Master Thesis

"EVALUATION OF DATA MINING METHODS TO SUPPORT DATA WAREHOUSE ADMINISTRATION AND MONITORING IN SAP BUSINESS WAREHOUSE"

Narasimha Raju Alluri

Faculty of Business Applications of Computer Science

UNIVERSITY OF APPLIED SCIENCES - FURTWANGEN
Furtwangen, Germany

SAP AG

Walldorf, Germany

20th April, 2005.

Primary supervisor: **Prof. Dr. Wolfram Reiners**
Secondary Supervisor: **Mr. Kai Willenborg**

Declaration

I hereby declare in lieu of oath that I composed this thesis independently and without inadmissible help from outside.

The sources used are quoted in full.

Walldorf, 20th April, 2005.

Signature

Acknowledgement

A journey is easier when you travel together. Interdependence is certainly more valuable than independence. This thesis is the result of six months of work whereby, I have been accompanied and supported by many people. It is a pleasant aspect that I have now the opportunity to express my gratitude for all of them.

This work was written at Development – Product line BI at, SAP AG, Walldorf, Germany. The development of this thesis was a time of personal growth and that development didn't always take place without pain.

I would like to thank Mr. Hoeltke Rainer and Mr. Willenborg Kai for showing this way of opportunity and letting me know several important things. They always listened to mepatiently and guided throughout the path with regular feedback and suggestions. I would also thank Prof. Dr. Wolfram Reiners for guiding me in this work and providing me necessary tips through out my thesis.

I am very much grateful to Dr. Weidner Jens from CRM Analytics for his supervision and constant source of encouragement, the thoughts he has offered have enriched my thesis without which, this work would not have been materialized in the present form.

I thank all the WHM (Warehouse Management) colleagues- BI Product Line Management for being kind enough to respond to my countless queries.

This list goes on and to make an end at some point. I thank one and all who have rendered help, directly or indirectly in completion of this work

Abstract

Data mining is not new. People who first discovered how to start fire and that the earth is round, also discovered knowledge, which is the main idea of Data mining. Even before technology was used for Data mining, statisticians were using probability and regression techniques to model historical data. Since several years the buzzword "Data Mining" creates a boom in many areas around Data mining. In the 1960s, Management Information Systems and later, in the 1970s, Decision Support Systems were praised for their great potential to supply executives with mountains of data needed to carry out their jobs. But the problem was that they simply supplied too much data and not enough information to be generally useful. Advances in data collection and the computerization of many business transactions flood us today with information, and generate an urgent need for new techniques and tools that can intelligently and automatically assist us in transforming this data into useful knowledge. Today, there is a huge amount of information locked up in the mountains of data in companies' databases – information that is potentially important but has not yet discovered. This is the time were Data mining comes in front and helps to identify this valuable information.

Data mining tools can predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. Today Data mining is primarily used by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. The list of this companies using Data mining technologies looks like a Fortunes-500 Who's who. The tasks of Data mining can be classified in different groups. The two main tasks are the prediction and description of data. Supporting these different tasks of Data mining, there is a variety of Data mining methods. This paper concentrates on evaluation of data mining methods to support Data Warehouse administration and monitoring in SAP BW. Firstly, get to know the product specific knowledge of SAP BW and then investigation in to the possible areas of SAP BW. As well, the analysis of key figures that are available as part of the SAP BW and then looking at the available Data mining methods that are part of SAP BW and a final analysis through a clustering scenario. The detailed screens of the data model, transformations and the results are presented concluding with the testing and analysis of the same data set with IBM Intelligent Miner.

List of Figures

List of Abbreviations

APD: Analysis Process Designer

BI: Business Intelligence

BW: Business Warehouse

CRM: Customer Relationship Management

DSS: Decision Support System

DW: Data Warehouse

DM: Data Mining

ERP Enterprise Resource Planning

IGS: Internet Graphics Server

IT: Information Technology

KDD: Knowledge Discovery in Databases

KM: Knowledge Management

KPI: Key Performance Indicator

MIS: Management Information Systems

MDM: Master Data Management

OLAP: Online Analytical Processing

SAP: System Application and Products (in Data Processing)

XI: Exchange Infrastructure

XML: Extensible Markup Language

Table of Contents

1. Introduction

1.1. The extent of the work

This work is done within the framework of a Master Thesis, in fulfilment of a partial requirement for the award of the degree: Master of Computer Science in Business Consulting by the University of Applied Science (Fachhochschule) Furtwangen, Germany.

It is aimed at "Evaluation of Data Mining Methods to Support Data Warehouse Administration and Monitoring in SAP BW", as to investigate the areas in BW where Data mining methods could be implemented and then to horizontally cut in to the Query perspective and analyse the available key figures. As to accomplish this, what data mining method would be used? Considering the efforts and resources that are involved in known methods, could there possibly be another way same goals could be achieved with less investment in efforts and resources? This work explores how this could be achieved with the available functionally of SAP Analysis Process Designer (APD) and Data mining work bench and at the end, use of some sophisticated algorithms that are part of IBM intelligent miner.

1.2. Document Outline

The documentation of the Master's Thesis is outlined in the following way. Chapter Two presents all about SAP Business Information Warehouse including the history, the evolution, the architecture and the features. Chapter Three presents all about Data Mining – A general introduction running through Knowledge Discovery in Databases, the common uses and the process of Data mining. Chapter four describes about the Data Mining methods available as part of SAP Data mining workbench. Chapter 5 provides with the AS-IS Analysis describing the technical content of SAP BW, Administration and monitoring in SAP BW ending up with the possible areas of data mining with a way forward for the TO-BE analysis. Chapter 6 goes through the motivation for the cluster analysis, the data model concluding with the results of the cluster analysis and a conclusion and the future outlook in chapter 7.

2. The my SAP BW- A Data Warehousing solution from SAP

2.1. History and Evolution

Although it seems that the availability of data warehousing applications has exploded in the 1990s, the recognition of the need for these applications is not at all new. The need for data warehousing originated in the mid-to-late 1980s with the fundamental recognition that information systems must be distinguished into operational and informational systems. [Delvin 1997] Operational systems support the day-to-day conduct of the business, and are optimized for fast response time of predefined transactions, with a focus on update transactions. Operational data is a current and real-time representation of the business state. In contrast, informational systems are used to manage and control the business. They support the analysis of data for decision making about how the enterprise will operate now and in the future. They are designed mainly for ad-hoc, complex and mostly read-only queries over data obtained from a variety of sources. Informational data is historical, i.e., it represents a stable view of the business over a period of time.

Over the past several years, the concept of a data warehouse has undergone many changes. The first data warehouses were created to answer the demands of business managers and executives who wanted to be able to extract, from the volumes of data produced by their in-house operational applications, the key historical and summary data that would allow them to better plan, analyze, and control their business enterprises. The initial data warehouses answered this need by providing integrated historical and summary data.

The next phase in the data warehouse evolution was the creation of the data mart. Independent data marts were appealing in part because they were small, relatively cheap, and quick to implement. Like the spread of the PC before them, these independent data marts created "islands of information" as each user created and stocked ones own data mart. To address the challenges raised by the spread of multiple independent data marts, the concept of the dependent data mart has emerged. In a dependent data mart architecture there is a central data warehouse that contains the "corporate view of the data" and supplies the departmental data marts with the specific data they require.

2.1.1. What is Data Warehousing?

For most large companies today, Enterprise Resource Planning (ERP) serves within the core set of day-today transaction processing applications. Even when data is efficiently captured and stored in ERP systems, it may remain relatively useless for reporting and decision making purposes. From General Ledger to Human Resources, ERP systems do the central work of running, tracking and reporting on business data processing.

Just as ERP is critical to business transactions, so is data warehousing central to analysis of those transactions. Through data warehousing product managers, marketing managers, communications specialists, human resources recruiters, financial executives, CIO's and CEO's formulate analytical queries and obtain reports; with these, they are able to make tactical and strategic business decisions — faster and better than ever before, thanks to their analytical resources.

According to Ralph Kimball, "data warehouse is a copy of transaction data specifically structured for query and analysis." [Kimball 1996]

“A Data Warehouse is a repository of integrated information, available for queries and analysis. Data and information are extracted from heterogeneous sources as they are generated.... This makes it much easier and more efficient to run queries over data that originally came from different sources." –Stanford University [SU 2002]

According to Bill Inmon, known as the father of Data Warehousing, a data warehouse is a subject oriented, integrated, time-variant, non-volatile collection of data in support of management decisions.

- **Subject-oriented** means that all relevant data about a subject is gathered and stored as a single set in a useful format;
- **Integrated** refers to data being stored in a globally accepted fashion with consistent naming conventions, measurements, encoding structures, and physical attributes, even when the underlying operational systems store the data differently;
- **Non-volatile** means the data warehouse is read-only: data is loaded into the data warehouse and accessed there;
- **Time-variant** data represents long-term data--from five to ten years as opposed to the 30 to 60 days time periods of operational data. . [W.H.Inmon 1999]

2.1.2. Operational Systems Vs Data Warehouse Systems

The fundamental difference between operational systems and data warehousing systems is that operational systems are designed to support transaction processing whereas data warehousing systems are designed to support online analytical processing (or OLAP, for short).
Operational systems are generally designed to support high-volume transaction processing with minimal back-end reporting and are generally process-oriented or process-driven, meaning that they are focused on specific business processes or tasks. Example tasks include billing, registration, etc.
Data warehousing systems are generally designed to support high-volume analytical processing (i.e. OLAP) and subsequent, often elaborate report generation and are generally subject oriented, organized around business areas that the organization needs information about. Such subject areas are usually populated with data from one or more operational systems. As an example, revenue may be a subject area of a data warehouse that incorporates data from operational systems that contain sales data, promotion data, costs data, etc.

Operational systems are generally concerned with current data and are generally updated regularly according to need and are optimized to perform fast inserts and updates of relatively small volumes of data. Data warehousing systems are generally concerned with historical data and are non-volatile, meaning that new data may be added regularly, but once loaded; the data is rarely changed, thus preserving an ever-growing history of information. In short, data within a data warehouse is generally read-only and optimized to perform fast retrievals of relatively large volumes of data.

To facilitate complex analyses and visualization, the data in a warehouse is typically modelled multi-dimensionally. The structural difference of the Operational and Data Warehouse models can be viewed in Figure 1 and Figure 2

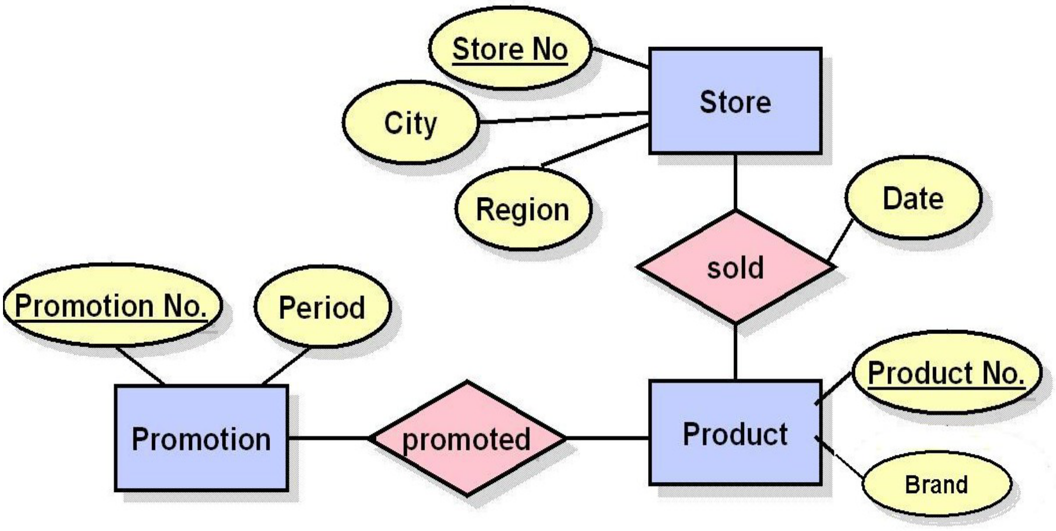


Figure 1: The Operational Data Model
Source: Biao Fu and Henry Fu, SAP BW: A Step-by-step Guide

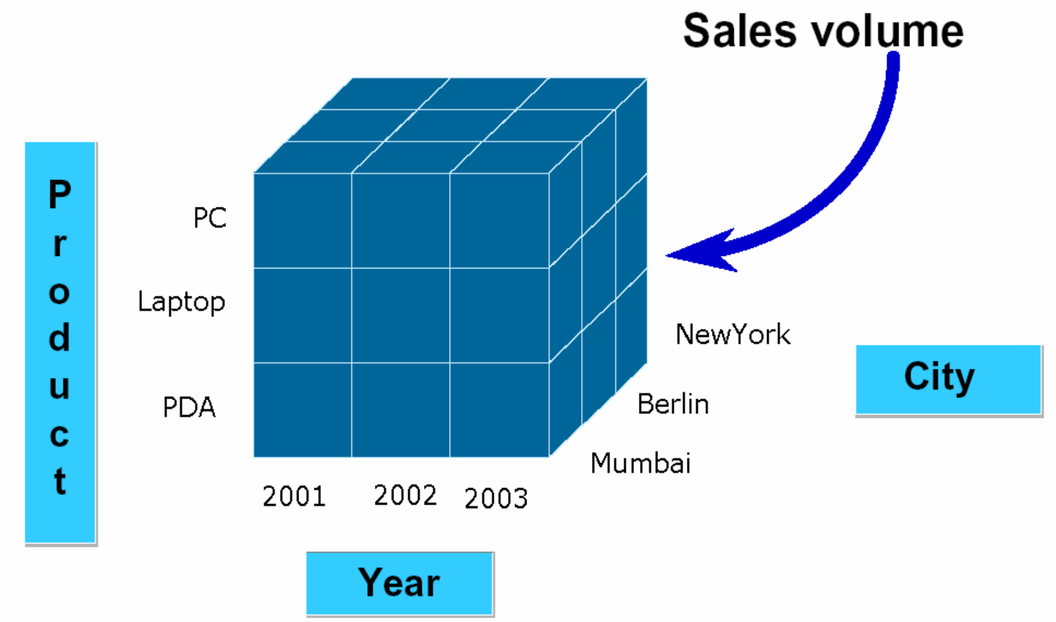


Figure 2: The Dimensional Model
Source: Biao Fu and Henry Fu, SAP BW: A Step-by-step Guide

Data warehousing is a concept. It is a set of hardware and software components that can be used to better analyze the massive amounts of data that companies are accumulating to make better business decisions. Data Warehousing is not just data in the data warehouse, but also the architecture and tools to collect, query, analyze and present information.

Today, there are a lot of data warehouse vendors. Most well known data warehouse vendors are: SAP AG, SPSS Inc., IBM Corporation, Oracle Corporation, SAS Institute, etc. Among them there are also ERP vendors that are offering their data warehouse products to go along with their ERP applications (e.g., Oracle's Data Applications Data Warehouse, SAP's Business Information

Warehouse, etc.). Until recently, ERP vendors focused mostly on enhancing the infrastructure to deliver high-performance OLTP solutions. Very little attention was given to providing applications to analyze massive amounts of data collected, or "jailed" within the ERP data repositories. Customers were left to building data warehousing and reporting solutions without any help from ERP vendors. [Inmon 1999]

Today, ERP vendors recognized that customers need data warehouse tools for their ERP applications. While traditional data warehouse vendors were offering tools that were intended for general purposes, ERP vendors created data warehousing solution that were focused on their own - specific - ERP application as well as external sources. For instance, SAP created "BW - Business Information Warehouse, which has the best interface to extract the data out any Source system.

2.2. SAP BW Architecture

SAP did not invent the software components of the SAP Business Intelligence solution overnight. The business intelligence capabilities found in SAP software evolved in parallel to the CIF and other informational processing frameworks. The SAP BI solution evolution has been quite rapid since the first generally available release of the SAP BW software in 1998. In fact organizations interested in implementing the SAP BW software component to realize a CIF will find themselves licensing the SAP Business Intelligence Solution rather than the SAP BW component. [McDonald et al. 2003]

In 1997, SAP launched an initiative to extend the reporting and analysis capabilities in the R/3 OLTP environment. This initiative, once called the Reporting Server, became the largest development project in the history of SAP after the SAP R/3 development. SAP selected five companies to pilot SAP Business Information Warehouse (BW) in 1997. In 1998, SAP launched a so-called Early Customer Program (ECP) with six customers to gather requirements and to do a proof of concept at customer sites. Release 1.2A of BW was made available to the public in September 1998 [Hashmi 2003, 42].

SAP BW Releases: A brief history of the SAP BW release pattern looks like

- BW 1.2b - introduction of InfoCubes and Business Content
- BW 2.0b - introduction of ODS, s mySAP.com interface
- BW 2.1c - analytical component
- BW 3.0 - further enhancement of ODS into a data warehouse, along with the creation of analytical applications, partnerships, and so forth. .[Inmon 2005]
- BW 3.5 - designed to deliver seamless integration capabilities into all of the SAP NetWeaver components like Information Broadcasting, Universal Data Access, Embedded BI–Integration in to SAP NetWeaver, Business Planning and Simulation, Unicode and so on [SAPNET 2005-1]

SAP NetWeaver provides an open integration and application platform and permits the integration of the Enterprise Services Architecture. You can unify business processes across technological boundaries, integrate applications for your employees as needed, and access and edit simple information easily and in a structured manner. [SAPHELP 2005-1]

The following figure 3 shows the overall architecture and the position of SAPBW in NetWeaver

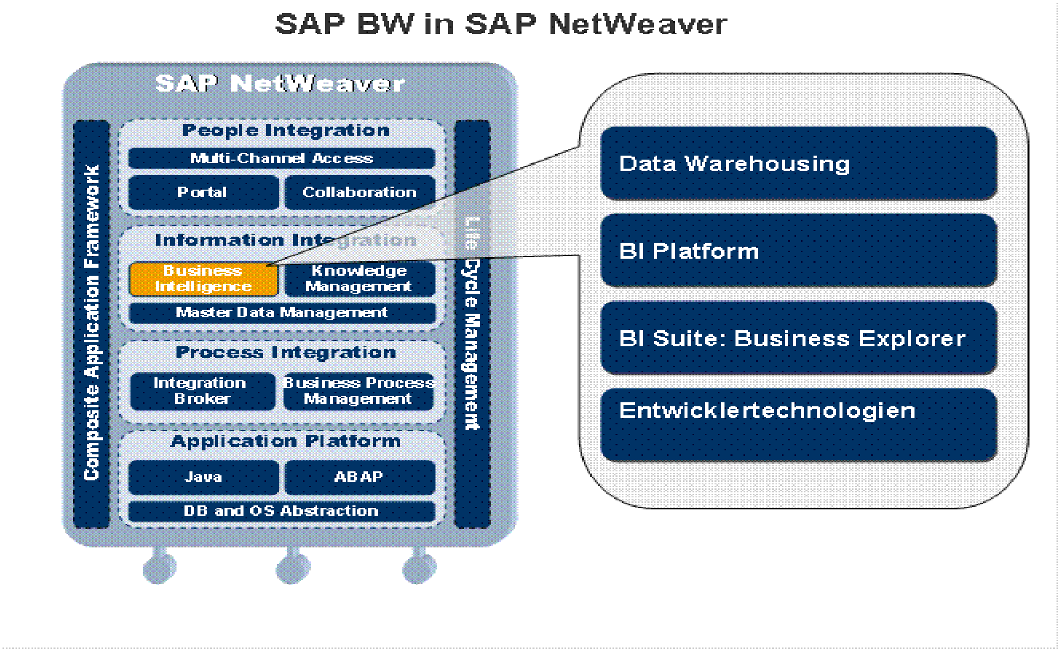


Figure 3: SAP NetWeaver Components
Source: SAP Help Portal, 2005

As of April 12, 2005 Figures BW has 9368 installations world wide with 5307 from EMEA, 2293 Americas and 1768 EAP. [SAPNET 2005-2]

The SAP Business Information Warehouse uses an integrated set of powerful components for data collection, storage, analysis and administration to meet all the requirements for ready-to-go data warehousing. SAP BW is completely based on an integrated Meta data concept, with Meta data being managed by Meta data services.

The SAP BW Meta Data Services components provide both an integrated Meta Data Repository where all Meta data is stored and a Meta Data Manager that handles all requests for retrieving, adding, changing, or deleting Meta data. The Meta Data Repository is integrated into the Administrator Workbench, with a list of all Meta data objects available there. Figure 4 shows the layered architectural structure and components of SAP BW. The SAP BW architecture can be divided into five main layers.

- Administration
- Extraction, Loading and transformation services
- Data Storage and Management
- Analysis and Access Services
- Presentation

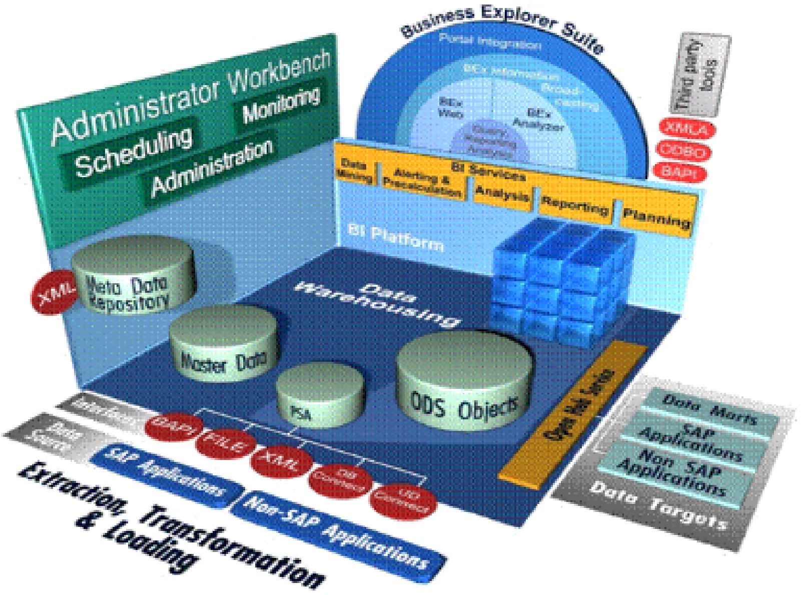


Figure 4: SAP BW Architecture
Source: SAP Help Portal, 2005

2.2.1. Extraction, Transformation and Loading Services

The extraction, transformation, and loading (ETL) services layer of the SAP BW architecture includes services for data extraction, data transformation, and loading of data and serves as a staging area for intermediate data storage for quality assurance purposes. [McDonald et al]. The core part of the ETL services of SAP BW is the staging Engine, which manages the staging process for all the data received from several types of source systems and is supported by the DataSource Manager. The DataSource Manager manages the definitions of the different sources of data known to the SAP BW system and supports five different types of interface which include BAPI, File Interface, XML interface, DB connect interface and UD connect interface as shown in Figure 4.

The BW includes pre-configured, ready-to-go extractors for R/3 applications, slashing the time required to set up extraction routines. At the same time, the BW is not restricted to R/3. It is possible to extract data from diverse SAP, non-SAP and legacy systems. In many cases, for example for products from complementary software partners, SAP has already defined business application interfaces (BAPIs) that guarantee quick implementation of efficient extraction routines. It is even possible to incorporate data from flat files. In other words, the BW can be easily extended to include external data of many kinds, for instance from content providers, demographic surveys or even syndicated POS (Point-of-sale) data reports. The BW also supports Delta extracts, i.e. an extract only of that data which is new or changed, minimizing transfer overhead and system load.

