5 DataJoiner Classic Connect and CrossAccess Gateway

COZUMEL is the third Windows NT Server and the second DataJoiner system. With the implementation of the Cross Access ODBC driver, this system provides access to the nonrelational data residing on the VSE system. The Intelligent Miner is implemented here for access to these datasources.

5.1.1.6 SUN Solaris Database Server

POSSUM is a SUN Ultra SPARC system running DB2 UDB for Solaris. The system provides database and mining services. POSSUM itself is not part of the ITSO BI environment, but has been added to be able to document the IM installation and configuration on this platform.

5.1.1.7 AS/400 Database Server

AS01 is an AS/400 running OS/400 with the native database services. In addition it provides mining services for direct database access. AS01 provides TCP/IP access to its local databases.

5.1.1.8 OS/390 Database Server

SC53 is the OS/390 system providing TCP/IP and APPC communication and VSAM, IMS and DB2 data sources. It is installed as a guest on WTSCPLX1 VM system in Poughkeepsie. Intelligent Miner is installed here to get direct access to the local databases.

5.1.2 Networking Configuration

All systems are connected to a token ring network. The host systems are connected through Open Systems Adapter (OSA), and the main domain for all host systems is SCG20.

This environment provides multiple layers of distributed data access from single client access to relational and nonrelational data through data replication and data warehousing, up to data mining and online analytical processing. The communication setup is based on TCP/IP and APPC. Whenever possible, TCP/IP was used to define the communication to the host systems, but some host applications still require APPC. To provide APPC for the workstations, two systems (AZOV and KHANKA) have been supplied with the IBM Communication Server to act as an APPC gateway.

5.1.3 The Data

The largest DB2 table has more than 3,500,000 rows with an average length of about 80 bytes. Together, all DB2 tables occupy about 700 MB of disk space. The VSAM KSDS cluster has more than 740,000 636 bytes records, which is more than 450 MB of data. The file occupies more than 727MB of disk space.

The next sections provide more detailed descriptions about the available data. We use three different kinds of data sources: DB2, DL/I and VSAM. All data is placed in DB2 tables on the workstation side. The VSAM and DL/I data is referenced on the host side as a relational table through CrossAccess. Therefore, the data descriptions in the following sections do not describe the data sources themselves, but the usage of the data in our environment.

5.1.3.1 Sales Information

The sales information is contained in a table called "SELL_DAYS", which holds data from the stores scanner cashiers. You find a summary of the columns in Table 3.

Table 3. Sales Information

Column Name	Description
DATE	date the article was sold
TRANS_ID	transaction ID
BASARTNO	basic article number
LOCATION	store number
COMPANY	company identification

Column Name	Description
DEPTNO	department number, only if several in one store
PIECES	number of sold units
IN_PRC	purchase price times number of units
OUT_PRC	retail price including tax
TAX	sales tax
NO_CUST	number of customers
WGRNO	product grouping number
SUPPLNO	supplier identification
TRANSFER_ DATE	transfer date of the data from the cashier
PROCESS_ DATE	processing date in database

5.1.3.2 Article Information

The article information tables contain all data about the articles, their hierarchical structure, and their relationship to suppliers and supplier data. These tables are mostly sourced by DB2 tables, except for the supplier data, which is derived from VSAM files.

Table 4 lists the names and main content of these tables.

Table 4. Article Information

Table Name	Description
BASART	article base data
ARTTXT	more article description
STRUCART	article structuring information
STRARTDAT	more article structuring data
DEPOT	relationship to product line, brand and supplier
SUPPLART	relationship to suppliers and order information

5.1.3.3 Organization Information

The information about the organization consists of data from all data sources, enriched with local data on the data warehouse servers. The information about the business lines is taken from a new DB2 table, the company data is derived from the DL/I database, and the information about the stores comes from the VSAM data set. Table 5 summarizes the available tables.

Table 5. Article Information

Table Name	Description
STORES	stores information from VSAM
COMPANY_DB	company information from DL/I
BUSINESS_LINES	business line definitions

5.1.4 Table Relationships

Figure 28 shows how the previously mentioned tables are related to each other.

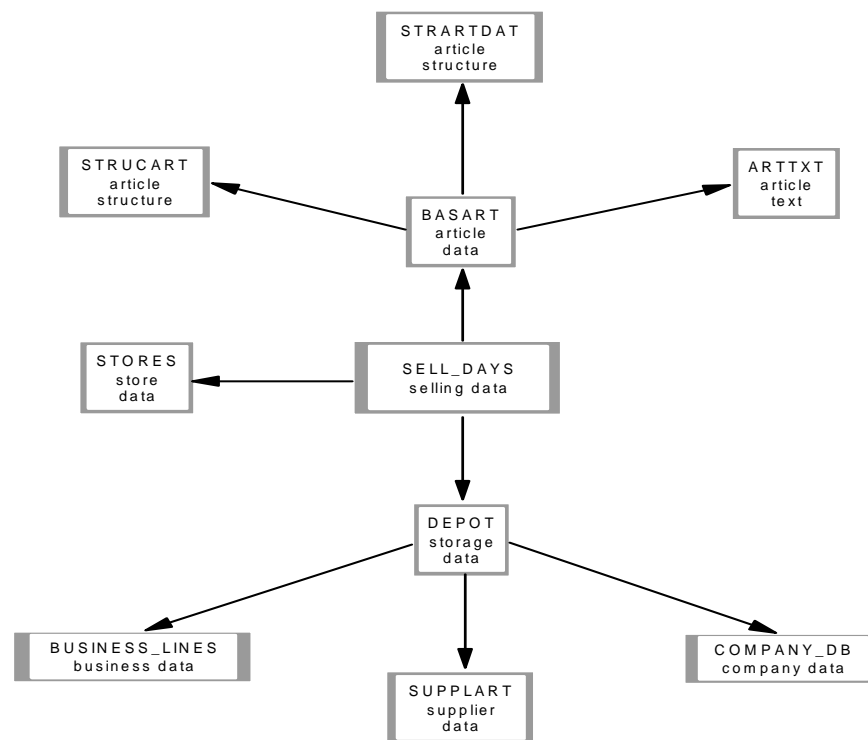


Figure 28. Table Relationships

5.2 Implementing Data Mining Techniques

The following sections cover the implementation of the different data mining techniques mentioned in 4.6, “Mining Functions” on page 56. The sections are structured according to Figure 2 on page 5, so we first describe an example of a business issue. The solution to this issue will need information based on the data we have available. The data is then mined for information that can then support the decision to be made.

5.2.1 Associations

Usually in a retail company, the store manager must decide what products should be on sale and how to organize the products within the store.

To make such decisions, the association technique can help the manager by finding associations among the products in the store so that he knows which products sell together.

In order to run the association technique, the transactional data source must be transformed as shown in Figure 29. The input table has at least two fields: one is the transaction identification and the other is the product. The STRUCART table can be used additionally, if the manager wants to determine associations on a higher level, such as product groups or departments.

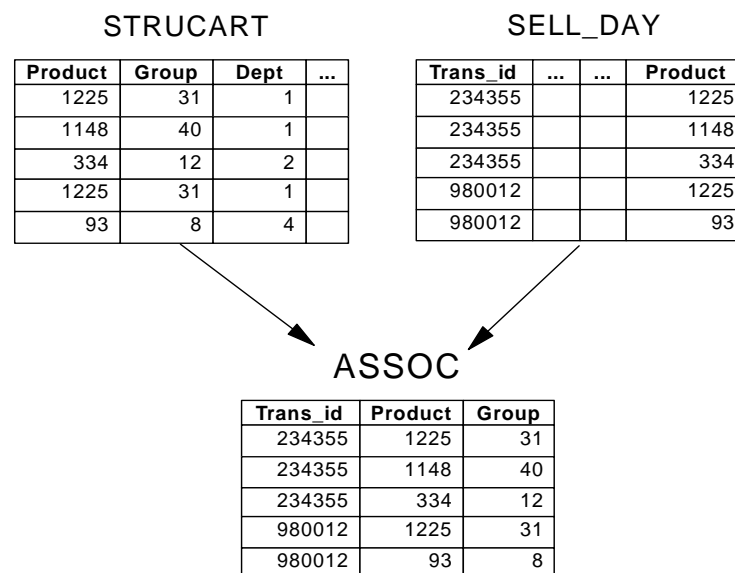


Figure 29. Data Transformation for an Associations Model

With the association table generated, you can choose the *Trans_id* and the *Product* column to run the model. Or, if you want to aggregate on more than the product level, you can choose the *Trans_id* and the *Group* columns to understand the associations among groups of products. The result of this mining run is a summary of the events and the set of rules that associates the products (items):

```

Number of Transactions = 267
Number of Items = 80 (48 large)
Items per Transaction: Maximum = 26, Average = 3.73034
Support: Minimum = 7 = 2.800%, Maximum = 267 = 100.000%
Confidence: Minimum = 30.0%, Maximum = 100.0%
----- Rules -----
Group  Support  Conf  Tp  Lift  Body          ==> Head
6      3.371    42.9  +   7.6  [Brandy]       ==> [Orange juice]
6      3.371    60.0  +   7.6  [Orange juice] ==> [Brandy]

7      2.996    80.0  +   8.5  [Gouda Cheese] ==> [Crackers]
7      2.996    32.0  +   8.5  [Crackers]     ==> [Gouda Cheese]

8      7.491    43.5  +   2.5  [Toy car]      ==> [Cream]
8      7.491    42.5  +   2.5  [Cream]        ==> [Toy car]

```

You can interpret these rules based on the support, the confidence, and the lift:

Support Support represents the percentage of transactions in which the mentioned items were sold together. It shows whether this rule is relevant in the total number of transactions that have happened.

Confidence Confidence gives the percentages of those transactions that contain the first item and also contain the second item. You can interpret this number as the probability that the second item will be present in a transaction if the first one is present (*conditional* probability).

Lift Lift is defined as the actual confidence factor divided by the expected confidence. It is a measure for the deviation of the prediction of a rule from the expected. If, for example, everybody buys bubble gum at the cash register, there will be many rules that associate all kinds of products with bubble gum, and therefore a high support and confidence may not mean anything. In our example, lift tells how much higher the confidence level is (compared to the total percentage of transactions) when the second item is present. A high lift means that the connection between the items is stronger. You use lift to qualify the confidence as being relevant or not.

Usually a customer buys more than one product from a company. If you want to study the behavior of the customer within one company, you should use the product groups. If you want to study just the behavior of the customer for a certain product of this company, then you use product codes.

By using these rules and associations, the manager can decide which products he should put on sale. If the first product is put on sale, the other has a high chance of selling along with it.

5.2.2 Sequential Patterns

If the marketing analyst knows the sequence consumption behavior of customers, he can plan for a direct mail campaign to customers to repeat this sequence.

The sequential patterns technique is used to understand the purchasing sequence of products that define the customer's behavior. For example, some customers buy coffee, sugar, and milk and then, in another transaction, buy meat and potatoes.

For the sequential patterns technique, the input data must have a transaction identifier, an item, and a customer number. Figure 30 shows the data transformation necessary, based on our sample input tables, to run the sequential patterns technique.

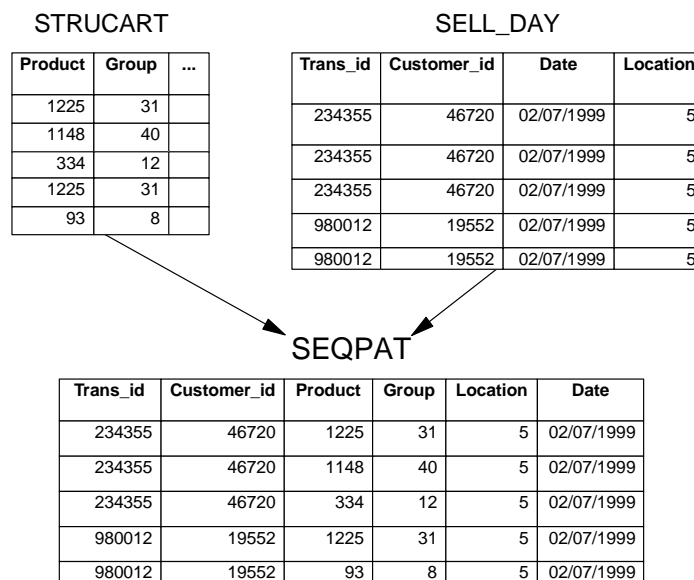


Figure 30. Data Transformation for a Sequential Patterns Model

The result of this model is a set of rules in a sequence and the summary of the events.

```
Number of Transactions = 267
Number of Items = 109 (29 large)
Items per Transaction: Maximum = 42, Average = 3.72659
Support: Minimum = 133 = 50.000%, Maximum = 267 = 100.000%
----- Sequences -----
1367 sequences:
Group  Support  Sequence
2      87.50    << [Misc. Toys] >> << [Baby products] >>
2      87.50    << [Baby products] >> << [Beers] >>
2      87.50    << [Spirits] >> << [Baby products] >>
2      87.50    << [Baby products] >> << [Baby products] >>
2      87.50    << [Beers] >> << [Baby products] >>
2      83.33    << [Baby products] >> << [Car accessories] >>
```

The interpretation of these sample sequences is that, with an 87.50% value for support, customers buy miscellaneous toys, and later buy baby products. The definition of support, explained previously for the associations technique, shows that the higher the support, the more frequently this sequence occurred.

Now the marketing analyst can target a direct mail campaign, offering certain baby products to those customers that already bought toys.

5.2.3 Clustering

Assume the marketing analyst needs to select customers, or prospects, to send direct mail, in order to create new product sales based on existing sales information. If the historical and demographic data of the customers is available in the database, the analyst can run a clustering technique to create customer profiles.

Clustering is used to group records that are similar, but to separate the ones that are very different. That means that the number of clusters is related to the number of different patterns in your data, and each cluster only contains records that are as similar as possible.

The example in Figure 31 shows the data preparation we used for clustering. In this example we created variables that represent time-related patterns for products, such as the amount of items sold in the first week of each month. This data is supplemented by additional details about the product.

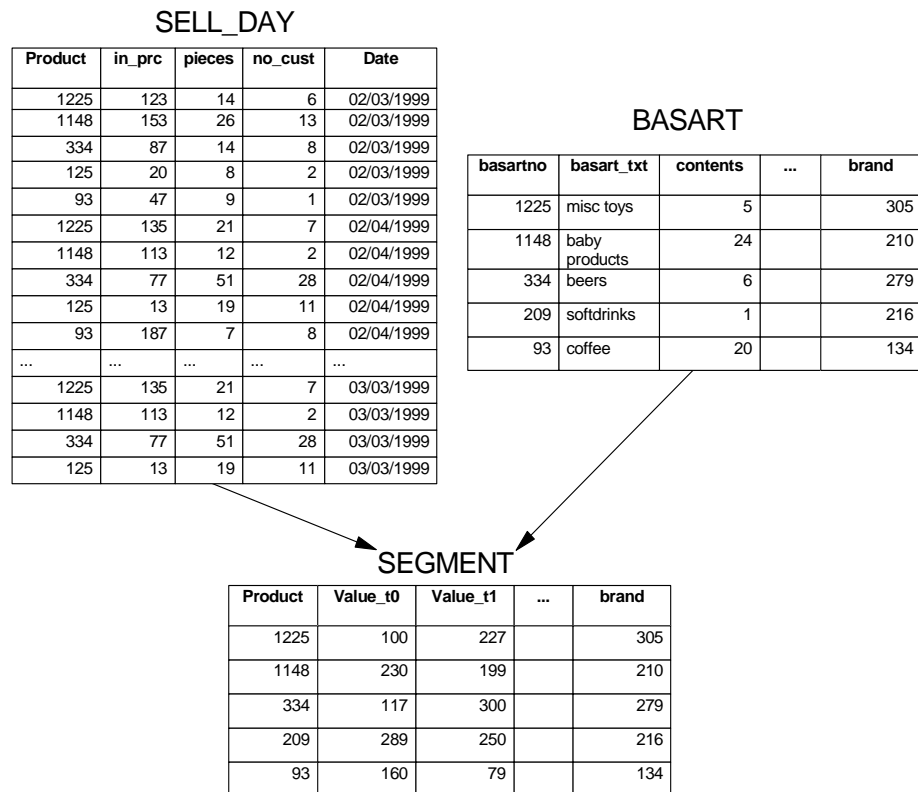


Figure 31. Data Transformation for a Clustering Model

The output of this technique is the cluster visualization of IM, as shown in Figure 32. We can then add the *cluster number* and the *score* variable to each record in the SEGMENT table. The score represents how well the record fits in the cluster to which it was assigned.

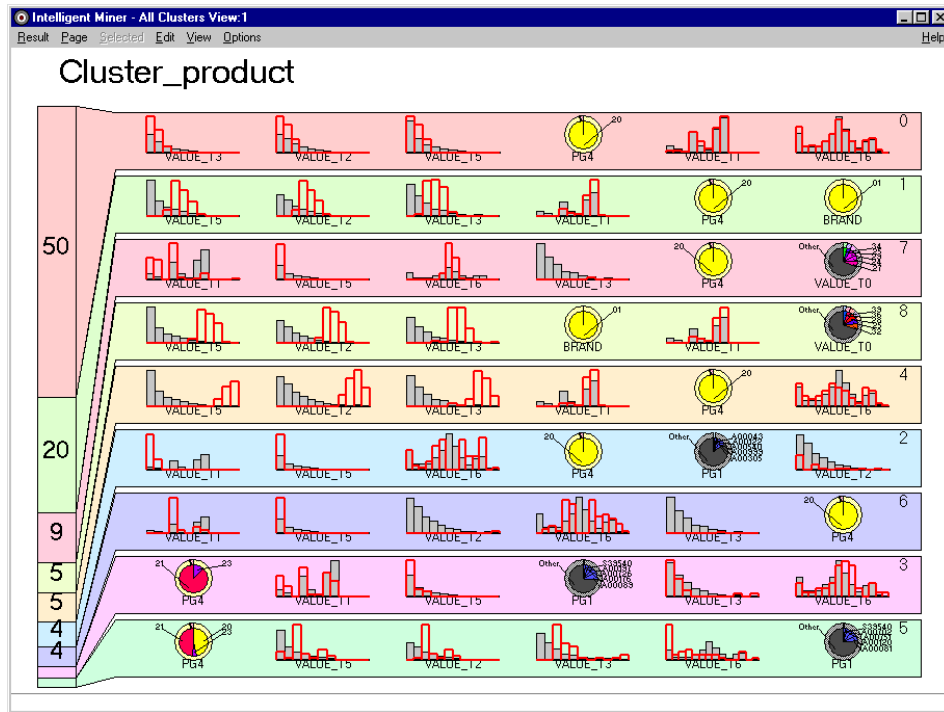


Figure 32. Clustering Result

With the clustering information, the marketing analyst can select which cluster, or combination of clusters, he wishes to target with his campaign. He uses the visualization to create a high-level description of the clusters. The result data that was appended to the original records can be used in any query or OLAP tool to assist the analyst later in making selections.

5.2.4 Classification

The clustering can help provide insight into a product set, but for a business user, it is even better to have rules stating why certain products are classified as such.

When you use classification to predict the assigned clusters, you get rules that allow you to evaluate new products by their sales patterns. In the next example, we analyzed the clusters from the previous section and assigned each cluster a class of “Good” or “Bad”. We then added this evaluation to the product records to try to find rules that predict whether a new product would be good or bad.

We used the SEGMENT table as a basis for our classification model, as shown in Figure 33.

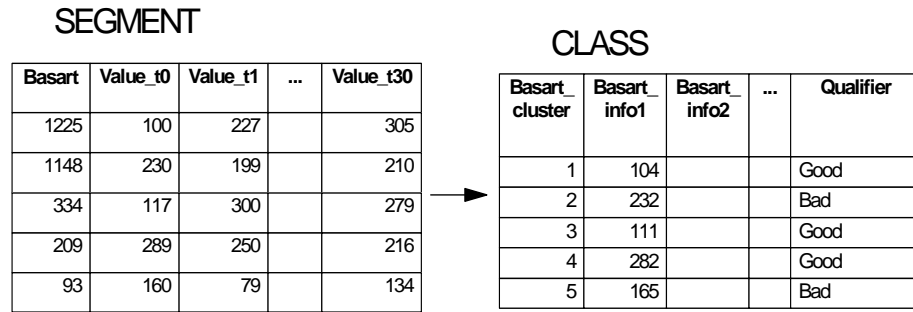


Figure 33. Data Transformation for a Classification Table

The quality results of the classification run are shown below.

Clf Training				
Number of classes = 2				
Errors = 33 (13.75%)				
Confusion Matrix				
Predicted Class	Good		Bad	
Good	84	19	total = 103	
Bad	14	123	total = 137	
	98	142	total = 240	

The result shows that the model classifies the products as profitable (Good) in almost 80% of the cases, and classifies them as non profitable (Bad) with more than 80% of efficiency (error rate = 13.75%).

The visualization output of classification is a *tree*, as shown in Figure 34, that represents the rules found. Each *node* in the tree represents a test, so for each *leaf* (end point) there is a sequence of tests that, together, form the rule about the records that are classified at the leaf.

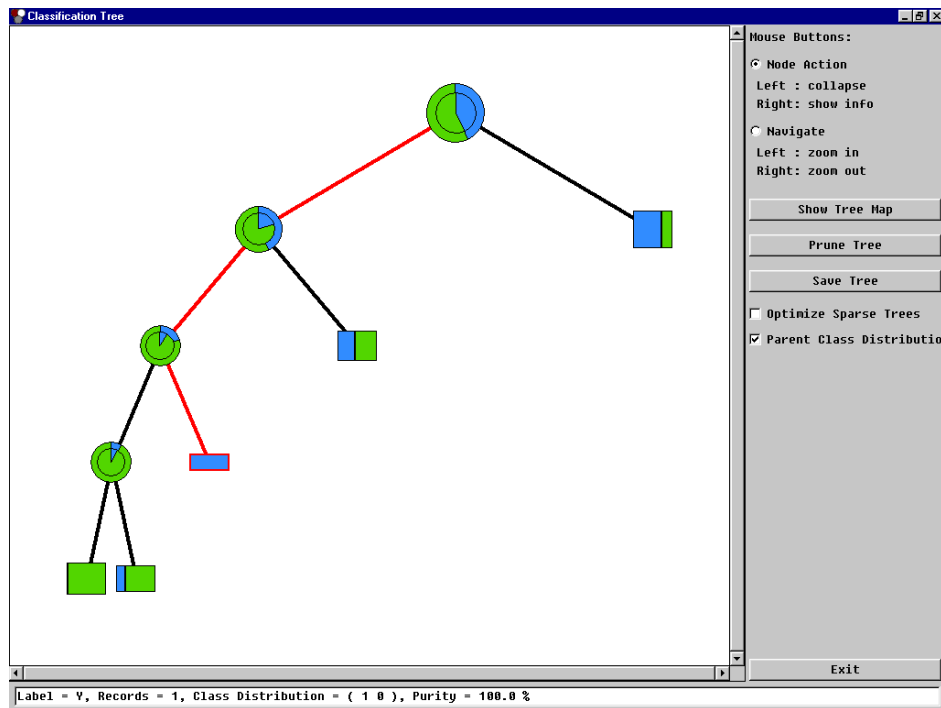


Figure 34. Classification Tree

The tree also provides information about the important factors that determine whether a product is good or bad, because those factors will end up higher in the tree.

Using the classification rules together with the associations that we found, we can optimize our product set. We might even be able to find products that cannot be profitable themselves, but have a high association with a profitable product. We would normally discontinue selling such "bad" products, but then we might also lose sales on the profitable product.

5.2.5 Prediction

The classification rules do not allow you to select, for example, the 100 most profitable products. Therefore, we assigned each product a value of 0 for "Bad" and a value of 1 for "Good", as shown in Figure 35.

SEGMENT					PREDICT				
Basart	Value_t0	Value_t1	...	Value_t30	Basart_cluster	Value_t0	Value_t1	...	Qualifier
1225	100	227		305	1	104	297		1
1148	230	199		210	2	232	145		0
334	117	300		279	3	111	450		1
209	289	250		216	4	282	257		1
93	160	79		134	5	165	77		0

Figure 35. Data Transformation for a Prediction Model

Prediction will assign each unknown product a number between 0 and 1, depending on how close the product is to an existing good or bad product. This number will allow you to select the 100 products with the highest numbers for further analysis.

Figure 36 shows the results of a prediction run.

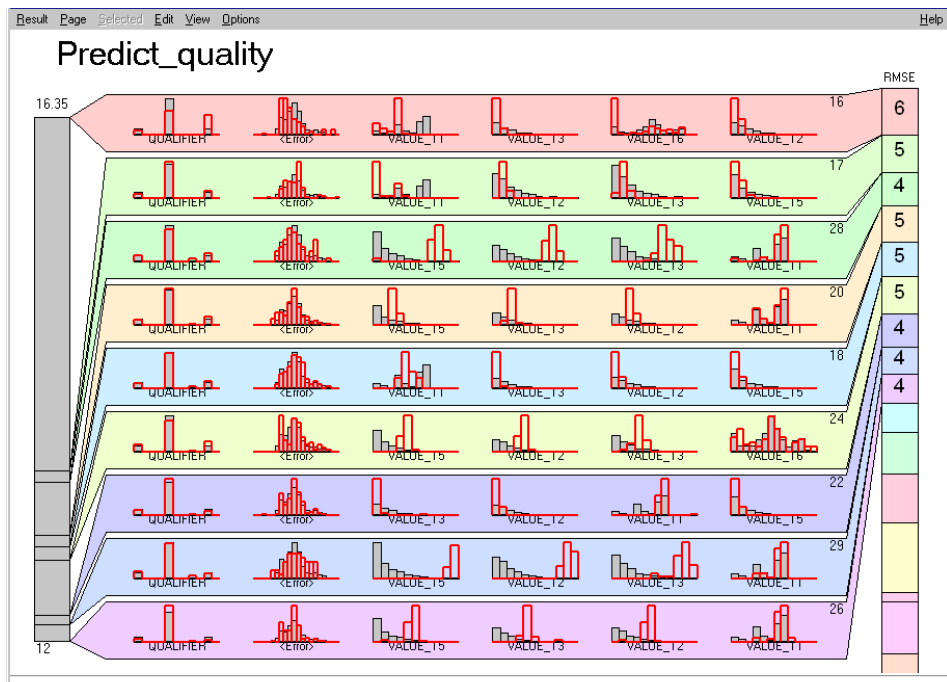


Figure 36. Prediction Result

The results look similar to the one shown for clustering, except that the first value shows the prediction, and the second the accuracy for the group that is predicted. You can use this chart to assess the quality of the prediction. The actual predicted value can be added to each record as a score. You can then use this value to analyze the products using traditional queries or OLAP.

5.2.6 Similar Time Sequences

If you are looking for time-related effects in your data, such as seasonality, normally you take the total sales and compare them to external market factors. It would be of much more help to see which products have similar sales trends, together with a number of external time-related factors, such as competitor's sales, economy indexes, or even the weather conditions.

Similar time sequences allow you to do just that. We converted the sales data to a vertical format, showing the sales numbers for each date and product (see Figure 37).

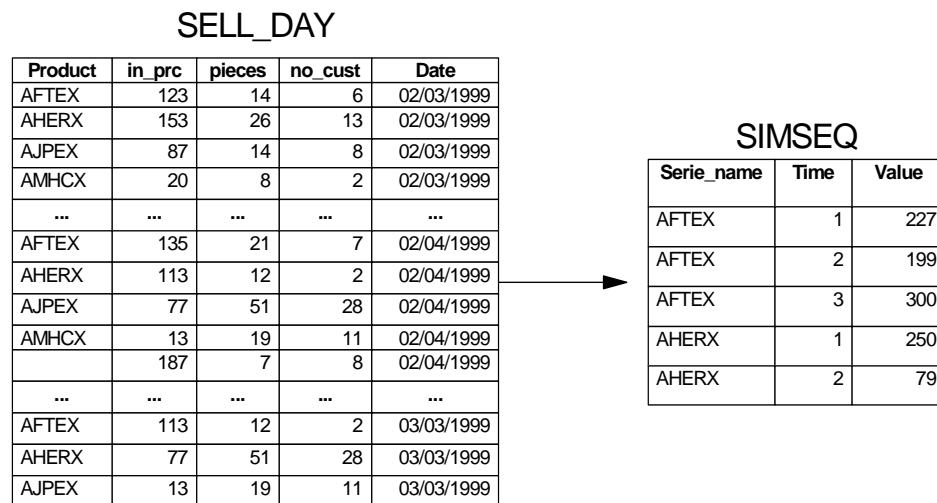


Figure 37. Data Transformation for a Similar Time Sequences Model

The result of this data mining technique is shown in Figure 38. The graphs show the matching fractions for products AFTEX and AHERX.