The key to any data warehouse is meta-data, i.e. additional data which describes the data in a way which makes it meaningful, which allows the user and the data warehouse to understand just what that data represents. A major advantage of the BW is that it employs the tried and trusted metadata models that are already part of R/3. In other words, these models do not have to be defined and built laboriously and expensively from scratch. Nevertheless, the BW is able to understand other metadata models permitting the seamless integration of data from legacy systems or external sources. In fact SAP cooperates closely with content providers for special knowledge to ensure the most efficient use is made of such sources. Synchronization means that changes made to the source system are automatically recognized by the BW and there is no need for additional administration so that there is no need for coding work. Transformation rules are employed to scrub, adjust and augment extracted data so that it matches the needs of the warehouse. Typical examples include adding the figures '19' to year dates, where required to create four-figure annotation. Geocoding is a special BW feature that allows data to be presented on easy-to-understand maps of countries or regions. Geocoding is performed just once for all data when it is loaded and is from then on available for all InfoCubes, increasing ease of use and cutting down on administrative overhead and system load. The geographical dimension need not be added by oneself, thus saving time and money. Validation ensures that data is intact beforeit is mapped to the InfoCube preventing problems further downstream.

Depending on the source systems and the type of data basis, the process of loading data into the SAP BW is technically supported in different ways. In the conception phase, the system firstly needs to detect the different data sources in order to be able to transform the data with the suitable tool afterwards. [BW310 2005] The following figure 5 gives a brief overview of the ETL process in SAP BW.

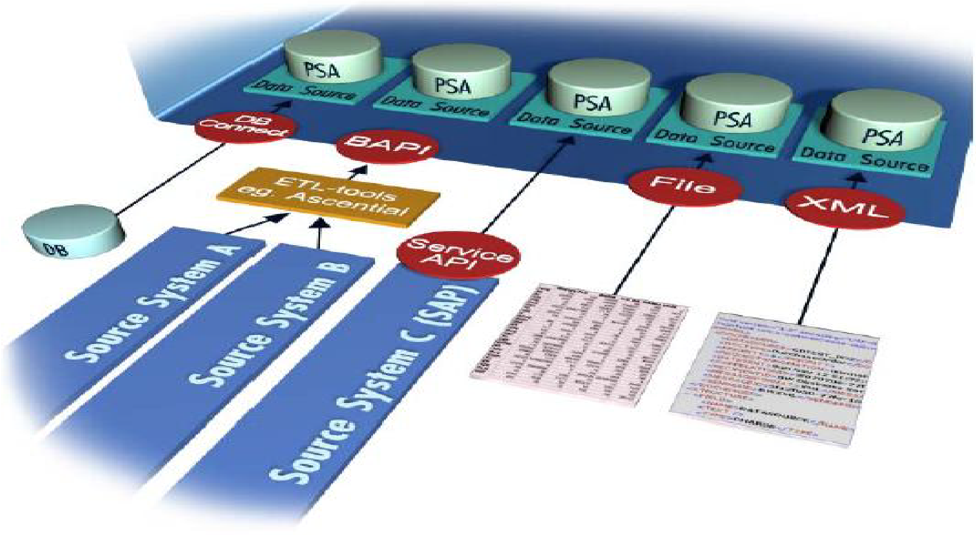


Figure 5: ETL: Extraction, Transformation and Loading in SAP BW
Source: SAPCOURSE, BW310

2.2.2. Data Storage and Management

The Data Storage and Management Layer also called as SAP BW Data Manager manages and provides access to the different data targets available in SAP BW, as well as aggregates stored in relational or multidimensional database management systems.

Data storage is based on an intelligent combination of InfoCubes (information data models) and master data that enriches the depth of knowledge available while ensuring high performance. Master data comprises non-volatile information on typical attributes for customers, companies, suppliers, etc. This can mean addresses, Regions or categories. Often this is drawn directly from R/3 applications, cutting down on maintenance and providing a perspective many other data warehouses simply cannot emulate. InfoCubes are the basis for multidimensional views. They comprise dimensions such as time, geographic region or product type, and key figures, i.e. actual volumes or quantities. A large number of prefabricated Info- Cubes are provided with the BW, allowing to begin exploring data immediately without having to spend time building InfoCubes. In addition, the InfoCubes can be modified or defined, as new needs arise.

Info Cubes provide a flexible means of aggregating data in accordance with one's needs. Many InfoCubes are pre-defined, reducing the time required to set up your warehouse. Aggregation mechanism is completely transparent to the end-user, and is designed to ensure high performance and zero down-time. Hierarchies are external to the InfoCubes, and therefore extremely flexible. Changes to the structure of your business are easy to model, and you can view data according to new or old structures without having to completely re-align that data. [McDonald et al.]

InfoCubes
Definition: InfoCubes are the central objects of the multi-dimensional model in SAPBW. Reports and analyses are based on these. An InfoCube describes a self-enclosed dataset for a business area from a reporting view, that is, for the reporting end user. Queries can be defined and/or executed in the basis of an InfoCube. [BW310]

There are following InfoCube types in SAP BW:

- BasicCube
- VirtualCube:
- RemoteCube
- SAP RemoteCube
- Virtual InfoCube with Services

Only BasicCubes physically contain data in the database. By doing so, they are also data targets. BW objects are data targets when data can be loaded into them. In contrast, virtual InfoCubes only represent logical views of a dataset. There is no difference between these InfoCube types as far as the reporting end user is concerned. Queries can be defined based on all InfoCube types. InfoCubes are thus Info Providers. BW objects are Info Providers when queries can be defined /executed based on them in SAP BW Reporting.

Master data tables
Additional information about characteristics is referred to as master data in the SAP BW system. A distinction is made between the following master data types:

- Attributes
- Texts
- (External) Hierarchies

Master data information is stored in separate tables, which are independent of the dimension tables, in what are called master data tables (separately for attributes, texts and hierarchies). When a master data-carrying characteristic is activated, master data tables (attributes, text, hierarchies) are generated in the characteristic maintenance depending on the settings in the respective tab strip. [BW310]

ODS objects
The ODS objects are flat data structures used to support reporting, analysis, and data integration in SAP BW. The BW data warehouse itself resides on its own dedicated server, with its own data pool, creating a robust, high-performance platform for analysis, reporting and exploration, and keeping the load on the operational environment to an absolute minimum.

Definition: An Operational Data Store object (ODS object) is used to store consolidated and cleansed data (transaction data or master data for example) on a document level (atomic level).

It describes a consolidated dataset from one or more InfoSources. You can analyze this data with a BEx query. [BW310]

2.2.3. Analysis and Access Services

The analysis and access layer provides access to analysis services and structured and unstructured information stored in the SAP BW. The primary components of this layer are OLAP Engine, OLAP BAPI, XML for Analysis, Business Explorer API, Open Hub Service etc.

Based on a powerful Online Analytical Processing (OLAP) engine, the BW offers in depth analysis of information in many different ways. The BW allows to move from a bird's eye perspective to one offering more detail (slicing or drilldown), or change perspective entirely, based on a completely new criterion (dicing or changing view.) The BW builds on the capabilities of R/3 to support complex financial reporting and analysis needs. It allows, for instance, varying fiscal years and financial periods, and simultaneously supports the euro and national currencies. The currency conversion can also be performed instantly at the latest exchange rates. The absolute figures can beseen, then view the same statistics as a percentage or as a quotient. [McDonald et al.]

The OLAP BAPI provides and open interface for accessing any kind of information available through the OLAP engine. The XML for Analysis is an XML API based on SOAP designed for standardized access to an analytical data provider (OLAP and data mining) over the web and the Business Explorer API connects the Business Explorer (BEx) - the SAP BE reporting and analysis front-end solution – to the OLAP engine, allowing access to all available queries. [BW305 2005]

The SAP BW Open Hub Service allows to easily provide application data within a BW system available for downstream systems like non-SAP data marts, Analytical Applications, and other external applications. InfoCubes, ODS objects, and Master Data tables can be data sources for the Open Hub Service (Refer to figure 6). This ensures the controlled distribution of consolidated data and information among several systems, whereby SAP BW serves as a central hub for information. Various extraction options, detailed scheduling and monitoring, and delta capability mainly characterize the SAP BW Open Hub Service. [OHS-Release 2002]

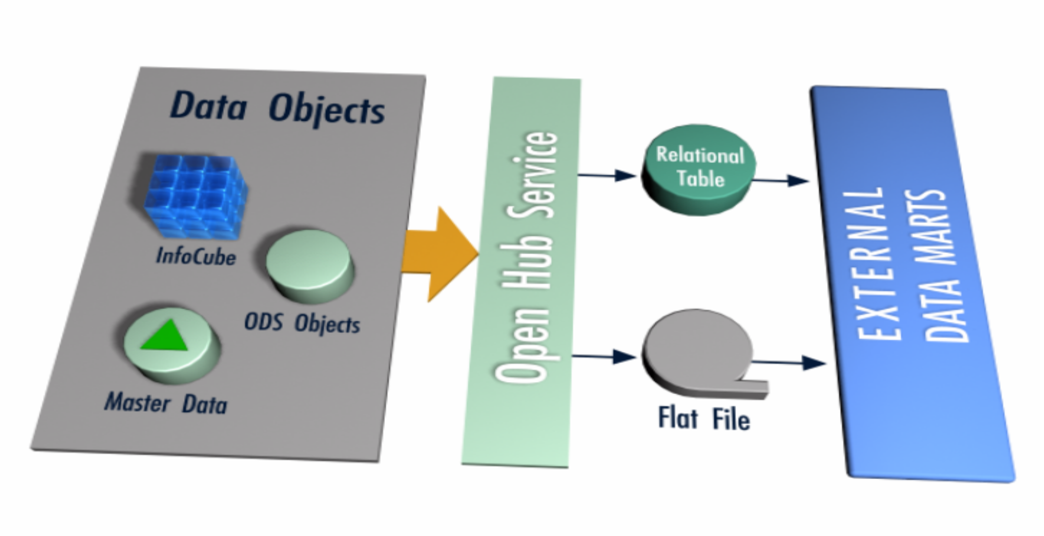
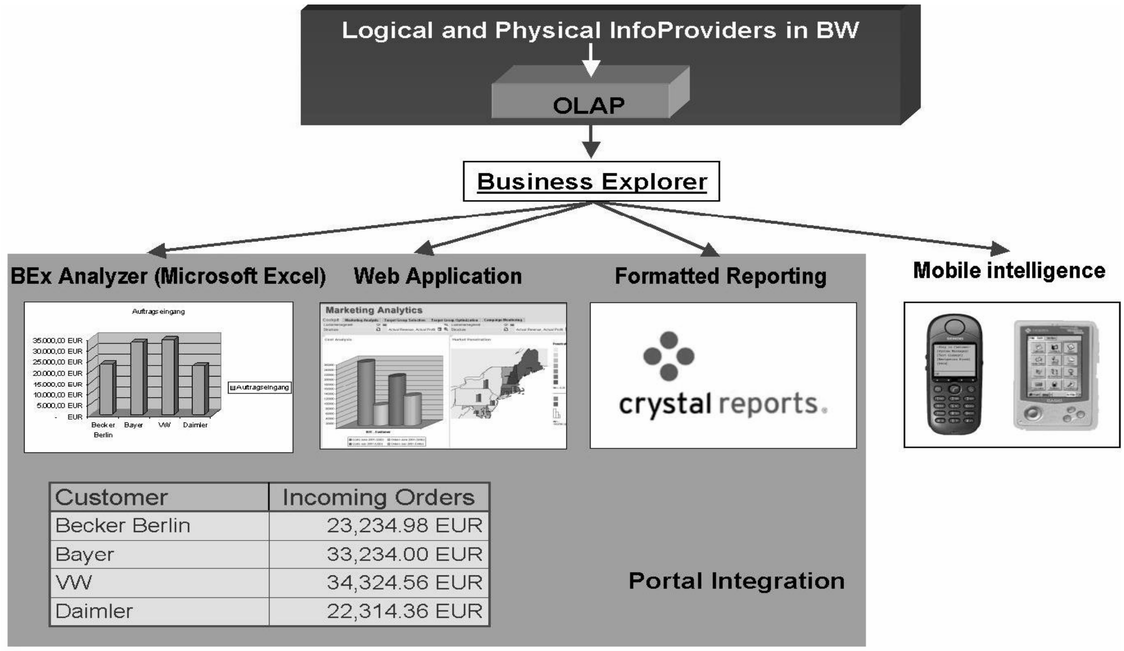


Figure 6: Open Hub Service in SAP BW
Source: SAP Help Portal, 2005

2.2.4. Presentation Services

As a top layer in the SAP BW architecture, the Business Explorer (BEx) serves as the reporting environment for the end users. It consists of the BEx Analyzer, BEx Query Designer, BEX Web Application Designer, BEx Browser, BEx Formatted Reporting etc. The Presentation Layer includes all components required to present information available on the SAP BW server in the traditional Microsoft Excel-based business Explorer Analyzer (BEx Analyzer), in the BEx Web environment, or third party applications.

The Business Explorer allows a large spectrum of user's access to the information in SAP BW. Using the Enterprise Portal (for example, through an iView that can call up alongside the applications from which the data is extracted), using the Internet (Web Application Design) or using mobile devices (WAP or iMode-enabled mobile telephones, Personal Digital Assistants). [BW305] The Business Explorer (BEx) component provides users with extensive analysis options as depicted in the figure 7.



Customer	Incoming Orders
Becker Berlin	23,234.98 EUR
Bayer	33,234.00 EUR
VW	34,324.56 EUR
Daimler	22,314.36 EUR

Figure 7: Reporting in SAP BW
Source: SAPCOURSE, BW305

The SAP BW presents information in a user-friendly, easy-to-understand fashion. It comes complete with a wide variety of standard reports that can be accessed with the click of a mouse; allowing knowledge workers to utilize the facts and figures stored in BW from the word go. Standard reports are supplied for the needs of particular departments, such as human resources, of particular industries, such as manufacturing, and even for individual roles, such as account managers, product managers, regional managers or financial controllers. Reports can also be adapted or custom-designed, or used to initiate ad-hoc queries with new parameters. Figure8 shows the Web Application Framework in SAP BW Presentation Layer.

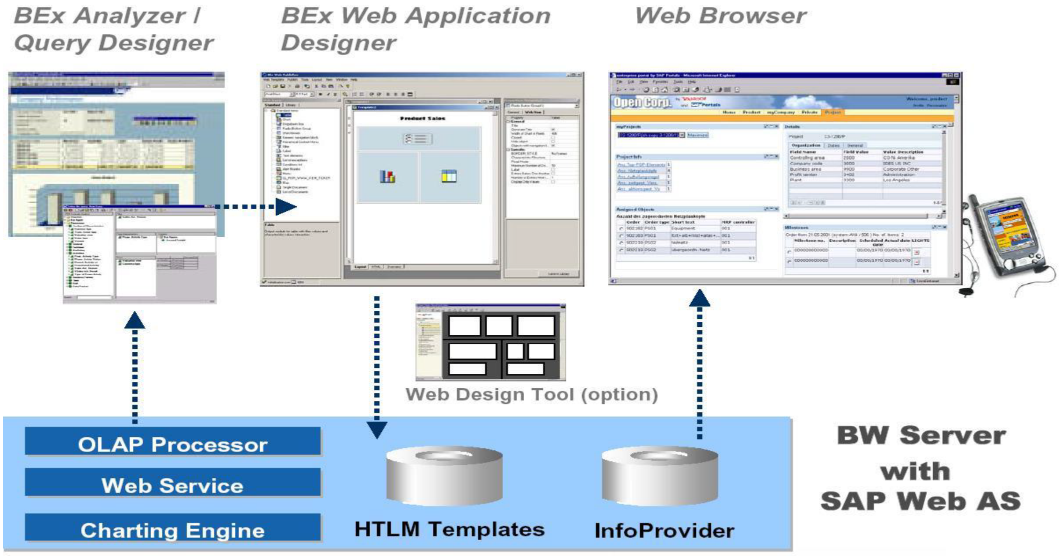


Figure 8: Web Application Framework
Source: SAP Help Portal, 2005

The catalogue browser provides a graphical overview of available reports, and allows point-and-click previews and selection. Results re displayed in the familiar environment of Microsoft Excel, here users can leverage their existing PC skills to analyze information, to reformat or process data further, or to distribute it to others, for example, as an email attachment. Favourite reports can be grouped together in clusters on a graphical desktop for even faster access. It includes geographical data visualization, enabling information to be shown on maps for even greater clarity and understanding. The following shows the functional overview of Business Explorer. [SAPHELP 2005]

Portal Integration Collaboration and Distribution

- Single point of Access
- Role-based data retrieval
- Personalization, collaboration and Profile Generation
- Integration of unstructured data

Query, Reporting and Analyses

- Query Design using the Desktop or web
- Multidimensional (OLAP) Analyses (Web based or MS Excel)
- Geographical Analysis
- Ad-hoc Reporting
- Alert
- Publishing iViews
- Seamless integration of web and Excel-based Analyses

Web Application Design

- Web Application Design
- Interactive analytical Content via the Web
- Information Cockpits and Dashboards
- Basis for Creating Analytical Applications
- Creation of iViews for the Portals
- Wizard based visualization

- APIs for additional, highly individual web design

Formatted Reporting

- Precise layouts to one pixel
- Wizard-based layout definition
- Static, Formatted Reports
- Form based Reports
- Predefined Crystal Reports in Business Content
- Publishing on the Web
- Practical printing options

Mobile Intelligence

- Online and offline Scenarios
- WAP- device and PDA Support
- Automatic device recognition
- Publishing using the Web Application Designer
- Device specific output
- Alerts, charts
- Integration into the mobile portal.

2.2.5. Administration Services

The Administrator work bench (AWB) is the primary administration, controlling, and monitoring tool in SAP BW. AWB is the data warehouse manager of the SAP BW system. You use AWB to manage, control, and monitor all the objects and processes in the SAP BW system. The AWB is where you create Meta objects. It is also the place where you use the scheduler to plan data uploads and where you track them using the monitor. Assistants enable you to analyze the data-loading processes closely. The assistants also help you to quickly identify the cause of any errors. [BW310]

The administration layer includes all services required to administer an SAP BW system. These services are available through Administrator Workbench. As the most prominent architectural component, the AWB includes a Modelling, Monitoring, Reporting Agent, Transport Connection, Documents, Business Content, Translation and Metadata Repository. [McDonald et al.] One could perform tasks in AWB in the following function areas: [BW310]

- Modelling
- Monitoring
- Reporting Agent
- Transport Connection
- Documents
- Business Content
- Translation
- Metadata Repositoryt he following figure depicts an overview of ADW tool.

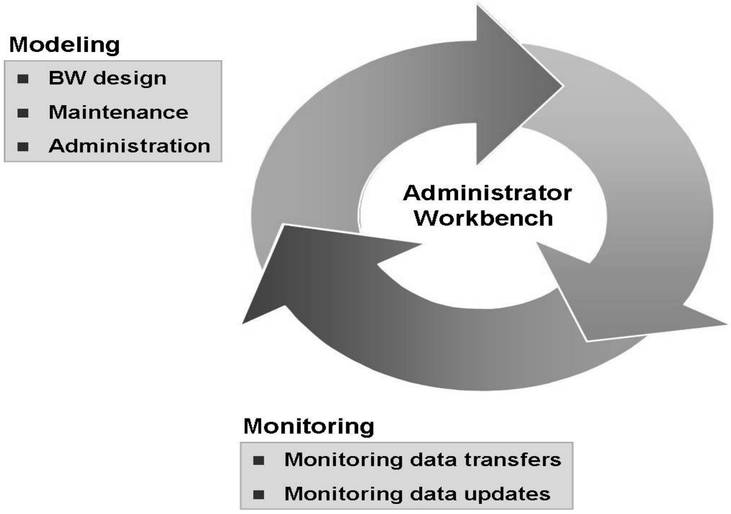


Figure 9: The Central Tool in SAP BW: AWB
Source: SAPCOURSE, BW310

The SAP BW includes effective, user-friendly tools for all aspects of data warehouse administration. The schema designer allows one to create InfoCubes, InfoSources, mapping and transformation rules with point-and-click simplicity. The BW infrastructure is also ideal for realignment tasks, i.e. the redefinition of aggregates in accordance with new categories. The use of attributes to describe master data means one could, for instance, reorganize sales regions, and customers would be automatically re-assigned to the right region thanks to the underlying address attributes. The aggregates need not have to be re-built from scratch, and to trawl through entire data warehouse making time-consuming changes to all data collected for all customers. Re-alignment is immediate, accurate and automatic. Data replication opens up the possibility of creating data marts serving the particular needs of a specific group of users, such as foreign based offices or purchasing department.

System administration tasks are equally easy to understand and simple to perform. The administrator's workbench provides a user-friendly graphical interface for scheduling data extracts, mapping, and aggregation routines, and for defining InfoCubes and reports. Moreover, extremely useful tools are provided for monitoring and planning, helping to get the best possible performance and greatest possible benefit out of the warehouse. The load monitor shows what's been loaded and when. It highlights problems and describes their root cause with messages such as 'Unable to access source XYZ' so that difficulties can be resolved quickly.

2.3. Features of SAP BW

The following section describes the features of SAP BW 3.5 and the positioning of SAP BW 3.5 in NetWeaver. The following figure depicts the roadmap – timeline and focal points of SAP BW

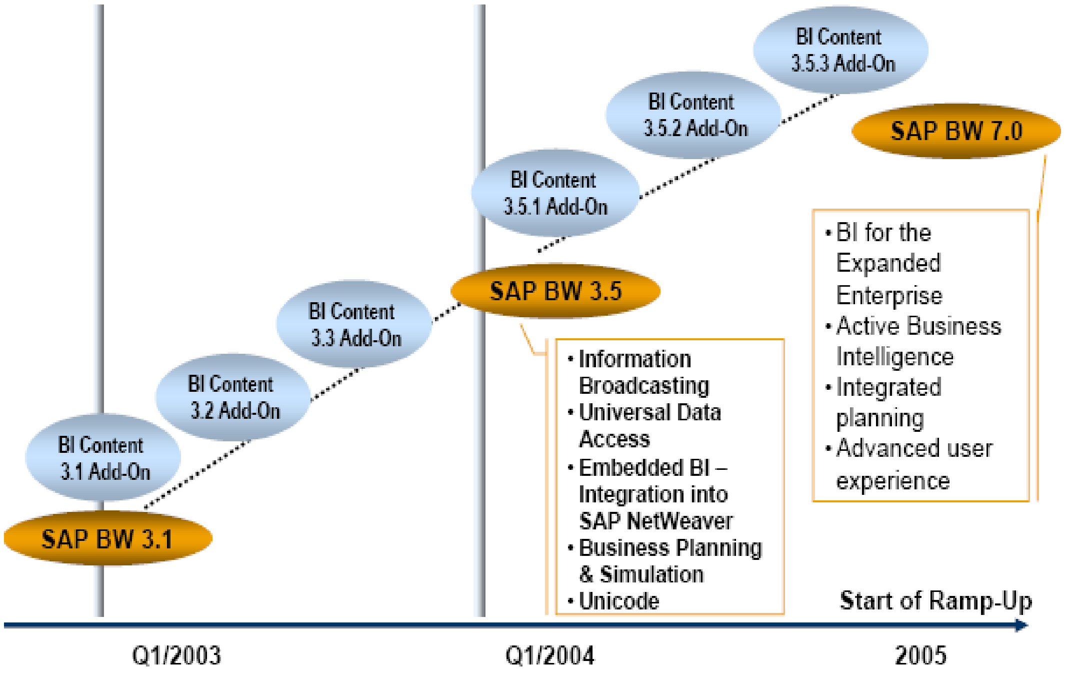


Figure 10: Roadmap - Timeline and Focal Points
Source: SAPNET, Features List SAP BW 3.5

SAP BW 3.5 is designed to deliver seamless integration capabilities into all of the SAP NetWeaver components, as well as offering new capabilities in the Business Intelligence platform and suite. The following section describes some important features of SAP BW 3.5[SAPNET 2005-3]

Information Broadcasting via Business Explorer (BEx) Broadcaster

- Share and disseminate insights to support decision-making processes
- Access the complete BI information portfolio via the SAP Enterprise Portal (SAP EP 6.0)
- Single, web-based wizard to broadcast personalized BI information portfolios to various end-users (pre-calculated for optimized query response time)
- Leverages SAP NetWeaver knowledge management features such as subscription, feedback, discussion, collaboration, rating, enterprise search, etc.
- Offers broadcasting services such as different scheduling options (ad-hoc, based on data loads, time scheduling), pre-calculation of queries and workbooks, sending pre-calculated queries and web templates as email attachments
- Based on the Java Repository Manager, all SAP BW metadata, master data, and transactional documents, as well as pre-calculated queries/templates for KM Services are enabled.