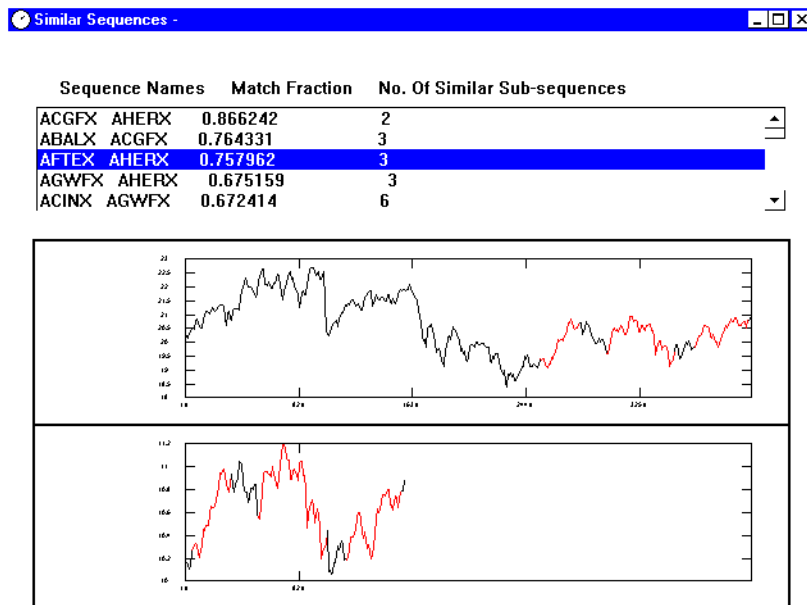


Figure 38. Similar Time Sequences Result

This result shows that the two products can be matched over about 75% of their pattern. The actual matches are highlighted in the display.

The analyst could look for unexpected matches between sequences, which could provide information about the influence of sales between products, or the importance of external factors.

Part 2. Installation and Configuration of Intelligent Miner for Data

Chapter 6. Implementation on Windows NT

This chapter describes the steps required to install Intelligent Miner for Data on Microsoft Windows NT. We installed the IM server on our systems, PALAU and COZUMEL, and the clients on several other systems. The installation below in Figure 39 describes both server and client installation.

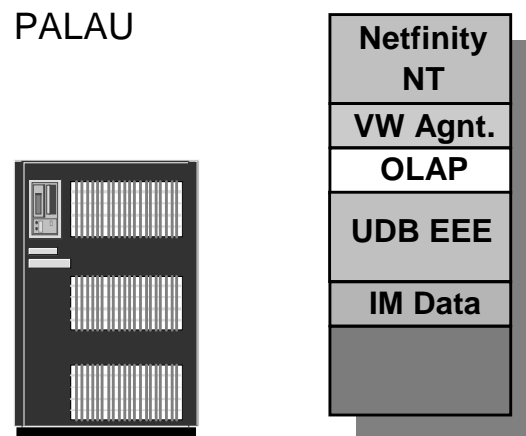


Figure 39. Intelligent Miner for Data on Windows NT

6.1 Prerequisites

This section lists the prerequisites necessary for the installation. It also explains how to verify whether your system is ready to run the IM server and client software.

6.1.1 Hardware Requirements

In general, there are three parameters you must check to insure the ability to install and run IM:

- Processor
- Memory
- Disk Space

The following sections describe the hardware requirements for both the IM server and IM client installation. We describe the actions required using Windows NT as a sample system.

6.1.1.1 Processor

For the IM server, you will need at least a fifth generation x86 system (for example, Pentium), running at 166 MHz. We recommend a Pentium running at 200 MHz or more.

The IM client should be a standard PC, for example, a Pentium 166.

If you do not know the type of processor used by your system, you can use a diagnostic tool in Windows NT, for example, the diagnostics tool located under **Programs --> Administrative Tools** in the Start Menu, to find out more about your system. Figure 40 shows an example of the **System** tab from the diagnostics utility. As mentioned before, the processor should be at least Family 5 at about 166 MHz. The processing speed listed in the window is an estimate.

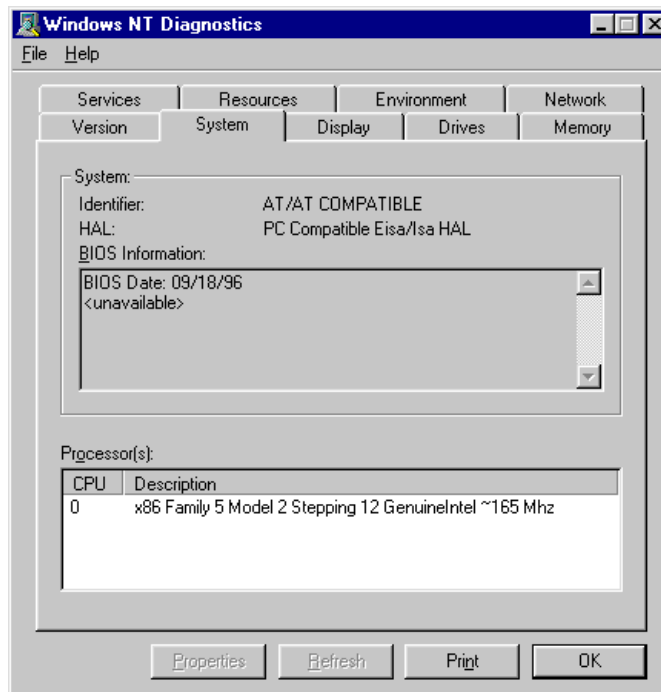


Figure 40. Windows NT System Configuration

6.1.1.2 Memory

To achieve appropriate performance, the IM server system should have at least 128 MB of RAM, however, 64 MB might be sufficient for the IM client system. For both server and client, the larger the amount of data to be mined, the more memory you will require.

If you do not know the amount of memory installed in the system, you can click the **Memory** tab of the diagnostics tool shown in Figure 40 on page 84 to find the amount of RAM installed. It is listed under **Physical Memory (K)** as **Total**.

You can also click **Display** to find the display settings for your system. The client software needs at least an 800 x 600 screen with 65536 display colors to display results. A resolution of 1024 x 768 is recommended. There are no display requirements for the server.

6.1.1.3 Disk Space

Table 6 lists the server storage requirements. These amounts further depend on the amount of data that you intend to process.

Table 6. Windows NT Server Storage Requirements

Storage type	Demonstration	Required	Recommended
Server disk	20 MB	40 MB	50 MB
Server & client disk	30 MB	60 MB	80 MB
Toolkit disk	40 MB	80 MB	100 MB

Table 7 lists the storage requirements for running a client system.

Table 7. Windows Client Storage Requirements

Storage type	Demonstration	Required	Recommended
Client disk	30 MB	35 MB	> 45 MB
Toolkit disk	40 MB	45 MB	> 55 MB
Client & Toolkit disk	55 MB	60 MB	> 70 MB

The amounts of free disk space on your system can be found by opening a window on **My Computer** and selecting **Details** from the **View** menu. The required disk space mentioned previously is based on the use of NTFS. For FAT partitions, additional storage is required, depending on the size of the partition and the amount of files stored. This is because larger FAT partitions generally use larger disk clusters. The cluster is the smallest unit of allocation, so this requires extra disk space, especially when large amounts of small files are involved.

6.1.2 Software Prerequisites

Before covering the prerequisite software, you should ensure that the versions of IM server and client that you are planning to install correspond. A

prerequisite for this is that the IM server has exactly the same version as the client. However, this is not checked during installation. For example, if you try to run a V2.1.2 client against a V2.1.3 server, you will receive an error message that prevents you from using the server.

The required dependent software for IM is related to the operating system and the IBM DB2 database software.

Table 8 lists the required and optional software for the server.

Table 8. Windows NT Server Software Prerequisites

Software	Version	Program number	Required
Microsoft Windows NT Server or Workstation	4.0 with SP3 applied	N/A	Yes
IBM DB2 for Windows NT	2.1.1	5622-664	One of these DB2 versions
IBM DB2 UDB	5.0	5648-A32	

Table 9 lists the required and optional software for the client.

Table 9. Windows Client Software Prerequisites

Software	Version	Required
Microsoft Windows 95	N/A	One of these operating systems.
Microsoft Windows NT Server or Workstation	4.0 with SP3 applied	
Web browser supporting frames		For online help

The following hints are directed to the system that is being installed as the IM server. The IM client system is straightforward and requires little verification work to be performed.

To check your Windows NT version, you can use the diagnostics utility again as shown in Figure 40 on page 84. You can also select **About** from the **Help** menu in any NT system tool. The result will look like the one shown Figure 41.

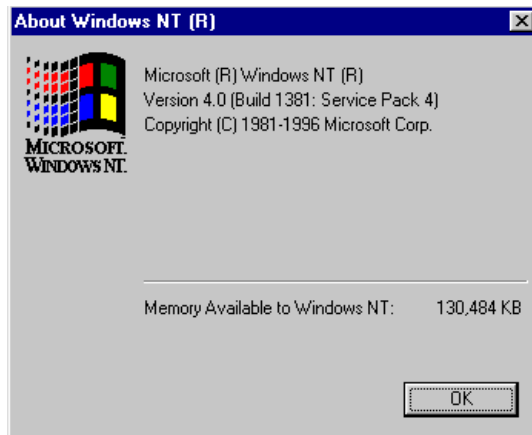


Figure 41. Windows NT version

You need at least Version 4 (Build 1381: Service Pack 3). This window also lists the amount of physical memory in your system.

If you will not use DB2 data stored locally on the IM server, you need to have DB2 CAE version 2.1.1 or higher installed. You can check your DB2 version by starting a command line processor from your Start Menu and selecting **Programs --> DB2 for Windows NT --> Command Window**. The output for DB2 UDB V5.2 is shown in Figure 42.

```

D:\SQLLIB\BIN>DB2.EXE
(<c> Copyright IBM Corporation 1993,1997
Command Line Processor for DB2 SDK 5.2.0)

You can issue database manager commands and SQL statements from the command
prompt. For example:
    db2 => connect to sample
    db2 => bind sample.bnd

For general help, type: ?.
For command help, type: ? command, where command can be
the first few keywords of a database manager command. For example:
    ? CATALOG DATABASE for help on the CATALOG DATABASE command
    ? CATALOG           for help on all of the CATALOG commands.

To exit db2 interactive mode, type QUIT at the command prompt. Outside
interactive mode, all commands must be prefixed with 'db2'.
To list the current command option settings, type LIST COMMAND OPTIONS.

For more detailed help, refer to the Online Reference Manual.

db2 => _
  
```

Figure 42. Windows NT Command Line Processor for DB2

Ensure that your DB2 library path is included in your path specification. You can check this by starting an NT Command Prompt and entering the **path** command. The output must contain the BIN directory under the directory in which you installed DB2, for example C:\SQLLIB\BIN. You can check this directory by starting a DB2 Command Window. It will have the BIN directory as current directory.

If this directory is not listed in your path, change the path by closing your command window and opening the NT **Control Panel** from your **Start** menu. Double-click **System** and select the **Environment** tab as shown in Figure 43.

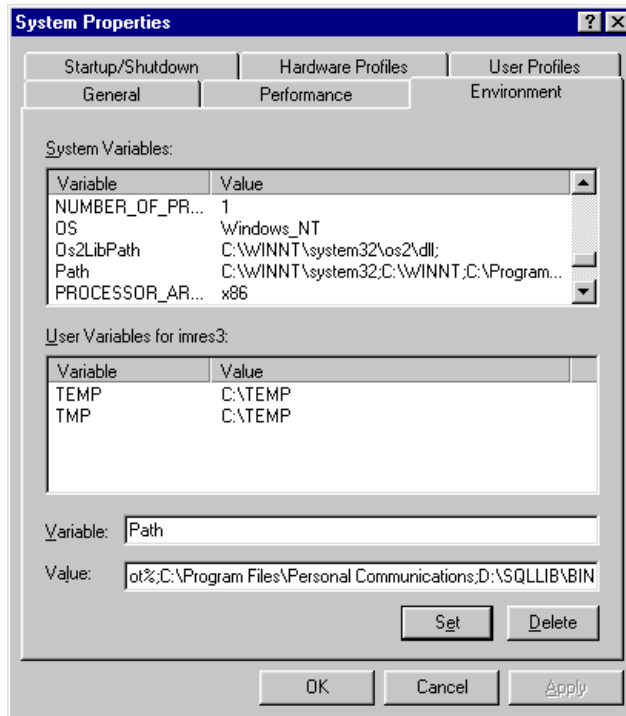


Figure 43. Setting the Path Variable

Find the variable **Path** in the upper list and click on it. Then add the location of your BIN directory to the end of the **Value** field. In the example above, this is D:\SQLLIB\BIN. Make sure you enter a semicolon to separate it from the preceding directory name. Press **OK** and open a new Command Prompt to check the path contents.

6.1.3 Networking Requirements

Intelligent Miner uses TCP/IP as the protocol for all communication between client and server. We describe all actions in this section using Windows NT as a sample system. Using Windows 95 as the operating system for the IM client might have different screen names but is similar to the procedure shown here for Windows NT.

To check whether TCP/IP is configured on your system, double-click the **Network** icon in the NT **Control Panel**. Select the **Protocol** tab, click **TCP/IP** and then **Properties**, as shown in Figure 44. The window must show an adapter and either have DHCP selected or an IP address specified.

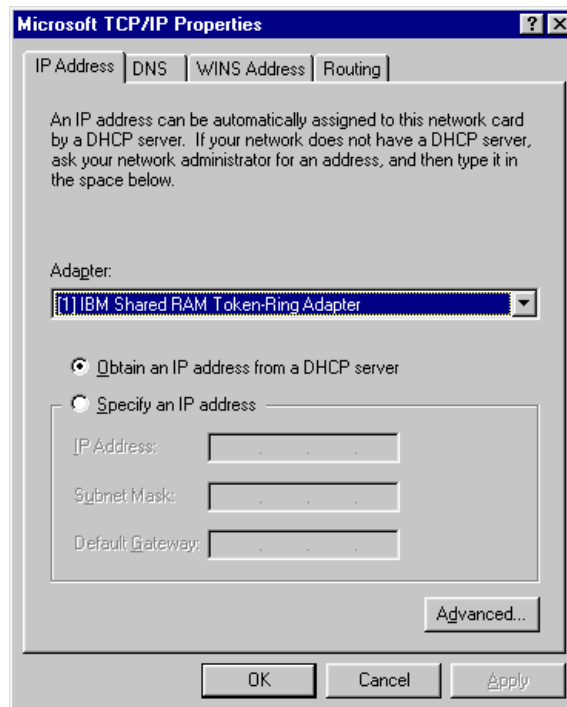


Figure 44. Windows NT TCP/IP Properties

Before the installation of IM Server, try to determine whether you can reach the server machine through the TCP/IP protocol from the machine you intend to use as the client. To test this, open an NT Command Prompt window and type `ping hostname` for the desired hostname.

If this fails, find the IP address of the host. Normally, you can find it in the window shown in Figure 44. Ping the address rather than the hostname. If

this also fails, your network set up is incorrect, and you should contact your network administrator. Otherwise, you will need to edit the file that is used to translate hostnames into IP addresses. This file, called HOSTS, is located in the directory WINNT\SYSTEM32\DRIVERS\ETC (normally found on drive C:). Add a line to that file with any text editor as shown in the example below for `examplehost` with address `9.9.9.9`.

```
#
# This is a sample HOSTS file used by Microsoft TCP/IP for Windows NT.
#
# This file contains the mappings of IP addresses to host names. Each
# entry should be kept on an individual line. The IP address should
# be placed in the first column followed by the corresponding host name.
# The IP address and the host name should be separated by at least one
# space.
#
# Additionally, comments (such as these) may be inserted on individual
# lines or following the machine name denoted by a '#' symbol.
#
# For example:
#
#       102.54.94.97       rhino.acme.com       # source server
#       38.25.63.10       x.acme.com          # x client host
#
127.0.0.1       localhost
9.9.9.9         examplehost
```

Never change the entry or address for `localhost`. Test your new set up by saving your hosts file and then `ping examplehost`.

You can use DHCP on your clients without any problem. If you need to use DHCP on your IM server, the server's hostname must be registered with a DNS server, because you cannot specify a fixed IP address in the client's hosts file.

6.2 Product Installation

Intelligent Miner for Data consists of three components:

- IM Server is the processing engine that runs the mining, data processing, and statistical algorithms.
- IM Client is the graphical user interface to IM Server.
- The IM Application Development Toolkit contains the header files and examples to build your own programs that use the IM API.

To install IM components on Windows NT:

1. Log on to the system with administrator authority.
2. Insert the IM server CD-ROM in the appropriate drive.
3. Run the **Setup** program from the appropriate language directory. For US English, this would be D:\EN\Setup.exe, if your CD-ROM drive letter is D. If you want to install a client only, you can find it on the client CD. Run the program, D:\WIN32\EN\Setup.exe to install.
If you try to install IM on a system with less than 65536 display colors, you will get the message shown in Figure 45.

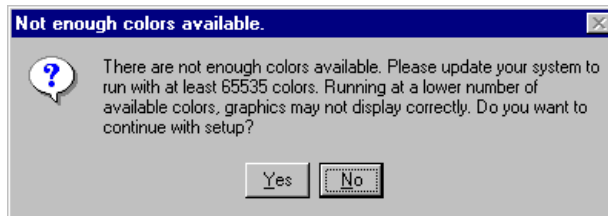


Figure 45. Message on Available Colors

If you plan to install only the IM server, confirm this message by selecting **Yes**. You can also reconfigure the display settings of your system after the installation. You will now see the window shown in Figure 46.

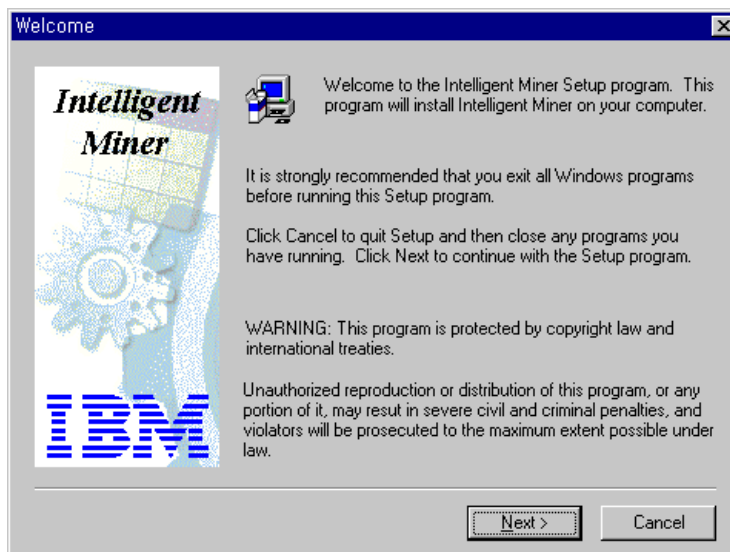


Figure 46. IM for Windows NT, Welcome Screen

- Click **Next >** on the **Welcome** window to select the components shown in Figure 47.

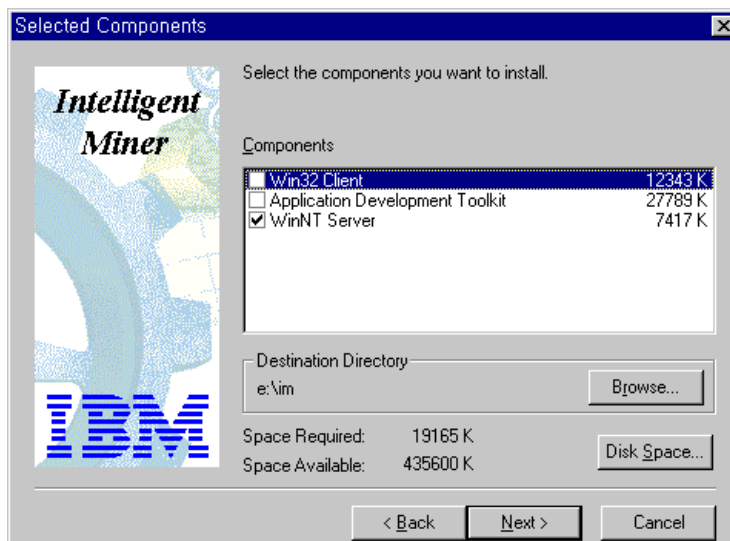


Figure 47. Selected Components

- Select the components you want to install, then click the **Next >** button. You can always add other components in a later stage by running the **Setup** program again. If you want to change the installation directory, click **Browse** and select another directory. Installation on a network drive is not recommended for reasons of performance and availability, and because you would not be able to share installations with other servers. Figure 48 shows the window that allows to select a user to run the server.

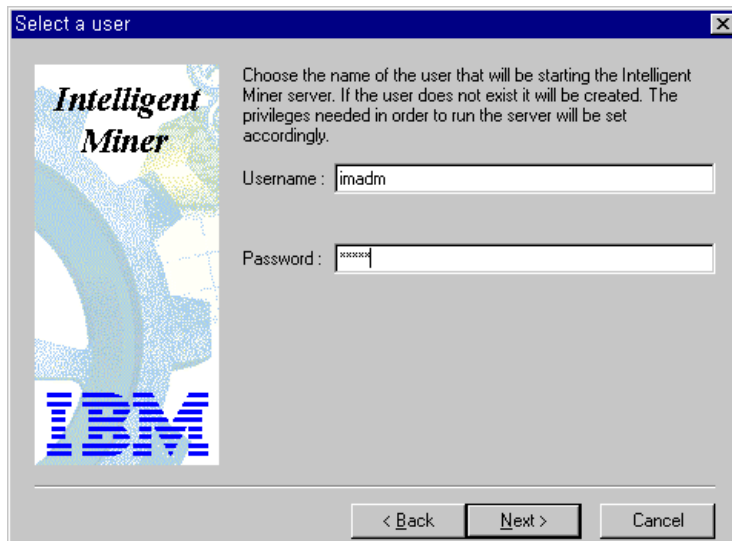


Figure 48. User Selection

6. Enter the user name and the password that you plan to use to run the server process and click **Next >**. Figure 49 shows the window to select the multi-user setup.

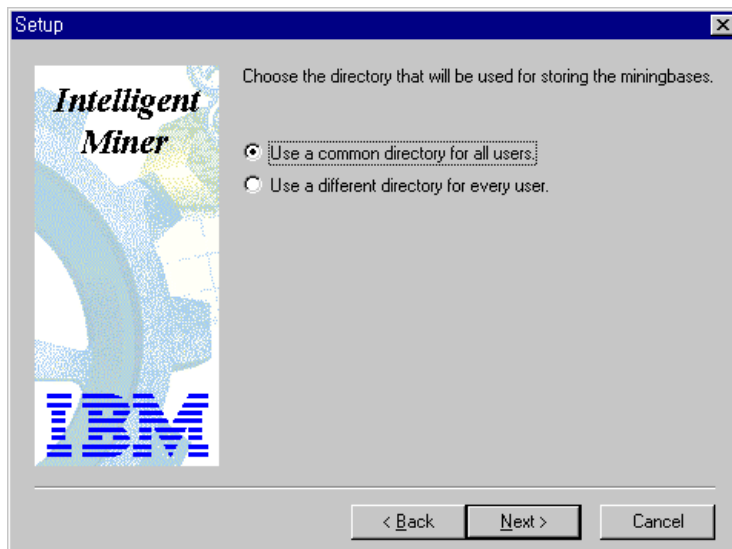


Figure 49. Multi-User Setup

7. Select the way in which you want to store your miningbases. If you choose to use a common directory, note that only one user at a time can access a mining database. You can override these settings at any time before starting the server by setting the environment variables as described at the end of this section. Clicking **Next >** will lead you to the window to select the directory where you want miningbases to be stored, as shown in Figure 50.

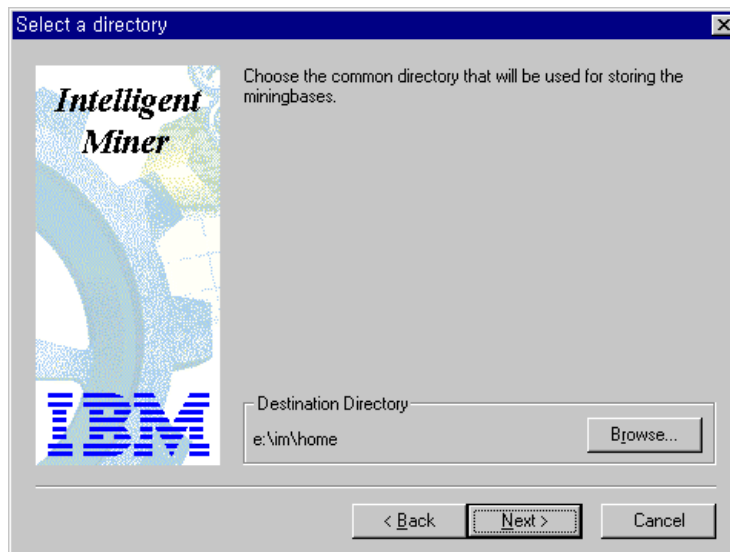


Figure 50. Directory for Miningbases

8. Select the directory where you want IM Server to store the miningbases by default and click **Next >** to select the program folder as shown in Figure 51.

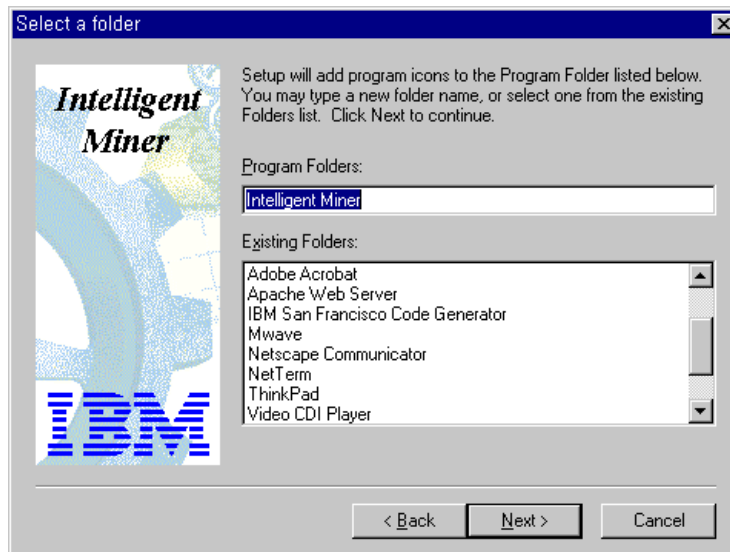


Figure 51. Select Start Menu Folder

9. Select the folder you want to create under **Programs** in your Start Menu and click **Next >**. The installation will now start. When finished, the window shown in Figure 52 will be displayed.



Figure 52. Start Menu Folder on Desktop

10. The Installation is now complete. You can close the window shown in Figure 52, view and print the README file and reboot your system. Logging off and logging on is not sufficient, because the rights of the user, specified during installation, will only be changed with a full restart.

The client installation program also installs the Java Runtime Environment (JRE) which is required. This installation uses its own environment and will not influence any existing Java installation such as the Sun JDK or a Web browser.

To view the online help, you will need to install a Web browser which supports frames such as Microsoft Internet Explorer Version 3.0 or Netscape Navigator for Windows 95/NT Version 3.01.