Universal Data Integration: The new Universal Data Integration significantly extends SAP BW data access capabilities to diverse data sources.

- BI Java Connectors: Several hundred of connectors provide access to all data sources that support JDBC, XMLA, OLE DB for OLAP and SAP Query
- UDConnect (Universal Data Connect): Out-of-the-box connectivity for additional data sources that can be accessed by the BI Java Connectors. UDConnect supports staging and remote scenarios to this data. For instance, extraction from /remote access to a relational database via JDBC or extraction from /remote access to an OLAP source using OLE DB for OLAP, and extraction from an OLAP source using XML for Analysis.

- BI Java software development Kit (BI Java SDK) for custom-built Java Applications accessing SAP BW or non-SAP BW data via the BI Java Connectors which is easy to use and learn and is based on open and accepted standards for interoperability

Embedded BI - Integration into SAP NetWeaver

- Web Application Server:
  Integration with new Internet Graphics Server (IGS) and WAS Alert Framework
    - Connecting BI alert framework to the SAP NetWeaver alert repository to streamline alert message processing
    - Platform independence for graphical rendering (charts, maps), improved usability and new chart designer in BEx Web Application Designer Inbound Message Processing

  Integration with SAP Exchange Infrastructure (SAP XI) to support real-time data acquisition
    - The data warehouse and/or operational data store is simply another subscriber to the real-time data being distributed by the Integration Broker
    - Data is active, event-driven that's available to the Business Intelligence system in "real time"
- Reporting on harmonized master data:
  Integration with SAP Master Data Management (MDM) helps to improve the quality of decisions made
    - Create consolidated views on customers, vendors and products
    - Enhance master data with global attributes for company-wide analysis (i.e. spend analysis)
- BI Web Services: The following BI web services can be accessed via open standards
    - XML Data Load, XML for Analysis, XML Query Result Set
    - Leveraging the Web Application Server 6.40 technology infrastructure
- Seamless deployment of BI web applications:
    1. into SAP Enterprise Portal roles for instant information delivery
    2. into SAP Enterprise Portal collaboration rooms
    3. into SAP Enterprise Portal KM folders, which Allows to search through BI applications in the context of unstructured information as well gain improved query response times through cached application retrieval

3. Data Mining and its Economic use

3.1. A general introduction to Data Mining

Data mining is not new. People who first discovered how to start fire and that the earth is round also discovered knowledge which is the main idea of Data mining. Even before technology were used for Data mining, statisticians were using probability and regressing techniques to model historical data.[Groth 1988, 19] Today technology allows to capture and store vast quantities of data. Finding and summarizing the patterns, trends, and anomalies in these data sets is one of the big challenges in today's information age. [Witten and Frank 2000] "With the unprecedented growth rate at which data is being collected and stored electronically today in almost all fields of human endeavour, the efficient extraction of useful information from the data available is becoming an increasing scientific challenge and a massive economic need."[Zaki and Ho 2000]

In the 1960s, Management Information Systems (MIS) and later, in the 1970s, Decision Support Systems (DSS) were praised for their great potential to supply executives with mountains of data needed to carry out their jobs. While these systems have supplied some useful information for managers, they have not lived up to their proponents expectations. One reason was that they simply supplied too much data and not enough information to be generally useful. [Müller and Lemke 2003] Advances in data collection, the widespread use of bar codes for most commercial products, and the computerization of many business transactions have flooded us with information, and generated an urgent need for new techniques and tools that can intelligently and automatically assist us in transforming this data into useful knowledge. [Fayyad 1996]

Today, there is a huge amount of information locked up in the mountains of data in companies' databases – information that is potentially important but has not yet discovered. The idea is to build computer programs that shift through databases, seeking regularities or patterns. [Witten and Frank] With the development in technology the Data mining process can be supported very well through powerful Data mining tools. So Data mining becomes a very hot topic. According to a study of the GARTNER GROUP, more than half of the companies in the Fortune-1000-companies use Data mining technologies for several purposes. [Dastani 2005]

3.1.1. The importance of Data in Data Mining

The term Data mining implies that data play an important role in the Data mining process. They are the foundation for all analysis and all Data mining techniques. In computing, data is information that has been translated into a form that is more convenient to store, move or process. Relative to today's computers and transmission media, data is information converted into binary digital form. [Techtarget 2005] On this data computer programs can base its Data mining techniques.

Definition of data / information -- (a collection of facts from which conclusions may be drawn; "statistical data"). [Princeton 2005]
The nature of the data sets: A data set is a set of measurements taken from some environment or process. In the simplest case, it has a collection of objects and for each object we have a set of same p measurements. In this case, one can think of the collection of the measurements on n objects as a form of n*p data matrix. The n rows represent the n objects on which the measurements were taken (for example medical patients, credit card customers and so on). Such rows may be referred to as individuals, entities, cases, objects, or records depending on the context.

The other dimension of the data matrix contains the set of p measurements made on each object. Typically one could assumethat the same p measurements are made on each individual although this need not be the case (for example, different medical tests could be performed on different patients). The p columns of the data matrix may be referred to as variables, features, attributes, or fields again the language depends on the research context. In all situations the idea is the same: these names refer to the measurement that is represented by each column. [Hand et al. 2004]

ID	Age	Sex	Marital Status	Education	Income
248	54	Male	Married	High school graduate	100000
249	77	Female	Married	High school graduate	12000
250	29	Male	Married	Some college	23000
251	9	Male	Not married	Child	0
252	85	Female	Not married	High school graduate	19798
253	40	Male	Married	High school graduate	40100
254	38	Female	Not married	Less than 1st grade	2691
255	7	Male	??	Child	0
256	49	Male	Married	11th grade	30000
257	76	Male	Married	Doctorate degree	30686

Figure 11: Examples of data in Public Use Micro data Sample data sets.
Source: Hand et al, Principles of Data Mining.

Data come in many forms and this paper is out of the scope to develop a complete taxonomy. Indeed, it is not even clear the complete taxonomy can be developed, since an important aspect of data in one situation may be unimportant in another.

There are certain basic distinctions to which one should draw attention. One is the difference between quantitative and categorical measurements (different names are sometimes used for these). A quantitative variable is measurements (different names are sometimes used for these). A quantitative variable is measured on a numerical scale and can, at least in principle, take any value. The columns Age and Income in figure 11 are examples of quantitative variables. In contrast, categorical variables such as Sex, Marital Status and Education in figure11 can take only certain, discrete values. The common three point severity scale used in medicine (mild, moderate, severe) is another example. Categorical variables may be ordinal (possessing a natural order, as in the Education scale) or nominal (simply naming the categories, as in the Marital Status case). A data analytic technique appropriate for one type of scale might not be appropriate for another. For example, were marital status represented by integers (e.g., 1 for single, 2 for married, 3 for widowed, and so forth) it would generally not be meaningful or appropriate to calculate the arithmetic mean of a sample of such scores using this scale. Similarly, simple linear regression (predicting one quantitative variable as a function of others) will usually be appropriate to apply to quantitative data, but applying it to categorical data may not be wise; other techniques, that have similar objectives (to the extent that the objectives can be similar when the data types differ), might be more appropriate with categorical scales.[Hand et al.]

Measurement scales, however defined, lie at the bottom of any data taxonomy. Moving up the taxonomy, one could find that data can occur in various relationships and structures. Data may arise sequentially in time series, and the data mining exercise might address entire time series or particular segments of those time series. Data might also describe spatial relationships, so that individual records take on their full significance only when considered in the context of others.

3.1.2. Definitions of Data Mining

Translating Data mining word by word means, the mining or digging in data with the purpose of finding information or respectively knowledge. Coming to the more abstract and very well known definition of Frawley, Data mining is defined as "The nontrivial extraction of implicit, previously unknown, and potentially useful information from data". [Frawley 1992]

Groth mentions another interesting aspect of Data mining. He describes it as "the process of automating information discovery". [Groth] Today Data mining is a term that covers a broad range of techniques to analyze data. The techniques use specific algorithms to identify and extract patterns and establish unknown relationships in order to discover hidden and valuable information in a huge amount of data. Most companies already collect massive quantities of data. Data mining techniques can be implemented on existing software and hardware platforms to enhance the value of existing information resources. [Thearling 2005]

In the words of Moxon: "Data mining is the process of discovering meaningful new correlation, patterns and trends by sifting through large amounts of data, using pattern recognition technologies as well as statistical and mathematical techniques." Data mining is a "knowledge discovery process of extracting previously unknown, actionable information from very large databases." [Moxon 1996]

According to their final goal, data mining techniques can be considered to be descriptive or predictive "Descriptive data mining intends to summarize data and to highlight their interesting properties, while predictive data mining aims to build models to forecast future behaviours". [Han and Kamber 2001]

3.2. KDD - Knowledge Discovery in Databases

Knowledge Discovery in Databases, also often used with the abbreviation KDD, "is the concept of extracting previously unknown and potentially useful information from large sets of data". [Witnessminer 2005] So KDD is only the concept of a multistage process that identifies pattern in data in order to find new information. Data mining is only one stage in the KDD process concerned with applying computational techniques to find patterns in data. This step consists of algorithms which delivers patterns in an acceptable time out of a defined database. Other stages in the KDD process are the comprehensibility and the validity of the discovered patterns. In theory and practice the expressions KDD and Data mining are often mixed. But it is important to understand that KDD is the whole concept and Data Mining is only a step in this concept of extracting data. Simplified, KDD is the concept and Data Mining is the tool. [Witnessminer]

The five main processes that are common in almost all of the methods are: Task Analysis, Pre-processing, Data Mining, Post-processing and Deployment. This is diagrammatically expressed and explained as could be seen in Figures 12 and 13.

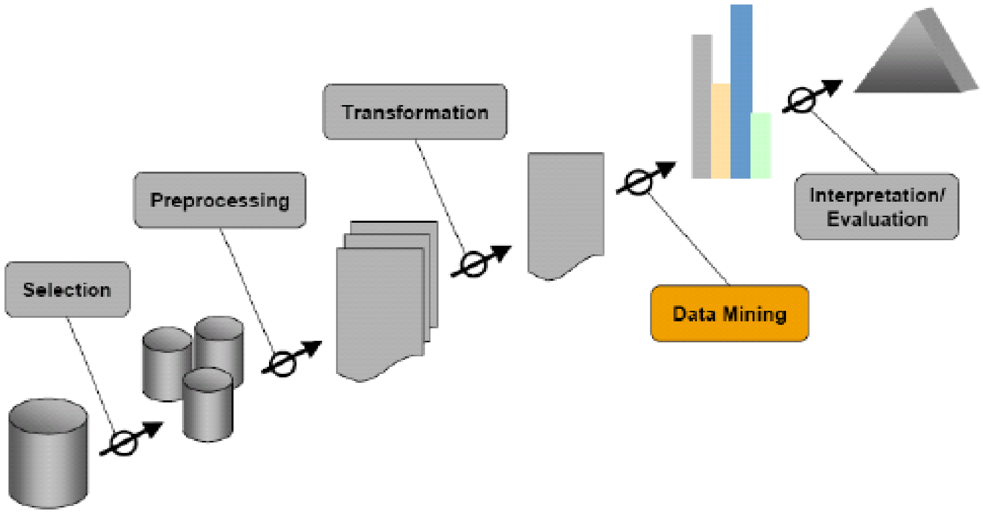


Figure 12: Knowledge discovery in Databases
Source: [Lesley 2004]

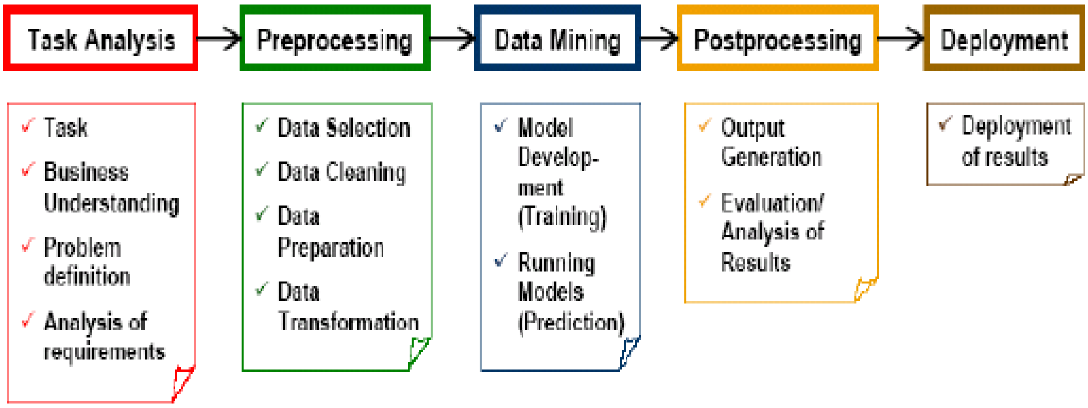


Figure 13: Knowledge discovery in Databases
Source: [Lesley 2004]

3.3. Data Mining and Data Warehouse

The evolution of database technology is an essential prerequisite for understanding the need of knowledge discovery in databases (KDD). Data mining is a pivotal step in the Knowledge Discovery in Database process- the extraction of interesting patterns from a set of data sources (relational, transactional, object-oriented, spatial, temporal, text, and legacy databases, as well as data warehouses and the World Wide Web). The patterns obtained are used to describe concepts, to analyze associations, to build classification and regression models, to cluster data, to model trends in time-series, and to detect outliers. Snce the patterns, which are present in data are not all, equally useful, interestingness measures are needed to estimate the relevance of the discovered patterns to guide the mining process.

The first step toward building a productive data mining program is to gather data. Most businesses already perform these data gathering tasks to a very high extent. [Chapple 2005]
Very often a data warehouse is used to manage and store that gathered data. Because of that huge amount of stored data, the key is to locate the data critical to the business. So companies use Data

Mining tools with the purpose to discover new information out of the data stored in the data warehouse. The data warehouse is the data foundation for all the analyses of the Data Mining tools. Data Mining helps companies focus on the most important information in their data warehouses. [Thearling] The major analysis of this work is done within the framework of SAP BW 3.5 which includes a couple of data mining methods available as part of it, which are described in detail in the next chapter.

3.4. Common uses of Data Mining

Data mining tools can predict future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move far beyond the analyses of past events provided by retrospective tools typical offered by decision support systems.

Data mining tools can give answers to business questions that traditionally were time consuming to resolve[Thearling] Today Data mining is primarily used by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables them to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. Furthermore, it enables these companies to determine the impact on sales, customer satisfaction, and corporate profits and it enables them to "drill down" into summary information to view detail transactional data. [Palace 2005]

As described in 3.1 “A general introduction in Data Mining” a large number of companies use Data Mining today. And the list of this companies looks like a Fortunes 500 *Who’s Who.* [Groth] So different the companies are, so different are the purposes of the use of Data Mining. Here are a few areas in which companies use Data mining to achieve a strategic benefit:-

Direct Marketing

The idea here is to find out who is most likely or most desirable to buy certain produces. This information can be used for several marketing activities.

Trend Analysis

With Trend analysis companies are able to predict trends in the marketplace. Using this information can lead to a strategic advantage because it is useful in reducing costs and timeliness to market.

Fraud Detection

Companies use Data mining techniques to model which business transactions are likely to be fraudulent. So this is used for insurance claims, cellular phone calls or credit card purchases.

Forecasting in Financial Markets

There are many possibilities to model financial markets with Data mining methods. For example neural networks can be used for financial gain. [Groth]

Apart from this applications, companies use Data mining also for: [Bao 205]

- **Business information**

  Investment analysis

  Loan approval
- **Manufacturing information**

  Controlling and scheduling

  Network management

  Experiment result analysis

- **Scientific information**

  Sky survey cataloguing
  Bio sequence Databases
  Geosciences: Quake finder

Performance and monitoring of standard software systems
The main purpose of this paper is find out how to introduce data mining functionality to support the SAP BW administrator in the areas of data loading, reporting, planning etc in order to proactively discover the error situations. From these investigations it is quite clear that the companies would like to come up with the product functionalities that would assist the system administrators. For instance, how data mining methods could make the work of the SAP BW administrator ease in order to perform his day-to-day activities.

3.5. The process of Data Mining

Data mining should be regarded as a strategic and competitive move. So before the Data mining process starts, the goal which is in focus of the analysis should be clarified. Otherwise it's not possible to search for new valuable information if the necessary parameters can not be defined as there are different models for the data mining process based on the task at hand. The following description is based on the model of Fayyad. [Fayyad]

Step 1: Data selection
Out of a data base the needed data were selected according to its objects and characteristics.

Step 2: Pre-Processing
In this step happens a cleaning of the selected data. This means for example the filling of missing values.

Step 3: Transformation
In the transformation phase the data are transformed in new formats, if necessary.

Step 4: Data Mining
In this step of the process identifies the patterns and relationships between the data.

Step 5: Interpretation and Evaluation
In the last step the result has to be interpreted and evaluated to come up with suitable actions.

The following picture shows the process in a graphical representation.

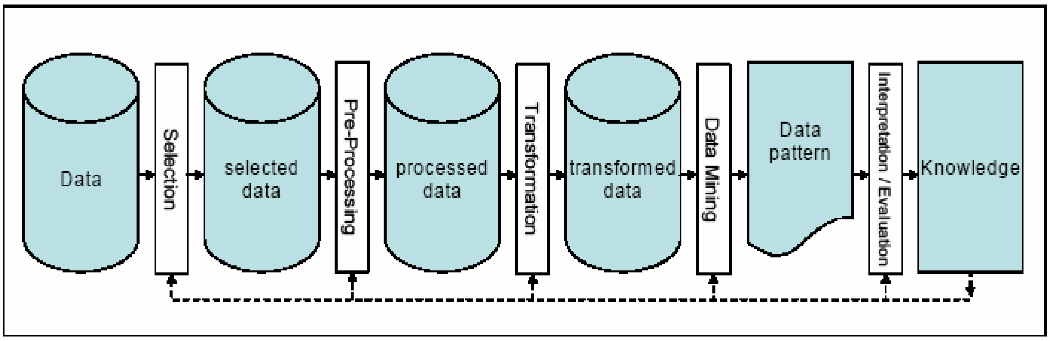


Figure 14: The process of Data mining
Source: SPSS, Clementine 7.0 user's guide.

Crossindustry process for Data mining, CRISP

The CRISP method is one of the several available learning methods. It encompasses all the facets of learning, beginning from the conception to the realization and deployment of the gained information. It begins, as could be seen from Figure 3.4 below, with an analysis or a business understanding of the problem. Questions on the relationship between the operating factors are asked at this stage. The dependence of one on another (or several others) is also stipulated at this stage.

After a business understanding is laid down, understanding the data then becomes the next task, according to the CRISP model. What tables has to be created? How would the tables be made available? Would a single data instance be enough or would several data instances be needed? What about the quality of the data? Based on the understanding of the data, the business understanding may have to be adjusted or additional inputs made to the data, e.g. creation of additional tables, as to be able to realize the desired business objective Data preparation then follows. Based on the analysis desired, columns might have to be filtered out, or data aggregated, merged, etc. The modelling process could then be done at this stage. As could be seen from the figure 14, additional data preparation needs may have to be done as to realize the desired model.

An evaluation of the whole process follows. In some cases, a supervised form of learning might be very helpful in this case. Interim results would be checked against the historical data as to ascertain the level of conformity, which also will serve in the evaluation of the entire process.
The gained information or intelligence could now be deployed. The destination could be another system, say, ERP system like the SAP CRM, or stored in a database system. Such could be final reports, presentations, action plans, etc. It could also be used for further analysis. Moreover, feedback could be made to the initial business understanding for the purpose of further analysis, after which the entire process would be repeated.

The overall process involved in the CRISP-Model could be summarized as follows: [CRISP, 2005]

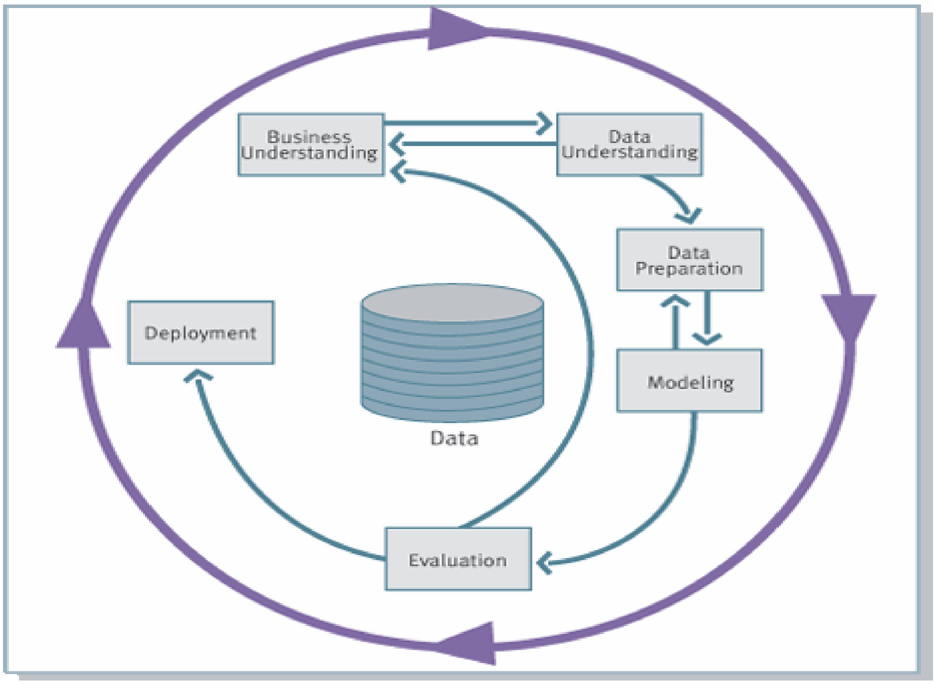


Figure 15: Phases of the CRISP-DM Process Model
Source: CRISP, 2005.

- **Business Understanding:** Description of the Business Objective and Data Mining Goals/Success
- **Data Understanding:** Selection of the data and exploratory analysis (quality, problems, description of selected data)
- **Data Preparation**: Cleaning, transformation, integration, formatting of the selected data
- **Modelling**: Selection, building, testing and running different models
- **Evaluation:** Approval of the models and assessment of the results (in accordance with the defined objectives), review of the process
- **Deployment**: Preparation of final reports, presentation, action plans and deployment of results

4. Methods of Data Mining

4.1. An overview of Data Mining Methods

In the last chapter the overview and the tasks of data mining were discussed. But how to realize these task, it is still needed to describe the data mining methods. “Data mining methods detect patterns in large amounts of data, and use these patterns to detect future instances in similar data.” [Zadok and Stolfo 2005]

There are many kinds of data mining methods. “Some are well founded in mathematics and statistics, whereas others are used simply because they produce useful results.”[Lidal and Dingsoyr 2005] Because data mining has emerged from many different fields, different kinds of methods can be used in different areas. “Researchers have approached the knowledge discovery process from different angels, with different algorithms, based on their scientific interests and backgrounds.” [Lidal and Dingsoyr] But no one method can solve all data mining problems. Some of them have several tasks at the same time; figure 16 gives a short conclusion about the tasks and different methods.

Tasks	Methods
Prediction & Description	Decision tress, Market basket analysis (Association analysis), Time series analysis, Neural networks, Agent network technology
Classification	Market basket analysis (Association analysis), Decision tress, Neural networks, Sorting
Regression	Linear regression, Logistic regression, Multinomial Regression.
Clustering	Cluster Analysis, Neural networks
Summarization	Genetic algorithms
Dependency modelling	Analysis of variance; Link Analysis
Change and deviation detection	Fuzzy Logic

Figure 16: Data mining tasks and methods

The following section will introduce some data mining methods that are available as part of SAP BW which are used normally in reality. Not in very detail, but to have a fundamental understanding of them.

4.2. The SAP data mining workbench

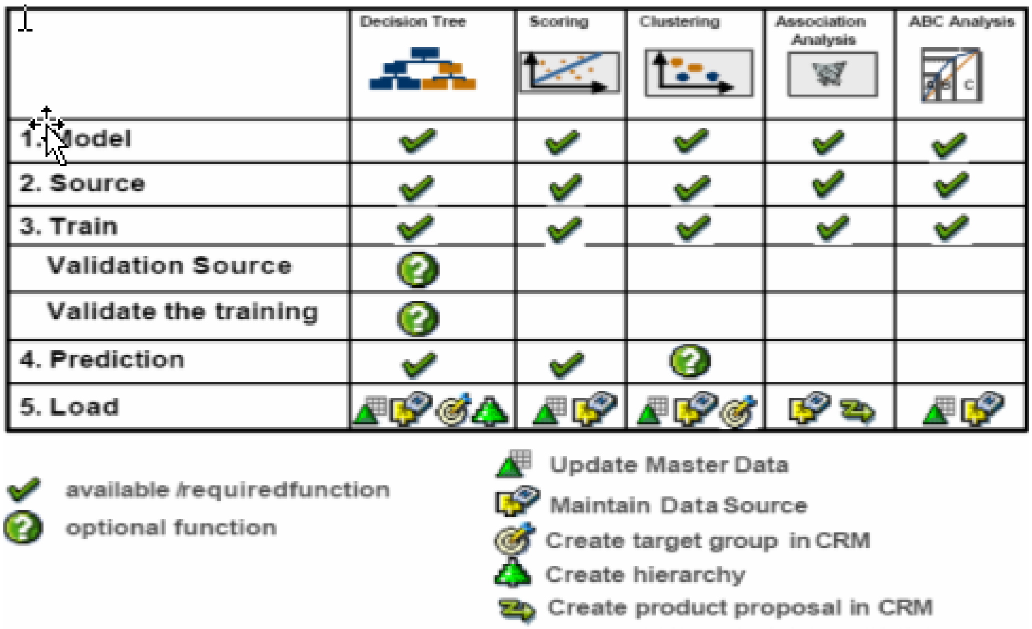
The SAP Business Information Warehouse is a complete suite of application, i.e. a solution which includes the activities of data collection and storage, decision support systems, query and reporting, online analytical processing, statistical analysis, foreasting, and data mining. In SAP BW, data from disparate database(s) of all systems in the enterprise are collected, consolidated, administered and provided for analysis and planning purposes. This data often provides further valuable potential.

Even with sophisticated analysis tools, new information presenting itself in the form of meaningful relationships between the data, is often hidden or too complex to be uncovered through pure observation or intuition. With the assistance of the SAP BW, it is now possible to easily investigate and identify these hidden or complex relations between the data. For this discovery process, several methods are provided (e.g. Statistical and Mathematical calculations, data cleansing and restructuring methods, etc.) The intelligence gained could be uploaded automatically into the SAP BW database or redirected into an operational system like the SAP CRM. In either case, the intelligence is made available for all decision-making and/or application processes and can thus be of significant importance: strategically, tactically, and operationally.

The SAP Data Mining Workbench offers a single point of entry for access to available data mining models namely

- Decision trees
- Clustering
- Association analysis (Market Basket analysis)
- Approximation (Regression and Weighted score tables)
- ABC classification

It also provides an option to connect with the third party data mining modals. For each model type a wizard guides the user through the process of creating the model, thus enabling users interested in analytical results to setup data mining models easily. The following figure shows the process steps for the analytical models available as part of the SAP data mining workbench.



	Decision Tree	Scoring	Clustering	Association Analysis	ABC Analysis
1. Model	✓	✓	✓	✓	✓
2. Source	✓	✓	✓	✓	✓
3. Train	✓	✓	✓	✓	✓
Validation Source	?				
Validate the training	?				
4. Prediction	✓	✓	?		
5. Load	Update Master Data Maintain Data Source Create target group in CRM Create hierarchy	Update Master Data Maintain Data Source	Update Master Data Maintain Data Source Create target group in CRM	Maintain Data Source Create product proposal in CRM	Update Master Data Maintain Data Source

✓ available /requiredfunction

? optional function

Update Master Data

Maintain Data Source

Create target group in CRM

Create hierarchy

Create product proposal in CRM

Figure 17: Process steps for applying analytical methods
Source: SAPCOURSE, CR900, my SAP CRM Analytics.

There are two basic broad classifications of data mining methods. These are the supervised and the unsupervised learning. In supervised learning, a sample data is first selected and with it, the system is 'trained' as to understand the dynamics involved in it. This is then weighed against the known historical data as to see the extent to which the system's output corresponds to the known output. Further learning might have to be applied, and as much as would simply be needed, until the system turns out an answer that largely (mostly 99.99%) reflect the decision already made on historical data. On the other hand is the unsupervised learning. This is, fundamentally, where data mining plays a great role. A heap of data is "mined" as to discover the complex, hidden and unexpected relationships and correlations that may exist in it. In as much as the system could be made to run the process as much as it is wished, it is basically done with no form of bias, as the case is in a supervised learning.

Supervised learning is mostly predictive while unsupervised learning is overly informative. This is so for in supervised learning, the interim result is weighed against historical data with known output to see if the result corresponds with known cases. The following chapter will introduce some data mining methods that are used normally and are part of SAP's offering. Not in very detail, but to have a fundamental understanding of them.

4.2.1. Approximation

Statistics orientation is a main way which makes sense to analyze data. The purpose of approximation (scoring) is to valuate the data records. SAP offers weighted score tables and regression analysis namelylinear regression and non-linear regression (Logistic and Multinomial regression) to perform the valuation

4.2.1.1. Regression Analysis

Regression is a function that maps a data item to a real-valued prediction variable. So it is predicting a value of a continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency. [Kumar and Joshi] There are many regression applications in practice, e.g., predicting the amount of bio-mass present in a forest given remotely-sensed microwave measurements, estimating the probability that a patient will die given the results of a set of diagnostic tests, predicting consumer demand for a new product as a function of advertising expenditure, and time series prediction where the input variables can be time-lagged versions of the prediction variable. [Bao]

Regression analysis is the technique which used to inter- and extrapolate the observations which can be classified in to Linear and Non-linear regression. Linear Regression is a statistical technique which attempts to build a model to the observed data, and though this line to predict future data. It quantifies the relationship between two continuous variables: "the dependent variable or the variable you are trying to predict and the independent or predictive variable". [Rud 2001] It works by finding a line through the data that minimizes the squared error from each point. The formula of linear regression is: [Whitehead 2005]

$Y = a + bX + c$
Y: a dummy dependent variable, =1 if event happens, =0 if event doesn't happen,
a: the coefficient on the constant term,
b: the coefficient(s) on the independent variable(s),
X: the independent variable(s),
c: the error term.

For instance, figure18 shows the relationship between sales and advertising along with the regression line. The goal is to be able to predict sales based on the amount spent on advertising.

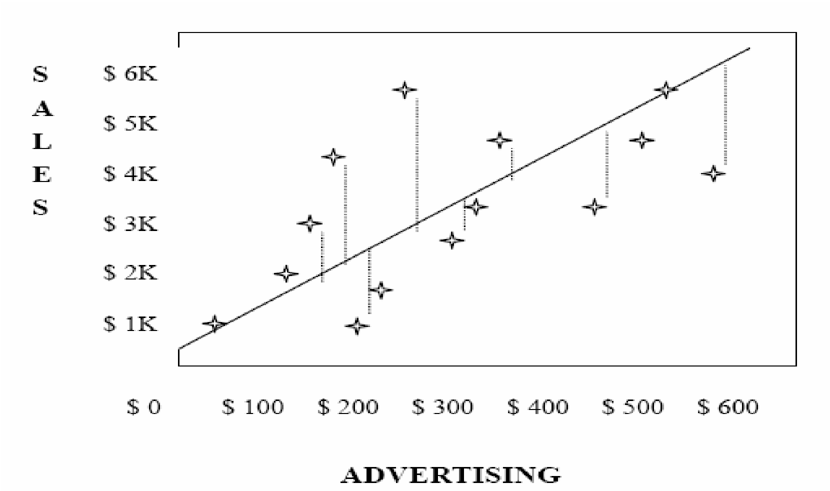


Figure 18: Simple linear regression
Source: Rud, Data Mining Cookbook

It is also possible that the relationship between the two variables is not linear. The relationship also can be curvilinear or multiple linear. Logistic Regression is very similar to linear regression. "The Logistic Regression model is simply a non-linear transformation of the Linear Regression." [Whitehead] It uses sigmoid function instead of linear function to fit the data. The main difference between them is that in the logisticr egression model the dependent variable is discrete or categorical, not continuous. So it is very useful in the marketing area because it can be used to predict a discrete action such as response to an offer or a default on a loan. [Rud] Logistic regression model can be described as following: [Whitehead]

ln[p/(1-p)] = a + bX + c
p: the probability that the event Y occurs, p(Y=1)
b: the coefficient(s) on the independent variable(s),
c: the error term
p/(1-p): the "odds ratio"
ln[p/(1-p)]: the log odds ratio, or "logit".

Logistic Regression like Linear Regression, also base on a statistical distribution. But the "logistic" distribution is an S-shaped distribution function which is similar to the standard normal distribution (which results in a profit regression model), but easier to work with in most applications because the probabilities are easier to calculate. "The logistic distribution constrains the estimated probabilities to lie between 0 and 1." [Whitehead] A graphical comparison of the Linear Regression and Logistic Regression models is illustrated in figure 19

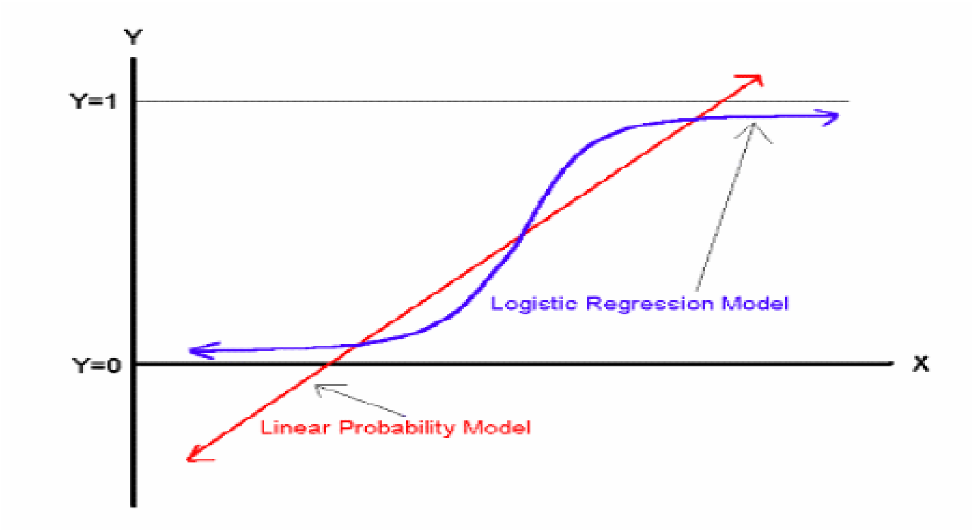


Figure 19: Comparison of Linear and logistic Regression
Source: Whitehead, an Introduction to Logistic Regression.

Multinomial Regression Before, what discussed in Linear Regression and Logistic Regression is only referred to two variables. When the nominal response variables are more than two categories, another regression method can be used: the so called Multinomial Regression. “Multinomial logit models are multiequation models” [GSE&IS 2005] For example, a response variable with n categories will generate (n-1) equations. This breaks the regression up into a series of binary regressions comparing each group to a baseline (reference) group. “For example, wife work has 3 values, 0=not working, 1=part time, 2=full time. If choosing not working (0) as the baseline group, multinomial logistic regression will assess the odds of working part time vs. not working, and working full time vs. not working.” [UCLA 2005] Multinomial logistic regression simultaneously estimates the (n-1) logits. “Further, it is also the case, that the model tests all possible combinations among the n groups although it only displays coefficients for the (n-1) comparisons.” [GSE&IS]

4.2.1.2. Weighted score tables

A weighted score table is a method of evaluating alternatives when the importance of each criteria differs. In a weighted score table, each alternative is given a score for each criteria. These scores are then weighted by the importance of each criterion. All of an alternative's weighted scores are then added together to calculate that alternative's total weighted score. The alternative with the highest total score should be the best alternative you can use weighted score tables to make predictions about future customer behaviour. You create a model in the data mining application to make predictions. After a model has been created based on historical data, it can then be applied to new data to make prediction s. The prediction, that is, the output of the model is called a Score. You can create a single score for your customers by taking into account different dimensions. SAP’s weighted score tables method allows you to define your own valuation function by first assigning weights to the individual model fields and then creating a weighted total from these model fields. The algorithm of weighted score tables: [SAPDOCS 2005]

A function f that is defined by weighted score tables is a linear combination of functions of a variable.

$$f\,(X_{1\,...}\,X_n) = W_1 * f_1(X_1) + ... + W_n * f_n(X_n)$$

The weights W1 ...W n are arbitrary numbers. Each of the functions f1... f n is mapped to exactly one model field. The arguments X1… X n of these functions are those values that the model fields can take.

For discrete model fields, the score table of the model field is used to directly assign a function value f i (X i) to individual values X i of the model field. A common function value can be assigned to values that are not listed explicitly in the table.
For continuous model fields, the score table of the model field is also used to directly assign a function value x i to individual values f i (X i) of the model field. Either a linear interpolation is made between two points, or the function value from the left or right point is taken. Respectively, either a polygon line or a piecewise constant function is defined. Depending on the option selected by the user, the function is continued as linear or continuous beyond the outer points.

4.2.2. Clustering

Clustering is a common descriptive task of Data mining where one seeks to identify a finite set of categories or clusters to describe the given data. Based on a given set of data points, each having a set of attributes, and a similarity measure among them, the identified clusters should guarantee that: [Kumar and Joshi]

- Data points in one cluster are more similar to one another,
- Data points in separate clusters are less similar to one another.

The identified clusters may be mutually exclusive and exhaustive, or consist of a richer representation such as hierarchical or overlapping clusters. Examples of clustering in a Data mining context include discovering homogeneous sub-populations for consumers in marketing databases and identification of sub-categories of spectra from infrared sky measurements. [Bao] According to Jain and Dubes "Cluster analysis organizes data by abstracting underlying structure either as a grouping of individuals or as a hierarchy of groups. The representation can then be investigated to see if the data group according to preconceived ideas or to suggest new experiments". [Jain and Dubes 1988] In brief, cluster analysis group's data objects into clusters such that objects belonging to the same cluster are similar, while those belonging to different ones are dissimilar.

"The term cluster analysis (first used by TRYON, 1939) actually encompasses a number of different classification algorithms." [STATSOFT 2005] A general question facing researchers in many areas of inquiry is how to organize observed data into meaningful structures, that is, how to classifying. Cluster analysis is an exploratory data analysis tool for solving classification problems. Its objective is to sort cases (people, things, events, etc) into groups, or clusters, so that the degree of association is strong between members of the same cluster and weak between members of different clusters. The feature of Cluster Analysis is there is no classes to be predicted but there are different ways in which the result of clustering can be expressed. "The groups that are identified may be exclusive, so that any instance belongs in only one group; or they may be overlapping, so that an instance may fall into several groups; or they may be probabilistic, whereby an instance belongs to each group with a certain probability; or they may be hierarchical, such that there is a crude division of instance into groups at the top level, and each of these groups is refined further- perhaps all the way down to individual instance." [Witten and Frank] Cluster analysis is thus a tool of discovery. It may reveal associations and structure in data, though not previously evident, but sensible and useful rule.

The most common used method of Cluster Analysis is K- Means clustering. Firstly, decide how many clusters will be sorted, it is the parameter K. Second the mean of all the instances in each cluster is calculated. These means are taken to be new centre value for their respective clusters. "Finally the whole process is repeated within the new cluster centres. The iteration continues until the same points are assigned to each cluster in consecutive rounds, at which point the cluster centre have stabilized and will remain the same thereafter." [Witten and Frank]

The major part of this thesis work concentrates on how to utilize cluster analysis and to come up with the patterns using K-means as well as the sophisticated algorithms (Demographic, Neural net methods) which are part of IBM Data Mining engine based on the statistical data available as part of SAP BW statistics content, which will be dealt in chapter 6.

4.2.3. Association analysis

Association Analysis (also known as Market Basket Analysis) uncovers the hidden patterns, correlations or casual structures among a set of items or objects. For example, Association Analysis enables you to understand what products and services customers tend to purchase at the same time. By analyzing the purchasing trends of your customers with Association Analysis, you can predict their future behaviour. It is also commonly referred to as "association discovery". [SAPDOCS]
These patterns may be expressed in the form of association rules such as:

- 72% of the customers who buy milk also buy bread and eggs. You can find that this rule applies to 20% of the transactions.
- 80% of the time that a specific brand of toaster is sold, customers also buy a set of kitchen gloves and matching cover sets

Customers who purchase pizza bases are three times more likely to purchase cheese than those not buying the pizza bases.

“Market Basket Analysis is an algorithm that examines a long list of transactions in order to determine which items are most frequently purchased together.” [Goransson 2005] It uses the information about ‘What’ customers purchased to give researchers insight into ‘Who’ they are and ‘Why’ they make such certain purchases. It also gives the information about the merchandise by telling which products tend to be purchased together and which are most amenable to promotion. [Berry and Linoff 1997] Finally this information is actionable: “It can suggest new store layout; it can determine which products to put on special; it can indicate when to issue coupons, and so on.” [Berry and Linoff] Because Market Basket Analysis is used to determine which products sell together, the input data to a Market Basket Analysis is normally a list of sales transactions, where each has two dimensions, one represents a product and the other represents either a sale or a customer (depending on whether the goal of the analysis is to find which items sell together at the same time, or to the same person). The cells of the data normally contain only 1 (bought product) or 0 (did not buy product) values, though poly-analyst can work with other data in the cells, such as quantity or revenue. [Goransson]

Market Basket Analysis is often used as a starting point when transaction data is available but the researcher doesn’t know what specific patterns to look for. It can be applied to many areas such like: [Albion 2005]

- Analysis of credit card purchases.
- Analysis of telephone calling patterns.
- Identification of fraudulent medical insurance claims. (Consider cases where common rules are broken).
- Analysis of telecom service purchases.

4.2.4. Decision Trees

“A decision tree is used as a classifier for determining an appropriate action or decision (among a predetermined set of actions) for a given case. A decision tree helps you to effectively identify the factors you must consider and how each factor has historically been associated with different outcomes of the decision”. [SAPDOCS] Decision trees have become one of the most popular data mining tools. Their visual presentation makes the decision trees very easy to read, understand and assimilate information from it. They are called decision trees because the resulting model is presented in the form of a tree structure. Decision trees are most commonly used for classification, that is, predicting to which group a particular case belongs. A decision tree is constructed from a training set. A training set contains historical data, which is used to predict the possible outcomes such as aspects of customer behaviour. For example, one can predict if a customer churns or remains loyal to the company.

Decision Trees are powerful and popular data mining tools for classification and prediction. It is "a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a classification or decision." [Berry and Linoff] It has rules that "can readily be expressed in English so that we humans can understand them or in a database access language like SQL so that records falling into a particular category may be retrieved." [Berry and Linoff] Decision Trees are normally drawn with the root at the top and the leaves at the bottom. A record enters the tree at the root node where a test is applied to determine which sub node the record will go next. There are different algorithms for choosing the initial test, but the goal is always the same: "To choose the test that best discriminates among the target classes." [Berry and Linoff] This process is repeated until the record arrives at a leaf node. All the records that end up at a given leaf of the tree are classified the same way. But from the root to each leaf there is a unique path that is an expression of the rule used to classify the data records. The following Decision Tree is one example that is to help a financial institution decide whether a person should be offered a loan. [Wilson 2005]

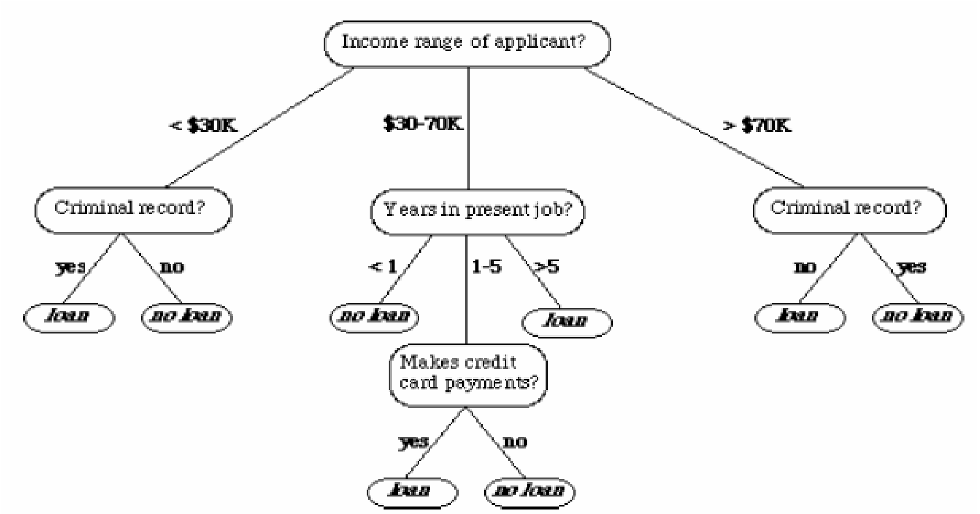


Figure 20: Decision Tree of deciding whether a person should be offered a loan
Source: Wilson, Introduction of Decision Trees.