6.3 Verifying the Installation

After rebooting your IM server system, log on as the user that you created or configured to start IM Server. Open your Control Panel and double-click **Services**. This will show the screen in Figure 53.

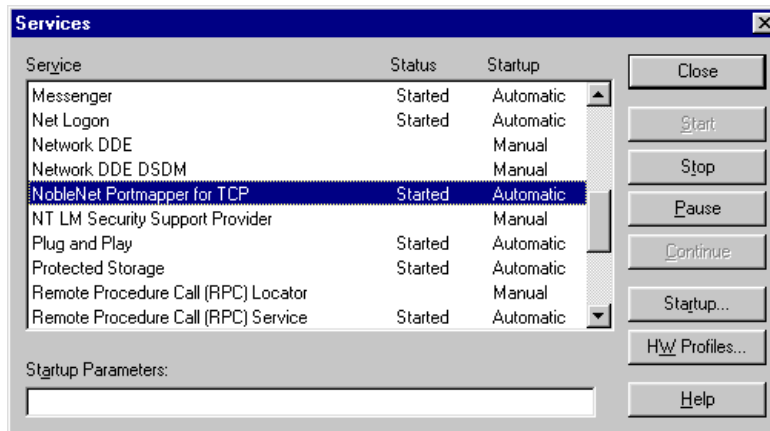


Figure 53. Windows NT Services Panel

The list will show the NobleNet PortMapper for TCP, which must be configured as **Started** and **Automatic**.

Before you start the server, you should check whether IDM_RES_DIR and IDM_MNB_DIR contain the values you wish to use for all users using the following commands:

```
set idm_res_dir
set idm_mnb_dir
```

These settings depend on the choices you made during installation. You can change the values by adding an '=' sign followed by the required directory as shown below:

```
set idm_res_dir=d:\results
set idm_mnb_dir=d:\mnbases
```

If you want users that will be connecting to the system to have their own directory, set the value of `IDM_HOME_DIR` to the parent directory where subdirectories for each user will be created by the server. Clear the other two variables with these commands:

```
set idm_home_dir=d:\im\home
set idm_res_dir=
set idm_mnb_dir=
```

6.4 Running the Server

You can now start the IM Server. To do this, open an NT Command Prompt window and enter `idmstart -d`. This will start the server with message output going to the current window. You can stop the server by pressing **Ctrl-C** or by closing the server window.

Note that you *must* be logged on as the user that you specified during installation. If you want another user to start the IM server, you must register that user by using the command:

```
idminit <username> <password> <domainname>.
```

This will create a user with the required rights in the domain that you specified. If you are working on a stand-alone server or workstation, the domain name is the local machine name.

You can create an entry in your Start Menu if you wish to start the server more easily. If you create this entry in the **Startup** folder, the server will be restarted after each reboot of the system. The entry must refer to `idmstart`.

When the server is up and running, you can start a client from the Start Menu. As a test, try connecting to server **localhost** with the appropriate `username` and password. Create a data source from a database table and perform a test run.

You can further verify your installation by following the steps described in Appendix A of the "Using the Intelligent Miner for Data" documentation that is shipped together with the product. You can use the `imdemo` command to start a stand-alone demonstration system with both client and server, or `idmstartdemo` to prepare a demonstration server that will accept incoming clients.

Chapter 7. Implementation on AIX

The following chapter describes the tasks to install IM on AIX. Figure 54 shows the system that was used.

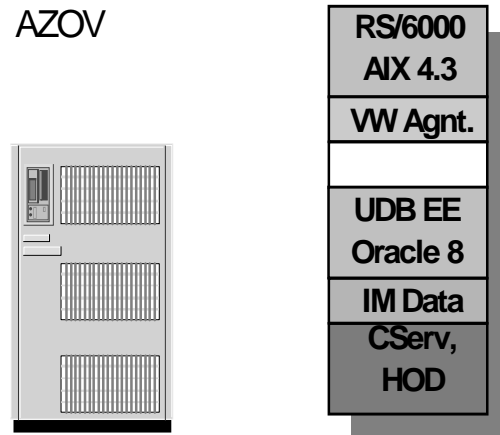


Figure 54. Intelligent Miner for Data on AIX

7.1 Prerequisites

In this section, we describe the prerequisites for hardware, software, and network configuration that should be verified before installing IM.

7.1.1 Hardware Requirements

Table 10 and Table 11 show the hardware requirements for the IM server, and IM client respectively.

Table 10. AIX Server Hardware Requirements

Storage type	For demonstration	Required	Recommended
RAM	64 MB	128 MB	512 MB to 2GB
Disk space for AIX server	75 MB	100 MB	200% of data
Disk space for AIX server including client	105 MB	130 MB	200% of data
Additional disk space for toolkit	4 MB	4 MB	4 MB

Table 11. AIX Client Hardware Requirements

Storage type	For demonstration	Required	Recommended
RAM	32 MB	64 MB	64 MB
Disk space	85 MB	100 MB	over 100 MB
Additional disk space for toolkit	4 MB	4 MB	4 MB

To check the amount of your RAM size, enter the command `lsattr -E -l sys0 -a realmem` on the AIX command line. For example, the following screen shows the result of this command.

```
root@sky > lsattr -E -l sys0 -a realmem
realmem 1048576 Amount of usable physical memory in Kbytes False
```

To check for the available disk space, enter the command `df -k` on the AIX command line. The free disk space is shown using the kilobyte unit.

7.1.2 Software Prerequisites

The following tables (Table 12 and Table 13) show the software requirements for server and client respectively.

Table 12. AIX Server Software Prerequisites

Software	Version	Program number	Required/ Optional
IBM AIX	4.1.5 (or higher)	5765-C34	Required
One of the following DB2 systems: IBM DB2 for AIX IBM DB2 Universal Database Enterprise Edition IBM DB2 Parallel Edition IBM DB2 Universal Enterprise Extended Edition	2.1.1 5.0 1.2 5.0	5765-454 5648-A32 5765-328 5648-A34	Required

Note

To use a DB2 database on a server that is different from the IM server, you must install DB2 Client Application Enabler (CAE) Version 2.1.1, or higher on your IM server rather than the complete DB2 server software.

If you plan to use the IM server with DataJoiner, please refer to Appendix A, “Using DB2 V 2.1 or DataJoiner on AIX” on page 175.

Table 13. AIX Client Software Prerequisites

Software	Version	Program number	Required/ Optional
IBM AIX	4.1.5 (or higher)	5765-C34	Required
Java Runtime Environment from JDK for AIX	1.1.4 (or higher)	N/A	Required
Netscape Navigator for AIX	3.0 (or higher)	N/A	Optional (required to display online help)

Note

The AIX client was built using the JDK V1.1.4, 09April98 build. Only this Java build or later builds are supported. You can check which Java version is installed by entering the command:

```
java -fullversion
```

If this command returns an earlier build date than 'a114-19980409', you must download the latest fixpack for JDK V1.1.4 or upgrade your system to JDK V1.1.6. It is recommended to use JDK V1.1.6. JDK for AIX which can be downloaded from the URL

```
http://www.ibm.com/java/jdk/download
```

Intelligent Miner assumes that JRE for AIX is installed under the directory /usr/jdk_base. If JRE for AIX is installed under a different directory on your system, you need to set the environment variable JAVA_HOME to point to this directory before starting the graphical user interface (GUI).

To check your AIX version, enter the command `oslevel` on the AIX command line.

To check the version of DB2 and JDK Runtime Environment that is installed, use the SMIT panels as shown in Figure 55 and Figure 56. Use the following path from the `smit` start window:

System Management -> Software Installation and Maintenance -> List Software and Related Information -> List Installed Software

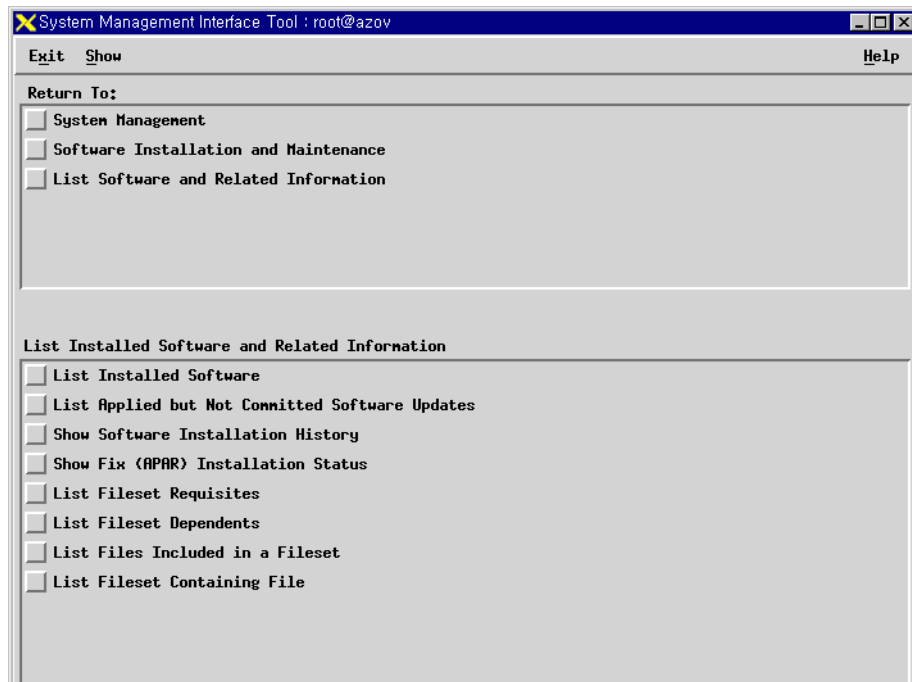


Figure 55. List Installed Software on AIX



Figure 56. Software Selection on AIX

Select the software from the list that will be displayed when you click on the **List** button. If you keep the default value (all), this will show all the software that is installed .

Figure 57 on page 104 and Figure 58 on page 105 show that the packages of JDK and DB2 are installed.

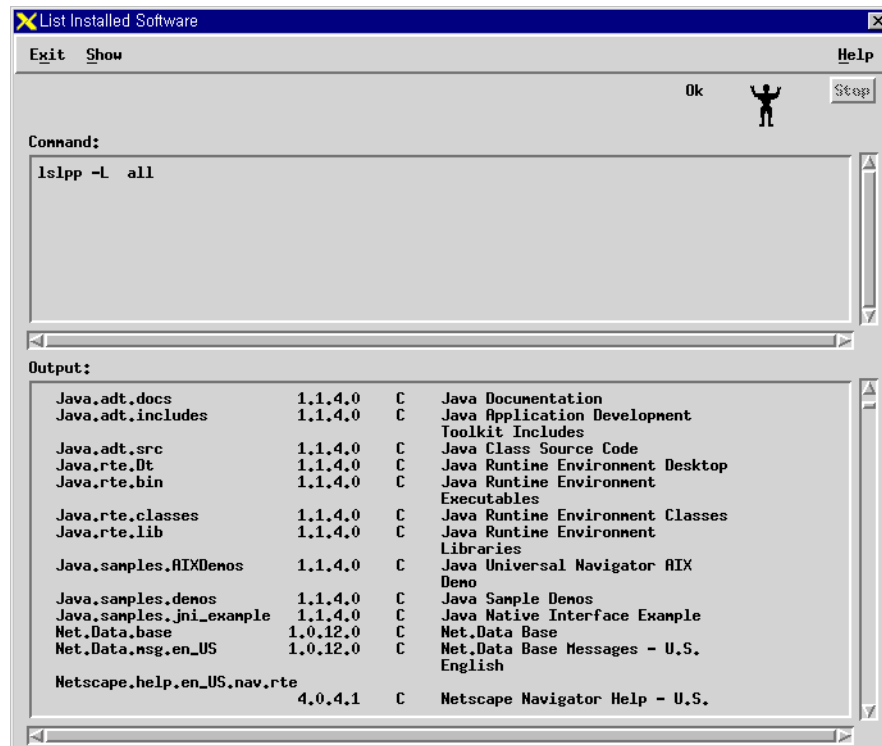


Figure 57. Display Installed JAVA Software on AIX

If you run the IM client on AIX, check that *Java.rte.Dt*, *Java.rte.bin*, *Java.rte.classes*, and *Java.rte.lib* are installed.

Note

You must ensure that DB2 UDB is installed BEFORE installing the Intelligent Miner for Data server. If you reinstall a new DB2 UDB after installing Intelligent Miner for Data, you must first uninstall Intelligent Miner for Data, and then reinstall the product.

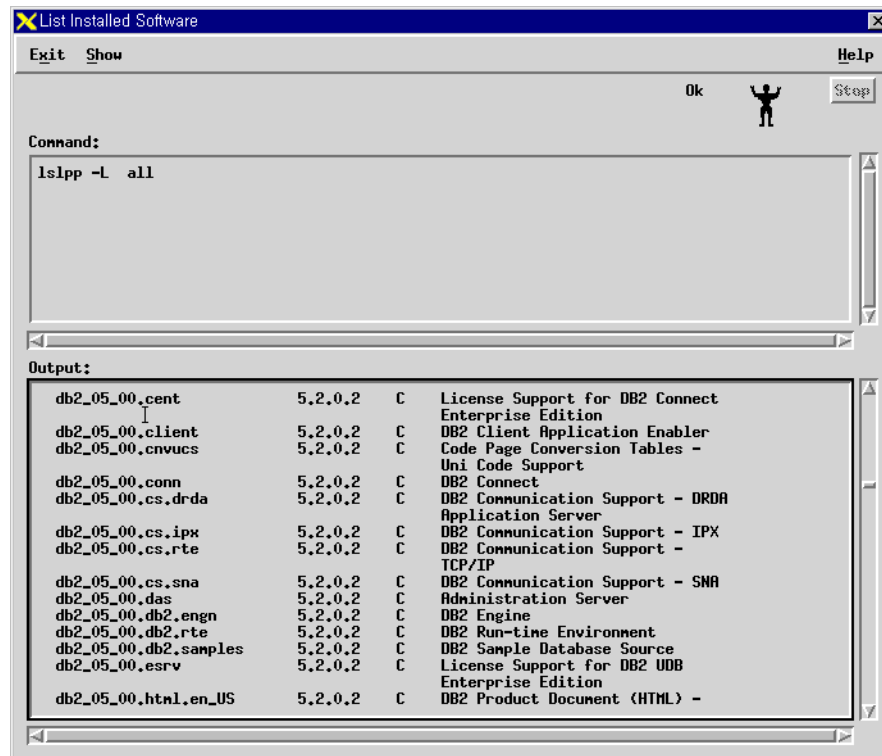


Figure 58. Display Installed DB2 Software on AIX

To install the IM server, one of following products must also be installed.

```
db2_02_01.client 2.1.1.0
db2_05_00.conn 5.0.0.0
db2_05_00.client 5.0.0.0
```

7.1.3 Networking Requirements

If you plan to run the IM in stand-alone mode on the server, you don't need to configure TCP/IP. But if you run the IM in Client/Server mode, TCP/IP must be configured to allow access from remote clients to the IM server on AIX. To check the configuration:

1. Enter the `smit` command on the AIX command line. Use the following path to get to the screen shown in Figure 59.

System Management -> Communication Applications and Services -> TCP/IP

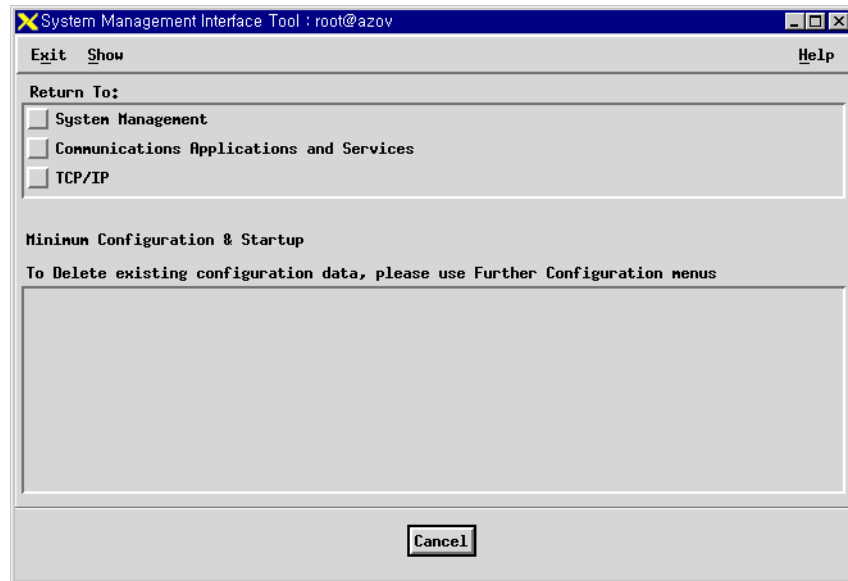


Figure 59. AIX TCP/IP SMIT Panel

2. Select the defined network interface (Figure 60). If you don't know which network interface to use, ask your network administrator.

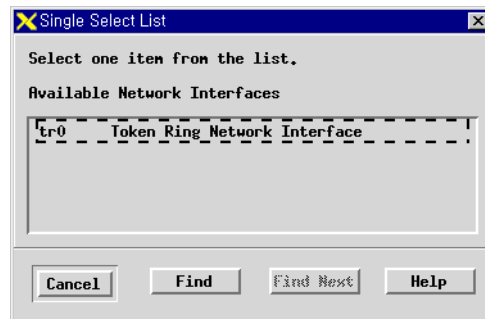
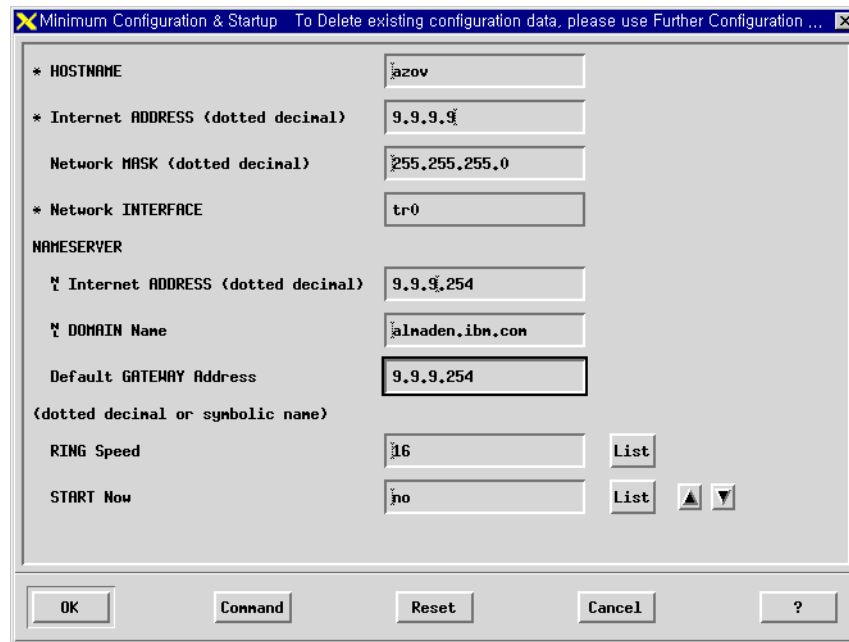


Figure 60. Defined Network Interface on AIX

3. Check for the host name, domain name, and so on, as shown in Figure 61. You will need to know the hostname or IP address in order to configure the connectivity on the IM client.



The image shows a window titled "Minimum Configuration & Startup" with a subtitle "To Delete existing configuration data, please use Further Configuration ...". The window contains several input fields for network configuration:

- * HOSTNAME: jazov
- * Internet ADDRESS (dotted decimal): 9.9.9.9
- Network MASK (dotted decimal): 255.255.255.0
- * Network INTERFACE: tr0
- NAMESERVER:
 - Internet ADDRESS (dotted decimal): 9.9.9.254
 - DOMAIN Name: alnaden.ibm.com
 - Default GATEWAY Address: 9.9.9.254
- (dotted decimal or symbolic name):
- RING Speed: 16 (with a "List" button)
- START Now: no (with a "List" button and up/down arrow icons)

At the bottom of the window are buttons for "OK", "Command", "Reset", "Cancel", and a help icon "?".

Figure 61. AIX TCP/IP Minimum Configuration

4. After completing steps 1 through 3, you can check the TCP/IP status on the IM server system.

To check all TCP/IP related subsystems, use the `lssrc -g tcpip` command on the AIX command line. Be sure that the `inetd` daemon is active. You can also check for `inetd` only with `lssrc -s inetd` command.

5. Add the dotted decimal address and host name into the `/etc/hosts` file of the IM client system (Figure 62).

```

# network. This file is used to resolve a hostname into an Internet
# address.
#
# At minimum, this file must contain the name and address for each
# device defined for TCP in your /etc/net file. It may also contain
# entries for well-known (reserved) names such as timeserver
# and printserver as well as any other host name and address.
#
# The format of this file is:
# Internet Address      Hostname      # Comments
# Items are separated by any number of blanks and/or tabs. A '#'
# indicates the beginning of a comment; characters up to the end of the
# line are not interpreted by routines which search this file. Blank
# lines are allowed.
#
# Internet Address      Hostname      # Comments
# 192.9.200.1           net0sample    # ethernet name/address
# 128.100.0.1           token0sample  # token ring name/address
# 10.2.0.2              x25sample     # x.25 name/address
#
127.0.0.1              loopback localhost datajoiner # loopback (lo0) name/address
192.9.9.9              azov

```

Figure 62. AIX Hosts File

6. Verify the TCP/IP connection from the client.

You can check the connection from the IM client to the IM server by typing `PING HOSTNAME` on the command line.

7.2 Product Installation

If you already have an Intelligent Miner Version 1 or Intelligent Miner Version 2.1.0 installed on your system, it will be upgraded to Version 2.1.3, if you use the smit installation. To do this:

1. Log in as a `root`.
2. Use the following path to create a group for IM users (for example, `imgroup`) using the `smit` command.

System Management -> Security & Users -> Add a Group

Figure 63 and, Figure 64, show the SMIT panels for creating user groups.

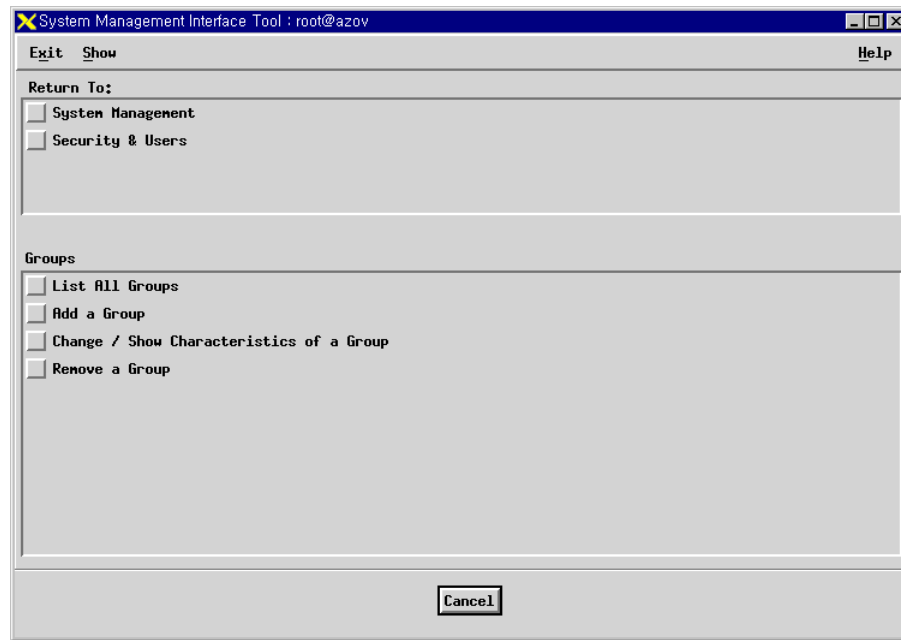


Figure 63. Create Group SMIT Panel on AIX

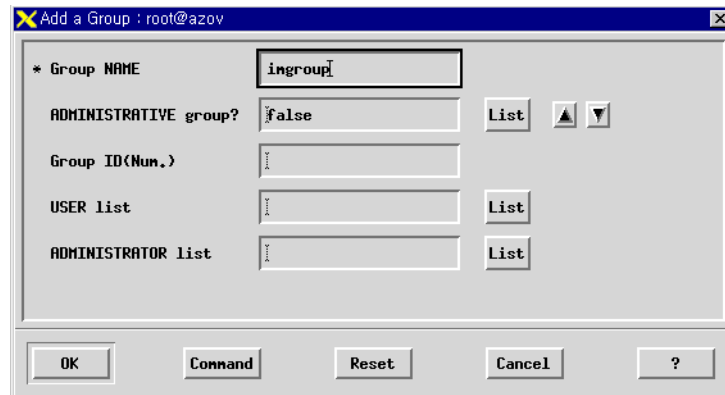


Figure 64. Create Group on AIX

Enter a Group Name and click **OK** to create the group.

3. Create a user for IM (for example, *imadm*) using the `smit` command. Figure 65 shows the SMIT panels for creating an AIX user. Use the following path from the SMIT main panel:

System Management -> Security & Users -> Add a User

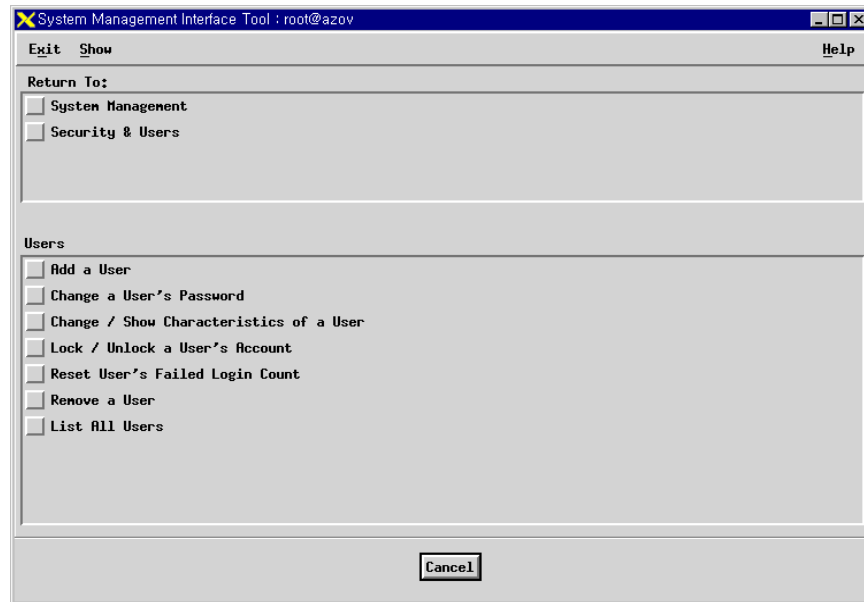
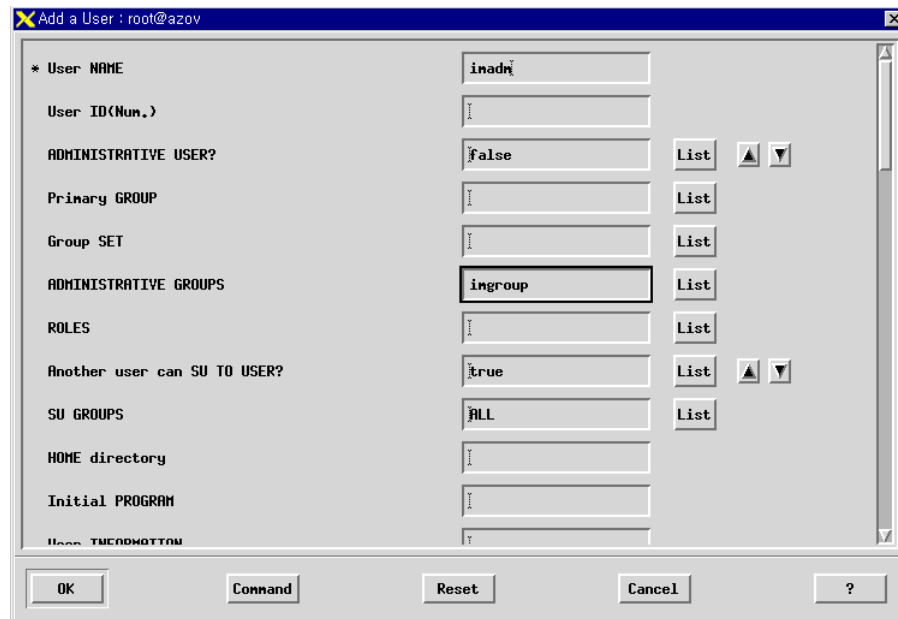


Figure 65. Create User SMIT Panel on AIX

Enter a User Name (*imadm*) and the User Group (*imgroup*) defined previously, as shown in Figure 66, then click **OK** to create the IM user.