4.2.5. ABC classification

Classification is a function that maps a data item into one of several predefined classes. So, the goal is that previously unseen records should be assigned to a class as accurately as possible. [Kumar and Joshi 2004] Examples of classification methods used as part of knowledge discovery applications include classifying trends in financial markets and automated identification of objects of interest in large image databases. It is not possible to separate the classes perfectly using a linear decision boundary. A bank might wish to use the classification regions to automatically decide whether future loan applicants will be given a loan or not. [Bao]

The ABC classification is a frequently used analytical method to classify objects (Customers, Products or Employees) based on a particular measure (Revenue or Profit). For example, you can classify your customers into three classes A, B and C according to the sales revenue they generate. "ABC classification allows you to classify your data based on specified classification rules. The data to be classified is generated by a query in the SAP BW. The classification rules refer to a single key figure value in your data and implicitly specify which absolute or relative key figure values map to which classes." [SAPDOCS]One should specify the following for the ABC classification: [SAPDOCS]

- The Characteristic for which the classification is to be performed. This entails specifying the characteristic values to be classified (such as Customer).
- The Key figure that is to form the basis for classifying the characteristic values (such as profit made from that customer)
- The attribute of the characteristic that should receive the result (the ABC Class)
- The Query for determining the data (such as profitability data from the customer)
- The Threshold value for the individual ABC classes. For example, all customers generating a profit of 0 to 20,000 belong to class C, those generating a profit between 20,001 and 80,000 to class B, and those generating more than 80,001 to class A.

4.3. The SAP Analysis process designer workbench

"The Analysis Process Designer (APD) is a workbench with an intuitive visual interface that enables you to visualize, transform, and deploy your data from SAP business warehouse. It combines all these different steps into a single data process that you can easily interact with" [SAPPRE 2004]. The following figure illustrates the architecture of APD:

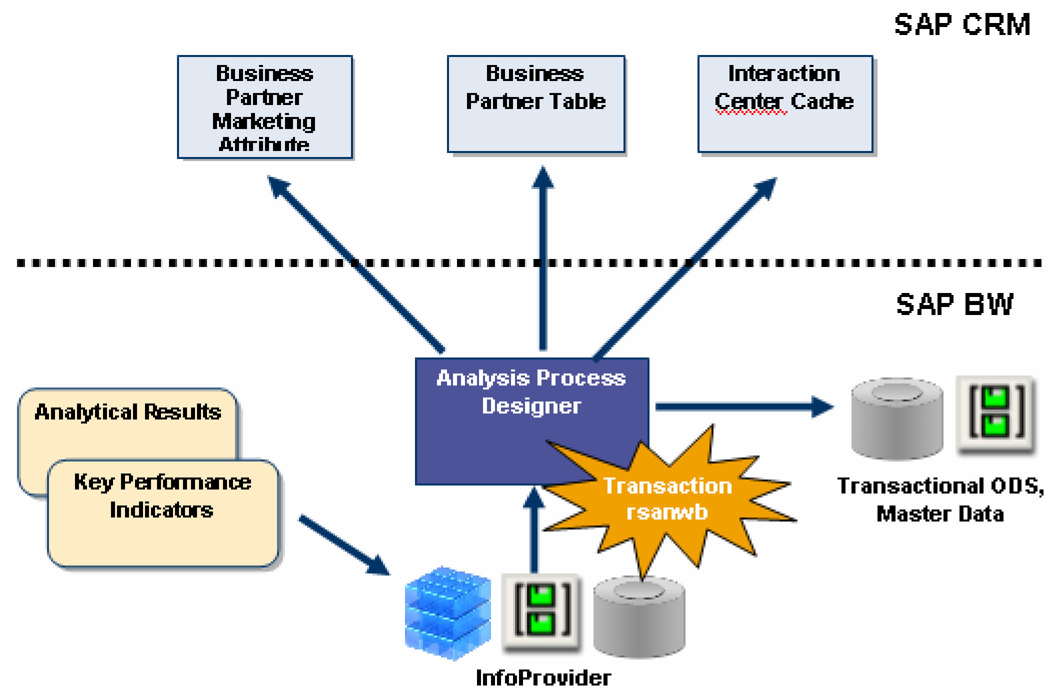

SAP CRM
Business Partner Marketing Attribute
Business Partner Table
Interaction Center Cache
SAP BW
Analytical Results
Key Performance Indicators
Analysis Process Designer
Transaction rsanwb
Transactional ODS, Master Data
InfoProvider

Figure 21: The Analysis process desiger (APD) a rchitecture.
Source: SAPNET, Analysis Process Designer

The Analysis Process Designer is the interface in the my SAP BW suite where, according to business need or questions at hand the designer has the possibility to connect to the stored data, modify the data, analyze the data (as the case may be) with the aim of getting results that would be used as answers to the questions and deliver these to an operational system where it might be used for further decision-making purposes. It is the application environment for the SAP data mining solution, from SAP BW Release 3.5 the data mining functions are fully integrated into the APD. The following functions could be performed in the APD:

- Creating and changing data mining models
- Training data mining models with SAP BW data (data mining model as data target in the analysis process)
- Execution of data mining methods such as prediction with decision tree, with cluster model and integration of data mining models from third parties (data mining model as a transformation in the analysis process)
- Visualization of data mining models

"By being fully integrated into SAP's data warehousing solution, SAP BW and the APD (including Data Mining features) realize the benefits of single database access instead of different data tables in a variety of source systems. This significantly decreases interfacing problems as well as related issues with data integrity, data quality and system performance". [SAPPRE] The figure 22 below shows a high level overview of how the APD is integrated into the SAP BW and other applications (for instance with SAP CRM).

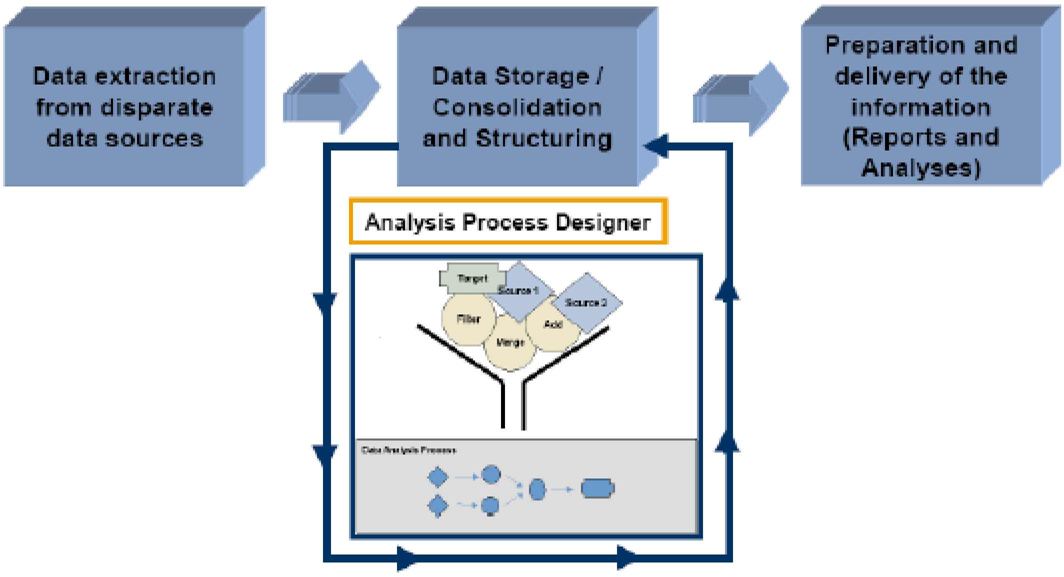

Data extraction from disparate data sources
Data Storage / Consolidation and Structuring
Preparation and delivery of the information (Reports and Analyses)
Analysis Process Designer
Target
Source 1
Source 2
Filter
Merge
Add
Data Analysis Process

Figure 22: APD integration with BW and other applications
Source: SAPNET, Analysis Process Designer

The data is first extracted from where it is stored. This could be a single instance database with several tables or several database instances with one or several tables. This data is then introduced into the SAP BW where it would be again stored, consolidated and structured. This has to be so because the APD deals basically with data within the SAP BW suite, already prepared in a form that it understands. Afterwards, the APD then manipulates the data as the case might be, interim results gained in the course of the APD process might become interesting for further analysis. This is then plugged back into the BW system and saved. Finally, the end result gained (Reports and/or Analysis) would then be prepared and delivered to where it is needed. This could be the BW system itself or an ERP system like the SAP CRM, SCM or a flat file.

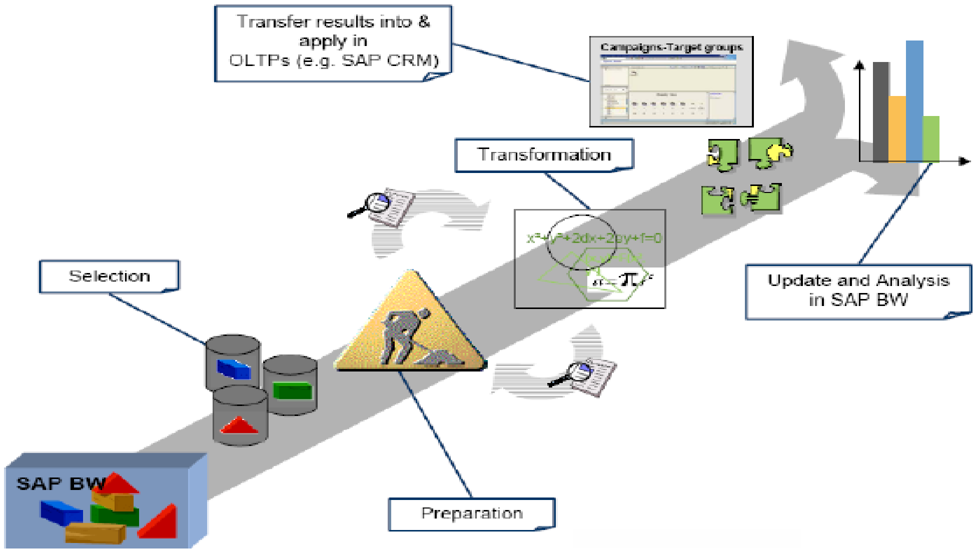

Transfer results into & apply in OLTPs (e.g. SAP CRM)
Campaigns-Target groups
Transformation
Selection
Update and Analysis in SAP BW
Preparation
SAP BW

Figure 23: Process description of the APD
Source: SAPNET, Analysis Process Designer

The above figure is process description of the APD. The system is primarily designed to extract only data that has first being uploaded into the data warehouse area of the SAP BW Suite. After the extraction process is completed, the data fields needed for the specific process is selected. The selected data fields, sets or tables are then prepared. The interim result of the preparation process might be plugged back into the system for further preparation or used for further analysis. The transformation process then follows after the preparation. The algorithm required, is at this stage, introduced into the system. It is after this that the result is discovered. This result is either stored/displayed in the SAP BW system in form of graphs, tables etc. or transferred/stored in an OLTP system (for instance SAP CRM).

This is the process that is of most importance as far as this work is concerned. Based on the perceived business need, the analysis process would be designed as to give answers to questions an organization might have. Moreover, the scenarios that are discussed in chapter 6 would be extensively explored and used in showing all the aspects involved in a typical APD modelling process.

5. AS-IS ANALYSIS: Current Situation of SAP BW Administration

5.1. The technical content of SAP BW

Implementing a Data Warehouse presents the administrator with challenges of a constantly changing nature. "Even in a productive system, in which no new InfoCubes are created, new data, for example, is always being loaded. This results in an increase in the quantity of data, or in a change to its structure. In addition to this, there are recreated or ad-hoc queries, which change the way that accessing data is seen as a whole. This not only influences the load times, but also the execution times for queries. On the other hand, it is a good idea to have an optimum work in order to minimize the response time of the Data Warehouse". [SAPDOCS] From these few points, it is already clear that an overview of the processes in the Business Information Warehouse is not only advantageous but also necessary.

SAP BW provides in the technical content of the Business Information Warehouse. For the user of the Business Information Warehouse, the most important of these sub-areas is BW statistics. The following sub-areas are delivered as per the technical content: [SAPHELP 2005-2]

- BW Statistics
- BW Data Slice
- BW Features Characteristics
- BW Formula Builder
- BEx Personalization
- Reporting Authorizations

BW Statistics: The BW statistics is of most importance as far as this work is concerned. Moreover, the clustering scenario that is discussed in chapter 6 is based on the statistics data. BW statistics is a tool for analyzing and optimizing the processes in the Business Information Warehouse. The implementation and day-to-day use of the BW leads to an increase in the overall amount of data being processed and to changes to the structure of this data. There are also new or ad-hoc queries that change the way in which the data is accessed. This affects not only the amount of time it takes to load queries, but also the amount of time it takes to execute queries. Ideally, processes should be run in such a way that the response time of the Business Information Warehouse is made as short as possible. To achieve this you need to be able to get an overview of the processes that are running in the Business Information Warehouse and be able to make any necessary changes in the system as and when required. The data that is required for the BW is provided for: [SAPHELP 2005-3]

- InfoCubes
- Queries
- InfoSources
- Aggregates

The data in BW statistics is saved and managed in the Business Information Warehouse. When a query is executed, data is specified for the OLAP server and for access to the database. This data is saved temporarily once the navigation step has been completed. This is also the case when the ODBO (OLE DB for OLAP) interface is used. Additional data is collected when the aggregates are filled and rolled up after loading data into warehouse management. It does not take long to calculate and save BW statistics data. However, the dataset can be considerable with larger installations. For this reason, the data input for each Info Provider in each area of OLAP and warehouse management can be activated and deactivated individually. It's possible to delete stored data. The following figure gives an overview of the dataflow in BW statistics:

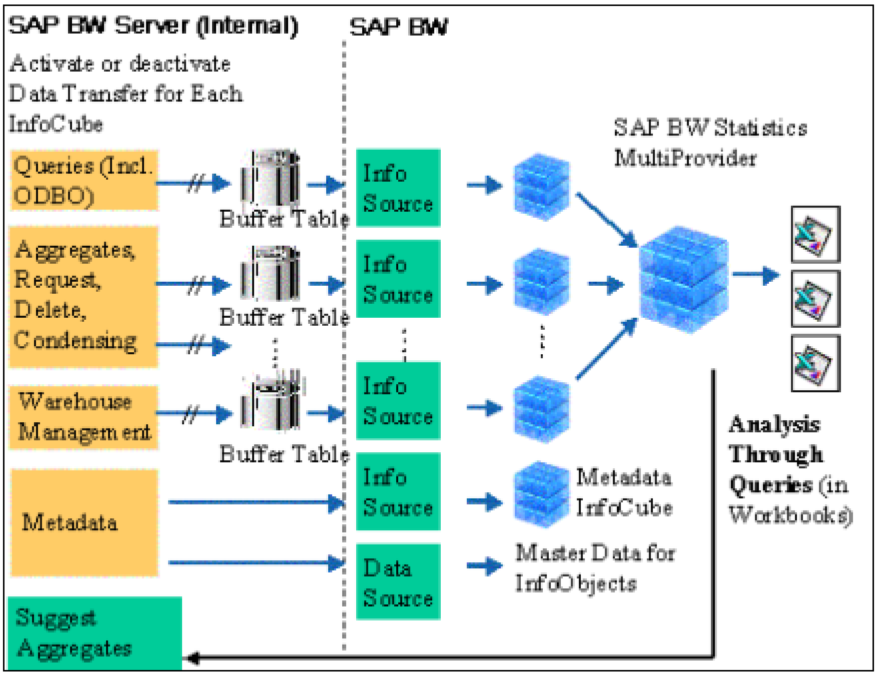


Figure 24: Overview of the dataflow in BW statistics
Source: SAP Help portal

Finally, to summarize the BW statistics helps to answer some important questions as follows:

- Which InfoCubes, info objects, info sources, source systems, queries, aggregates, and so on, are currently being used in the system? How frequently? Which datasets are being moved? Who is currently using the system?
- Are there queries, whose run time is over the allowed fast value for online processing? Are tasks, such as batch printing or loading data, executed in times of less work?
- How does the data flow through the Data Warehouse, from where and where to?

5.1.1. Statistical content cubes

Within the framework of the technical content SAP BW provides the following cubes which store the statistical content data. "A Multi Provider (MultiCube) in BW does not contain any data itself. Instead, data is stored in the relevant Info providers" [SAPDOCS]. To start with, SAP BW provides a BW Statistics Multi Provider which does not contain any data itself. Instead, data is stored in the relevant basic cubes. The relevant BasicCubes are:

- **BW Statistics – OLAP** (This Info Cube contains the data that is generated as a result of executing the queries)

- **BW Statistics - OLAP, Detail Navigation** (This Info Cube contains the data that is generated as a result of executing a query. The details correspond to the definition of the aggregate. This Info Cube is used by the BW system for the proposal of aggregates.)

- **BW Statistics – Aggregates** (This Info Cube contains not only general data but also data that appears in an aggregate after data is filled and rolled up)

- **BW Statistics – WHM** (This Info Cube contains the data that arises from the execution of a process in Warehouse Management. This Info Cube allows you to see how data requests are processed for the process concerned -for example, from which source system are they, which Info Source is used with which transfer method, and in what time frame)

- **BW Statistics – Metadata** (This cube contains metadata from the Metadata Repository. It does not contain any transaction data and no data is loaded. The Info Cube also does not contain any special key figures. It reveals the information about the existing Objects and structures in the OLAP, WHM and BEx areas, and about the BW Metadata Repository and hierarchies to be displayed)

- **BW Statistics: Condensing InfoCubes** (This Info Cube contains data that is created when an Info Cube's data requests are compressed. It reveals information on the number of edited data records for condensing or compressing an InfoCube and the runtime of the condenser, which is the program that compresses the fact table contents of an Info Cube)

- **BW Statistics: Deleting Data from InfoCubes** (This Info Cube contains the data that results from deleting data from an Info Cube)

The Info Cube BW Statistics – OLAP is the most important cube as far as this work is concerned. As the major part of the analysis is on Query performance and optimization, this Cube contains some important characteristics (Info Cube, BW System , User , Query , Time and so on) and Key figures (OLAP times, Data manager times and so on). The basic idea was to use these key figures for Cluster analysis to come up with some useful patterns.

5.1.2. Brief overview of some Characteristics and key figures

As mentioned before, the major analysis of this work is on performance of the queries and the BW Statistics – OLAP cube contains some of the important key figures used for analysis. The following section gives an overview of the Characteristics, Time Characteristics and Key figures available as part of this cube:

Characteristics

- InfoCube
- Navigation Step (current numbers within the session)
- OLAP Reading On / Off
- Runtime Category (1, 2, 3, ... 10, 20, 30, ... Seconds)
- BW System
- User
- OLAP Processor Method
- Navigation Step (GUID)
- Front-end Session (GUID)
- Statistical Data (GUID)
- Object Version (for example, 0TCTIFCUBE)
- Type of Data Read
- UTC Time Stamp
- Time
- Query

Time Characteristics

- Calendar Day
- Calendar Year
- Calendar Year / Month
- Calendar Year / Quarter
- Calendar Year / Week

Key figures

- Start Date
- Frequency
- Start Time
- Number of Database Selects
- Number of Navigations
- Number of Front-end Sessions
- Number of Texts Read
- Cells Transferred to the Front-end
- Records Selected on the Database
- Records transferred from the Database to the Server
- ODBO: Size of the Internal Buffer
- ODBO: External Calls for the Function Module
- Total (OLAP)
- Read Cycles (Fetch) OLAP Processor
- Formatting Transferred to the Front-end
- Number of Texts Read
- Time, Authorization Check
- Time, Reading on the Database
- Time, Data Manager InfoCube Access
- Time, Data Manager Reading from Basic Cube
- Time, Data Manager Reading from ODS
- Time, Data Manager Reading from Remote Cube
- Time, Data Manager Auth orizations for Non-Cumulative
- Time, Data Manager Determining SIDs for Remote Cube
- Time, Front-end
- Time Between Navigation Steps
- Time, General ODBO
- Time, ODBO: Axes Preparation
- Time, ODBO: Data Records Preparation
- Time, ODBO: Conversion into Flat Table Form
- Time, ODBO: Initialization
- Time, ODBO: Data Requests
- Total Time (OLAP)
- Time, OLAP Processor Initialization
- Time, Reading Texts/Master Data
- Time That the System Was Unable to Assign
- Time, Inputting Variables

- Time, OLAP Processor

The major objective is to analyse these key figures, their importance in performance of the query. Since, it has got huge number of key figures it's always not so easy to decide which key figures need to be considered for the cluster analysis. Finally the OLAP times and Data Manager times and other key figures were considered for analysis, which is described in chapter 6.

5.2. SAP BW administration and monitoring

Data warehousing is indeed becoming common place in large organizations. According to a Forrester Research survey of executives at large firms 62% percent have data in, on average, three data warehouses or data marts. The same survey indicates that the pace of data warehousing will increase before it slows down; with the average growth showing the number of data warehouses and marts to double to nearly six by 2004 and increase in size from approximately 130 GB to approximately 260 GB. [IDUG 2005]

A data warehouse is a completely different beast from the operational OLTP. Its problems and the tools needed to solve them are different. Form these it is quite clear that administrators are very much concerned with warehouse availability and performance during access. Coming to the SAP BW, The Administrator Workbench (AWB) is the main tool for tasks in the data warehousing process. "The AWB provides data modelling functions as well as functions for control, monitoring and maintenance of all processes in SAP BW having to do with data procurement, data retention, and data processing". [SAPHELP 2005-4] The following functions are provided as part of AWB:

- Modelling
- Monitoring
- Reporting Agent
- Transport connection
- Documents
- Business Content
- Translation
- Metadata Repository

To summarize, it becomes quite clear that the administration of complex enterprise data warehouses plays a pivotal role in today's IT landscapes and how one could use Data Mining methods to support the administration of data warehouses considering the Performance and system stability that eventually motivated to analyse the query performance with cluster analysis.

5.3. Possible business scenarios for data mining

During the initial phase of the investigation several issues were considered as to how and in which areas of the BW, the Data mining methods could be useful. It was not always easy to find out as there are several other qualitative aspects that could influence the performance and stability of SAP systems for instance:

- The number of the application servers available
- The underlying database technology
- The number of work processes available at a particular moment of time

- The number of parallel processes available at a moment of time, These are a few qualitative factors which are not easy to measure and may be in future further research might help to even measure such kind of qualitative aspects of these typical SAP systems.

Finally, after asking several experts in these areas, the following areas are identified where Data mining methods might be useful to support data warehouse administration in SAP BW:

5.3.1. Data loads and Process chains

The execution of data load processes in Warehouse Management - for e.g. how data requests are processed for a particular process. Presently, as of BW 3.5 the BW administrator could monitor the data load processes arising out of the data loads using the statistics content cube named 'BW Statistics – WHM' (Technical name: 0BWTC_C05). This InfoCube helps the administrator to see how data requests are processed for the process concerned (for example, from which source system are they, which Info Source is used with which transfer method, and in what time frame). Presently there are a few key figures as part of this cube. The important ones are:

- Records (WHM Process) for a particular processing step when loading data
- Time (WHM Process) for a particular processing step when loading data

It seems as if more key figures might be needed and which and how these new key figures could be derived is out of the scope of this paper. But, further investigation could be made in this regard such that the new key figures are used for Data mining purpose in future.

5.3.2. Queries

The term 'Query' is the much talked about buzz-word of the available objects in SAP BW. The Query is of utmost importance since it is the object through which the data available in the Data warehouse (for instance SAP BW) is presented using the front end tools (BEx), based on the typical reporting requirements of the users.

Several factors determine how well a query performs, some with greater influence than others. Presently, as of BW 3.5 the SAP BW administrator could monitor the queries using the statistics content Cube BW Statistics – OLAP (Technical name: 0BWTC_C05). There are a lot of key figures which could help in analysing the query performance as part of this which are documented in the above section (Pease refer to the section on Brief overview of some Characteristics and key figures). With the available key figures namely OLAP key figures the administrator could find out reasons for the performance of the queriesF or e.g. the administrator could look at the various OLAP times:

- Time, OLAP Processor Initialization
- Time, OLAP Processor
- Time, Reading from the Database
- Time, Front-end
- Time, Authorization Check
- Time, Reading Texts/Master Data

From the above OLAP key figures the administrator can check which key figures are responsible for the high OLAP times, as the case may beand perform the necessary action steps. Taking this idea in to consideration these OLAP key figures are used for Cluster analysis. The algorithm used is known as K-means cluster analysis, which is available as part of SAP's Data mining workbench. The details of the analysis are documented in chapter 6.

5.3.3. Dormant data

Dormant data is data that is seldom or never used. "Studies show that much of the data loaded into data warehouses and analytical application databases is dormant, that is, it is infrequently used or never used." [ITTOOLBOX 2005] Unlike OLTP databases, data warehouses like SAP BW continuously collect and store detailed and summary historical information for business analysis. Frequently data warehouses include information to satisfy unknown requirements and data is included that may or may not be used. These databases expand significantly over time as new information is added from internal and external data sources.

Bill Inmon, a noted data warehouse expert, states that "dormant data typically increases as a percentage of total data as warehouses grow. He asserts that dormant data may be as much as 65% - 70% of data warehouses that are a terabyte or greater in size." [FILETEK 2005] He recommends a simple formula for calculating the data dormancy ratio "the number of queries per year times the average amount of data per query divided by total data warehouse space." [FILETEK] While this ratio may be high since it does not consider that some queries inevitably use the same data, it does provide a rule of thumb for making ballpark estimates. But, how can one actually identify dormant data? Bill Inmon writes, "Understanding that there is dormant data in a data warehouse is one thing. Finding the dormant data is another matter altogether. The best way to find the dormant data is to monitor the end users query activity against the data warehouse ... the monitor sits between the end-users query activity and the data warehouse server." [FILETEK]

From the above section it is quite clear that as part of Data warehouse tools like SAP BW, there is a desperate need for some kind of monitor to say that a particular data could be archived for a certain moment of time, presentlywe don't ha ve any monitor in SAP BW. Once a product like SAP BW offers such kind of monitor, these key figures could be further used for data analysis and further investigation for the possibility of any Data mining methods could be realized in the future. To summarize, minimizing dormant data reduces system costs and improves performance, service levels and IT staff productivity and this paper strongly recommends coming up with some sort of monitor in the near future.

5.3.4. Table spaces and buffers

Another interesting aspect of a typical data warehouse tool like SAP BW which is directly related to performance and system stability are table spaces and buffers at the data base layer.
There are number of factors that are responsible for the performance of table spaces and buffers from the SAP BW perspective, some of them are:

- The size of the Info Cube size,
- The number of partitions of an Info Cube
- The number of CPU's and their respective times
- The Database specific settings and so on.

It's quite clear that there are monitors for SAPsystems for e.g. Database Performance Analysis (Transaction code: ST04), Database Tables and Index Monitor (Transaction code: DB02) and it does make sense to take into to derive some key figures like Number of table spaces / buffers, Amount and time of table space / buffers, CPU times etc, which could be eventually used for Data mining purposes

5.4. A way forward

To summarize the AS-IS analysis, the current situation of SAP BW administration as described in the previous sections, itseems quite clear that the companies like SAP are looking forward to bundle some sophisticated Data Mining features as part of their products (Here SAP BW) to easily administer and monitor the complexities of data ware houses that eventually will lead to make ease the day to day activities of the SAP BW administrators. After successfully knowing the needs and the possible areas it's quite obvious to pick up a Business scenario that would help in the realization of the TO-BE analysis

At this stage of this work, after identifying the possible areas in SAP BW namely Data load processes, Queries, Dormant data and table spaces, the strategy is to cut horizontally - taking in to account the time and technical constraints which led to the idea of Cluster analysis for the Queries, which is further described in the TO-BE Analysis.

6. TO-BE ANALYSIS – A Scenariowith Cluster Analysis

6.1. Motivations forcluster Analysis

6.1.1. Technical drivers

The major technical driver for the cluster analysis is obviously, the availability of clustering algorithm as part of the SAP's Data Mining work bench. The algorithm for the clustering (k-means is implemented as part of the work bench) is already pre-configured into the system and simply made available for use. Nevertheless, effort would be made here to describe the fundamental principle behind the concept. As clustering is used to group records together according to an algorithm or mathematical formula that attempts to find centroids or centres, around which similar records gravitate. This method initially takes the number of components of the population equal to the final required number of clusters. In this step itself the final required number of clusters is chosen such that the points are mutually farthest apart. Next, it examines each component in the population and assigns it to one of the clusters depending on the minimum distance. The distance measure used is the Euclidean metric. It simply is the geometric distance in the multidimensional space. It is computed as: [SAPDOCS 2005]

$$\mathbf{Distance\ (x,\ y) = \{\ \Sigma\ (x_i - y_i)_2\ \}^{1/2}}$$

After every input record is assigned to some cluster or the other, the centroid's position is recalculated based on the records assigned to it. With the new centroids means, the assignments are checked again and this continues until a all the stopping conditions are reached (i.e., maximum number of iterations reached or cluster assignments do not change much between iterations)

The availability of APD functionality for creating and changing data mining models, training data mining models with is integrated in SAP BW , the availability of data transformations functions and visualization of data mining models is also technical motivator for the analysis.

6.1.2. Business drivers

The main business driver is to provide some kind of monitor, proactively for the administration of SAP BW. With the use of statistical content data the objective is to track down Query behaviour as to divide the queries into segments based on key figures namely OLAP times. Since, these times are responsible for the performance of queries. As well this is confirmed in the AS-IS analysis phase from several experts at SAP from different areas. The major key figures used for the cluster analysis are various OLAP times namely:

- Time, OLAP Processor Initialization
- Time, OLAP Processor
- Time, Reading from the Database
- Time, Front-end
- Time, Authorization Check
- Time, Reading Texts/Master Data
- The Number of times a Query is executed

The ultimate objective is to cluster the queries in to different groups for a certain time period and throw to the BW administrator those clusters, which might seem to be peculiar such that, the results might help the BW administrator to perform the necessary follow up actions, that would eventually help in performance and monitoring of these Queries in the future.
Looking, at the present developments (to bundle new features as part of their products) with reference to SAP BW, It does make sense to have performance and proactive monitors as part of SAP BW which is the main motivation from the Business / product management perspective.

6.2. Analysis of Queries with cluster analysis

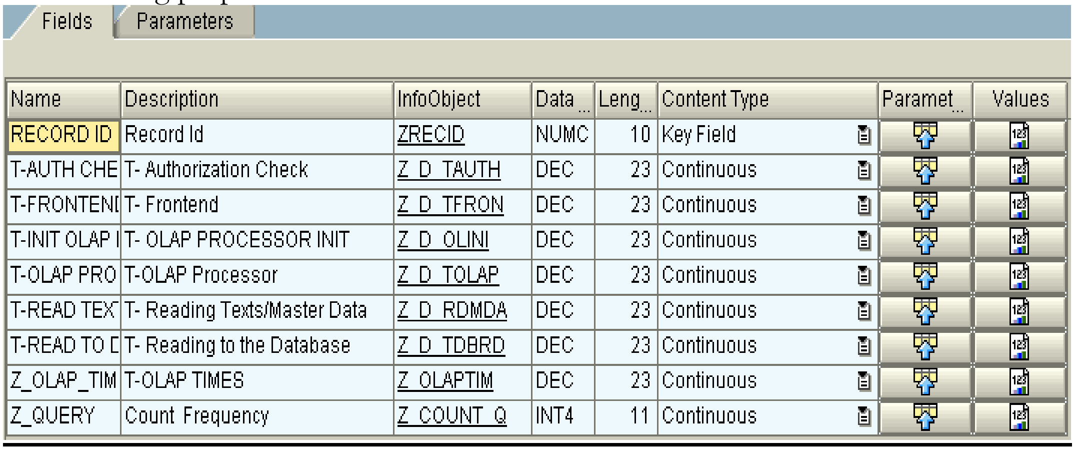
The main Objective is to track the behaviour of queries and divide them into segments based on the OLAP times. The Key figures considered for the analysis are OLAP Times and Data manager times. For this particular data model the Total time (OLAP) is taken in to consideration, which enables the valuation of query runtimes which includes the following times:

- Initialization of OLAP Processor
- OLAP Processor
- Reading on the Database
- Front End
- Reading Texts/Master Data
- Authorization Check
- A new Key figure Count Frequency (Number of times a query is executed)
- The total OLAP time (Which is an aggregated key figure of all the above OLAP times)

Finally the data model consists of 9 key figures (Excluding Record id used for record identification). All the activities are performed on AB5 and Q50 on internal SAP test systems

6.3. The Data Model:

To come up with this meaningful data model, much investigation is done by consulting experts in these areas. As far as this work is concerned several attempts have been made to come up with this meaningful data model. The following figure depicts the modal attributes which are used for the clustering purpose.



Name	Description	InfoObject	Data ...	Leng...	Content Type	Paramet...	Values
RECORD ID	Record Id	ZRECID	NUMC	10	Key Field		
T-AUTH CHE	T- Authorization Check	Z_D_TAUTH	DEC	23	Continuous		
T-FRONTENI	T- Frontend	Z_D_TFRON	DEC	23	Continuous		
T-INIT OLAP I	T- OLAP PROCESSOR INIT	Z_D_OLINI	DEC	23	Continuous		
T-OLAP PRO	T-OLAP Processor	Z_D_TOLAP	DEC	23	Continuous		
T-READ TEX	T- Reading Texts/Master Data	Z_D_RDMDA	DEC	23	Continuous		
T-READ TO D	T- Reading to the Database	Z_D_TDBRD	DEC	23	Continuous		
Z_OLAP_TIM	T-OLAP TIMES	Z_OLAPTIM	DEC	23	Continuous		
Z_QUERY	Count Frequency	Z_COUNT_Q	INT4	11	Continuous		

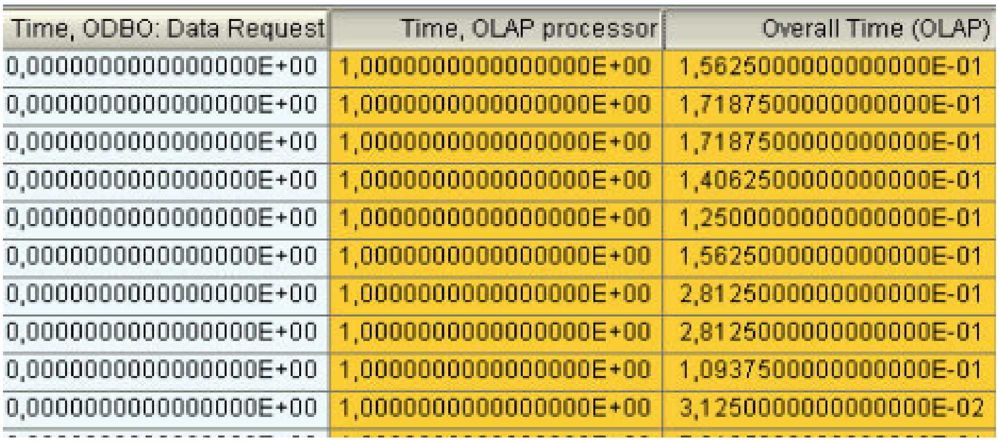
Figure 25: The Data Model

The models are created and evaluated on the SAP internal Test systems (AB5 and Q50). The technical name of this model is C_IWP_20_1

6.4. Data preparation

The data preparation is one of the major tasks and much time is devoted to this part of the work. It has been known that the data for Data mining entirely depends on the data distributions and the amount of source data. Initially attempts were made based on the data used on the test systems and ironically the results don't show up any patterns. Then after consulting with the experts it was known that it makes sense to work with the data from a productive system. Finally, the data used for analysis is from the productive internal BW system.

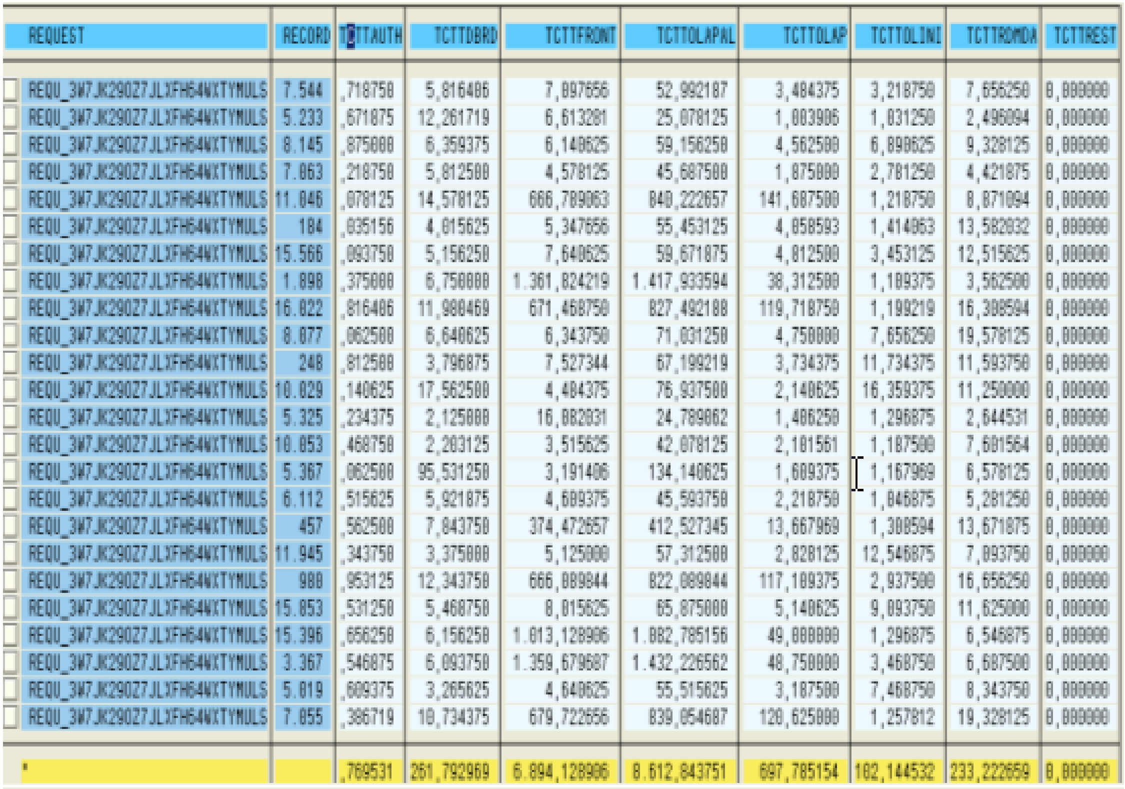
Several attempts were made from the concerned colleagues of the productive system to load the data in to test system, but the data load process in to the info cube was not successful due to some technical constraints for instance the data loaded in to the Cube has data quality problems where the time of the OLAP processor is always filled with the value 1 and the time of the OLAP processor is greater than the Overall time of OLAP, Which should not be the case as the Overall time of OLAP is atotal of all OLAP times (**Overall OLAP time = Time, OLAP Processor Initialization + Time, OLAP Processor + Time, Reading from the Database + Time, Front-end + Time, Authorization Check + Time, Reading Texts/Master Data + ODBO times**) as shown in the figure26



Time, ODBO: Data Request	Time, OLAP processor	Overall Time (OLAP)
0,0000000000000000E+00	1,0000000000000000E+00	1,5625000000000000E-01
0,0000000000000000E+00	1,0000000000000000E+00	1,7187500000000000E-01
0,0000000000000000E+00	1,0000000000000000E+00	1,7187500000000000E-01
0,0000000000000000E+00	1,0000000000000000E+00	1,4062500000000000E-01
0,0000000000000000E+00	1,0000000000000000E+00	1,2500000000000000E-01
0,0000000000000000E+00	1,0000000000000000E+00	1,5625000000000000E-01
0,0000000000000000E+00	1,0000000000000000E+00	2,8125000000000000E-01
0,0000000000000000E+00	1,0000000000000000E+00	2,8125000000000000E-01
0,0000000000000000E+00	1,0000000000000000E+00	1,0937500000000000E-01
0,0000000000000000E+00	1,0000000000000000E+00	3,1250000000000000E-02

Figure 26: Data Preparation - Data quality problems

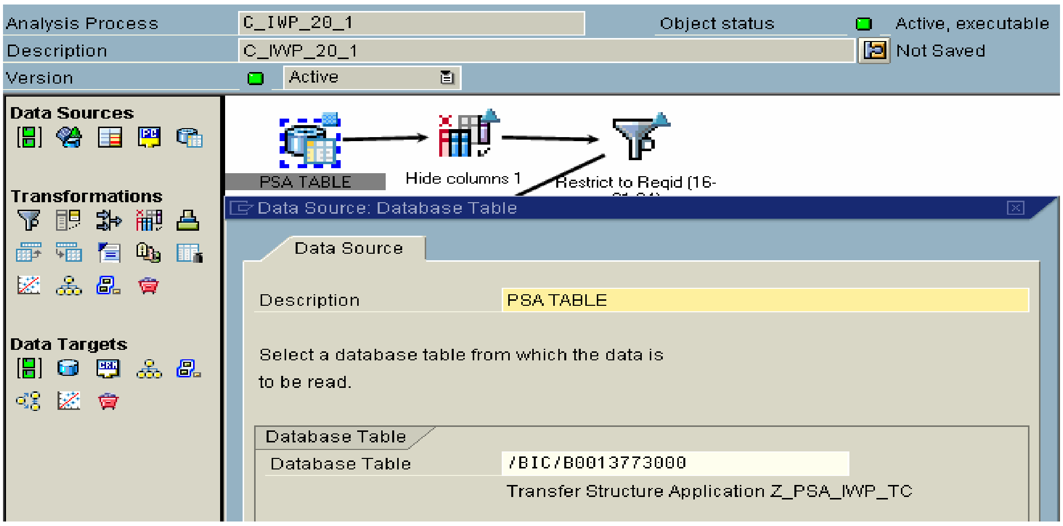
But fortunately, after several attempts the data could be loaded in to the PSA (An intermediate data store before loading in to the cube and its known as one type of upload type in SAP BW). As to the outcome, the PSA table is used for the analysis and correspondingly checked for the data consistency by manually totalling the key figures as shown in the figure 27.



	REQUEST	RECORD	TCTTAUTH	TCTTDBRD	TCTTFRONT	TCTTOLAPAL	TCTTOLAP	TCTTOLINI	TCTTROMDA	TCTTREST
□	REQU_3W7JK29027JLXFH64NXTYMULS	7.544	,718750	5,816406	7,097656	52,992187	3,484375	3,218750	7,656250	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	5.233	,671875	12,261719	6,613281	25,078125	1,003906	1,031250	2,496094	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	8.145	,875000	6,359375	6,140625	59,156250	4,562500	6,890625	9,328125	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	7.063	,218750	5,812500	4,578125	45,687500	1,875000	2,781250	4,421875	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	11.046	,078125	14,578125	666,789063	840,222657	141,687500	1,218750	8,871094	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	184	,035156	4,015625	5,347656	55,453125	4,058593	1,414063	13,582032	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	15.566	,093750	5,156250	7,640625	59,671875	4,812500	3,453125	12,515625	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	1.898	,375000	6,750000	1.361,824219	1.417,933594	38,312500	1,109375	3,562500	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	16.822	,816406	11,980469	671,468750	827,492188	119,718750	1,199219	16,308594	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	8.077	,062500	6,640625	6,343750	71,031250	4,750000	7,656250	19,578125	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	248	,812500	3,796875	7,527344	67,199219	3,734375	11,734375	11,593750	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	10.029	,140625	17,562500	4,484375	76,937500	2,140625	16,359375	11,250000	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	5.325	,234375	2,125000	16,082031	24,789062	1,406250	1,296875	2,644531	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	10.053	,468750	2,203125	3,515625	42,078125	2,101561	1,187500	7,601564	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	5.367	,062500	95,531250	3,191406	134,140625	1,609375	1,167969	6,578125	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	6.112	,515625	5,921875	4,609375	45,593750	2,218750	1,046875	5,281250	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	457	,562500	7,843750	374,472657	412,527345	13,667969	1,308594	13,671875	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	11.945	,343750	3,375000	5,125000	57,312500	2,828125	12,546875	7,093750	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	980	,953125	12,343750	666,089844	822,089844	117,109375	2,937500	16,656250	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	15.053	,531250	5,468750	8,015625	65,875000	5,140625	9,093750	11,625000	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	15.396	,656250	6,156250	1.013,128906	1.082,785156	49,000000	1,296875	6,546875	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	3.367	,546875	6,093750	1.359,679687	1.432,226562	40,750000	3,468750	6,687500	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	5.019	,609375	3,265625	4,640625	55,515625	3,187500	7,468750	8,343750	0,000000
□	REQU_3W7JK29027JLXFH64NXTYMULS	7.055	,386719	10,734375	679,722656	839,054687	120,625000	1,257812	19,328125	0,000000
	*		,769531	261,792969	6.894,128906	8.612,843751	697,785154	102,144532	233,222659	0,000000

Figure 27: Data Preparation - Consistency check

As a result the PSA table is used as the Data Source instead of Info Cube or Query as shown in the figure28 .



Analysis Process C_IWP_20_1 Object status 🟩 Active, executable

Description C_IWP_20_1 Not Saved

Version 🟩 Active

Data Source: Database Table

Data Source

Description PSA TABLE

Select a database table from which the data is to be read.