Add a User : root@azov

* User NAME	inadm	
User ID (Num.)		
ADMINISTRATIVE USER?	false	List ▲ ▼
Primary GROUP		List
Group SET		List
ADMINISTRATIVE GROUPS	ingroup	List
ROLES		List
Another user can SU TO USER?	true	List ▲ ▼
SU GROUPS	ALL	List
HOME directory		
Initial PROGRAM		
Users: THE BOTTOM		

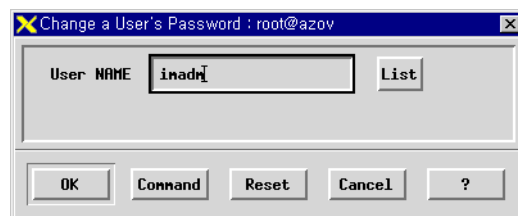
OK Command Reset Cancel ?

Figure 66. Create User on AIX

4. Change IM user password.

After the IM user has been created, the password must be set for this user. With the first logon the IM user must change its password.

Figure 67 and Figure 68 show the *smit* panel for changing the user password and the first logon as the IM user.

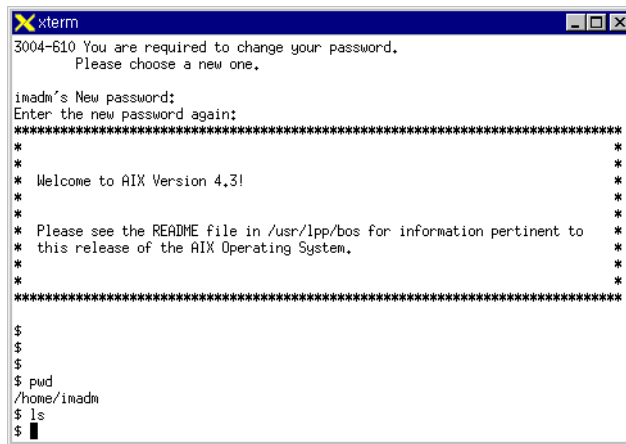


Change a User's Password : root@azov

User NAME	inadm	List
-----------	-------	------

OK Command Reset Cancel ?

Figure 67. Change User Password on AIX

The image shows a terminal window titled 'xterm'. The text inside the window is as follows:

```
3004-610 You are required to change your password.  
Please choose a new one.  
  
imadm's New password:  
Enter the new password again:  
*****  
*                                                                 *  
*                                                                 *  
* Welcome to AIX Version 4,3!                                   *  
*                                                                 *  
*                                                                 *  
* Please see the README file in /usr/lpp/bos for information pertinent to *  
* this release of the AIX Operating System.                       *  
*                                                                 *  
*****  
  
$  
$  
$  
$ pwd  
/home/imadm  
$ ls  
$
```

Figure 68. Log in as a IM User

5. Configure the IM user to access a database if you are using DB2.

The database administrator should grant the connect and select privileges to the IM user to allow access to the required databases and tables.

6. Log in as *root*.

7. Insert the IM CD-ROM in the CD-ROM drive.

The IM for Data CD-ROM for AIX contains the following software packages:

- Installation package for the AIX server
This package contains file sets (installp images) for the AIX server.
- Installation package for the AIX client
This package contains file sets (installp images) for the AIX client.
- Installation packages for text files supporting the AIX client
This package contains file sets (installp images) that provide online help and other texts in U.S English to assist users of AIX clients.
- Additional installation packages for text files in various languages
Each of these packages contains file sets (installp images) that provide online help and other texts in the supported languages. For a list of the supported languages, see the README file.
- Installation package for the Application Development Toolkit (ADK)
This package contains header files required to write programs using the Intelligent Miner Application Programming interfaces. Before you

install the ADT, make sure that either the client or server software is already installed.

8. Enter the `smit` command on the AIX command line to start software installation. To start the installation, use the following path from the SMIT main panel:

System Management -> Software Installation and Maintenance -> Install and Update from LATEST Available Software

Figure 69 shows SMIT software installation panel.

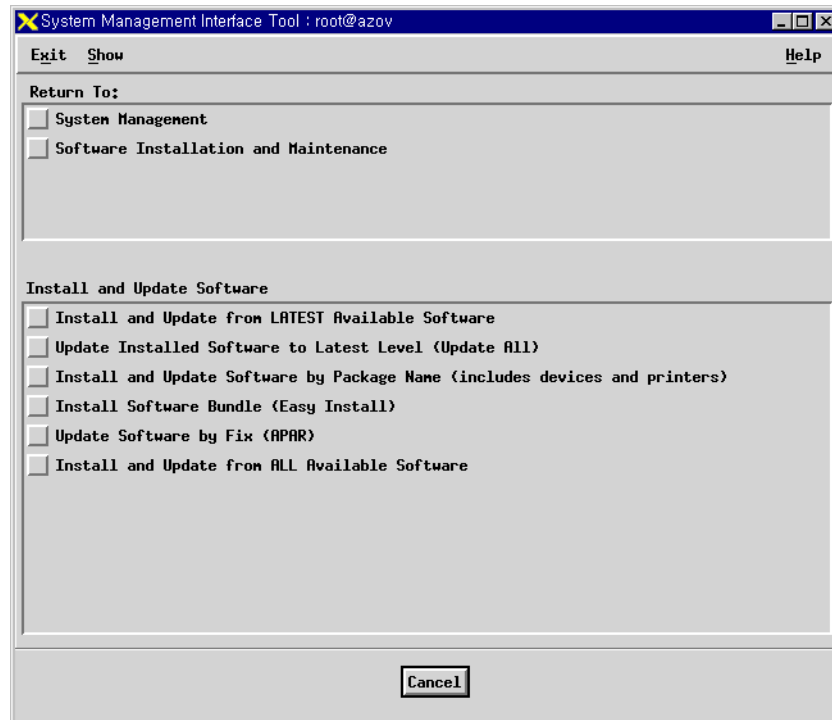


Figure 69. Software Installation Smit Panel on AIX

Select the CD-ROM drive from the list of input devices and click **OK** on the screen shown in Figure 70.

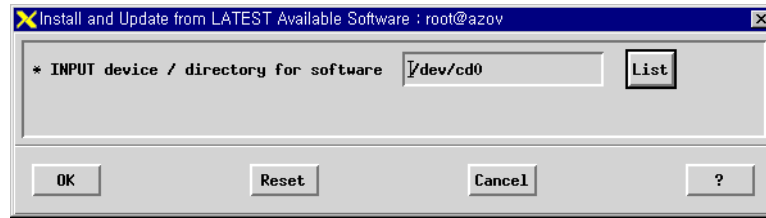


Figure 70. Installation Device Selection on AIX

9. Select the installation package.

The list contains the following software packages:

- IMiner.server
Intelligent Miner for Data - Serial Server
- IMiner.common
Intelligent Miner for Data - Common Parts
- IMiner.client
Intelligent Miner for Data - Client
Intelligent Miner for Data - Images
- IMiner.html.en_US
Intelligent Miner for Data - Help Navigation - U.S. English
Intelligent Miner for Data - Examples & concepts - U.S. English
Intelligent Miner for Data - TaskGuide Texts - U.S. English
Intelligent Miner for Data - Glossary Terms - U.S. English
Intelligent Miner for Data - Valid Values - U.S. English
- Additional installation packages containing translated text files for each language supported.
- The names of these packages are in the format IMiner.html.<lang>, where <lang> is the language identifier.
- IMiner.toolkit
Intelligent Miner for Data - Toolkit (Header Files)

When you click one of the group headings to select it for installation, all the components under this heading are installed automatically. You do not need to select the components individually.

10. Install the AIX client software.

This step is not required unless you use the Intelligent Miner in stand-alone mode. However, should you install the client software to verify that the Intelligent Miner server is installed properly.

11. Create links to the Intelligent Miner executables.

You can use one of the following methods to invoke the executables.

- Create links in the directory /usr/bin to the executables in the directory /usr/lpp/IMiner/bin.

To create links to executables, log in as the user and invoke the following command:

```
/usr/lpp/IMiner/bin/imln
```

- Add the path containing the executables to your environment.

To add the path to your environment, append the /usr/lpp/IMiner/bin directory to the PATH of your profile and export PATH (Figure 71).

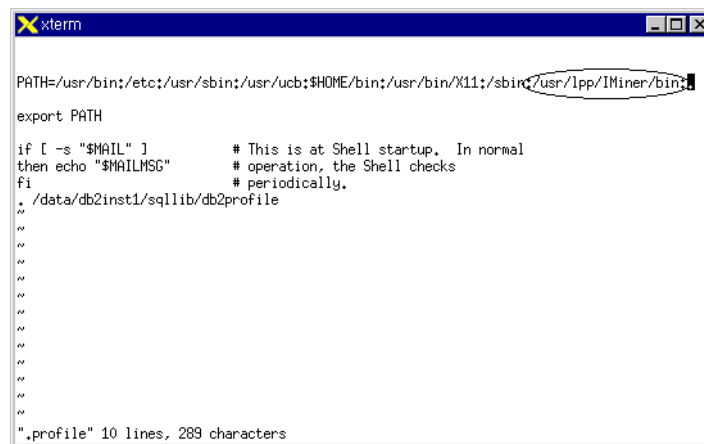
The image shows a terminal window titled 'xterm'. The command prompt shows the current PATH variable: 'PATH=/usr/bin:/etc:/usr/sbin:/usr/ucb:\$HOME/bin:/usr/bin/X11:/sbin:'. The path '/usr/lpp/IMiner/bin:' is being appended to the end of the PATH variable. Below this, the command 'export PATH' is entered. The terminal also displays a shell startup script snippet with comments about mail checking and a profile line count: 'if [-s "\$MAIL"] # This is at Shell startup. In normal then echo "\$MAILMSG" # operation, the Shell checks fi # periodically. /data/db2inst1/sqllib/db2profile'. At the bottom, it says '._profile" 10 lines, 289 characters'.

Figure 71. Append IM Command Directory

7.3 Installation Verification

IM uses the environment variables IDM_MNB_DIR and IDM_RES_DIR, which point to the directories containing the mining databases and result files respectively.

If you do not set these variables and use the default values, directories named *idmmnb* and *idmres* are created in each user's home directory.

Before starting the IM server, you can set the environment variable *IDM_MNB_DIR*, *IDM_RES_DIR* to point to one common directory, if multiple users are to share the same mining bases and result files.

For example, if you want all users to share the mining bases and result files in the *IDMMNB* and *IDMRES* subdirectory respectively, type the following commands in the Korn shell:

```
export IDM_MNB_DIR=/miningbases
export IDM_RES_DIR=/resultfiles
```

To verify that the IM server was successfully installed, enter the command *idmstartdemo* on the command line and access the IM server from the client, then try to run the sample data mining functions.

7.4 Running the Server

There are two modes to start the IM AIX server using DB2 UDB.

Client/Server Mode

1. Log on as *root*.
2. Set the DB2INSTANCE variable to the DB2 UDB instance.
3. Type one of the following commands on the command line:

```
idmstart -d    This command also shows tracing information of the AIX
                server.
```

```
idmstart       This command starts the AIX server without showing tracing
                information.
```

4. Log out.
5. Call *im* on the client to test your installation.

Stand-alone Mode

1. Log on as user *root*.
2. Set the DB2INSTANCE variable to the DB2 UDB instance.
3. Type one of the following commands on the command line:

```
idmstart -d    This command also shows tracing information of the AIX
                server.
```

```
idmstart       This command starts the AIX server without showing tracing
                information.
```

4. Log out.

5. Log on as Intelligent Miner user.
6. Set the DB2INSTANCE variable to the DB2 UDB instance.
7. Call `im` on the client to test your installation.

Note

The DB2 UDB 5.0 fixpack 3 with PTF number U453782 is prerequisite for working with Intelligent Miner in the following cases:

- Running Intelligent Miner parallel with UDB EEE.
- Running Intelligent Miner in stand-alone mode.
- Running Intelligent Miner in client/server mode with an AIX client and `IDM_CLI_USED` set on the client.

If you have this PTF applied and you want to connect remotely to a DB2/MVS or DB2/390 database from an Intelligent Miner AIX server, ensure that you also have one of the following PTFs applied:

UQ13906 - DB2/MVS V3.1

UQ13907 - DB2/MVS V4.1

UQ13908 - DB2 for OS/390 V5.1

Connecting from an Intelligent Miner AIX server to a DB2/MVS or DB2/OS390 database is not fully supported. You can use DB2/MVS or DB2/OS390 database tables as input for mining runs but you cannot write output tables. Connecting to an AS/400 database from an Intelligent Miner AIX server is not supported at all.

Chapter 8. Implementation on Sun Solaris

The following chapter describes the hardware requirements, software requirements, and the installation process for the IM on Sun Solaris servers. Figure 72 shows the system used.

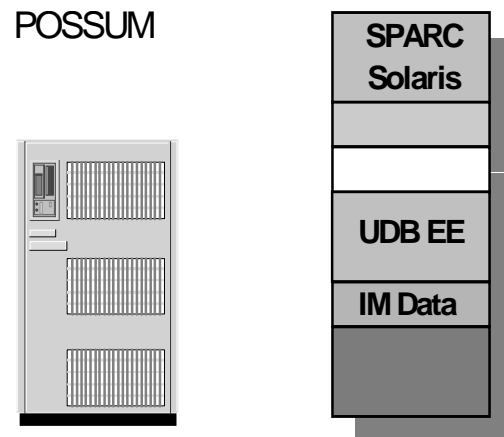


Figure 72. Intelligent Miner for Data on SUN Solaris

8.1 Prerequisites

In this section, we describe the prerequisites for hardware, software and network configuration.

8.1.1 Hardware Requirements

The IM for Sun Solaris is runs on systems based on the SUN SPARC processor.

Intelligent Miner for Data is available only as a server for the Sun/Solaris platform. To use this server, you need to install one of the supported clients:

- AIX client
- OS/2 client
- Windows client

Table 14 lists the hardware requirements for the IM server to be installed on Sun/Solaris:

Table 14. Sun Solaris Server Hardware Requirements

Storage type	For demonstration	Required	Recommended
RAM	64 MB	128 MB	512 MB to 2GB
Disk space for Sun Solaris	40 MB	50 MB	200% of data
Additional disk space for toolkit	4 MB	4 MB	4 MB

Figure 75 on page 123 shows how to check for the available amount of RAM in your system.

To check the available disk space, enter the command `df -k` on the command line to see the free disk space which is shown in kilobytes.

8.1.2 Software Prerequisites

Table 15 lists the required and optional software needed on the Sun Solaris server.

Table 15. Sun Solaris Server Software Prerequisites

Software	Version	Program number	Required/Optional
Sun Solaris	2.5 (or higher)		Required
One of the following DB2 systems:			Required
IBM DB2 for Sun Solaris	2.1.1	5765-578	
IBM DB2 Universal Database	5.0	5648-A32	

Note

1. To use a DB2 database on a server that is different from the IM server, you must install the DB2 Client Application Enabler (CAE) Version 2.1.1 or higher on your IM server rather than the complete DB2 server software.
2. The DB2 library path must be included in your LD_LIBRARY_PATH specification.

To check the prerequisite software and its version, type `admintool` on the command line. This command starts a graphical user interface (GUI) from which you can then select the **Browse -> Software**. Figure 73 shows the results of this command. You can see more detailed information about the software on your system by clicking **Show Details**.

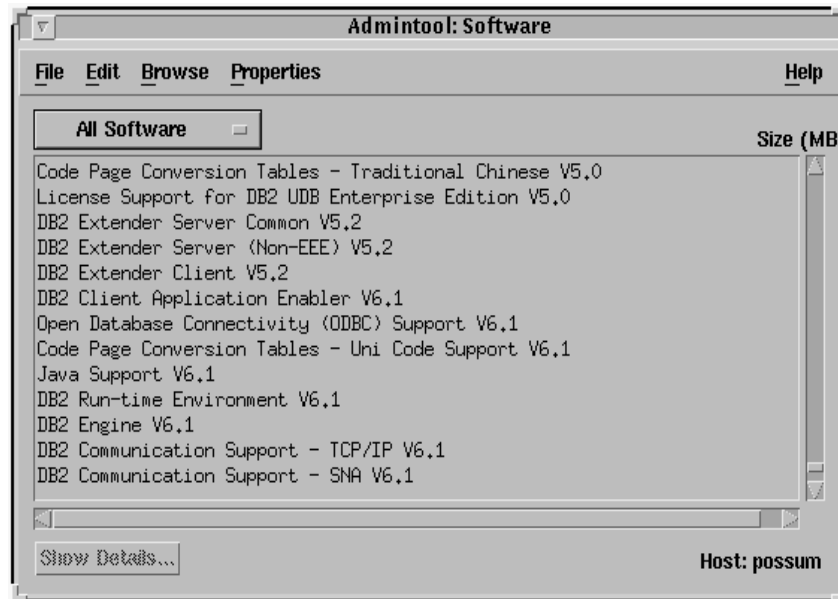


Figure 73. Installed Software list on Sun Solaris

Figure 74 shows the detailed software information about the DB2 UDB version installed on the system. You can see the software name, product name, package instance, vendor, installed directory, and so forth.

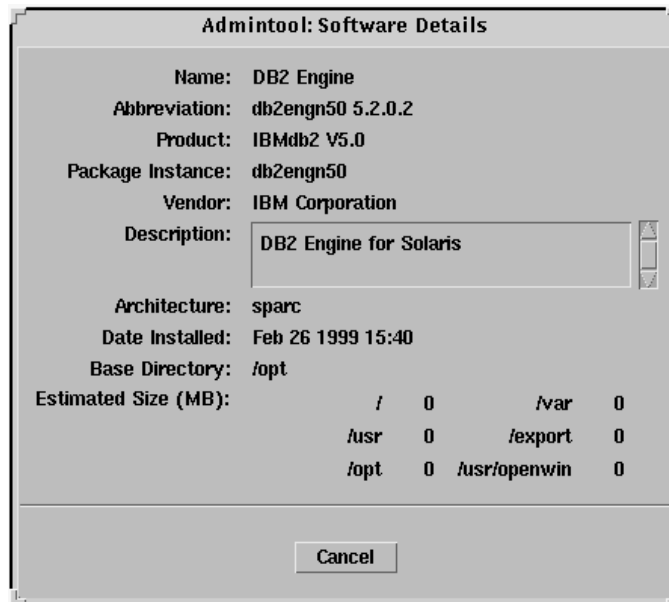


Figure 74. Detail Software Information on Sun Solaris

8.1.3 Networking Requirements

TCP/IP must be configured to allow access between remote clients and the IM server. To check this configuration:

1. Log in to the OpenWindow of Sun Solaris as a *root*.
2. Click the right mouse button, then click the **Workstation Info** from the **WorkSpace** panel.

This shows the general workstation information as shown in Figure 75. Remember the workstation name and the internet address because you will need this information to configure the TCP/IP connectivity from the IM client to this server.

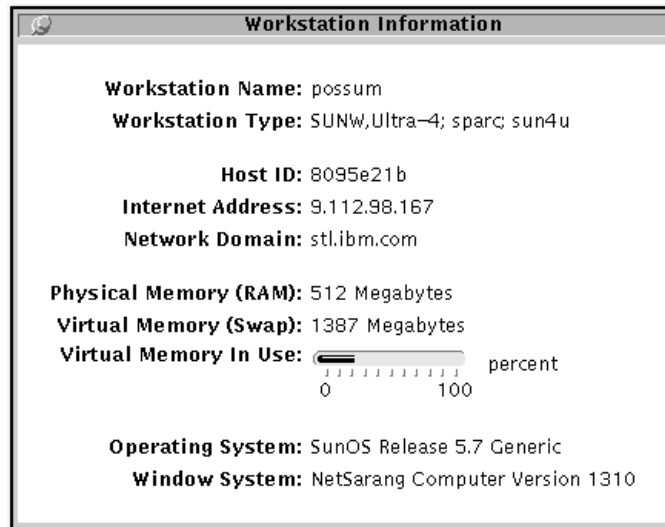


Figure 75. Sun Solaris Workstation Information

You can also check the host name and internet address by entering the command `admintool` on the command line. This command displays a graphic user interface as shown in Figure 76.

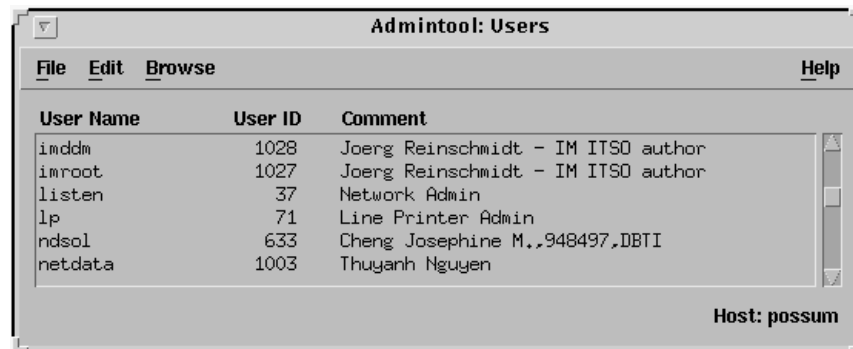


Figure 76. Admintool on Sun Solaris

The **Browse -> Hosts** menu selection is shown in Figure 77. You can add, delete modify from the **EDIT** menu.

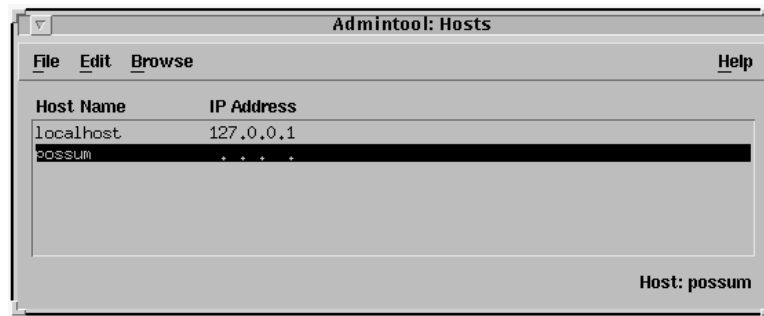


Figure 77. Host Information on SUN Solaris

3. You can check the TCP/IP status of the system. To do this, enter the `ps -ef | grep inetd` command from the command line. You must see the running *inetd* process as shown below.

```
#ps -ef | grep inetd
root 147 1 0 Feb 28 ? 0:01 /usr/sbin/inetd -s
```

4. Verify the TCP/IP connection from the client.

You can check the connection from the client by entering the command `PING HOSTNAME` on the command line.

8.2 Product Installation

To install the IM server:

1. Log in as a root.
2. Create a IM user group using the `admintool` command.

Select **Browse -> Groups** and then select the **Edit -> Add** menu. Figure 78 shows the add group panel. Enter your IM group name (for example, imgroup), group ID and member lists, then click **OK**.



Figure 78. Add Group on Sun Solaris

3. Create the IM user.

Select **Browse -> Users** from the Admintool panel and then select **Edit -> Add** from the menu to get to the Add User panel as shown in Figure 79.

The screenshot shows the 'Admintool: Add User' window. It has three main sections: 'USER IDENTITY', 'ACCOUNT SECURITY', and 'HOME DIRECTORY'. In the 'USER IDENTITY' section, 'User Name' is empty, 'User ID' is '1006', 'Primary Group' is '0', 'Secondary Groups' is empty, 'Comment' is empty, and 'Login Shell' is 'Bourne /bin/sh'. In the 'ACCOUNT SECURITY' section, 'Password' is 'Cleared until first login', 'Min Change', 'Max Change', and 'Max Inactive' are all empty with 'days' as a unit, 'Expiration Date' is 'None', and 'Warning' is empty with 'days' as a unit. In the 'HOME DIRECTORY' section, 'Create Home Dir' is unchecked and 'Path' is empty. At the bottom are buttons for 'OK', 'Apply', 'Reset', 'Cancel', and 'Help'.

Figure 79. Add User on Sun Solaris

Type your IM user name (for example, *imadm*), user ID, primary group (for example, *imgroup*), home directory, and so forth. Select your login shell and password option.

Password specifies the means by which a user sets up a password. The followings describe the password options.

- **Cleared until first login** -The Account will not have a password. The user must manually set a password, using the `passwd` command, after first logging in.

- **Account is locked** - The account is locked. The user will not be able to log in until the administrator assigns a password.
- **No password--setuid only** - The account cannot be logged in to, but account programs are allowed to run.
- **Normal password** - The administrator assigns a password to the account when adding the user.

If you created an IM user with **Cleared until first login** option then log in as an IM user and change your password with `passwd` command manually before either running IM Server or connecting from the client site. If you created an IM user with **Normal password** option then you are required to log in as an IM user to verify your new password before either running IM Server or connecting from the client site.

4. Insert the server CD-ROM in the CD-ROM drive.

If the Volume Manager is installed, the CD-ROM is automatically mounted as `/cdrom/IMDSS213`

If the Volume Manager is not installed, mount the CD-ROM by entering the commands:

```
mkdir -p /cdrom/IMDSS213
mount -f ufs -r /dev/dsk/c0t60s2 /cdrom/IMDSS213
```

5. Change directory to `/cdrom/IMDSS213` with the following command:

```
cd /cdrom/IMDSS213
```

6. Type the following command and press enter to install both the IM server and the toolkit:

```
./imininstall
```

To install only the IM server, enter the command:

```
pkadd -a ./admin -d . IMiner.
```

To install only the IM toolkit (ADT), enter the command:

```
pkadd -a ./admin -d . IMinerTK.
```


Note

The IM server for Sun Solaris includes demonstration data. You can use the demonstration data when you start the IM in demo mode.

The English demonstration data is installed with the IM server package (IMiner). You can replace the English demonstration data with data in one of the supported national languages. For example, to replace the English demonstration data with Korean demonstration data, enter the following command:

```
pkgadd -a ./admin -d . IMdemoKR
```

7. Create links to the IM executables.

You can use one of the following methods to invoke the executables:

- Create links in the /usr/bin directory to the executables in the /opt/IMiner/bin directory.
- Add the path /opt/IMiner/bin to your system environment.
- Change to the /opt/IMiner/bin directory before you invoke the executables.

To create links to the executables, log in as user root and enter the following command:

```
/opt/IMiner/bin/imln
```

To add the path to your environment, set the /opt/IMiner/bin directory to the search path of your profile. Your profile is one of the following, depending on the shell that you use on your Solaris system:

- For sh and ksh: \$HOME/.profile
- For csh and tcsh: \$HOME/.login

8.3 Installation Verification

IM uses the environment variables IDM_MNB_DIR and IDM_RES_DIR, which point to the directories containing the mining bases and result files respectively.

If you do not set these variables and use the default values, directories named idmmnb and idmres are created in each user's home directory.

Before starting the IM server, you can set the environment variable `IDM_MNB_DIR`, `IDM_RES_DIR` to point to one common directory if multiple users share the same mining bases and result files.

To specify a common directory for mining bases on the Korn shell, enter the following command before starting the IM server:

```
export IDM_MNB_DIR=/miningbases
```

Note

Only one user at a time can have write access to a shared mining base.

To specify a common directory for result files on the Korn shell, enter the following command before starting the IM server:

```
export IDM_RES_DIR=/resultfiles
```

To verify that the IM server was successfully installed, enter the command `idmstartdemo` on the command line and access the IM server from the client, then run the sample data mining functions.

8.4 Running the Server

To start Sun Solaris server:

1. Log on as user root or as another user.
2. Set the `DB2INSTANCE` variable to the DB2 UDB instance.
3. Type one of the following commands on the command line:

```
idmstart -d
```

This command also shows tracing information of the AIX server.

```
idmstart
```

This command starts the AIX server without showing tracing information.

4. Log out.
5. Call `'im'` on the client to test your installation.

Chapter 9. Implementation on OS/400

This chapter describes the steps required to install Intelligent Miner for Data on an OS/400 system. We installed the server on our system AS01 and clients on several other systems. Figure 80 shows the system used.

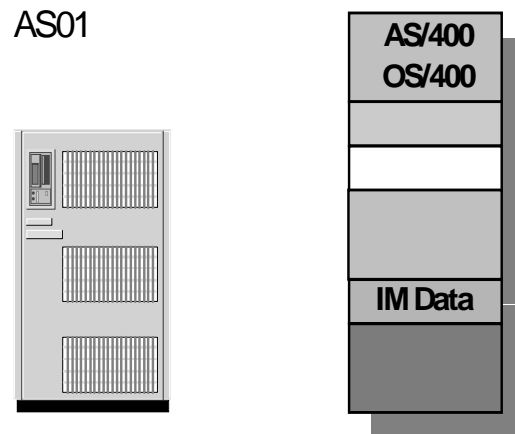


Figure 80. Intelligent Miner for Data on OS/400

9.1 Prerequisites

This section lists the prerequisites necessary for installation. It also describes how to verify whether your system is ready to run the server and client software.

9.1.1 Hardware Requirements

The Intelligent Miner for AS/400 requires an AS/400 Advanced Series RISC System. We advise using a system with 256 MB or more of main storage and 250 MB of available DASD. You also need a CD-ROM device to perform the installation.

9.1.2 Software Prerequisites

IBM OS/400 Version 4 Release 2 (Program Number 5769-SS1) is required. To check your OS/400 version, enter the `DSPSFWRSC` command.

The output from the command lists the resources on your system.

Display Software Resources				System:	AS01
Resource					
ID	Option	Feature	Description		
5769999	*BASE	5050	AS/400 Licensed Internal Code		
5769SS1	*BASE	5050	Operating System/400		
5769SS1	*BASE	2924	Operating System/400		
5769SS1	1	5050	OS/400 - Extended Base Support		
5769SS1	1	2924	OS/400 - Extended Base Support		
5769SS1	2	5050	OS/400 - Online Information		
5769SS1	2	2924	OS/400 - Online Information		
5769SS1	3	5050	OS/400 - Extended Base Directory Support		
5769SS1	3	2924	OS/400 - Extended Base Directory Support		
5769SS1	7	5050	OS/400 - Example Tools Library		
5769SS1	7	2924	OS/400 - Example Tools Library		
5769SS1	8	5050	OS/400 - AFP Compatibility Fonts		
5769SS1	9	5050	OS/400 - *PRV CL Compiler Support		
5769SS1	9	2924	OS/400 - *PRV CL Compiler Support		
				More...	
Press Enter to continue.					
F3=Exit F11=Display libraries/releases F12=Cancel					
F19=Display trademarks					

When the panel above is displayed, place the cursor on the line Operating System/400 and press **F11** to see the version information for the product mentioned in that line.

Display Software Resources						System:	AS01
Resource			Feature				
ID	Option	Feature	Type	Library	Release		
5769999	*BASE	5050	*CODE	QSYS	V4R3M0	L00	
5769SS1	*BASE	5050	*CODE	QSYS	V4R3M0	L00	
5769SS1	*BASE	2924	*LNG	QSYS	V4R3M0	L00	
5769SS1	1	5050	*CODE	QSYS2	V4R3M0		
5769SS1	1	2924	*LNG	QSYS2	V4R3M0		
5769SS1	2	5050	*CODE	QHLPSSYS	V4R3M0		
5769SS1	2	2924	*LNG	QHLPSSYS	V4R3M0		
5769SS1	3	5050	*CODE	QSYSDIR	V4R3M0		
5769SS1	3	2924	*LNG	QSYSDIR	V4R3M0		
5769SS1	7	5050	*CODE	QUSRTOOL	V4R3M0		
5769SS1	7	2924	*LNG	QUSRTOOL	V4R3M0		
5769SS1	8	5050	*CODE	QFNICPL	V4R3M0		
5769SS1	9	5050	*CODE	QSVSV4R2M0	V4R3M0		
5769SS1	9	2924	*LNG	QSVSV4R2M0	V4R3M0		
						More...	
Press Enter to continue.							
F3=Exit	F11=Display descriptions		F12=Cancel	F19=Display trademarks			

In the example shown above, the operating system is Version 4 Release 3.

Table 16 lists the PTFs required for IM when installed on different versions of OS/400.

Table 16. OS/400 Required PTFs

OS/400 Version and Release	PTF required
V4R2	SF50975
V4R3	SF51008

To check whether the PTFs are installed on your system, use the `DSPPTF` command.

The command displays the following output:

```
Display PTF Status                                     System:  AS01

Product ID . . . . . : 5769999
IPL source . . . . . : ##MACH#B
Release of base option . . . . . : V4R3M0 L00

Type options, press Enter.
  5=Display PTF details  6=Print cover letter  8=Display cover letter

PTF                                     IPL
Opt ID      Status                     Action
TL99054     Temporarily applied        None
TL98349     Permanently applied         None
TL98279     Permanently applied         None
TL98230     Permanently applied         None
TL98202     Superseded                  None
MF21462     Temporarily applied         None
MF21052     Temporarily applied         None
MF21001     Temporarily applied         None
MF20989     Permanently applied         None

F3=Exit  F11=Display alternate view  F17=Position to  F12=Cancel  More...
```

9.1.3 Networking Requirements

The intelligent Miner needs a working TCP/IP connection from the client to the server. Before starting the installation, check your server configuration by entering the `CHGTCPDMN` command and pressing **F4**.

```

Change TCP/IP Domain (CHGTCPDMN)

Type choices, press Enter.

Host name . . . . . 'AS400'

Domain name . . . . . 'ibm.com'


Host name search priority . . . *REMOTE      *REMOTE, *LOCAL, *SAME
Domain name server:
  Internet address . . . . . '127.0.0.1'
                               '9.9.9.9'


F3=Exit   F4=Prompt   F5=Refresh   F10=Additional parameters   F12=Cancel
F13=How to use this display   F24=More keys
Bottom

```

If there is no configuration information, contact your system administrator. Check your TCP/IP setup by pinging your server from one of the supported clients. In the case shown above, you would enter the `ping as400.ibm.com` command.

You also need the RPC server for client/server communication. To start it, you need `*IOSYSCFG` authority. You can check for this authorization by using the `WRKUSRPRF` command. Select your own username, choose Option 5 then check the second page of the profile settings under special authority. You can also, check whether your username is listed in the system distribution directory using the `WRKDIR` command. You can add your profile to this directory with the `ADDIRE` command if you are not listed. Start the RPC server with the `STRNFSSVR *RPC` command.

9.1.4 Relational Databases

Check your system configuration with the `DSPNETA` command. The entries listed as current system name, local control point name, and default local location must be the same. In the example below this name is `AS400`.

```
Display Network Attributes

Current system name . . . . . : AS400
Pending system name . . . . . :
Local network ID . . . . . : ITSCNET
Local control point name . . . . . : AS400
Default local location . . . . . : AS400
Default mode . . . . . : BLANK
APPN node type . . . . . : *NETNODE
Data compression . . . . . : *NONE
Intermediate data compression . . . . . : *NONE
Maximum number of intermediate sessions . . . . . : 200
Route addition resistance . . . . . : 128
Server network ID/control point name . . . . . : *LCLNETID *ANY

System: AS01

Press Enter to continue.

F3=Exit F12=Cancel

More...
```

To check whether local databases are configured for access as relational databases, use the `WRKRDBDIRE` command.

```
Work with Relational Database Directory Entries

Position to . . . . .

Type options, press Enter.
  1=Add  2=Change  4=Remove  5=Display details  6=Print details

      Relational      Remote
Option Database      Location      Text

      AS400          *LOCAL

F3=Exit  F5=Refresh  F6=Print list  F12=Cancel
(C) COPYRIGHT IBM CORP. 1980, 1998.

Bottom
```

There must be an entry with the same name as the current system name from the previous screen with `*LOCAL` listed as the remote location.

9.2 Product Installation

To install the IM server, sign on to the AS/400 system with a user profile that has similar authority as `QSECOFR` and ensure that the system operator message queue is in break mode. You can do this by entering the following command:

```
CHGMSGQ QSYSOPR *BREAK
```

Insert the distribution CD into the CD-ROM drive and check whether your drive is active using the following command (assuming your drive is `OPT01`):

```
WRKCFGSTS *DEV OPT01
```

```
Work with Configuration Status          AS01          03/02/99 12:03:59
Position to . . . . . Starting characters

Type options, press Enter.
  1=Vary on   2=Vary off   5=Work with job   8=Work with description
  9=Display mode status   13=Work with APPN status...

Opt  Description      Status      -----Job-----
OPT01                                ACTIVE

Parameters or command                                     Bottom
===>
F3=Exit  F4=Prompt  F12=Cancel  F23=More options  F24=More keys
```

To install IM using the primary language of your system, use the following command:

```
RSTLICPGM LICPGM(5733IM2) DEV(OPT01)
```

To install in any other language, use the command:

```
RSTLICPGM LICPGM(5733IM2) DEV(OPT01) LNG(xxxx)
```

Where `xxxx` is the desired language. Language codes follow the AIX convention, for example, `en_US` for US English. An overview of codes can be found in the `README` member of `QYDMREADME` in library `QIDM` after installation.

To install an additional language after installation of the server, use the following command:

```
RSTLICPGM LICPGM(5733IM2) DEV(OPT01) LNG(xxxx) RSTOBJ(*LNG)
```

The installation procedure creates a library, *QIDM*, on your system. Any user profile or job that uses IM must have this library in its library list. You can check this with the *DSPLIBL* command. If necessary, add the library using the *ADDLIBL QIDM* command.

9.3 Installation Verification and Starting the Server

Start the Intelligent Miner server with the *STRIDM* command. This will start a job called *QYDMIDMD* in the *QSYSWRK* subsystem. The job runs *IDMD*, using the job description *QGPL/QDFTJOB*. The user starting this job will need **ALLOBJ* authority in the home directory of the client user profile. You can check whether the server is running using the *WRKUSRJOB* command

```
Work with User Jobs                                AS01                                03/04/99 18:28:05

Type options, press Enter.
  2=Change   3=Hold   4=End   5=Work with   6=Release   7=Display message
  8=Work with spooled files   13=Disconnect

Opt  Job          User          Type      -----Status-----  Function
    QYDMIDMD      IMADM        BATCH      OUTQ
    QYDMIDMD      IMADM        BATCH      OUTQ
    QYDMIDMD      IMADM        BATCH      ACTIVE                PGM-IDMD
    QYTCNSLD      IMADM        BATCH      ACTIVE                PGM-QYTCNSLD
    QZSHSH        IMADM        BATCHI     OUTQ
    QZSHSH        IMADM        BATCHI     OUTQ

                                                                 Bottom

Parameters or command
====>
F3=Exit      F4=Prompt   F5=Refresh   F9=Retrieve   F11=Display schedule data
F12=Cancel   F21=Select assistance level
```

This output above shows a job *QYDMIDMD* with status *ACTIVE* .

After starting the server, try to connect and run a data mining job from one of the supported clients to verify your installation.

If you want the server and system to start together, your system administrator must add the *STRIDM* command to the startup program. Make sure that the

RPC server is started before the IM server, as described in 9.1.3 on page 131.

Start the server in demo mode with the command:

```
CALL QIDM/QYDMDEMO PARM(<language> <directory>)
```

Substitute your language identifier for <language> (using AIX conventions such as *en_US* for US English as shown in section 9.2 on page 134) and the installation directory of the demo data for <directory> (for example, *'/home/<user>/idmnmb'*). The server will now display the demonstration mining bases and data to all connecting clients.

If you need to stop the server, use the `ENDIDM *CNTRL` command.

Chapter 10. Implementation on OS/390

This chapter contains information about the prerequisites, the installation procedures, and the customizing of Intelligent Miner for Data for OS/390 V2. Figure 81 shows the system used with the software installed.

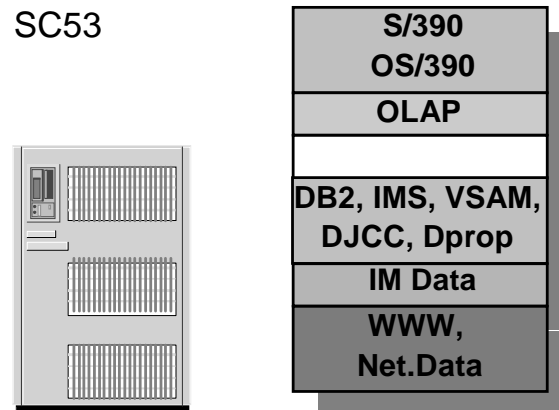


Figure 81. Intelligent Miner for Data on OS/390

The following list of publications in Table 17 will help you to implement your IM on OS/390.

Table 17. OS/390 Publications Useful During Installation

Publication Title	Form Number
OS/390 SMP/E Users Guide	SC28-1740
OS/390 SMP/E Commands	SC28-1805
OS/390 SMP/E Reference	SC28-1806
OS/390 Security Server (RACF) Command Language Reference	SC28-1919
OS/390 V2R5.0 OpenEdition User's Guide	SC28-1891-04
Accessing OS/390 OpenEdition MVS from the internet	SG24-4721
DB2 for MVS/ESA V4 Command Reference	SC26-3267

Note

Do not use this redbook section if you are installing IM server with an MVS Custom-Built Installation Process Offering (CBIPO), SystemPac, or ServerPac. When using these offerings, use the jobs and documentation supplied.

If you install Intelligent Miner for Data V2 using the CBPDO (5751-CS3), your CBPDO contains a softcopy preventive service planning (PSP) upgrade for this product. All service and HOLDDATA for IM are included on the CBPDO tape.

10.1 Installation Planning

This section describes the IBM support available and prerequisite values while you are installing IM.

The following terminology is used in this chapter.

- *Driving system*: the system used to install the program.
- *Target system*: the system on which the program is installed.

10.1.1 IBM Service Information

Before installing IM, you should review the current Preventive Service Planning(PSP) information. If you obtained IM as part of a CBPDO, there is HOLDDATA and PSP information included on the CBPDO tape. If you obtained IM on a product tape, or if the CBPDO is more than two weeks old when you install it, you should contact the IBM Support Center or SoftwareXcel to obtain the current PSP Bucket. PSP Buckets are identified by UPGRADEs, which specify product levels, and SUBSETs, which specify the function modification identifiers (FMIDs) for a product level. The UPGRADE and SUBSET values for IM are shown in Table 18.

Table 18. OS/390 PSP Upgrade and Subset ID

UPGRADE	SUBSET	Description
INTMINOS390	H258100	IM Base

Table 19 identifies the component IDs (COMPID) for IM.

Table 19. OS/390 Component IDs

FMID	COMPID	Component Name	RETAIN Release
H258100	5655IM200	DATA MINER FOR OS/390	100

The following program temporary fix (PTFs) of IM have been incorporated into this release and are listed by FMID below:

- FMID H258100
UN99732 UN99936 UQ00572 UQ01555 UQ02555
UQ02704 UQ03220 UQ05849 UQ05878 UQ09930
UQ10659 UQ11197 UQ12410 UQ12434

Note

IM for Data OS/390 V2.1 (Program Number : 5647A01) requires the IM client to be on the matching service level where:

IM Client 2.1.2 requires IM for Data OS/390 PTFs: UQ21485 and UQ21852

IM Client 2.1.3 requires IM for Data OS/390 PTFs: UQ23830, UQ23887, and UQ23938 in addition to the 2.1.2 PTFs.

We do recommend, however, that you use the latest level (Version 2.1.3), because it contains the latest available fixes.

10.1.2 Installation Planning Values

Before performing the installation, you must decide on the values for several variables, if they are not already defined . Table 20 shows the values that should be assigned before implementing IM.

Table 20. OS/390 Installation Planning Table

Name	Your Value	Example Value
<i>Driving system</i> SYSAFF		SC53
<i>Target system</i> SYSAFF		SC52
IM server host name (OpenEdition)		wtsc53oe
IM server IP address (OpenEdition)		
OMVS group name		OMVSGRP

Name	Your Value	Example Value
OMVS GID	0	0
Default group name		SYS1
Default group ID		0
IM group name		IMGROUP
IM group ID	0	0
IM user name (Administrator)		IMSERV
IM user id (Administrator)	0	0
IM user (Administrator) home directory		/u/imserv
IM user name (Client)		IMUSER
IM user id (Client)		250
IM user (Client) home directory		/u/imuser
BPXPRMxx member name		BPXPRM01
STEPLIBLIST		/system/steplib

10.2 Understanding the Prerequisites

This section describes the hardware and software requirements for IM for OS/390.

10.2.1 Hardware Requirements

The *driving system* can run in any hardware environment that supports the required software shown in Table 21.

Table 21. OS/390 Driving System Software Prerequisites

Program Number	Product Name and Minimum VRM/Service Level
5645-001	OS/390 Version 1.3

The *target system* requires the following hardware environment in order for you to install and use IM.

- A processor that supports OS/390 V1.3
- A nine-track tape unit or a cartridge unit to install the product

IM libraries can reside on any DASD supported by OS/390 V1.3 (see Table 22).

Table 22. OS/390 Total DASD Space Required

Library Type	Total Space required
Target	84MB
Distribution	87MB
HFS	14MB

Note

If you have a previous release of Intelligent Miner for Data OS/390 product installed in these libraries, the installation of this release will delete the old installation and reclaim the space used by the old release and any service that had been installed. You can determine whether or not there is enough space for these libraries by deleting the old release with a dummy function, compressing the libraries, and comparing the space requirements with the remaining free space in the libraries.

10.2.2 Software Prerequisites

The following software in Table 23 is required on the target system:

Table 23. OS/390 Target System Software Prerequisites

Software	Version	Program Number	Required/Optional
IBM OS/390	1.3.0 (or higher)	5647-A01	Required
TCP/IP	3.1	5647-A01	Required
TCP/IP OpenEdition MVS Applications	3.1.0	5655-HAL	Required
Security Server or RACF	2.1	5695-039	Optional
DB2 for MVS/ESA	5.1	5695-DB2	Optional

Please check the version information with your system administration or contact the IBM Support Center.

10.2.3 Networking Requirements

The following sections identify the networking system requirements for IM. In many cases, the same system can be used as both a *driving system* and a *target system*. However, you might want to set up a clone of your system to use as a target system by making a separate IPL-able copy of the running system. The clone system should include copies of all system libraries that SMP/E updates, copies of the SMP/E CSI data sets that describe the system libraries, and your PARMLIB and PROCLIB data sets.

If OpenEdition MVS and TCP/IP are being used for the first time, we highly recommend that a local OS/390 specialist with skills in OpenEdition MVS and TCP/IP be enlisted to assist. The redbook, *Accessing OS/390 OpenEdition MVS from the Internet*, SG24-4721, provides guidance on configuring TCP/IP with emphasis on accessing OpenEdition.

If TCP/IP V3R1 for MVS is already installed on your system, it will, in many cases, be useful to set up a separate TCP/IP started task for OpenEdition MVS services.

Note

You can run one or more TCP/IP address space copies on a single MVS system. To concurrently connect more than one TCP/IP instance to OpenEdition MVS, you must configure the OpenEdition MVS to use Common I-Net (C-INET) AF_INET PFS rather than integrated sockets AF_INET PFS, which does not support multiple TCP/IP instances connected to OpenEdition MVS. When DB2 uses TCP/IP, however, you can have only one TCP/IP instance connect to OpenEdition MVS on the same MVS system, because DB2 uses asynchronous I/O services, and C-INET AF_INET PFS does not support asynchronous I/O. Therefore the integrated sockets AF_INET PFS must be used. In other words, you must configure OpenEdition MVS to use the integrated sockets AF_INET PFS when using DB2 with TCP/IP. Refer to the *TCP/IP for MVS: Customization and Administration Guide*, SC31-7134 for further consideration of multiple copies of TCP/IP.

10.2.4 Verifying the TCP/IP Configuration

Normally the name of the transport providers implemented as started tasks are TCPIPMVS (the default transport provider for traditional MVS process) and TCPIPOE (the default transport provider for OpenEdition MVS process). Contact your system administrator to determine the implemented transport

provided on your system. For example, you can view the started tasks for the TCP/IP on the **System Display and Search Facility (SDSF)**.

Enter the command `DA` on the *COMMAND INPUT* line as shown below.

```
Display Filter View Print Options Help
-----
HGX1900----- SDSF PRIMARY OPTION MENU -- INVALID COMMAND
COMMAND INPUT ===> DA                                SCROLL ===> CSR

LOG      - Display the system log
DA       - Display active users in the sysplex
I        - Display jobs in the JES2 input queue
O        - Display jobs in the JES2 output queue
H        - Display jobs in the JES2 held output queue
ST       - Display status of jobs in the JES2 queues
PR       - Display JES2 printers on this system
INIT     - Display JES2 initiators on this system
MAS      - Display JES2 members in the MAS
LINE     - Display JES2 lines on this system
NODE     - Display JES2 nodes on this system
SO       - Display JES2 spool offload for this system

Licensed Materials - Property of IBM

5647-A01 (C) Copyright IBM Corp. 1981, 1997. All rights reserved.
US Government Users Restricted Rights - Use, duplication or
F1=HELP   F2=SPLIT   F3=END     F4=RETURN  F5=IFIND   F6=BOOK
F7=UP     F8=DOWN    F9=SWAP    F10=LEFT   F11=RIGHT  F12=RETRIEVE
```

The following screen lists all the currently active jobs on the system.

Display Filter View Print Options Help											

SDSF DA SC53 SC53 PAG 0 SIO 0 CPU 12/ 6 LINE 69-79 (79)											
COMMAND INPUT ==>> SCROLL ==>> CSR											
NP	JOBNAME	STEPNAME	PROCSTEP	JOBID	OWNER	C	POS	DP	REAL	PAGING	SIO
	DBC2SPAS	DBC2SPAS	IEFPROC	STC23276	STC		NS	FE	730	0.00	0.00
	DB2RES4	IKJACCN	SCGPVM14	TSU24491	DB2RES4		IN	F9	845	0.00	0.00
	TCPIPOE	TCPIPOE	TCPIP	STC24509	TCPIPOE		NS	FE	1793	0.00	0.00
	KARRAS	IKJACCN	SCGSA065	TSU24503	KARRAS		LO	FF	775	0.00	0.00
	KMT2	IKJACCN	TCP66006	TSU24500	KMT2		LO	FF	702	0.00	0.00
	DB2IMSTR	DB2IMSTR	IEFPROC	STC24517	STC		NS	FE	807	0.00	0.00
	MVSINFSC5	MVSINFSC5	MVSCLNT	STC22804	STC		NS	FE	5398	0.00	0.00
	DB2IDIST	DB2IDIST	IEFPROC	STC24521	STC		NS	FE	1878	0.00	0.00
	DB2IDEM1	DB2IDEM1	IEFPROC	STC24520	STC		NS	FE	4721	0.00	0.00
	IRLIPROC	IRLIPROC		STC24518	STC		NS	FE	216	0.00	0.00
	TCPIPMVS	TCPIPMVS	TCPIP	STC24508	TCPIPOE		NS	FE	2287	0.00	0.00
F1=HELP			F2=SPLIT		F3=END		F4=RETURN		F5=IFIND		F6=BOOK
F7=UP			F8=DOWN		F9=SWAP		F10=LEFT		F11=RIGHT		F12=RETRIEVE

You can see the multiple transport driver support environment (**TCPIPOE**, **TCPIPMVS**) on the sample screen above.

After completing the steps above, you should verify your system's host name and address configuration for TCP/IP OpenEdition (**TCPIPOE**) as follows:

You can check your home IP address from the TSO command panel by entering the command `NETSTAT HOME TCP TCPIPOE`, which is for TCP/IP OpenEdition.

The status of a TCP/IP transport provider is monitored with the TSO `NETSTAT` command. You can specify the name of the transport providers' started task to be monitored if you have multiple transport providers installed

```
Menu List Mode Functions Utilities Help

                                ISPF Command Shell
Enter TSO or Workstation commands below:

====> NETSTAT HOME TCP TCPIPOE

Place cursor on choice and press enter to Retrieve command

=>
=>
=>
=>
=>
=>
=>
=>
=>
=>

F1=Help      F3=Exit      F10=Actions  F12=Cancel
```

The following screen shows the result of the NETSTAT HOME TCP TCPIPOE command.

```
MVS TCP/IP NETSTAT CS/390 V2R5      TCPIP NAME: TCPIPOE      19:32:24
Home address list:
Address      Link      Flg
-----
9.9.9.18    OSAL2160    P
127.0.0.1    LOOPBACK
***
```

Alternatively, you can check for the host name and IP address on the **OpenMVS ISPF Shell** panel.

```

File Directory Special_file Tools File_systems Options Setup Help
-----
OpenMVS ISPF Shell

Enter a pathname and do one of these:

- Press Enter.
- Select an action bar choice.
- Specify an action code or command on the command line.

Return to this panel to work with a different pathname.
More:      +

=====
=====
=====

Command ==> sh cat /etc/hosts
F1=Help      F3=Exit      F5=Retrieve  F6=Keyshelp  F7=Backward  F8=Forward
F10=Actions  F11=Command  F12=Cancel

```

Enter the shell command `sh cat /etc/hosts` or `ex cat /etc/hosts` to view the hosts list. The following screen shows the result of the `sh cat /etc/hosts` command.

```

BROWSE -- /tmp/DB2RES4.15:00:01.121320.ishell ----- Line 00000000 Col 001 080
Command ==>                                     Scroll ==> PAGE
***** Top of Data *****
-----
Set up environment variables for Java and Servlets for OS/390 -
-----
PATH reset to /usr/lpp/java/J1.1/bin:/bin:.
-----
CLASSPATH reset to ./usr/lpp/java/J1.1/lib/classes.zip:/usr/lpp/internet/server
-----
--> Path set for JAVA Servlet support
--> LD PATH set for JDBC support
9.9.9.18      wtsc53oe wtsc53oe.itso.ibm.com
***** Bottom of Data *****

F1=HELP      F2=SPLIT      F3=END      F4=RETURN      F5=RFIND      F6=RCHANGE
F7=UP        F8=DOWN       F9=SWAP     F10=LEFT       F11=RIGHT     F12=RETRIEVE

```

10.2.5 Networking Between the IM Server and Client

Before you try to use the IM for Data Client, verify the connection between the client and the server using the `PING` command or equivalent tools. If you cannot ping the server from the client, the IM for Data Client will not be able to communicate with the server. A good test of the TCP/IP and OpenEdition installation is to use the file transfer protocol (FTP) a hierarchical file system (HFS) file from the server to the client.

To verify that the PORTMAP server has been configured and started, enter the `sh netstat` command on the OpenMVS ISPF Shell panel.

The screen below shows the results of this command. The state of **Listen** means that the defined daemon is running.

```
BROWSE -- /tmp/DB2RES4.16:23:34.852586.ishell ----- Line 00000000 Col 001 080
Command ==>                                         Scroll ==> PAGE
***** Top of Data *****

-----
Set up environment variables for Java and Servlets for OS/390 -
-----
PATH reset to /usr/lpp/java/J1.1/bin:/bin:.
-----
CLASSPATH reset to ./usr/lpp/java/J1.1/lib/classes.zip:/usr/lpp/internet/server
-----
--> Path set for JAVA Servlet support
--> LD PATH set for JDBC support
MVS TCP/IP onetstat CS/390 V2R5          TCPIP Name: TCPIPOE          16:23:38
User Id  Conn  Local Socket          Foreign Socket          State
-----
DBC2DIST 00024 9.12.2.18..33330      0.0.0.0..0             Listen
DB2EDIST 00025 9.12.2.18..33309      0.0.0.0..0             Listen
DB2IDIST 053FB 0.0.0.0..33320        0.0.0.0..0             Listen
DB2JDIST 00022 9.12.2.18..33323      0.0.0.0..0             Listen
FTPDOEL 0002F 0.0.0.0..21           0.0.0.0..0             Listen
INETDL 00020 0.0.0.0..23           0.0.0.0..0             Listen
PMAPOEL 00036 0.0.0.0..111          0.0.0.0..0             Listen
TCPIPOE 0000B 0.0.0.0..1025         0.0.0.0..0             Listen
TCPIPOE 0000F 127.0.0.1..1025       127.0.0.1..1026        Establish
PMAPOEL 00035 0.0.0.0..111          *.*                     UDP
***** Bottom of Data *****

F1=HELP      F2=SPLIT     F3=END       F4=RETURN    F5=RFIND     F6=RCHANGE
F7=UP        F8=DOWN      F9=SWAP      F10=LEFT     F11=RIGHT    F12=RETRIEVE
```

10.3 Product Installation and Customization

This section describes the installation method and step-by-step procedures to activate the functions of IM. As you begin the installation process, consider these SMP/E suggestions:

- To install IM into its own SMP/E environment, consult the SMP/E publications for instructions on creating and initializing the SMPCSI and the SMP/E control data sets.
- Sample jobs have been provided to help perform some or all of the installation tasks. The SMP/E jobs assume that all DDDEF entries required for SMP/E execution have been defined in the appropriate SMP/E zones.
- The SMP/E dialogs can be used instead of the sample jobs to accomplish the SMP/E installation steps.

10.3.1 Installation Procedure

To install IM on the system, follow these steps:

1. Log in to the *driving system* as a system administrator with OpenEdition MVS authorization uid=0 (super user).
2. Ensure that the SMP/E environment is pointing to the correct zones and that the libraries are set up correctly.

IM for Data V2 is installed using the SMP/E RECEIVE, APPLY and ACCEPT commands. The SMP/E dialogs can be used to accomplish the SMP/E installation steps.

The SMP/E installation jobs provided assume that all necessary DD statements for the execution of SMP/E are defined using DDDEFs. Sample jobs are provided to assist you in installing IM. After the RECEIVE step has completed, these sample jobs can be found in SMPTLIB in the data sets:

hlq.IBM.H258100.F7

Copy these jobs into your own library and modify them so they can be used during the installation of IM. Table 24 shows the sample jobs.

Table 24. OS/390 Sample Job List

JOB Name	Description
IDMRECEV	Sample RECEIVE job
IDMALLOC	Sample job to allocate target and distribution libraries

JOB Name	Description
IDMDDDEF	Sample job to define SMP/E DDDEFs
IDMHFS	Sample job to initialize HFS libraries
IDMAPPCK	Sample APPLY CHECK job
IDMAPPLY	Sample APPLY job
IDMDB2	Sample job to bind the DB2 plan (optional)
IDMDEMO	Sample job to install the Demo Package
IDMVERIFY	Sample installation verification job
IDMACCCK	Sample ACCEPT CHECK job
IDMACCEP	Sample ACCEPT job
IDMCFLD	Sample job to activate customer defined computed fields (optional)
IDMSECUR	Sample job to verify the security environment
IDMSTART	Sample job to start the IM server

The recommended values for some SMP/E CSI subentries are shown in Table 25. Using values lower than these might result in a failure during the installation process. DSSPACE is a subentry in the GLOBAL options entry and PEMAX is a subentry of GENERAL in the GLOBAL options entry. For further information, refer to the SMP/E publications for instructions on updating the global zone.

Table 25. OS/390 SMP/E Options Subentry Values

Subentry	Value	Comment
DSSPACE	(200,100,200)	
PEMAX	9999	The SMP/E default is larger than what can be specified here

IM uses the CALLLIBS function provided in SMP/E Release 8 to resolve external references during the installation. When IM is installed, ensure that you have done the following:

- Verify that the SMP/E SMPLTS data set has been allocated. Refer to the *SMP/E Reference* for information on allocating the SMPLTS data set.
- Provide DDDEFs for the SCEELKED library.

Note

The DDDEFs mentioned above are only used to resolve the link-edit for IM using CALLLIBS. These data sets are not updated during the installation of IM.

3. Unload the sample JCL from the product tape.

Sample installation jobs are provided on the distribution tape to help you install IM. The following sample JCL will copy the IM jobs from the tape. Add your appropriate job card and change the parameters shown in boldface to uppercase values to meet your site's requirements before submitting the job.

```
// STEP EXEC PGM=IEBCOPY
// SYSPRINT DD SYSOUT=A
// IN DD DSN=IBM.H258100.F7,UNIT=tunit, VOL=SER=258100,
// LABEL=(8,SL),DISP=(OLD,KEEP)
// OUT DD DSNAME=jcl-library-name,
// DISP=(NEW,CATLG,DELETE),
// VOL=SER=dasdvol, UNIT=dunit,
// DCB=*.STEP1.IN,SPACE=(8800,(100,10,5))
// SYSUT3 DD UNIT=SYSDA, SPACE=(CYL,(1,1))
// SYSIN DD *COPY INDD=IN, OUTDD=OUT
/*
```

Where

- **tunit** is the unit value matching the product tape or cartridge
- **jcl-library-name** is the name of the data set where the sample jobs will reside
- **dasdvol** is the volume serial number of the DASD where the data set will reside
- **dunit** is the DASD unit type of the volume.

If the SMS product is installed, the parameters **VOL=SER** and **UNIT** should be deleted.

4. Perform the SMP/E RECEIVE.

Edit and submit the sample job IDMRECEV to perform the SMP/E RECEIVE for IM. For more information, refer to Appendix B.1, “IDMRECEV” on page 179.

Note

If you obtained IM for data V2 as part of a CBPDO, you can use the RCVPDO job found in the CBPDO RIMLIB data set to RECEIVE the IM for Data V2 FMIDs as well as any service, HOLDDATA, or preventive service planning (PSP) information included on the CBPDO tape. For more information, refer to the documentation included in the CBPDO.

This job should complete with a return code of 0.

You can also access the sample installation jobs by performing an SMP/E RECEIVE job for *FMID*, then copying the jobs from data set *hlq.IBM.H258100.F7* to a work data set for editing and submission.

5. Receive the cumulative service tape. (optional)

This step is bypassed if receiving the product from a CBPDO.

This job should complete with a return code of 0.

6. Allocate SMP/E target and distribution libraries.

Edit and submit the sample job IDMALLOC to allocate the SMP/E target and distribution libraries for IM. For more information, refer to Appendix B.2, “IDMALLOC” on page 179.

This job should complete with a return code of 0.

7. Create DDDEF entries.

Edit and submit the sample job IDMDDEF to create DDDEF entries for the SMP/E target and distribution libraries for IM. For more information, refer to Appendix B.3, “IDMDDEF” on page 181.

This job should complete with a return code of '0' if the entries did not previously exist, or '4' if they did exist.

8. Initialize HFS libraries.

The IM server runs in the OpenEdition MVS environment that allows, for example, multitasking or security processing. For the server to run in OpenEdition MVS, all executables (load modules) must be known to OpenEdition MVS.

The header files to code using the APIs that are shipped with the product are stored in the HFS provided by OpenEdition MVS.

The sample job, IDMHFS, performs the following tasks:

- Creating the directory `/usr/lpp/iminer/bin` and others. The path `/usr/lpp` must exist. You need write authority for the directory to install the product. The directory is normally created by the OpenEdition administrator to contain all Licensed Program Products (LPP).
- Creating dummy files in the bin directory for all IM executables.
- Setting the security bits and the sticky bit to indicate to OpenEditions MVS that the real executable (load module) is stored in the STEPLIB. The STEPLIB must be used in the server start-up job.

Edit and submit this sample job to perform the tasks described above. For more information, refer to Appendix B.4, “IDMHFS” on page 182.

This job should complete with a return code of 0.

9. Perform `SMP/E APPLY CHECK`.

Edit and submit the sample job IDMAPPCK to perform an `SMP/E APPLY CHECK` for IM. For more information, refer to Appendix B.5, “IDMAPPCK” on page 182.

To receive the full benefit of the SMP/E Causer SYSMOD Summary Report, do not bypass the following on the APPLY CHECK: PRE, ID, REQ, and IFREQ. This is because the SMP/E root cause analysis identifies the cause only of ERRORS and not of WARNINGS (SYSMODs that are bypassed are treated as warnings, not errors, by SMP/E).

The `GROUPEXTEND` operand indicates that `SMP/E` accepts all requisite SYSMODs. The requisite SYSMODs might be applicable to other functions.

This job should complete with a return code of 0.

10. Perform `SMP/E APPLY`.

Edit and submit the sample job IDMAPPLY to perform an `SMP/E APPLY` for IM. For more information, refer to Appendix B.6, “IDMAPPLY” on page 183.

This job should complete with a return code of 0.

11. Bind the DB2 plan.

Run this job only if you want to work with DB2. You must have the privilege to run the bind job.

Edit and submit the sample job IDMDB2 to bind the DB2 plan for IM. For more information, refer to Appendix B.7, “IDMDB2” on page 183.

This job should complete with a return code of 0.

For this plan to run, grant the execute right to all users by using a **SPUFI** command as shown in following example:

```
GRANT execute ON PLAN IDM2PLAN To public;
```

Note

In a DB2 Data Sharing environment you can use the DB2 group attachment name rather than the DB2 subsystem ID for the variable #ssid. This way you can start the bind job on each OS/390 CEC independent of which DB2 subsystem is installed.

12. Install the demonstration package.

Edit and submit the sample job IDMDemo to install the DEMO Package in a user's home directory. For more information, refer to Appendix B.8, "IDMDemo" on page 184.

This job must be run for each user that wants to work with the demonstration package, but for at least one user, because one data file of this demonstration package is used for the installation verification job (IDMVERIFY).

This job should complete with a return code of 0.

13. Verify the installation.

Edit and submit the sample job IDMVERIFY to verify that IM was installed correctly. For more information, refer to Appendix B.9, "IDMVERIFY" on page 184.

This job should complete with a return code of 0.

14. Perform SMP/E ACCEPT CHECK.

Edit and submit the sample job IDMACCCK to perform an SMP/E ACCEPT CHECK for IM. For more information, refer to Appendix B.10, "IDMACCCK" on page 185.

To receive the full benefit of the SMP/E Causer SYSMOD Summary Report, do not bypass the following on the ACCEPT CHECK: PRE, ID, REQ, and IFREQ. This is because the SMP/E root cause analysis identifies the cause only of ERRORS and not of WARNINGS (SYSMODs that are bypassed are treated as warnings, not errors, by SMP/E).

The GROUPEXTEND operand indicates that SMP/E accepts all requisite SYSMODs. The requisite SYSMODs might be applicable to other functions.

This job should complete with a return code of 0.

15. Perform SMP/E ACCEPT.

Edit and submit the sample job IDMACCEP to perform an SMP/E ACCEPT for IM. For more information, refer to Appendix B.11, “IDMACCEP” on page 185.

Before using SMP/E to load new distribution libraries, we recommend that you set the ACCJCLIN indicator in the distribution zone. This will cause entries produced from JCLIN to be saved in the distribution zone whenever a SYSMOD containing inline JCLIN is ACCEPTed. For more information on the ACCJCLIN indicator, see the description of inline JCLIN in the SMP/E publications.

This job should complete with a return code of 0.

If PTFs containing replacement modules are being ACCEPTed, SMP/E ACCEPT processing will linkedit/bind the modules into the distribution libraries. During this processing, the linkage editor or binder may return messages documenting unresolved external references, resulting in a return code of '4' from the ACCEPT step. These messages can be ignored because the distribution libraries are not executable and the unresolved external references will not affect the executable system libraries.

10.3.2 Installation Customization

After successful installation of the Intelligent Miner program files, the environment needs to be customized for the specific environment. To do this, follow these instructions:

1. Log in to the *target system* as a system administrator with OpenEdition MVS authorization uid=0 (super user).
2. Create a IM group (for example, imgroup).

Add RACF profile definitions for new RACF groups or alter RACF profile definitions for existing RACF groups to define an OMVS segment containing a valid GID.

To add the IM group, use the RACF interactive panels as shown in below. Select option 3 of the RACF main menu.

```
RACF - SERVICES OPTION MENU
OPTION ==> 3
```

```
SELECT ONE OF THE FOLLOWING:
```

- 1 DATA SET PROFILES
- 2 GENERAL RESOURCE PROFILES
- 3 GROUP PROFILES AND USER-TO-GROUP CONNECTIONS
- 4 USER PROFILES AND YOUR OWN PASSWORD
- 5 SYSTEM OPTIONS
- 6 REMOTE SHARING FACILITY
- 7 DIGITAL CERTIFICATES
- 99 EXIT

Licensed Materials - Property of IBM
5695-039 (C) Copyright IBM Corp. 1983, 1994
All Rights Reserved - U.S. Government Users

```
F1=HELP      F2=SPLIT    F3=END      F4=RETURN    F5=RFIND    F6=RCHANGE
F7=UP        F8=DOWN     F9=SWAP     F10=LEFT     F11=RIGHT   F12=RETRIEVE
```

Select option 1 to add a new group and enter the group name, for example IMGROUP.

```
RACF - GROUP PROFILE SERVICES
OPTION ==> 1
```

```
SELECT ONE OF THE FOLLOWING.
```

- 1 ADD Add a group profile
- 2 CHANGE Change a group profile
- 3 DELETE Delete a group profile
- 4 CONNECT Add or change a user connection
- 5 REMOVE Remove users from the group

```
D or 8 DISPLAY    Display profile contents
S or 9 SEARCH     Search the RACF data base for profiles
```

```
ENTER THE FOLLOWING INFORMATION.
```

```
GROUP NAME        ==> IMGROUP
```

```
F1=HELP      F2=SPLIT    F3=END      F4=RETURN    F5=RFIND    F6=RCHANGE
F7=UP        F8=DOWN     F9=SWAP     F10=LEFT     F11=RIGHT   F12=RETRIEVE
```

Fill all the specifications and select the **OMVS PARAMETERS**.

```

RACF - ADD GROUP IMGROUP
COMMAND ===>

Enter the following information:

OWNER                ===> DB2RES4      Userid or group name
SUPERIOR GROUP       ===> OMVSGRP
USE TERMINAL UACC    ===> YES         YES or NO

Identify a model profile for group datasets (optional):

PROFILE NAME         ===>

To ADD the following optional information, enter any character:

_  INSTALLATION DATA
_  DFP PARAMETERS
S OMVS PARAMETERS
_  OVM PARAMETERS

F1=HELP      F2=SPLIT    F3=END      F4=RETURN    F5=RFIND     F6=RCHANGE
F7=UP        F8=DOWN     F9=SWAP     F10=LEFT     F11=RIGHT    F12=RETRIEVE

```

For a user to be able to request OpenEdition services and invoke the shell, the user's current RACF group (IMGROUP) **must** have an OpenEdition group ID (GID) assigned to it. All groups that an OpenMVS user belongs to should be assigned an OMVS GID. Also, the user's default group must have a GID assigned for POSIX standards conformance.

Add a valid OMVS GID for the group you are adding.

```

RACF - ADD GROUP IMGROUP                                OMVS PARAMETERS
COMMAND ===>

Enter OMVS segment information:

GROUP IDENTIFIER  ===> 1                                0 - 2147483647


F1=HELP    F2=SPLIT    F3=END    F4=RETURN    F5=RFIND    F6=RCHANGE
F7=UP      F8=DOWN     F9=SWAP    F10=LEFT    F11=RIGHT   F12=RETRIEVE

```

After the group has been added the PROFILE ADDED message will be returned as shown in the screen below.

```

RACF - GROUP PROFILE SERVICES                            PROFILE ADDED
OPTION ===>

SELECT ONE OF THE FOLLOWING.

1 ADD          Add a group profile
2 CHANGE       Change a group profile
3 DELETE       Delete a group profile
4 CONNECT      Add or change a user connection
5 REMOVE       Remove users from the group


D or 8 DISPLAY Display profile contents
S or 9 SEARCH  Search the RACF data base for profiles


ENTER THE FOLLOWING INFORMATION.

GROUP NAME      ===> IMGROUP


F1=HELP    F2=SPLIT    F3=END    F4=RETURN    F5=RFIND    F6=RCHANGE
F7=UP      F8=DOWN     F9=SWAP    F10=LEFT    F11=RIGHT   F12=RETRIEVE

```

3. Create an IM user, for example, IMSERV, to run the server startup job called IDMSTART as a regular OpenEdition MVS user.

Regular OpenEdition MVS users are those users who have a UID set to a value other than zero.

To add new users that can access OpenEdition MVS services via TSO and rlogin, select option 1 of the

RACF - SERVICES OPTION MENU and provide the user name (IMSERV) of the new user.

```
RACF - USER PROFILE SERVICES
OPTION ==> 1

SELECT ONE OF THE FOLLOWING:

      1  ADD          Add a user profile
      2  CHANGE       Change a user profile
      3  DELETE       Delete a user profile
      4  PASSWORD     Change your own password or interval
      5  AUDIT        Monitor user activity (Auditors only)

D or 8  DISPLAY      Display profile contents
S or 9  SEARCH       Search the RACF data base for profiles

ENTER THE FOLLOWING INFORMATION:

      USER      ==> IMSERV      Userid

F1=HELP      F2=SPLIT      F3=END      F4=RETURN      F5=RFIND      F6=RCHANGE
F7=UP        F8=DOWN       F9=SWAP     F10=LEFT      F11=RIGHT     F12=RETRIEVE
```

Fill in the blanks with all the information necessary. Remember to provide the user with a default group with a OMVS GID assigned to it.

RACF - ADD USER IMSERV
COMMAND ===>

ENTER THE FOLLOWING INFORMATION:

OWNER	===> DB2RES4	Userid or group name
USER NAME	===> IMSERV	
DEFAULT GROUP	===> IMGROUP	Group name
PASSWORD	===>	User's initial password
	===>	Re-enter password to verify
PASSWORD INTERVAL	===>	1 - 254 (days), NO, or blank

Press ENTER to continue.

F1=HELP	F2=SPLIT	F3=END	F4=RETURN	F5=RFIND	F6=RCHANGE
F7=UP	F8=DOWN	F9=SWAP	F10=LEFT	F11=RIGHT	F12=RETRIEVE

Specify **YES** for the optional information to add the TSO/E and OMVS segment information.

RACF - ADD USER IMSERV
COMMAND ===>

TO ASSIGN USER ATTRIBUTES, ENTER YES:

GROUP ACCESS	===> NO	SPECIAL	===> NO
ADSP	===> NO	OPERATIONS	===> NO
OIDCARD	===> NO	AUDITOR	===> NO
NO-PASSWORD	===> NO		

IDENTIFY THE MODEL PROFILE FOR USER DATA SETS (OPTIONAL):

MODEL PROFILE ===>

TO CREATE THE FOLLOWING, ENTER YES (OPTIONAL):

A GENERIC DATA SET PROFILE	===> NO
A MINIDISK PROFILE	===> NO

TO ADD OPTIONAL INFORMATION, ENTER YES ===> **YES**

F1=HELP	F2=SPLIT	F3=END	F4=RETURN	F5=RFIND	F6=RCHANGE
F7=UP	F8=DOWN	F9=SWAP	F10=LEFT	F11=RIGHT	F12=RETRIEVE

Select the parameters required for this user as shown in the example screen below. In this case, IMSERV requires both TSO/E and OMVS segments.

```
RACF - ADD USER IMSERV  
COMMAND ==>>
```

To ADD the following information, enter any character:

```
_ CLASS AUTHORITY  
_ INSTALLATION DATA  
_ GROUP AUTHORITY  
_ SECURITY LEVEL or CATEGORIES  
_ SECURITY LABEL  
_ LOGON RESTRICTIONS  
_ NATIONAL LANGUAGES
```

```
_ DFP PARAMETERS  
S TSO PARAMETERS  
_ OPERPARM PARAMETERS  
_ CICS PARAMETERS  
_ WORK ATTRIBUTES  
S OMVS PARAMETERS  
_ NETVIEW PARAMETERS  
_ DCE PARAMETERS  
_ OVM PARAMETERS
```

```
F1=HELP      F2=SPLIT    F3=END       F4=RETURN    F5=RFIND     F6=RCHANGE  
F7=UP        F8=DOWN     F9=SWAP      F10=LEFT     F11=RIGHT    F12=RETRIEVE
```

Enter all the information necessary for the TSO/E segment,

RACF - ADD USER IMSERV

TSO-RELATED INFORMATION

COMMAND ===>

ENTER THE FOLLOWING TSO-RELATED INFORMATION:

JOB CLASS ===>
MESSAGE CLASS ===>
HOLD CLASS ===>
SYSOUT CLASS ===>
ACCOUNT NUMBER ===> **ACCN#**
LOGON PROCEDURE NAME ===> **IKJACCN#**
REGION SIZE ===> **4096**
UNIT ===> **SYSDA**
DESTINATION ID ===>
MAXIMUM REGION SIZE ===>
USER DATA ===>
LOGON SECURITY LABEL ===>
COMMAND ===>
 ===>

F1=HELP F2=SPLIT F3=END F4=RETURN F5=RFIND F6=RCHANGE
F7=UP F8=DOWN F9=SWAP F10=LEFT F11=RIGHT F12=RETRIEVE

then enter a UID for the new IM user.

RACF - ADD USER IMSERV

1 of 3
OMVS PARAMETERS

COMMAND ===>

Enter User Identifier (UID) below, then press ENTER:

USER IDENTIFIER ===> **0** 0 - 2147483647

F1=HELP F2=SPLIT F3=END F4=RETURN F5=RFIND F6=RCHANGE
F7=UP F8=DOWN F9=SWAP F10=LEFT F11=RIGHT F12=RETRIEVE

Set the parameter to reflect the initial directory path name for the user (for example IMSERV). Remember that the directory must exist for the user to be able to access the OpenEdition MVS environment.

```
RACF - ADD USER IMSERV      2 of 3
                           OMVS PARAMETERS

COMMAND ===>

Enter Initial Directory Path Name (HOME), then press ENTER:

=> /u/imserv               <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
F1=HELP      F2=SPLIT    F3=END      F4=RETURN   F5=RFIND    F6=RCHANGE
F7=UP        F8=DOWN     F9=SWAP    F10=LEFT   F11=RIGHT   F12=RETRIEVE
```

The last step is to define the installed shell program path name.

```
RACF - ADD USER IMSERV      3 of 3
                           OMVS PARAMETERS

COMMAND ===>

Enter Program Path Name (PROGRAM), then press ENTER:

=> /bin/sh                 <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
=>                         <=
F1=HELP      F2=SPLIT    F3=END      F4=RETURN   F5=RFIND    F6=RCHANGE
F7=UP        F8=DOWN     F9=SWAP    F10=LEFT   F11=RIGHT   F12=RETRIEVE
```

Finally, you will receive the message **PROFILE ADDED**.

```
RACF - USER PROFILE SERVICES                                Profile added
OPTION ==>

SELECT ONE OF THE FOLLOWING:

      1  ADD          Add a user profile
      2  CHANGE       Change a user profile
      3  DELETE       Delete a user profile
      4  PASSWORD     Change your own password or interval
      5  AUDIT        Monitor user activity (Auditors only)

D or 8  DISPLAY      Display profile contents
S or 9  SEARCH       Search the RACF data base for profiles

ENTER THE FOLLOWING INFORMATION:

USER      ==> IMSERV      Userid

F1=HELP    F2=SPLIT    F3=END      F4=RETURN   F5=RFIND    F6=RCHANGE
F7=UP      F8=DOWN     F9=SWAP   F10=LEFT   F11=RIGHT   F12=RETRIEVE
```

4. Verify the defined IMSERV user.

IMSERV must be defined as a valid user to run the first job. Validate this in the following way:

- Log in to the *target system* as a IMSERV user.
- Open the OpenMVS ISPF shell panel.

Enter the directory (/usr) name as shown in the following screen.

```
File Directory Special_file Tools File_systems Options Setup Help
-----
                                OpenMVS ISPF Shell

Enter a pathname and do one of these:

- Press Enter.
- Select an action bar choice.
- Specify an action code or command on the command line.

Return to this panel to work with a different pathname.                More:      +

  /usr
  _____
  _____
  _____

Command ==> _____
F1=Help      F3=Exit      F5=Retrieve F6=Keyshelp F7=Backward F8=Forward
F10=Actions  F11=Command F12=Cancel
```

- List files in any path (for example, /usr).

You then should see the directory list, which will be like the one shown below.

Directory List

/usr/

Select one or more files with / or action codes.

Type	Filename	Row 1 of 12
_ Dir	.	
_ Dir	..	
_ Dir	adsm	
_ Dir	bin	
_ Dir	include	
_ Dir	lib	
_ Dir	lpp	
_ Dir	mail	
_ Dir	man	
_ Dir	sbin	
_ Dir	share	
_ Dir	spool	

Command ==>

F1=Help	F3=Exit	F4=Name	F5=Retrieve	F6=Keyshelp	F7=Backward
F8=Forward	F11=Command	F12=Cancel			

5. Create a IM client user (for example, IMUSER) as regular OpenEdition MVS user.

Create a IM user (IMUSER) to run the IM on the client site using the same steps as previously described in step 3 for the IMSERV user. The home directory (for example, /u/imuser) and the USER Identifier should be set accordingly. The OMVS UID of a client user, IMUSER in our example, must not be 0 !

6. Verify the defined IMUSER user.

Follow the same steps as described for step 4.

7. Create a RACF data set profile for *HLQ.SIDMLOAD* where HLQ is the high level qualifier of the product libraries.

You can create a *HLQ.SIDMLOAD* data set profile on the RACF panel.

The screen shown below shows the main RACF panel. Enter a '1' on the Option line.

```

RACF - SERVICES OPTION MENU
OPTION ==> 1

SELECT ONE OF THE FOLLOWING:

1  DATA SET PROFILES

2  GENERAL RESOURCE PROFILES

3  GROUP PROFILES AND USER-TO-GROUP CONNECTIONS

4  USER PROFILES AND YOUR OWN PASSWORD

5  SYSTEM OPTIONS

6  REMOTE SHARING FACILITY

7  DIGITAL CERTIFICATES
99 EXIT

      Licensed Materials - Property of IBM
      5695-039 (C) Copyright IBM Corp. 1983, 1994
      All Rights Reserved - U.S. Government Users

F1=HELP    F2=SPLIT    F3=END      F4=RETURN   F5=RFIND    F6=RCHANGE
F7=UP      F8=DOWN     F9=SWAP    F10=LEFT   F11=RIGHT   F12=RETRIEVE

```

Enter a '1' on OPTION line to add a profile.

```

RACF - DATA SET PROFILE SERVICES
OPTION ==> 1

SELECT ONE OF THE FOLLOWING:

1  ADD          Add a profile
2  CHANGE       Change a profile
3  DELETE       Delete a profile
4  ACCESS       Maintain the access lists
5  AUDIT        Monitor access attempts (for auditors only)

D or 8 DISPLAY  Display profile contents
S or 9 SEARCH   Search the RACF data base for profiles

F1=HELP    F2=SPLIT    F3=END      F4=RETURN   F5=RFIND    F6=RCHANGE
F7=UP      F8=DOWN     F9=SWAP    F10=LEFT   F11=RIGHT   F12=RETRIEVE

```

Enter a profile name, for example 'IDM.V2R1M0.SIDMLOAD'.

RACF - DATA SET PROFILE SERVICES - ADD
COMMAND ===>

ENTER THE FOLLOWING INFORMATION:

PROFILE NAME	====>	'IDM.V2R1M0.SIDMLOAD'
TYPE	====>	MODEL, TAPE, GENERIC, or blank
VOLUME SERIAL	====>	If a discrete profile and the data set is not cataloged
UNIT	====>	If you are adding a profile and specified VOLUME SERIAL
PASSWORD	====>	Data set password, if the data is password protected
	====>	Re-enter password to verify
USE A MODEL	====>	YES or NO

F1=HELP F2=SPLIT F3=END F4=RETURN F5=RFIND F6=RCHANGE
F7=UP F8=DOWN F9=SWAP F10=LEFT F11=RIGHT F12=RETRIEVE

Edit the parameters for the appropriate security level.

RACF - ADD DATA SET PROFILE
COMMAND ===>

PROFILE: 'IDM.V2R1M0.SIDMLOAD'

ENTER OR CHANGE THE FOLLOWING INFORMATION:

OWNER	====>	DB2RES4	Userid or group name
LEVEL	====>	0	0-99
FAILED ACCESSES	====>	FAIL	FAIL or WARN
UACC	====>	NONE	NONE, READ, UPDATE, CONTROL, ALTER or EXECUTE
AUDIT SUCCESSES	====>	NOAUDIT	READ, UPDATE, CONTROL, ALTER, or NOAUDIT
AUDIT FAILURES	====>	READ	READ, UPDATE, CONTROL, ALTER, or NOAUDIT
INDICATOR	====>	SET	SET, NOSET, or ONLY
NOTIFY	====>		Userid
ERASE ON DELETE	====>		YES or blank

TO ADD OPTIONAL INFORMATION, ENTER YES ===> **NO**

F1=HELP F2=SPLIT F3=END F4=RETURN F5=RFIND F6=RCHANGE
F7=UP F8=DOWN F9=SWAP F10=LEFT F11=RIGHT F12=RETRIEVE

If the profile is added successfully, a screen like the one shown below will be displayed.

```

RACF - DATA SET PROFILE SERVICES                                PROFILE ADDED
OPTION ==>

SELECT ONE OF THE FOLLOWING:

      1  ADD           Add a profile
      2  CHANGE        Change a profile
      3  DELETE        Delete a profile
      4  ACCESS        Maintain the access lists
      5  AUDIT         Monitor access attempts (for auditors only)

D or 8  DISPLAY       Display profile contents
S or 9  SEARCH        Search the RACF data base for profiles

F1=HELP   F2=SPLIT   F3=END   F4=RETURN  F5=RFIND   F6=RCHANGE
F7=UP     F8=DOWN    F9=SWAP   F10=LEFT   F11=RIGHT  F12=RETRIEVE

```

8. Permit RACF read access to *HLQ.SIDMLOAD* for the IMSERV and IMUSER IDs.

You can do this on the TSO command line by entering the
 permit '*HLQ.SIDMLOAD*' id(IMSERV, IMUSER) acc(read) command.

```

Menu List Mode Functions Utilities Help

                                ISPF Command Shell
Enter TSO or Workstation commands below:

==> permit 'IDM.V2R1M0.SIDMLOAD' id(IMSERV, IMUSER) acc(read)

Place cursor on choice and press enter to Retrieve command

=>
=>
=>
=>
=>
=>
=>
=>
=>
=>

F1=Help   F3=Exit   F10=Actions  F12=Cancel

```

You can also do this on the RACF panel by following the **DATA SET PROFILES -> ACCESS -> ADD** steps.

9. Permit RACF read access to BPX.DAEMON for the IMSERV. (optional)

If the RACF facility profile BPX.DAEMON is defined on your system, IMSERV must have read access to it. This facility is used to restrict the access rights of OpenEdition users. For more information, refer to the *MVS/ESA Open Edition DCE: PLANNING*, SC09-1484 publication.

You can do this using the TSO command line by entering the `permit BPX.DAEMON class(facility) id(IMSERV) acc(read)` command. As mentioned before, you can use the RACF panel by following **GENERAL RESOURCE PROFILE -> ACCESS -> ADD** steps.

10. Update the program-controlled libraries.

All libraries used for the execution of IM and related products must be, in RACF terms, program-controlled. These libraries are:

- SIDMLOAD for IM
- SCEERUN for IBM Language Environment for MVS
- SDSNLOAD, SDSNEXIT, SDSNLINK for DB2
- LINKLIB for any program that is called from this library

If this profile does not exist, create it using the `rdefine` command.

If the program-class profile '*' already exists, modify it according to the following example:

```
ralter program *  
  
addmem (  
  'HLQ.SIDMLOAD' /VOLxxx/nopadchk  
  'SYS1.SCEERUN' /VOLxxx/nopadchk  
  'SYS1.SCLBDLL' /VOLxxx/nopadchk  
  'SYS1.LINKLIB' /VOLxxx/nopadchk  
  'SYS1.SDSNLOAD' /VOLxxx/nopadchk  
  'SYS1.SDSNEXIT' /VOLxxx/nopadchk  
  'SYS1.SDSNLINK' /VOLxxx/nopadchk  
)  
  
uacc(read)
```

11. Permit RACF read access to the program-controlled libraries for the IMSERV, IMUSER IDs.

Enter the following command on TSO command line:

```
permit * class(program) id(IMSERV, IMUSER) acc(read)
```

or use the RACF panel by following the

GENERAL RESOURCE PROFILE -> ACCESS -> ADD steps.

12. Activate the new definitions.

Enter the following command on TSO command line.

```
setropts when(program) refresh
```

You can refresh the RACF panel by following the

SYSTEM OPTIONS -> REFRESH steps.

13. Edit an HFS file that includes a list of step libraries.

The member BPXPRMxx in the data set SYS1.PARMLib contains customization values for OpenEdition. In there, check the value of STEPLIBLIST. It contains the name of an HFS file that includes a list of step libraries that contain executables that are authorized to issue "set-user-ID" instructions. This is a threefold pointing: BPXPRMxx -> HFS file -> library list -> executables.

Add the following libraries to that list:

- HLQ.SIDMLOAD
- SYS1.SCEERUN
- SYS1.SCLBDLL
- SYS1.LINKLIB
- SYS1.SDSNLOAD
- SYS1.SDSNEXIT
- SYS1.SDSNLINK

14. Set the mode of the IMSERV home directory /u/imserv) to the execute permission (755), because the IMUSER must have the execute permission for the IMSERV's home directory. To do this, enter the command `sh chmod 755 /u/imserv` on the OpenMVS ISPF shell as shown in the following screen.

```
File Directory Special_file Tools File_systems Options Setup Help
-----
OpenMVS ISPF Shell

Enter a pathname and do one of these:

- Press Enter.
- Select an action bar choice.
- Specify an action code or command on the command line.

Return to this panel to work with a different pathname.
More:      +

-----
-----
-----

Command ==> sh chmod 755 /u/imserv _____
F1=Help      F3=Exit      F5=Retrieve  F6=Keyshelp  F7=Backward  F8=Forward
F10=Actions  F11=Command F12=Cancel
```

15. Grant DB2 access rights.

If DB2 tables are used for mining, all the client users and IMSERV need READ permission to the following DB2 system tables:

- SYSIBM.SYSTABLESPACE
- SYSIBM.SYSTABLES
- SYSIBM.SYSCOLUMNS
- SYSIBM.SYSINDEXES

You must have the privilege to run the following `grant` command on SFUPI panel.

```
GRANT read ON TABLE sysibm.systablespace, sysibm.systables,
sysibm.syscolumns, sysibm.sysindexes TO PUBLIC.
```

We recommend that you grant READ access for these tables either to PUBLIC or to a specific group (for example, IMGROUP) for all IM users.

16. Verify the security environment.

This job covers most of the security rules but not all. Therefore the successful completion of the job does not guarantee the completeness of the required security definitions.

Edit and submit the sample job `IDMSECUR` to verify the security definitions required to run IM. For more information, refer to Appendix B.12, “`IDMSECUR`” on page 185.

This job should complete with a return code of 0.

10.4 Running the Server

This section describes how to run the IM server and also covers some DB2-related considerations.

10.4.1 Starting the Server

If you want to run IM as a started task, you must run the following RACF commands:

```
RDEFINE STARTED IDMSERV ** UACC(NONE) STDATA(IMSERV)

GROUP(IMGROUP) TRUSTED(NO)

SETR RACLIST(STARTED) REFRESH
```

where `IMSERV` is the user ID that starts IM. These commands associate the user `IMSERV` and group `IMGROUP` with the started task `IDMSERV`.

Edit, then submit the sample job called `IDMSTART` to start the IM server in the OpenEdition environment. For more information, refer to Appendix B.13, “`IDMSTART`” on page 186. For the *TIME* and *REGION* parameters, we recommend 1440 and 0M, respectively. `IDMSTART` is a never-ending job waiting for requests sent from a workstation client. The job is stopped using TSO/E `CANCEL` command.

Note

For NLS support you must set the OpenEdition environment variables `NLSPATH`, `LANG` and `LC_ALL` in this job for your language.

As an example, for Korean the variable must be set as follows:

- `NLSPATH=/usr/lpp/nls/%L/%N`
- `LANG=Ko_KR`
- `LC_ALL=Ko_KR.IBM-933`

OpenEdition MVS parameters are defined in data set SYS1.PARMLIB, member BPXPRMxx. Become familiar with the meaning of these parameters. To determine what the appropriate values are, you must know about the system's total OpenEdition MVS workload. If possible, we recommend setting at least three parameters to values different from their default:

- MAXCPU TIME=2147483647

This is the maximum allowable value, which avoids the possibility that long mining runs might be cancelled.

- MAXASSIZE=2147483647

This is the maximum allowable value, which lets mining jobs allocate as much storage as available. Ensure that sufficient page data sets are allocated to support the storage size specified in this parameter. If multiple mining jobs run in parallel, the system might experience a shortage of auxiliary storage.

- MAXFILEPROC=256

This is the maximum number of open files per process.

10.4.2 DB2 Considerations

If you provide your own customer defined computed fields, edit then submit the sample job IDMCFLD to run a customization program checks for Customer Defined Computed Fields. For more information, refer to Appendix B.14, "IDMCFLD" on page 187. The results of the check are stored in the HFS member /usr/lpp/IMiner/bin/idmdfncf.dcl.

This job should complete with a return code 0.

For the association, sequential patterns, and time sequences algorithms some fields of the input data must be sorted and must also have a specific IM data type in order to get correct results. For DB2 tables, the sorting is done by an ORDER BY clause in the SELECT statement. The conversion from the DB2 data type to the IM data type is done by a DB2 scalar function. The syntax of a select statement does not allow the use of the scalar function for a column if the column is also used in the ORDER BY clause. Therefore, IM requires the following DB2 data types for the following fields:

- Association

```
Field:Transaction ID
IM data type:Categorical
DB2 data type:CHAR, VARCHAR
```

- Sequential Patterns

Field:Transaction Group ID, Transaction ID
IM data type:Categorical
DB2 data type:CHAR, VARCHAR

- Time Sequences

Field:Time
IM data type:Discrete Numeric
DB2 data type:FLOAT
Field:Sequence ID
IM data type:Categorical
DB2 data type:CHAR, VARCHAR

During the creation of DB2 table spaces, IM uses DB2 storage groups with names starting with the 5-character common prefix as specified in the `IDM_STOGROUP_PREFIX` environment variable. The last three characters have to be a sequential number where the value 000 is used to create non-partitioned DB2 table spaces, and values greater than 000 are used for partitioned table spaces.

Appendix A. Using DB2 V 2.1 or DataJoiner on AIX

If you have a DB2 UDB, together with a DB2 version different from UDB installed, for example, DB2 CS Version 2.1.1, or if you are planning to use DataJoiner to access different data sources, ensure that Intelligent Miner accesses the correct DB2 runtime library *libdb2.a*. If no LIBPATH is set or no link exists from /usr/lib to the library *libdb2.a*, then Intelligent Miner first tries to access /usr/lpp/db2_05_00/lib/libdb2.a

The following rules help you to ensure that the correct library is accessed.

A.1 Running with DB2 for AIX

If you want to run IM in client/server mode:

1. Log on as root on the Intelligent Miner server.
2. Check or set the DB2INSTANCE variable.
 - For DB2 version 2.1 it has to be the DB2 V2 instance name.
 - For DataJoiner it has to be the DataJoiner instance name.
3. Set the LIBPATH:
 - For DB2 version 2.1 it should be set to /usr/lpp/db2_02_01/lib
 - For DataJoiner it should be set to /usr/lpp/djx_02_01_01/lib
4. Type one of the following commands on the command line:

<code>idmstart -d</code>	This command also shows tracing information of the AIX server.
<code>idmstart</code>	This command starts the AIX server without showing tracing information.'
5. Log out.
6. Call 'im' on the client and test the connection.

If you want to run the IM with stand-alone mode:

1. Log on as user root.
2. Check or set the DB2INSTANCE variable:
 - For DB2 version 2.1 it has to be the DB2 V2 instance name
 - For DataJoiner it has to be the DataJoiner instance name
3. Set the LIBPATH:

- For DB2 version 2.1 it should be set to /usr/lpp/db2_02_01/lib
 - For DataJoiner it should be set to /usr/lpp/djx_02_01_01/lib
4. Type one of the following commands on the command line:

<code>idmstart -d</code>	This command also shows tracing information of the AIX server.
<code>idmstart</code>	This command starts the AIX server without showing tracing information.'
 5. Log out
 6. Log on as the Intelligent Miner user.
 7. Check or set the DB2INSTANCE variable.
 - For DB2 version 2.1 it has to be the DB2 V2 instance name.
 - For DataJoiner it has to be the DataJoiner instance name.
 8. Set the LIBPATH:
 - For DB2 version 2.1 it should be set to /usr/lpp/db2_02_01/lib.
 - For DataJoiner it should be set to /usr/lpp/djx_02_01_01/lib.
 9. Call 'im' and test the connection.

A.2 Working with Data Joiner (Version 2.1.1)

If you want to use Data Joiner together with IM, you must configure the IM users to be able to access the databases. To do this, you can add the db2profile in the IM user profile. Data Joiner needs another configuration, called user mapping, to access the mapped database. You must create a user mapping for each IM user.

To install IM you must have at least one of the following pre-installed software:

```
db2_02_01.client 2.1.1.0
db2_05_00.conn 5.0.0.0
db2_05_00.client 5.0.0.0
```

If you work with Data Joiner V2.1 you need the fixpack with the PTF number U459154. Otherwise the AIX system will display the following warning message when you try to install IM:

WARNINGS-----

Problems described in this section are not likely to be the source of any immediate or serious failures, but further actions may be necessary or desired.

Missing Requisites of Previously Installed Software

The following fileset updates are requisites of software that is already installed. They are not currently installed but should be to ensure that the system functions correctly. These updates are not on the current installation media.

```
djx_02_01_01.cs.sna p=U459154      # Fileset Update
djx_02_01_01.cs.drda p=U459154     # Fileset Update
<< End of Warning Section >>
```

If you have this PTF applied and you want to connect remotely to a DB2/MVS or DB2/02390 database from an Intelligent Miner AIX server, make sure that you also have one of the following PTFs applied:

```
UQ13906 - DB2/MVS V3.1
UQ13907 - DB2/MVS V4.1
UQ13908 - DB2 for OS/390 V5.1
```

Note that connecting from an Intelligent Miner AIX server to a DB2/MVS or DB2/OS390 database is not fully supported. You can use DB2/MVS or DB2/OS390 database tables as input for mining runs but you cannot write output tables. Connecting to an AS/400 database from an Intelligent Miner AIX server is not supported at all.

Appendix B. Intelligent Miner for Data Installation Sample JCL

This appendix describes the sample JCL scripts for installing Intelligent Miner for Data on OS/390.

B.1 IDMRECEV

This sample JCL reads SYSMODs and HOLDDATA for FMID=H258100 into the SMPPTS and the global zone to prepare the installation of the SYSMODs to your system. The SYSMODs and HOLDDATA can be on Expanded Service Option (ESO) tapes, a function tape, a cumulative service tape (CUM tape), a data set, or a custom-built product delivery offering tape (CBPDO tape).

Before running this job you must:

1. Update the job card as required for your installation.
2. Change:
 - *#HLQ* to the high level qualifier of the product libraries
 - *#UUUU* to the appropriate device type for the product tape

B.2 IDMALLOC

This sample JCL allocates and catalogs the target and distribution libraries for SMP/E. Table 26 and Table 27 show the target and distribution libraries

(data sets) and the attributes which must be specified to install IM for Data V2.

Table 26. Target Libraries

Library DDNAME	T Y P E	D S O R G	R E C F M	L R E C L	No of Blks	BLK SIZE	No of 3390 Trks	No of DIR Blks
SIDMLOAD	U	PO	U	0	15000	6144	1700	5
SIDMDBRM	U	PO	FB	80	150	8800	10	2
SIDMHPP	U	PO	VB	255	150	6220	10	2
SIDMH	U	PO	VB	255	100	6220	10	2
SIDMMSG	U	PO	VB	255	150	6200	10	2
SIDMSAM1	U	PO	FB	80	150	8800	10	5
SIDMSAM2	U	PO	VB	255	150	6220	10	2
SIDMDAT1	U	PO	FB	255	150	6120	10	2
SIDMSDEF	U	PO	FB	80	400	8800	60	5
SIDMDEMO	U	PO	VB	255	150	6220	10	2

Table 27. Distribution Libraries

Library DDNAME	T Y P E	D S O R G	R E C F M	L R E C L	No of Blks	BLK SIZE	No of 3390 Trks	No of DIR Blks
AIDMLOAD	U	PO	U	0	15000	6144	1700	5
AIDMDBRM	U	PO	FB	80	150	8800	10	2
AIDMHPP	U	PO	VB	255	150	6220	10	2
AIDMH	U	PO	VB	255	100	6220	10	2
AIDMMSG	U	PO	VB	255	150	6200	10	2
AIDMSAM1	U	PO	FB	80	150	8800	10	5
AIDMSAM2	U	PO	VB	255	150	6220	10	2
AIDMDAT1	U	PO	FB	255	150	6120	10	2

Library DDNAME	T Y P E	D S O R G	R E C F M	L R E C L	No of Blks	BLK SIZE	No of 3390 Trks	No of DIR Blks
AIDMSDEF	U	PO	FB	80	400	8800	60	5
AIDMDEMO	U	PO	VB	255	150	6220	10	2

Before running this job you must:

1. Update the job card as required for your installation.
2. Change:
 - *#HLQ* to the high level qualifier for the product libraries

Note

If SMS is used, the parameters VOL=SER and UNIT must be deleted.

B.3 IDMDDEF

This sample JCL is used to define target and distribution libraries to SMP/E. It will add the DDDEF entries for both the target and distribution libraries in the target zone.

UCLIN updates only entries in SMP/E data sets. It does not affect any elements or load modules in any product library. You must ensure that the appropriate changes are made to the libraries. Be sure to understand the relationships between the various entries before making any UCLIN changes. This helps ensure that any UCLIN changes you make are complete and consistent with one another. When SMP/E processes UCLIN, it checks only the specified entry. It does not check how the changes might affect other entries.

Before running this job:

1. Update the job card as required for your installation
2. Change:
 - *#globalcsi* to make it your CSI name.
 - *#tzone* to make it your target zone name.
 - *#dzone* to make it your distribution zone name.

- *#HLQ* to the high level qualifier for the product libraries.
3. Optionally, you may also change the PATH statements. If you do this, ensure that the paths in the other sample jobs are changed accordingly.

B.4 IDMHFS

This sample JCL creates HFS Directories, allocates dummy files for load modules, and changes attributes for dummy files.

Before running this job:

1. Update the job card as required for your installation.
2. Optionally change:
 - The parameter IDMDIR in the PROC statements below to your directory if you wish.
 - The path /usr/lpp/IMiner in step RMD1 to your directory if you wish.

Ensure that the paths in the job IDMDDDEF are changed accordingly.

Note

These dummy files are necessary to allow the load modules to run in Open Edition while they are stored in the STEPLIB. The server code **MUST** reside in the STEPLIB (not in HFS) to allow it to be APF authorized.

B.5 IDMAPPC

This sample JCL performs a trial run to check whether the SYSMODs will be installed correctly (**APPLY CHECK**). The purpose of the **APPLY CHECK** option is to perform a test run to inform you of possible errors, and to provide reports of the SYSMOD status, libraries that will be updated, regression conditions, and SYSMODs that will be deleted. The target system libraries are not permanently updated.

During this check processing, the list of target zone entries is maintained in storage, and data is written to the target zone as a temporary storage medium. Check processing deletes any data written to the target zone. Consequently, no permanent updates are made to the target zone.

Although you can choose whether or not to do an **APPLY CHECK**, it is best to do a trial run before installing a SYSMOD on your system.

Before running this job:

1. Update the job card as required for your installation.
2. Change:
 - *#HLQ* to the high level qualifier of the product libraries
 - *#TARGETZONE* to the target zone for your installation

B.6 IDMAPPLY

This sample JCL performs the actual APPLY. The APPLY is used to cause SMP/E to install the elements supplied by a SYSMOD into the operating (or target) system libraries. The APPLY process:

- Selects SYSMODs present in the global zone and applicable to the specified target system.
- Makes sure all other required SYSMODs have either been applied or are being applied concurrently.

Before running this job:

1. UPDATE the job card as required for your installation
2. Change:
 - *#HLQIM* to the high level qualifier of the product libraries
 - *#TARGETZONE* to the target zone for your installation

B.7 IDMDB2

This sample JCL binds the DB2 plan for the data access API, the pre-processing library, and the client/server code.

Before running this job:

1. Update the job card as required for your installation.
2. Change:
 - *#HLQDB* to the high level qualifier of the DB2 load library.
 - *#HLQIM* to the high level qualifier of the product libraries.
 - *#ssid* to your DB2 Subsystem ID.
 - *#xxxx* to the owner of your data base.

After running this job, make sure to GRANT the right to execute this plan to all users that want to work with this product.

B.8 IDMDEMO

This sample JCL installs the demo mining bases in a private home directory. This job must be run by the owner of the home directory where the files are to be copied to, because it will become the owner of the directories created by this job. This job calls a script file in the HFS which copies the file *idmdemo.tar* into the users home directory, expands the tar file, and changes the paths used in the demo files to the users home directory.

Before running this job:

1. Update the job card as required for your installation.
2. Change the following in the DEMO EXEC below:
 - *IDMDIR*= the directory where the product was installed
 - *TARGET*= the target home directory

B.9 IDMVERIFY

This sample JCL verifies the correct installation of the IM for Data V2 product.

Before running this job:

1. Update the job card as required for your installation.
2. Change:
 - *#HLQIM* to the high level qualifier of the product libraries
 - *#HLQLE* to the high level qualifier of the LE run time library
3. Edit the data member IDMV1PAR in SIDMSAM2 and do a global change of:
 - */u/xxx* to your home directory. This is the directory for the output files of this job.
 - */u/yyy* to the directory where the demo data was installed using job IDMDEMO -This is the directory for the input data of this job. The input data is a file previously installed with the IDMDEMO job.

This job creates a set of output files *HLQ.VERIFY.** in your home directory. The contents of these files is not relevant at this point. The file called IDMMSGEN is the message file used for trace messages

B.10 IDMACCCK

This sample JCL performs an ACCEPT CHECK. The intent of the ACCEPT CHECK option is to perform a test run informing you of possible error conditions, and providing reports of SYSMOD status, libraries that will be updated, regression conditions, and SYSMODs that will be deleted. During ACCEPT CHECK processing, the list of distribution zone entries is maintained in storage and the data is written to the distribution zone as a temporary storage medium. ACCEPT CHECK processing deletes any data written to the distribution zone. Consequently, no permanent updates are made to the distribution zone.

Before running this job:

1. Update the job card as required for your installation.
2. Change:
 - *#HLQ* to the high level qualifier of the product libraries
 - *#DLIBZONE* to the distribution zone for your installation

B.11 IDMACCEP

This sample JCL performs the ACCEPT, which is used to cause SMP/E to install the elements supplied by a SYSMOD into the distribution libraries (or DLIBs). The ACCEPT process:

- Selects SYSMODs present in the global zone that are applicable to the specified distribution libraries.
- Makes sure all other required SYSMODs have been accepted or are being accepted concurrently.

Before running this job:

1. Update the job card as required for your installation.
2. Change:
 - *#HLQ* to the high level qualifier of the product libraries
 - *#DLIBZONE* to the distribution zone for your installation

B.12 IDMSECUR

This sample JCL runs a test program which verifies the security definitions required to run the server program correctly.

Return code of this job must be 0, otherwise check SYSPRINT for errors.

Successful completion does not guarantee the completeness and correctness of the security definitions.

Before running this job:

1. Update the job card as required for your installation.
2. Change:
 - *#USERID* to the user ID of a client user
 - *#PASSWORD* to the password of a client user. (For security reasons do not forget to erase this password after this job was submitted).
 - *#HLQIM* to the high level qualifier of the product libraries.
 - *#HLQLE* to the high level qualifier of the LE run time library.

Submit this job under the user ID that will start the IM start-up job later.

Note

It is absolutely mandatory that ALL security guidelines described in the Program Directory are observed very carefully. Otherwise the client cannot communicate with the server because RACF (or equivalent products) will refuse the access to programs and datasets needed to run the IM server. For technical reasons this program can perform only a limited test. So correct execution of this test is necessary but not sufficient.

B.13 IDMSTART

This sample JCL starts the server program.

Before running this job:

1. Update the job card as required for your installation.
2. Change:
 - */usr/lpp/IMiner* to the path you selected.
 - *#HLQIM* to the high level qualifier of the product libraries.
 - *#HLQLE* to the high level qualifier of the LE run time library.
 - *#HLQCL* to the high level qualifier of the C Class Library.
 - *#HLQDB* to the high level qualifier of the DB2 load library.

- #xxx to the directory of IMSERV
- The environment variable *IDM_BIN_DIR* must be set to the path for the executables.
- The environment variable *IDM_DB_FILE* must point to the HFS file which contains the list of DB2 SSIDs in your installation. This file must contain one line for each SSID starting at column 1.

Other environment variables can be set here as well:

- **IDM_RES_DIR**= path of results files in HFS
- **IDM_MNB_DIR**= path of mining bases in HFS

During normal operation the server module *idmcserv* should be used. For debugging and problem analysis purposes the module *idmcser1* is provided. This module writes diagnosis messages into STDERR.

Use this module in the PARM field if you experience one of the following errors:

- Client/server connection problems
- Security problems
- DB2 problems
- Server abnormal terminations or hangs

Rerun the test and provide the STDERR file to the IBM Support Center.

B.14 IDMCFLD

This sample JCL runs a customization program that checks for customer defined computed fields. The result of the check is stored in the HFS member *idmdfncf.dcl* in the path defined by the environment variable *IDM_BIN_DIR*.

Before running this job:

1. Update the job card as required for your installation.
2. change:
 - */usr/lpp/IMiner* to the path you selected.
 - *#HLQIM* to the high level qualifier of the product libraries.
 - *#HLQLE* to the high level qualifier of the LE run time libraries.
 - *#xxx* to your user directory.

Submit this job under the user ID that will start the IM start-up job.

Note

You need to run this job only if you define your own computed fields. It can be run later when need might be.

Appendix C. Special Notices

This publication is intended to help business analysts and information system specialists understand the planning and implementation efforts required when adding the data mining functionality to an existing Business Intelligence environment. The information in this publication is not intended as the specification of any programming interfaces that are provided by Intelligent Miner for data.

References in this publication to IBM products, programs or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM product, program, or service is not intended to state or imply that only IBM's product, program, or service may be used. Any functionally equivalent program that does not infringe any of IBM's intellectual property rights may be used instead of the IBM product, program or service.

Information in this book was developed in conjunction with use of the equipment specified, and is limited in application to those specific hardware and software products and levels.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to the IBM Director of Licensing, IBM Corporation, 500 Columbus Avenue, Thornwood, NY 10594 USA.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact IBM Corporation, Dept. 600A, Mail Drop 1329, Somers, NY 10589 USA.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The information contained in this document has not been submitted to any formal IBM test and is distributed AS IS. The information about non-IBM ("vendor") products in this manual has been supplied by the vendor and IBM assumes no responsibility for its accuracy or completeness. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no

guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environments do so at their own risk.

Any pointers in this publication to external Web sites are provided for convenience only and do not in any manner serve as an endorsement of these Web sites.

Reference to PTF numbers that have not been released through the normal distribution process does not imply general availability. The purpose of including these reference numbers is to alert IBM customers to specific information relative to the implementation of the PTF when it becomes available to each customer according to the normal IBM PTF distribution process.

The following terms are trademarks of the International Business Machines Corporation in the United States and/or other countries:

AIX	AS/400
CICS	DATABASE 2
DB2	IBM ®
IMS	Information Warehouse
Intelligent Miner	MVS
MVS/ESA	OS/2
OS/390	OS/400
RS/6000	S/390
Visual Warehouse	

The following terms are trademarks of other companies:

Java and HotJava are trademarks of Sun Microsystems, Incorporated.

Microsoft, Windows, Windows NT, and the Windows 95 logo are trademarks or registered trademarks of Microsoft Corporation.

Pentium, MMX, ProShare, LANDesk, and ActionMedia are trademarks or registered trademarks of Intel Corporation in the U.S. and other countries.

UNIX is a registered trademark in the United States and other countries licensed exclusively through X/Open Company Limited.

SET and the SET logo are trademarks owned by SET Secure Electronic Transaction LLC.

Other company, product, and service names may be trademarks or service marks of others.

Appendix D. Related Publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this redbook.

D.1 International Technical Support Organization Publications

For information on ordering these ITSO publications see “How to Get ITSO Redbooks” on page 195.

- *Data Modeling Techniques for Data Warehousing*, SG24-2238
- *From Multiple Operational Data to Data Warehousing and Business Intelligence*, SG24-5174
- *Managing Multidimensional Data Marts with Visual Warehouse and DB2 OLAP Server*, SG24-5270
- *Discovering Data Mining*, SG24-2566
- *Intelligent Miner for Data Applications Guide*, SG24-5252
- *Mining Relational and Nonrelational Data with IBM Intelligent Miner for Data*, SG24-5278
- *MVS/ESA Open Edition DCE: PLANNING*, SC09-1484
- *Accessing OS/390 OpenEdition MVS from the Internet*, SG24-4721
- *TCP/IP for MVS: Customization and Administration Guide*, SC31-7134

D.2 Redbooks on CD-ROMs

Redbooks are also available on CD-ROMs. **Order a subscription** and receive updates 2-4 times a year at significant savings.

CD-ROM Title	Subscription Number	Collection Kit Number
System/390 Redbooks Collection	SBOF-7201	SK2T-2177
Networking and Systems Management Redbooks Collection	SBOF-7370	SK2T-6022
Transaction Processing and Data Management Redbook	SBOF-7240	SK2T-8038
Lotus Redbooks Collection	SBOF-6899	SK2T-8039
Tivoli Redbooks Collection	SBOF-6898	SK2T-8044
AS/400 Redbooks Collection	SBOF-7270	SK2T-2849
RS/6000 Redbooks Collection (HTML, BkMgr)	SBOF-7230	SK2T-8040
RS/6000 Redbooks Collection (PostScript)	SBOF-7205	SK2T-8041

CD-ROM Title	Subscription Number	Collection Kit Number
RS/6000 Redbooks Collection (PDF Format)	SBOF-8700	SK2T-8043
Application Development Redbooks Collection	SBOF-7290	SK2T-8037

D.3 Other Publications

These publications are also relevant as further information sources:

- *Data Mining*, ISBN 0-201-40380-3
- *Data Warehousing*, ISBN 0-07-041034-8

How to Get ITSO Redbooks

This section explains how both customers and IBM employees can find out about ITSO redbooks, redpieces, and CD-ROMs. A form for ordering books and CD-ROMs by fax or e-mail is also provided.

- **Redbooks Web Site** <http://www.redbooks.ibm.com/>

Search for, view, download or order hardcopy/CD-ROM redbooks from the redbooks web site. Also read redpieces and download additional materials (code samples or diskette/CD-ROM images) from this redbooks site.

Redpieces are redbooks in progress; not all redbooks become redpieces and sometimes just a few chapters will be published this way. The intent is to get the information out much quicker than the formal publishing process allows.

- **E-mail Orders**

Send orders via e-mail including information from the redbooks fax order form to:

	e-mail address
In United States	usib6fpl@ibmmail.com
Outside North America	Contact information is in the "How to Order" section at this site: http://www.elink.ibm.link.ibm.com/pbl/pbl/

- **Telephone Orders**

United States (toll free)	1-800-879-2755
Canada (toll free)	1-800-IBM-4YOU
Outside North America	Country coordinator phone number is in the "How to Order" section at this site: http://www.elink.ibm.link.ibm.com/pbl/pbl/

- **Fax Orders**

United States (toll free)	1-800-445-9269
Canada	1-403-267-4455
Outside North America	Fax phone number is in the "How to Order" section at this site: http://www.elink.ibm.link.ibm.com/pbl/pbl/

This information was current at the time of publication, but is continually subject to change. The latest information for customer may be found at <http://www.redbooks.ibm.com/> and for IBM employees at <http://w3.itso.ibm.com/>.

IBM Intranet for Employees

IBM employees may register for information on workshops, residencies, and redbooks by accessing the IBM Intranet Web site at <http://w3.itso.ibm.com/> and clicking the ITSO Mailing List button. Look in the Materials repository for workshops, presentations, papers, and Web pages developed and written by the ITSO technical professionals; click the Additional Materials button. Employees may also view redbook, residency, and workshop announcements at <http://inews.ibm.com/>.

IBM Redbook Fax Order Form

Please send me the following:

Title	Order Number	Quantity

First name	Last name
------------	-----------

Company

Address

City	Postal code	Country
------	-------------	---------

Telephone number	Telefax number	VAT number
------------------	----------------	------------

<input type="checkbox"/> Invoice to customer number	
---	--

<input type="checkbox"/> Credit card number	
---	--

Credit card expiration date	Card issued to	Signature
-----------------------------	----------------	-----------

We accept American Express, Diners, Eurocard, Master Card, and Visa. Payment by credit card not available in all countries. Signature mandatory for credit card payment.

Glossary

A

adaptive connection. A numeric weight used to describe the strength of the connection between two processing units in a neural network. The connection is called adaptive because it is adjusted during training. Values typically range from zero to one, or -0.5 to +0.5.

aggregate. To summarize data in a field.

application programming interface (API). A functional interface supplied by the operating system or a separate orderable licensed program that allows an application program written in a high-level language to use specific data or functions of the operating system or the licensed program.

architecture. The number of processing units in the input, output, and hidden layer of a neural network. The number of units in the input and output layers is calculated from the mining data and input parameters. An intelligent data mining agent calculates the number of hidden layers and the number of processing units in those hidden layers.

associations. The relationship of items in a transaction in which those items imply the presence of other items in the same transaction.

attribute. Characteristics or properties that can be controlled, usually to obtain a required appearance. For example, color is an attribute of a line. In object-oriented programming, a data element defined within a class.

B

back propagation. A general-purpose neural network named for the method used to adjust weights while learning data patterns. The Classification - Neural mining function uses such a network.

boundary field. The upper limit of an interval as used for discretization using ranges of a processing function.

bucket. One of the bars in a bar chart showing the frequency of a specific value.

C

categorical values. Discrete, nonnumerical data represented by character strings; for example, colors or special brands.

chi-square test. A test to check whether two variables are statistically dependent or not. Chi-square is calculated by subtracting the expected frequencies (imaginary values) from the observed frequencies (actual values). The expected frequencies represent the values that were to be expected if the variable question were statistically independent.

classification. The assignment of objects into groups or categories based on their characteristics.

cluster. A group of records with similar characteristics.

cluster prototype. The attribute values that are typical of all records in a given cluster. Used to compare the input records to determine whether a record should be assigned to the cluster represented by these values.

clustering. A mining function that creates groups of data records within the input data on the basis of similar characteristics. Each group is called a *cluster*.

confidence factor. Indicates the strength or the reliability of the associations detected.

continuous field. A field that can have any floating point number as its value.

D

DATABASE 2 (DB2). An IBM relational database management system.

database table. A table residing in a database.

database view. An alternative representation of data from one or more database tables. A view can include all or some of the columns contained in the database table or tables on which it is defined.

data field. In a database table, the intersection from table description and table column where the corresponding data is entered.

data format. There are different kinds of data formats, for example, database tables, database views, pipes, or flat files.

data table. A data table, regardless of the data format it contains.

data type. There are different kinds of Intelligent Miner data types, for example, discrete numeric, discrete nonnumeric, binary, or continuous.

discrete. Pertaining to data that consists of distinct elements such as character or to physical quantities having a finite number of distinctly recognizable values.

discretization. The act of making mathematically discrete.

E

envelope. The area between two curves that are parallel to a curve of time-sequence data. The first curve runs above the curve of time-sequence data, the second one below. Both curves have the same distance to the curve of time-sequence data. The width of the envelope, that is, the distance from the first parallel curve to the second, is defined as epsilon.

epsilon. The maximum width of an envelope that encloses a sequence. Another sequence is epsilon-similar if it fits in this envelope.

epsilon-similar. Two sequences are epsilon-similar if one sequence does not go beyond the envelope that encloses the other sequence.

equality compatible. Pertaining to different data types that can be operands for the = logical operator.

Euclidean distance. The square root of the sum of the squared differences between two numeric vectors. The Euclidean distance is used to calculate the error between the calculated network output and the target output in neural classification, to calculate the difference between a record and a prototype cluster value in neural clustering. A zero value indicates an exact match; larger numbers indicate greater differences.

F

field. A set of one or more related data items grouped for processing. In this document, with regard to database tables and views, *field* is synonymous with *column*.

file. A collection of related data that is stored and retrieved by an assigned name.

file name. (1) A name assigned or declared for a file. (2) The name used by a program to identify a file.

flat file. (1) A one-dimensional or two-dimensional array; a list or table of items. (2) A file that has no hierarchical structure.

formatted information. An arrangement of information into discrete units and structures in a manner that facilitates its access and processing. Contrast with *narrative information*.

F-test. A statistical test that checks whether two estimates of the variances of two independent samples are the same. In addition, the F-test checks whether the null hypothesis is true or false.

function. Any instruction or set of related instructions that perform a specific operation.

fuzzy logic. In artificial intelligence, a technique using approximate rules of inference in which truth values and quantifiers are defined as possibility distributions that carry linguistic labels.

I

input data. The metadata of the database table, database view, or flat file containing the data you specified to be mined.

input layer. A set of processing units in a neural network which present the numeric values derived from user data to the network. The number of fields and type of data in those fields are used to calculate the number of processing units in the input layer.

instance. In object-oriented programming, a single, actual occurrence of a particular object. Any level of the object class hierarchy can have instances. An instance can be considered in terms of a copy of the object type frame that is filled in with particular information.

interval. A set of real numbers between two numbers either including or excluding both of them.

interval boundaries. Values that represent the upper and lower limits of an interval.

item category. A categorization of an item. For example, a room in a hotel can have the following categories: Standard, Comfort, Superior, Luxury. The lower category is called the child item category. Each child item category can have several parent item categories. Each parent item category can have several grandparent item categories.

item description. The descriptive name of a character string in a data table.

item ID. The identifier for an item.

item set. A collection of items. For example, all items bought by one customer during one visit to a department store.

K

Kohonen Feature Map. A neural network model comprised of processing units arranged in an input layer and output layer. All processors in the input layer are connected to each processor in the output layer by an adaptive connection. The learning algorithm used involves competition

between units for each input pattern and the declaration of a winning unit. Used in neural clustering to partition data into similar record groups.

L

large item sets. The total volume of items above the specified support factor returned by the Associations mining function.

learning algorithm. The set of well-defined rules used during the training process to adjust the connection weights of a neural network. The criteria and methods used to adjust the weights define the different learning algorithms.

learning parameters. The variables used by each neural network model to control the training of a neural network which is accomplished by modifying network weights.

lift. Confidence factor divided by expected confidence.

M

metadata. In databases, data that describes data objects.

mining. Synonym for analyzing or searching.

mining base. A repository where all information about the input data, the mining run settings, and the corresponding results is stored.

model. A specific type of neural network and its associated learning algorithm. Examples include the Kohonen Feature Map and back propagation.

N

narrative information. Information that is presented according to the syntax of a natural language. Contrast with formatted information.

neural network. A collection of processing units and adaptive connections that is designed to perform a specific processing function.

Neural Network Utility (NNU.) A family of IBM application development products for creating neural network and fuzzy rule system applications.

nonsupervised learning. A learning algorithm that requires only input data to be present in the data source during the training process. No target output is provided; instead, the desired output is discovered during the mining run. A Kohonen Feature Map, for example, uses nonsupervised learning.

NP-complete. In the context on neurocomputing: A learning algorithm is NP-complete if it converges to a solution in time polynomial in size of the problem and the accuracy required.

O

offset. (1) The number of measuring units from an arbitrary starting point in a record, area, or control block, to some other point. (2) The distance from the beginning of an object to the beginning of a particular field.

operator. (1) A symbol that represents an operation to be done. (2) In a language statement, the lexical entity that indicates the action to be performed on operands.

output data. The metadata of the database table, database view, or flat file containing the data being produced or to be produced by a function.

output layer. A set of processing units in a neural network which contain the output calculated by the network. The number of outputs depends on the number of classification categories or maximum cluster value in neural classification and neural clustering, respectively.

P

pass. One cycle of processing a body of data.

prediction. The dependency and the variation of one field's value within a record on the other fields within the same record. A profile is then

generated that can predict a value for the particular field in a new record of the same form, based on its other field values.

processing unit. A processing unit in a neural network is used to calculate an output by summing all incoming values multiplied by their respective adaptive connection weights.

Q

quantile. One of a finite number of nonoverlapping subranges or intervals, each of which is represented by an assigned value.

Q is an *N%* -*quantile* of a value set *S* when:

- Approximately *N* percent of the values in *S* are lower than or equal to *Q*.

- Approximately (100-*N*) percent of the values are greater than or equal to *Q*.

The approximation is less exact when there are many values equal to *Q*. *N* is called the quantile label. The 50%.-quantile represents the median.

R

radial basis function (RBF). In data mining functions, radial basis functions are used to predict values. They represent functions of the distance or the radius from a particular point. They are used to build up approximations to more complicated functions.

record. A set of one or more related data items grouped for processing. In reference to a database table, *record* is synonymous with *row*.

region. (Sub)set of records with similar characteristics in their active fields. Regions are used to visualize a prediction result.

round-robin method. A method by which items are sequentially assigned to units. When an item has been assigned to the last unit in the series, the next item is assigned to the first again. This process is repeated until the last item has been assigned. The Intelligent Miner uses this method, for example, to store records in output files during a partitioning job.

rule. A clause in the form head<== body. It specifies that the head is true if the body is true.

rule body. Represents the specified input data for a mining function.

rule group. Covers all rules containing the same items in different variations.

rule head. Represents the derived items detected by the Associations mining function.

S

scale. A system of mathematical notation; fixed-point or floating-point scale of an arithmetic value.

scaling. To adjust the representation of a quantity by a factor in order to bring its range within prescribed limits.

scale factor. A number used as a multiplier in scaling. For example, a scale factor of 1/1000 would be suitable to scale the values 856, 432, -95, and /182 to lie in the range from -1 to +1, inclusive.

self-organizing feature map. See *Kohonen Feature Map*

sensitivity analysis report. An output from the Classification - Neural mining function that shows which input fields are relevant to the classification decision.

sequential patterns. Intertransaction patterns such that the presence of one set of items is followed by another set of items in a database of transactions over a period of time.

similar time sequences. Occurrences of similar sequences in a database of time sequences.

Structured Query Language (SQL). An established set of statements used to manage information stored in a database. By using these statements, users can add, delete, or update information in a table, request information through a query, and display results in a report.

supervised learning. A learning algorithm that requires input and resulting output pairs to be presented to the network during the training

process. Back propagation, for example, uses supervised learning and makes adjustments during training so that the value computed by the neural network will approach the actual value as the network learns from the data presented. Supervised learning is used in the techniques provided for predicting classifications as well as for predicting values.

support factor. Indicates the occurrence of the detected association rules and sequential patterns based on the input data.

symbolic name. In a programming language, a unique name used to represent an entity such as a field, file, data structure, or label. In the Intelligent Miner you specify symbolic names, for example, for input data, name mappings, or taxonomies.

T

taxonomy. Represents a hierarchy or a lattice of associations between the item categories of an item. These associations are called taxonomy relations.

taxonomy relation. The hierarchical associations between the item categories you defined for an item. A taxonomy relation consists of a child item category and a parent item category.

trained network. A neural network containing connection weights that have been adjusted by a learning algorithm. A trained network can be considered a virtual processor; it transforms inputs to outputs.

training. The process of developing a model which understands the input data. In neural networks, the model is created by reading the records of the input and modifying the network weights until the network calculates the desired output data.

translation process. Converting the data provided in the database to scaled numeric values in the appropriate range for a mining kernel using neural networks. Different techniques are used depending on whether the

data is numeric or symbolic. Also, converting neural network output back to the units used in the database.

transaction. A set of items or events that are linked by a common key value, for example, the articles (items) bought by a customer (customer number) on a particular date (transaction identifier). In this example, the customer number represents the key value.

transaction ID. The identifier for a transaction, for example, the date of a transaction.

transaction group. The identifier for a set of transactions. For example, a customer number can represent a transaction group that includes all purchases of a particular customer during the month of May.

V

vector. A quantity usually characterized by an ordered set of numbers.

W

weight. The numeric value of an adaptive connection representing the strength of the connection between two processing units in a neural network.

winner. The index of the cluster which has the minimum Euclidean distance from the input record. Used in the Kohonen Feature Map to determine which output units will have their weights adjusted.

List of Abbreviations

ADK	application development toolkit	DSS	decision support system
ANN	artificial neural network	EIS	executive information system
ANSI	American National Standards Institute	ESO	expanded service option
APPC	advanced program to program communication	FMID	function modification identifier
API	application programming interface	GID	group ID
ASCII	American National Standard Code for Information Interchange	GUI	graphical user interface
CAE	client application enabler	HFS	hierarchical file system
CBIPO	custom-build installation process offering	HLQ	high level qualifier
CBPDO	custom-build product delivery offering	HTML	Hypertext Markup Language
DML	data manipulation language	HLQ	high level qualifier
DAM	data access module	IBM	International Business Machines Corporation
DARM	data archive retrieval manager	IDS	intelligent decision support
DBMS	database management system	ISO	International Organization for Standardization
DCL	data control language	I/O	input/output
DDL	data definition language	IM	Intelligent Miner
DML	data manipulation language	IMS	Information Management System
DRDA	distributed relational database architecture	IT	information technology
DUW	distributed unit of work	ITSO	International Technical Support Organization
DW	data warehouse	JCL	job control language
		JDBC	java database connectivity
		JDK	java developers kit
		JRE	java runtime environment
		LIS	large item set

LOB	large object
LPP	licensed program product
MLP	multilayer perceptron
ODBC	Open Database Connectivity
OEM	original equipment manufacturer
OLAP	on-line analytical processing
OLTP	on-line transaction processing
OSA	open systems adapter
PSP	preventive service planning
RACF	resource access control facility
RAM	random access memory
RBF	radial basis function
RBFN	radial basis function network
RDBMS	relational database management system
ROI	return on investment
RUW	remote unit of work
SDSF	system display and search facility
SMP/E	system modification program/enhanced
SQL	structured query language
TCP/IP	Transmission Control Protocol/Internet Protocol
UDF	user-defined function
UDT	user-defined type
VSAM	Virtual Storage Access Method

Index

A

- access control 45
- aggregate values 54
- AIX
 - group 108
 - hardware requirements 100
 - IDM_MNB_DIR 115
 - IDM_RES_DIR 115
 - installation verification 115
 - networking requirements 105
 - password 111
 - product installation 108
 - software requirements 101
 - user 109
- algorithm categorization 57
- analysis 9
 - discovery-driven 9
 - factor 56
 - principal component 56
 - verification-driven 9
- application domains 41
- application mode 56
- applications 11
- association discovery 57, 69
 - data preparation 69
 - result 70
 - sample 71
- associations 57
- auditing 46

B

- bivariate statistic 56
- business analysis 21
 - cycle 22
 - objectives 22
 - requirements 22
 - roles 22
- business analyst 31
- business application
 - implementation 28
 - support 28
- business drivers 15
- business environment 17
- business feedback 29

- roles 29
- Business Intelligence 3
 - components 8
 - data structure 33
 - drivers 3
 - environment 8
 - evolution 4
 - introduction 3
 - sample environment 63
- business interaction 17
- business requirement 23

C

- calculate values 54
- classification 57, 59, 74
 - data preparation 75
 - result 75
 - sample 75, 76
 - visualization 75
- cluster visualization 73
- clustering 26, 57, 58, 72
 - data preparation 72
 - result 73
 - sample 74
- conditional probability 70
- confidence 70
- contents container 53
- correlations 57
 - chance 57
 - known 57
 - unknown and important 57
 - unknown but trivial 57
- create
 - data mining operations 61
 - results 60
- cross-fertilization 6
- cross-validation 26
- culture 17
- customer satisfaction 21

D

- data 24
 - aggregation 41
 - article information 67
 - central repository 8
 - clean 9

- cleansing 41
- consistency 39
- cross-validation 26
- DB2 table 66
- denormalization 41
- dimension reduction 26
- discovery 30
- imputation 25
- inconsistency 9
- integrated 9
- integration 41
- intercorrelation 26
- manipulation 26
- mapping 41
- missing values 25
- normalization 26
- organization information 68
- outliers 25
- placement 43
- preparation 25
- preparation function 54
- propagation 40
- quality 23
- relationships 4
- replication 39
- requirements 23
- sales information 66
- sample 66
- security 45
- sources 33
- splitting 26
- transformation 41
- validated 9
- visualization 27
- VSAM KSDS cluster 66
- data analysis 23
 - cycle 23
 - goal 23
 - roles 23
- data architect 31
- data considerations 33
- data copying 41
- data delivery 41
- data discovery 30
- data gathering 24
 - building a data mart 24
 - cycle 24
 - roles 24
- data manipulation 6, 26
- data mart 24
- data mining 8, 10
 - activities 22
 - applications 11, 13
 - basic principles 3
 - benefits 10
 - business feedback 29
 - data considerations 33
 - database design 48
 - definition 10
 - deliverables 22
 - drivers 15
 - enabler 16
 - expert 31
 - getting started 15
 - goals 22
 - implement techniques 69
 - inhibitors 17
 - input 9
 - iterations 20
 - maintenance 48
 - model validity 49
 - operations 11, 13
 - operations creation 61
 - organizational environment 17
 - outlook 6
 - output 10
 - parallelism 47
 - performance 46
 - process 20
 - questions 15
 - requirements 11, 22
 - result interpretation 27
 - result types 43
 - roles 27, 31
 - scalability 47
 - security 45
 - techniques 11
 - time 21
 - toolkit 51
 - update models 29
 - validity 49
- data modeler 31
- data preparation 25
 - process 33
- data propagation 40
 - apply 40
 - capture 40
- data quality 25

- data replication 39
 - architecture 39
 - products 39
- data repository 8, 24
- Data Sources 33
- data stores 39
- data transformation 41
 - stages 41
- data warehouse
 - for data mining 41
 - requirements 42
- Data Warehousing 6
 - concept 6
 - impacts 7
 - reasons 7
- database segmentation 13
- decision cycle 5
- decision making 5
- decision optimization 30
- decision tree 75
- demographic data 33
- deviation detection 14
- DFSORT 40
- DHCP 89, 90
- diagnostics tool 84
- dimension reduction 26
- discovery
 - association 57
 - sequential pattern 58
- discretize
 - quantiles 55
 - ranges 55
- disk space 83
 - AIX 100
 - OS/390 141
 - OS/400 129
 - Sun/Solaris 120
 - Windows
 - NT 85
- display settings 85
- DL/I database 68
- DPropNR 40
- DPropR 40
- DRDA 40

E

- external sources 33
- extract knowledge 3

F

- factor analysis 56
- field
 - filter 55
 - pivot 55
- filter
 - fields 55
 - records 55
- function
 - data mining 56
 - data preparation 54
 - statistical mining 56

G

- generate knowledge 4

H

- hostname 89, 107
- hostnames 90
- hosts file 90
- hypothesis verification 9

I

- imputation 25
- information discovery 9
- Intelligent Miner 51
 - block diagram 51
 - client 52
 - components 90
 - components on Windows NT 92
 - data preparation functions 54
 - modes 56
 - overview 51
 - server 52
 - task guide 53
 - user interface 53
 - Windows NT 83
 - working with databases 52
- intercorrelation 26
- interdependencies 9
- ITarchitect 31

J

- join data sources 55

K

knowledge
 creation 6
 generation 4
 requirements 5
knwoledge
 cycle 6

L

learning organization 6
library path 88
lift 70
linear regression 56
link analysis 13
localities 9

M

machine learning 3, 16
maintenance 48
memory 83, 100, 120, 129
metadata 43
mining
 algorithm 57
 function 56
mining base container 53
mining bases 94
mining parallelism 48
missing values 25
mode 56
 application 56
 test 56
 training 56

N

noise 9

O

OLAP 8
online analytical processing
 see OLAP 4
operational data 33
operational systems 7
operations 11
organizational culture 17
organizational environment 17
OS/309
 installation customization 154

OS/390

DB2 considerations 173
group 154
installation procedure 148
user 158

OS/400

hardware requirements 129
installation verification 135
networking requirements 131
product installation 134
relational databases 132
software requirements 129
user 134

outliers 25

overfitting 26

P

parallelism 47
 mining 48
 server 48
password 93
path 88
Performance 46
pivot field 55
portmapper 96
prediction 57, 76
 data preparation 76
 result 77
 sample 78
predictive modeling 13
principal component analysis 56
processor 83
project manager 31

R

RACF 154
refresh propagation 40
result
 create mining 60
 interpretation 70
 object 60
 visualize 60
result data 43
result interpretation 27
 cycle 27
 roles 27
result placament 43
 for analysts 43

- for applications 44
- for decision makers 44
- for resources 44

Roles 31

S

- security 45
- sequential patterns 57, 58, 71
 - data preparation 71
 - result 72
 - sample 72
- server parallelism 48
- similar time sequence 57, 60
- similar time sequences 78
 - data preparation 78
 - result 78
 - sample 79
- solution architect 31
- staging table 40
- statistical function 56
- statistics
 - bivariate 56
 - univariate 56
- Sun Solaris
 - hardware requirements 119
 - installation verification 127
 - networking requirements 122
 - product installation 124
 - software requirements 120
- Sun/Solaris
 - group 124
 - IDM_MNB_DIR 127
 - IDM_RES_DIR 127
 - password 125
 - user 125
- support 70

T

- task guide 53
- TCP/IP 89
- technical environment 17
- technology enablers 16
- test mode 56
- test set 26
- training mode 56
- training set 26
- transactional data 69

U

- univariate curve fitting 56
- univariate statistic 56
- user interface 53

V

- value
 - aggregate 54
 - calculate 54
 - missing 55
 - nonvalid 55
 - prediction 60
- values
 - encode missing 55
- visualize results 60
- VSAM cluster 66
- VSAM dataset 68

W

- Windows NT
 - administrative tools 84
 - hardware requirements 83
 - IDM_MNB_DIR 96
 - IDM_RES_DIR 96
 - installation verification 96
 - networking requirements 89
 - password 93
 - product installation 90
 - software requirements 85
 - user name 93
- workarea 53

ITSO Redbook Evaluation

Intelligent Miner for Data: Enhance Your Business Intelligence
SG24-5422-00

Your feedback is very important to help us maintain the quality of ITSO redbooks. **Please complete this questionnaire and return it using one of the following methods:**

- Use the online evaluation form found at <http://www.redbooks.ibm.com>
- Fax this form to: USA International Access Code + 1 914 432 8264
- Send your comments in an Internet note to redbook@us.ibm.com

Which of the following best describes you?

☐ **Customer** ☐ **Business Partner** ☐ **Solution Developer** ☐ **IBM employee**
☐ **None of the above**

Please rate your overall satisfaction with this book using the scale:
(1 = very good, 2 = good, 3 = average, 4 = poor, 5 = very poor)

Overall Satisfaction _____

Please answer the following questions:

Was this redbook published in time for your needs? Yes____ No____

If no, please explain:

What other redbooks would you like to see published?

Comments/Suggestions: (THANK YOU FOR YOUR FEEDBACK!)