Database Table

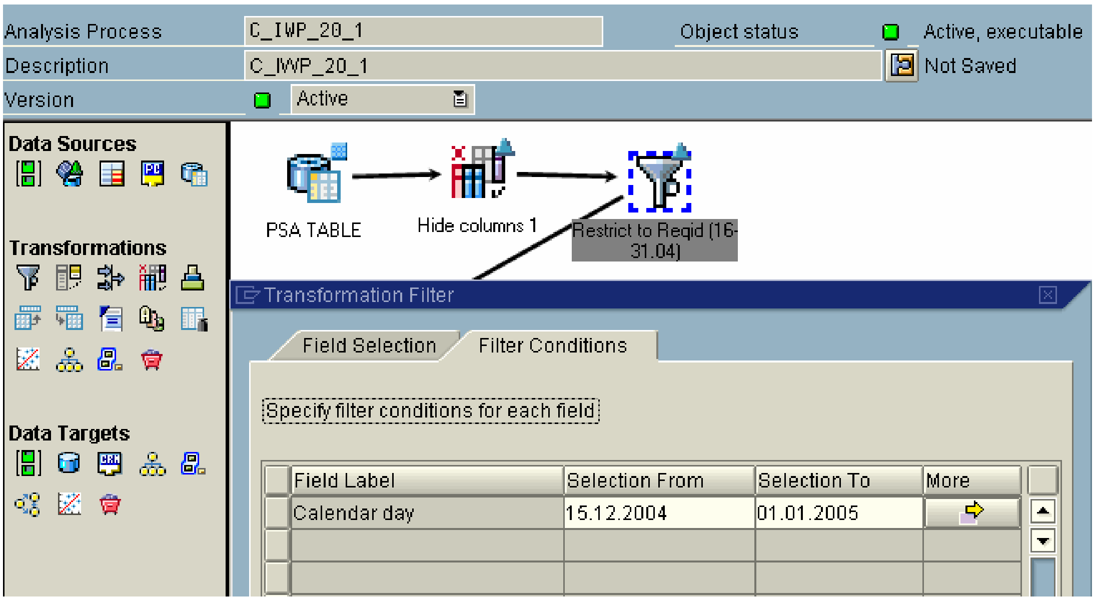
Database Table /BIC/B0013773000

Transfer Structure Application Z_PSA_IWP_TC

Figure 28: Data Selection - The PSA as a source table

Data Selection

The Data is filtered accordingly for a period of 15 days keeping in mind the objective is to provide the administrator some meaningful clusters for analysis. One of the main purposes of Data mining is to mine data on large data sets. So finally, it was decided to at least have 800 records and the least possible time period. So the data is selected for 15 days since its more than 800 records for this time period and after aggregation it consists 1116 records (available in the next screens). The PSA as a data source is shown in the following figure



Analysis Process	C_IWP_20_1
Description	C_IWP_20_1
Version	🟩 Active
Object status	🟩 Active, executable
	Not Saved

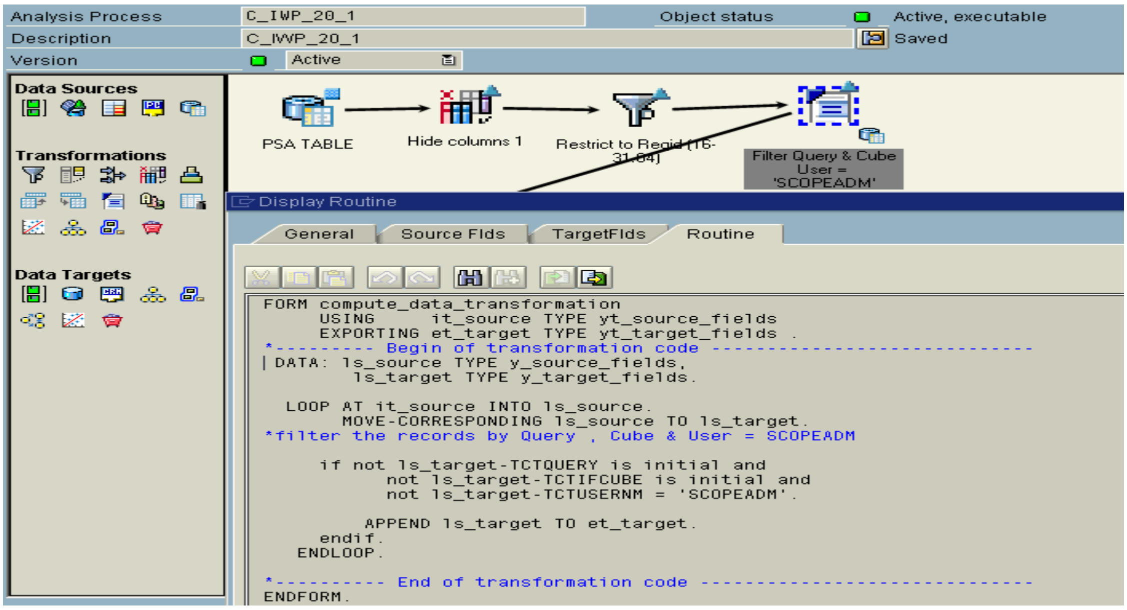
Specify filter conditions for each field

Field Label	Selection From	Selection To	More
Calendar day	15.12.2004	01.01.2005	

Figure 29: Data Selection - Time period of data

6.5. Data transformation

Several transformations are made to make the data meaningful after consulting the experts in these areas, which will be discussed in the following sections. In the first step the Records are further filtered for Query, Info provider to get rid of the initial values and the user ‘SCOPEADM’ since he is not the genuine user as in the figure 30.

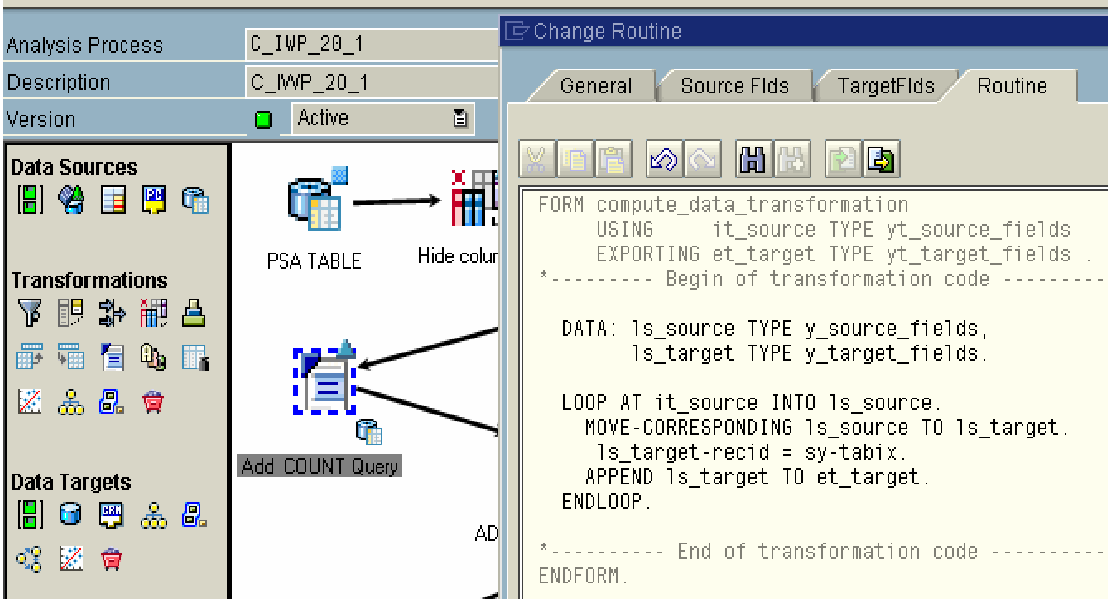


Analysis Process	C_IWP_20_1
Description	C_IWP_20_1
Version	🟩 Active
Object status	🟩 Active, executable
	Saved

```
FORM compute_data_transformation
     USING     it_source TYPE yt_source_fields
     EXPORTING et_target TYPE yt_target_fields .
*--------- Begin of transformation code ------------------------------
| DATA: ls_source TYPE y_source_fields,
        ls_target TYPE y_target_fields.

  LOOP AT it_source INTO ls_source.
       MOVE-CORRESPONDING ls_source TO ls_target.
*filter the records by Query , Cube & User = SCOPEADM

     if not ls_target-TCTQUERY is initial and
           not ls_target-TCTIFCUBE is initial and
           not ls_target-TCTUSERNM = 'SCOPEADM'.

          APPEND ls_target TO et_target.
     endif.
   ENDLOOP.

*---------- End of transformation code ------------------------------
ENDFORM.
```

Figure 30: Data transformations- Filter Query, Info Cube and User

One of the attributes considered for clustering is the number of times a Query is executed from the data set selected for the specified time period, so a new key figure called query frequency is added to the analysis process as in the figure 31.

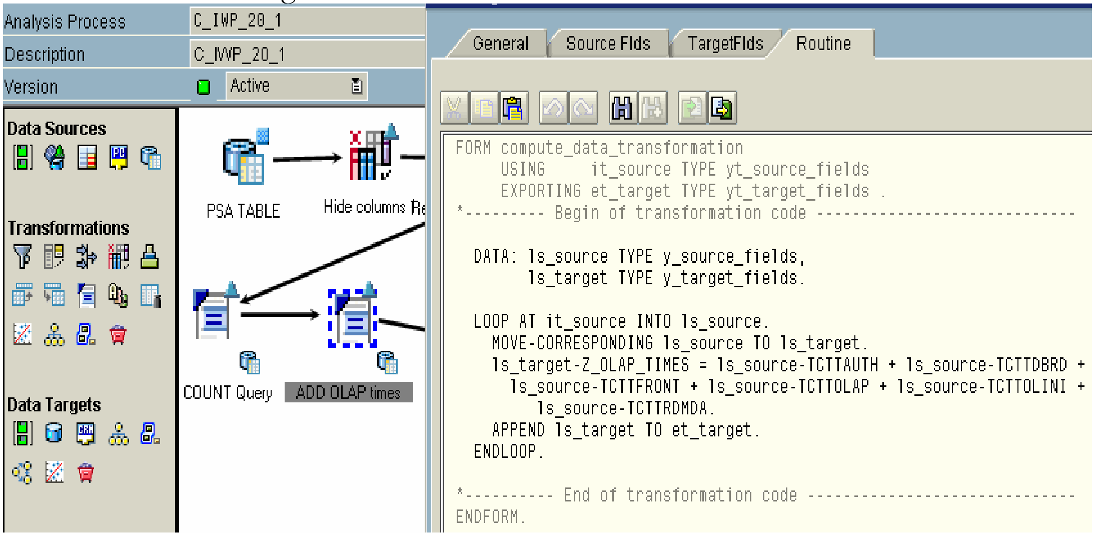


Analysis Process C_IWP_20_1
Description C_IWP_20_1
Version Active

```
FORM compute_data_transformation
     USING     it_source TYPE yt_source_fields
     EXPORTING et_target TYPE yt_target_fields .
*--------- Begin of transformation code ---------

  DATA: ls_source TYPE y_source_fields,
        ls_target TYPE y_target_fields.

  LOOP AT it_source INTO ls_source.
    MOVE-CORRESPONDING ls_source TO ls_target.
     ls_target-recid = sy-tabix.
    APPEND ls_target TO et_target.
  ENDLOOP.

*---------- End of transformation code ----------
ENDFORM.
```

Figure 31: Data Transformation - Adihg new key figure

The next step in the transformation is to transform the Total OLAP time from the corresponding OLAP times as in the figure 32



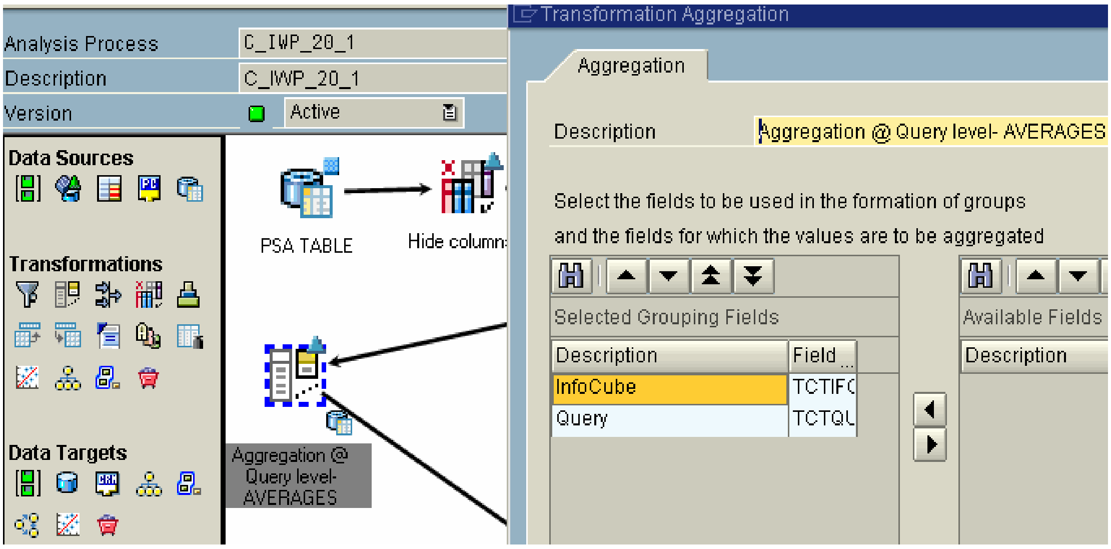
Analysis Process C_IWP_20_1
Description C_IWP_20_1
Version Active

```
FORM compute_data_transformation
     USING     it_source TYPE yt_source_fields
     EXPORTING et_target TYPE yt_target_fields .
*--------- Begin of transformation code ----------------------------

  DATA: ls_source TYPE y_source_fields,
        ls_target TYPE y_target_fields.

  LOOP AT it_source INTO ls_source.
    MOVE-CORRESPONDING ls_source TO ls_target.
    ls_target-Z_OLAP_TIMES = ls_source-TCTTAUTH + ls_source-TCTTDBRD +
      ls_source-TCTTFRONT + ls_source-TCTTOLAP + ls_source-TCTTOLINI +
         ls_source-TCTTRDMDA.
    APPEND ls_target TO et_target.
  ENDLOOP.

*---------- End of transformation code ----------------------------
ENDFORM.
```

Figure 32: Data Transformation - Transformation of OLAP times

6.5.1. Data Aggregation

The important step in data transformations is to get rid of the repeated queries and info provider values in the data set and to come up with the unique values, As a result the aggregation is performed at the Query and Info Provider level and the average values are used as the type of aggregation. The process of aggregation is shown in the figure 33



Analysis Process: C_IWP_20_1
Description: C_IWP_20_1
Version: Active

Transformation Aggregation

Aggregation

Description: Aggregation @ Query level- AVERAGES

Select the fields to be used in the formation of groups
and the fields for which the values are to be aggregated

Selected Grouping Fields

Description	Field ...
InfoCube	TCTIFC
Query	TCTQU

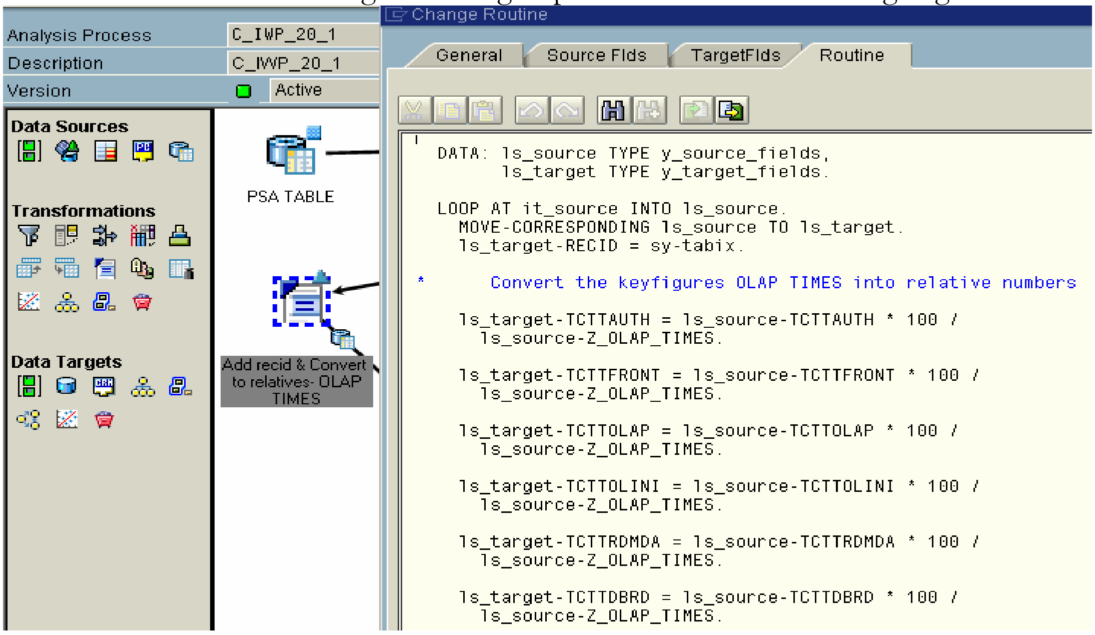
Available Fields

Description

Figure 33: Data Transformations - Aggregation of data

6.5.2. Relative numbers

Based on the expert advice, it makes sense to work on the relative values (the percentages) for the corresponding OLAP times with a transformation routine as in the figure 34, to get rid of the uneven data distributions and get meaningful patterns from the clustering engine.



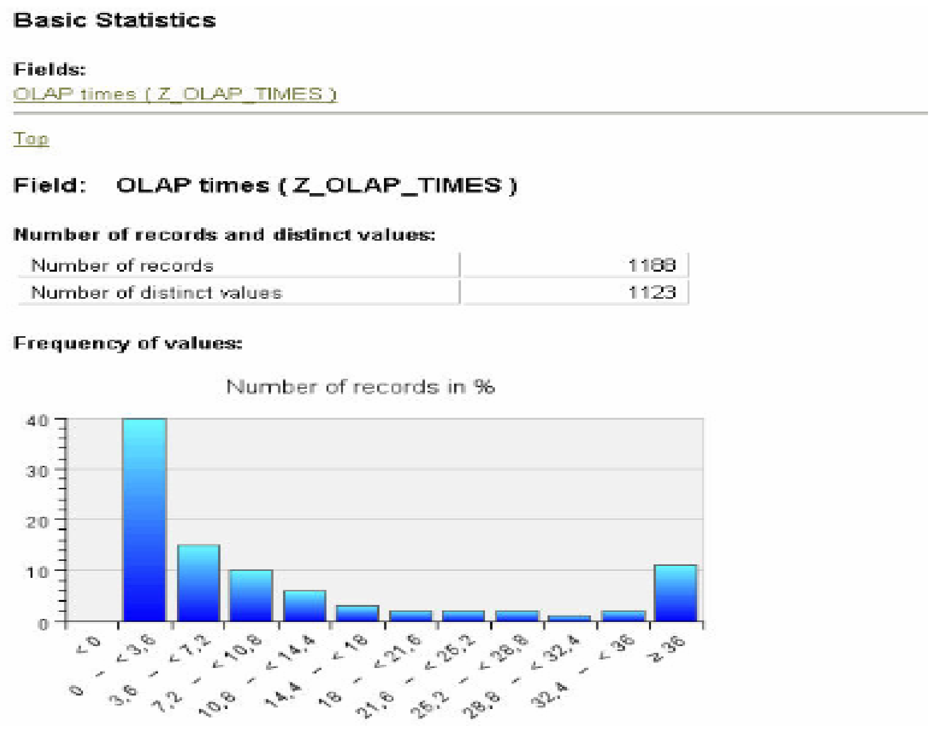
Analysis Process: C_IWP_20_1
Description: C_IWP_20_1
Version: Active

Change Routine

```
  DATA: ls_source TYPE y_source_fields,
        ls_target TYPE y_target_fields.

  LOOP AT it_source INTO ls_source.
    MOVE-CORRESPONDING ls_source TO ls_target.
    ls_target-RECID = sy-tabix.

*       Convert the keyfigures OLAP TIMES into relative numbers

    ls_target-TCTTAUTH = ls_source-TCTTAUTH * 100 /
      ls_source-Z_OLAP_TIMES.

    ls_target-TCTTFRONT = ls_source-TCTTFRONT * 100 /
      ls_source-Z_OLAP_TIMES.

    ls_target-TCTTOLAP = ls_source-TCTTOLAP * 100 /
      ls_source-Z_OLAP_TIMES.

    ls_target-TCTTOLINI = ls_source-TCTTOLINI * 100 /
      ls_source-Z_OLAP_TIMES.

    ls_target-TCTTRDMDA = ls_source-TCTTRDMDA * 100 /
      ls_source-Z_OLAP_TIMES.

    ls_target-TCTTDBRD = ls_source-TCTTDBRD * 100 /
      ls_source-Z_OLAP_TIMES.
```

Figure 34: Data Transformations - Conversion of OLAP key figures

Look at the data distributions for the Total OLAP time is in such a way as shown in the figure 35.



Basic Statistics

Fields:
OLAP times (Z_OLAP_TIMES)

Top

Field: OLAP times (Z_OLAP_TIMES)

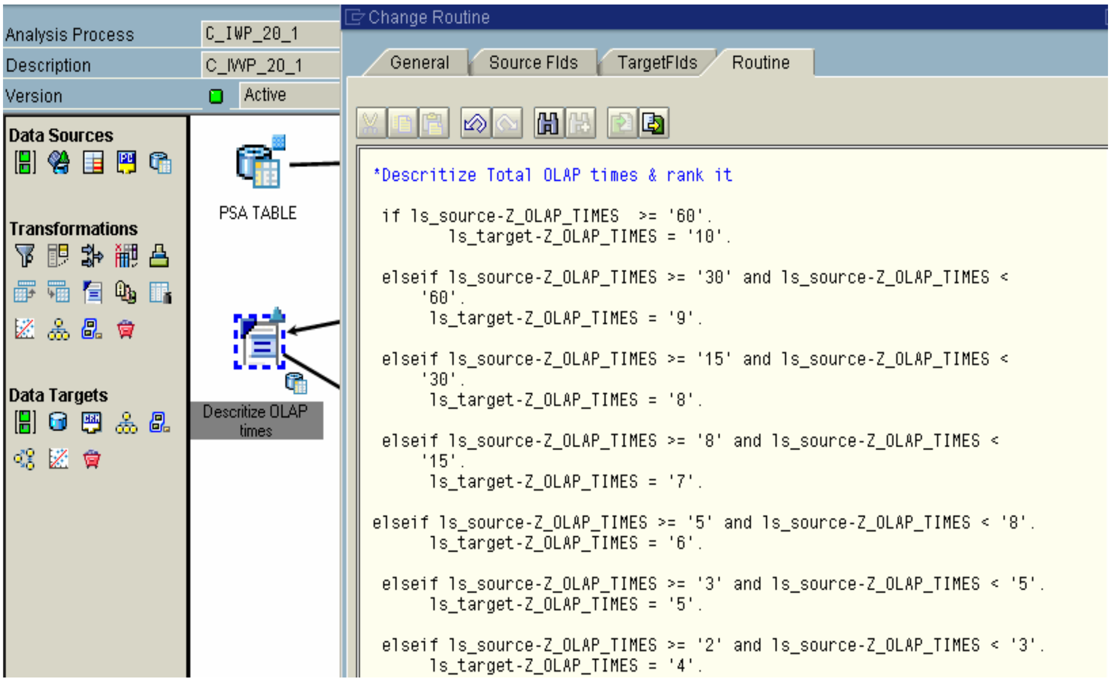
Number of records and distinct values:

Number of records	1188
Number of distinct values	1123

Frequency of values:

Figure 35: Data distribution of TOTAL OLAP time before Transformation

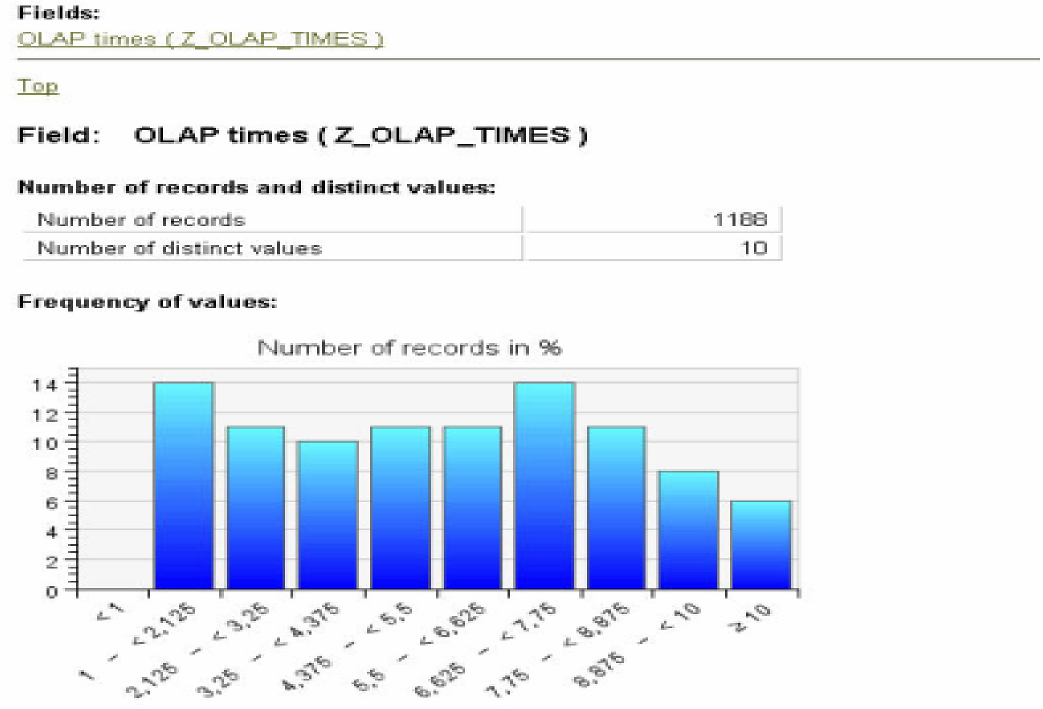
Based on the expert suggestion and to get rid of such uneven kind of data distribution the Total time OLAP is ranked so that that data records are a bit more evenly distributed that would help the clustering engine to distribute data evenly across various segments, as shown in the figure 36



```
*Descritize Total OLAP times & rank it

 if ls_source-Z_OLAP_TIMES  >= '60'.
        ls_target-Z_OLAP_TIMES = '10'.

 elseif ls_source-Z_OLAP_TIMES >= '30' and ls_source-Z_OLAP_TIMES <
     '60'.
      ls_target-Z_OLAP_TIMES = '9'.

 elseif ls_source-Z_OLAP_TIMES >= '15' and ls_source-Z_OLAP_TIMES <
     '30'.
      ls_target-Z_OLAP_TIMES = '8'.

 elseif ls_source-Z_OLAP_TIMES >= '8' and ls_source-Z_OLAP_TIMES <
     '15'.
      ls_target-Z_OLAP_TIMES = '7'.

elseif ls_source-Z_OLAP_TIMES >= '5' and ls_source-Z_OLAP_TIMES < '8'.
      ls_target-Z_OLAP_TIMES = '6'.

 elseif ls_source-Z_OLAP_TIMES >= '3' and ls_source-Z_OLAP_TIMES < '5'.
      ls_target-Z_OLAP_TIMES = '5'.

 elseif ls_source-Z_OLAP_TIMES >= '2' and ls_source-Z_OLAP_TIMES < '3'.
      ls_target-Z_OLAP_TIMES = '4'.
```

Figure 36: Data transformation - Discritizing Total OLAP

The following figure shows the basic statistics of the TOTAL OLAP times after transformation and it was suggested by the experts that K-means algorithm would better work on such data distributions to eventually come up with the meaningful patterns



Fields:

OLAP times (Z_OLAP_TIMES)

Top

Field: OLAP times (Z_OLAP_TIMES)

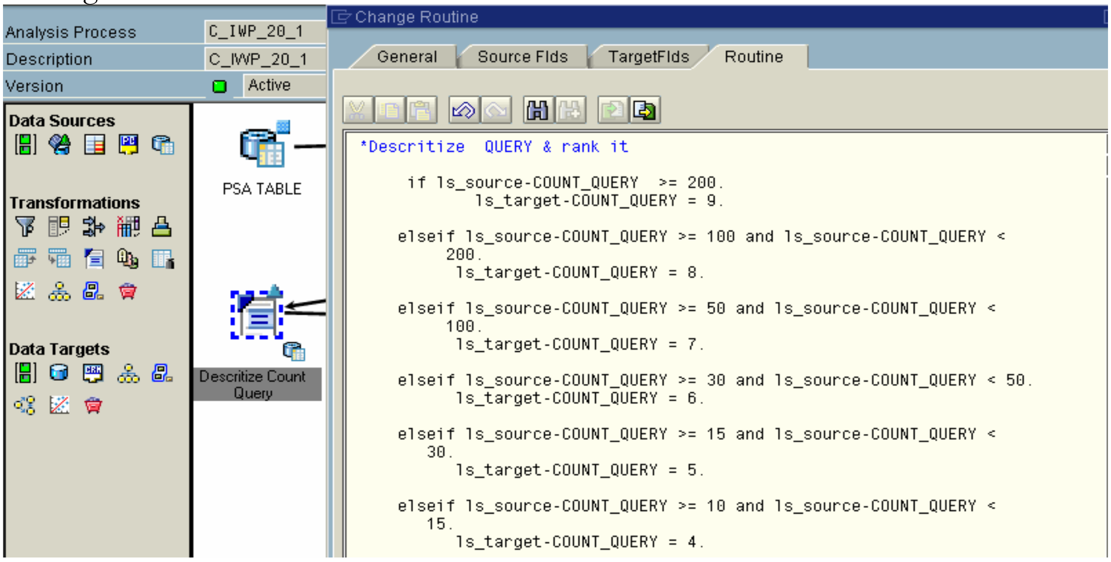
Number of records and distinct values:

Number of records	1188
Number of distinct values	10

Frequency of values:

Figure 37: Data distribution of TOTAL OLAP after transformation

In the same way the Query frequency (The number of times a Query is executed) is transformed as in the figure 38



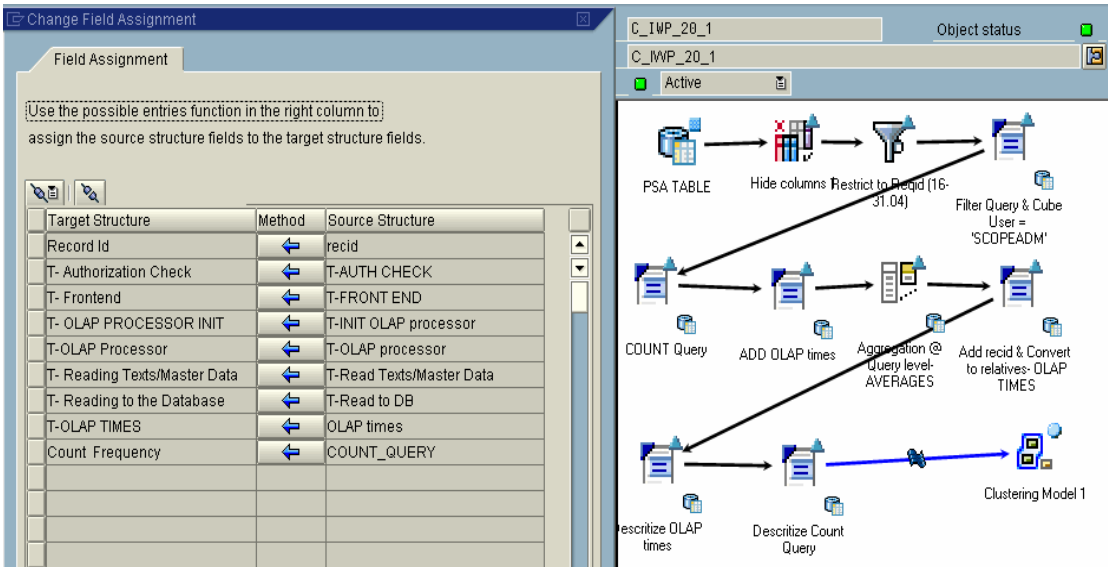
Analysis Process	C_IWP_20_1
Description	C_IWP_20_1
Version	Active

```
*Descritize  QUERY & rank it

     if ls_source-COUNT_QUERY  >= 200.
            ls_target-COUNT_QUERY = 9.

    elseif ls_source-COUNT_QUERY >= 100 and ls_source-COUNT_QUERY <
         200.
          ls_target-COUNT_QUERY = 8.

    elseif ls_source-COUNT_QUERY >= 50 and ls_source-COUNT_QUERY <
         100.
          ls_target-COUNT_QUERY = 7.

    elseif ls_source-COUNT_QUERY >= 30 and ls_source-COUNT_QUERY < 50.
          ls_target-COUNT_QUERY = 6.

    elseif ls_source-COUNT_QUERY >= 15 and ls_source-COUNT_QUERY <
       30.
          ls_target-COUNT_QUERY = 5.

    elseif ls_source-COUNT_QUERY >= 10 and ls_source-COUNT_QUERY <
       15.
          ls_target-COUNT_QUERY = 4.
```

Figure 38: Discritization Query frequencies

6.5.3. Mapping the Modal attributes

In this step the modal attributes (The Data model) is mapped to the attributes of the data source table which is depicted in figure 39



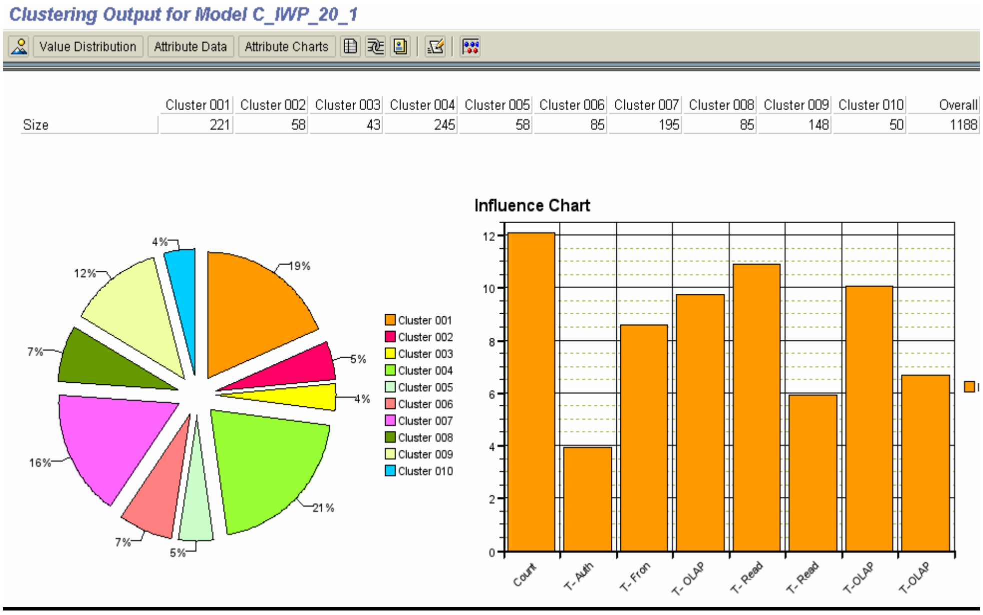
Use the possible entries function in the right column to assign the source structure fields to the target structure fields.

Target Structure	Method	Source Structure
Record Id	⇐	recid
T- Authorization Check	⇐	T-AUTH CHECK
T- Frontend	⇐	T-FRONT END
T- OLAP PROCESSOR INIT	⇐	T-INIT OLAP processor
T-OLAP Processor	⇐	T-OLAP processor
T- Reading Texts/Master Data	⇐	T-Read Texts/Master Data
T- Reading to the Database	⇐	T-Read to DB
T-OLAP TIMES	⇐	OLAP times
Count Frequency	⇐	COUNT_QUERY

Figure 39: Mapping the Modal attributes

6.6. Results of the cluster analysis

In this step the results of the clusters are analysed looking at the features of various clusters, The following figure shows the Influence of attributes which represents the relative importance of every attribute considered for clustering in the formation of clusters. The higher the index, higher is the influence in deciding which cluster an entity would get assigned to.



	Cluster 001	Cluster 002	Cluster 003	Cluster 004	Cluster 005	Cluster 006	Cluster 007	Cluster 008	Cluster 009	Cluster 010	Overall
Size	221	58	43	245	58	85	195	85	148	50	1188

Figure 40: Cluster Analysis - The influence chart

Analysis of Cluster segments

Cluster 1 - Contributes 19% to the data set This Cluster is characterized by queries that are frequently executed with High Total OLAP time when we look at the reason for the high Total

OLAP time by analysing the corresponding OLAP times, the time reading to the data base is characterised by this cluster.

The Administrator could quickly come to a conclusion that the queries in this cluster have high time reading to the database as well these are the queries that are more frequently executed and some follow up actions could be taken to reduce the time.

The following figure 41 and 42 gives a picture of the attributes namely Query frequency, Time reading to the database, total time OLAP the details of all the screens are illustrated in APPENDIX-1

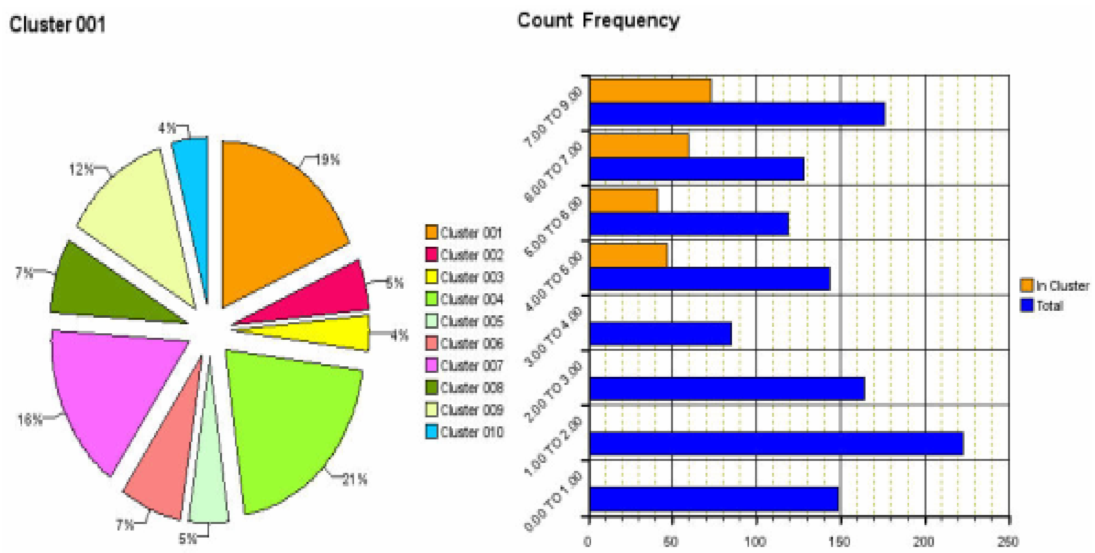


Figure 41: Cluster analysis - Results of cluster 1

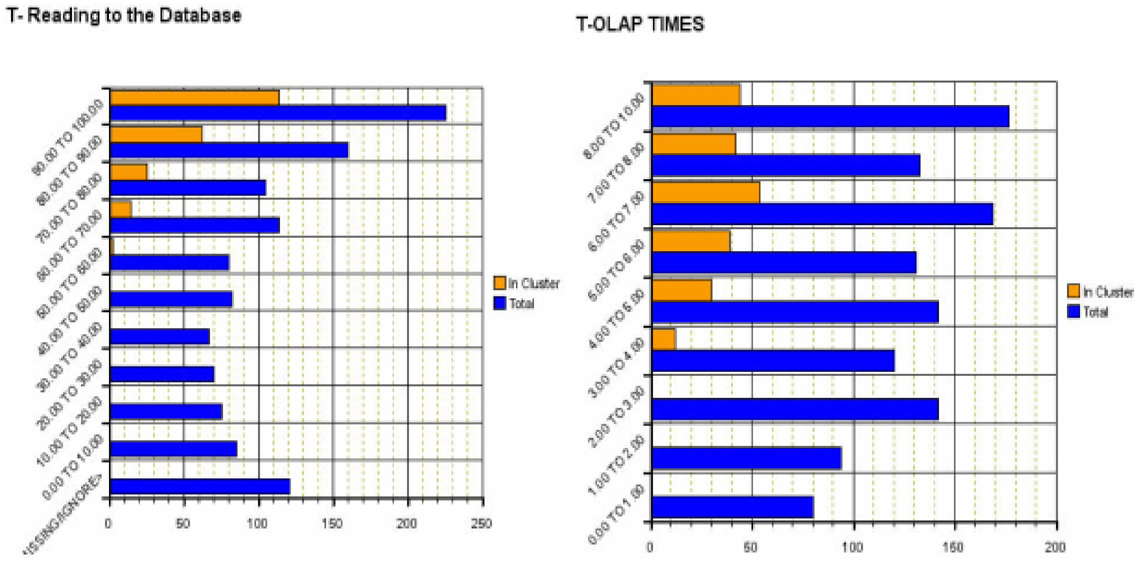


Figure 42: Cluster analysis - Results of cluster 1

Cluster 4 – This is a large segment of 22% and is characterized by queries that are less frequently executed with high Total OLAP time. The reason being time read to the data base as we have seen for the cluster one, but the interesting aspect is that this cluster contains the queries that are not so frequently executed when compared to cluster 1.

The quick impression for the administrator, could be this cluster is less important when compared to cluster 1

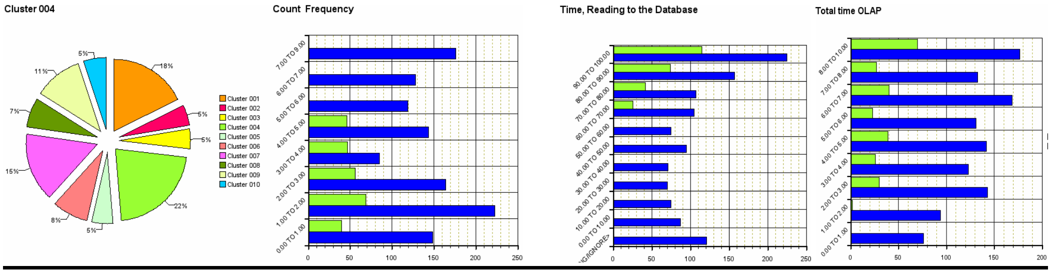


Figure 43: Cluster analysis - Results of cluster 4

Cluster 5 – This is a small segment of 5% which is characterized by the Queries with high TOTAL OLAP TIME and frequently executed. This reason being the High front-end times.

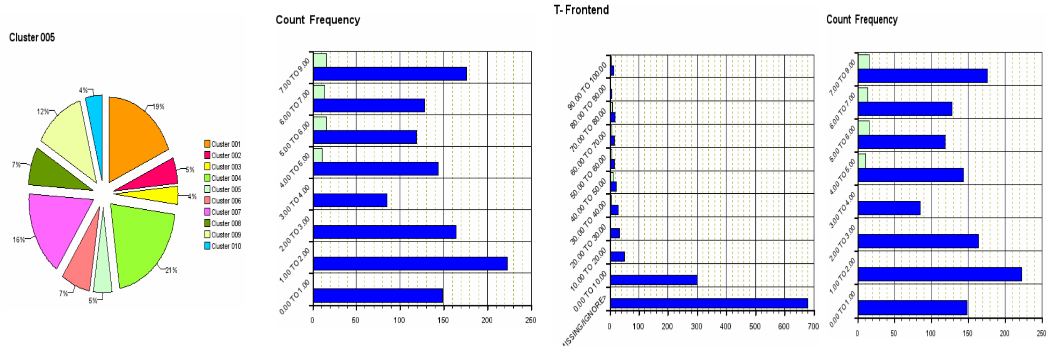


Figure 44: Cluster analysis - Results of cluster 5

Cluster 10 – This is a small segment of 5% to the total dataset and is characterized by the Queries with high TOTAL OLAP TIME and less frequently executed when we look at the corresponding key figures its due to High front-end times.

The administrator quickly comes up with questions –

- Why the users need to push so huge amount of data to the front end?
- Are these queries based on financial info providers which generally have high amounts of data?
- Try to find the user patterns like casual user, Information consumer or Analyst?

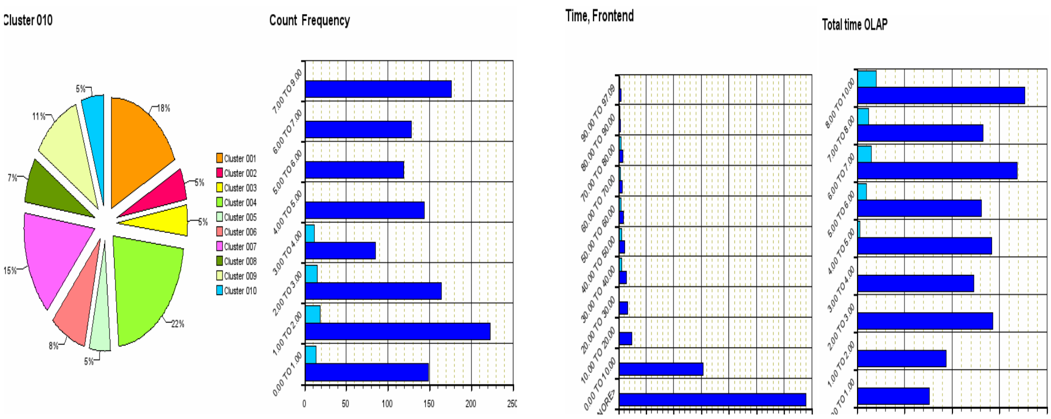


Figure 45: Cluster analysis - Results of cluster 10

Note: All the remaining screens are documented in APPENDIX-B

7. Conclusion and outlook

Data Mining provides many different techniques to extract "knowledge" from data. It is an exiting multidisciplinary field of research which has many extremely useful applications. At present the techniques are becoming more commonly used but have not been applied in all areas. As it has been shown, businesses will use data mining for a variety of applications (Here, the main objective being system monitoring and performance that would eventually lead to find out some interesting patterns form the System data). But primarily, the focus of data mining is to find useful trends in existing data. Companies can use Data mining to seek out changes in existing trends or, perhaps more importantly, discover new trends once unknown because of the huge task of analyzing large sums of data. As it's shown that this area of application for Data mining (for System performance and stability) is an emerging area as companies try to bundle the new features for their products.

Coming to the TO-BE part of this work, Using cluster analysis - The results help in analyzing the Query information from various OLAP times and it's possible to derive some strategies to optimize the query performance in future. The process of Data preparation - Cleaning, transformation and integration of the selected data plays a vital role to come up with the meaningful patterns and most of the time is devoted to this part. Coming to the Clustering scenario that has been discussed hee, the Business Meta data of the Queries, Info providers and users could be joined to further analyse the results and a kind of consensus has been reached in this regard by the people at SAP, to further investigate in this area. The interesting part of this work is that, these results will be implemented with BI-NetWaever 2006 as technical content Queries (The rules regarding the amount of data set, and the transformations and so on will be hard coded). As a result, the SAP BW administrator executes these queries and some peculiar clusters will be thrown out to the front end for further analysis for e.g. the clusters which are characterised by High total OLAP time correspondingly with high front end times, the queries that are frequently executed and so on.

The final analysis with respective to the cluster analysis method could be summed up as "advanced clustering algorithms could be useful to consider various data types and the number of clusters". To this effect the same data set is used for analysis with the IBM intelligent miner, which is equipped with sophisticated clustering algorithms (namely of type Demographic and Neural). The demographic clustering algorithm is used for analysis based on the expert suggestions, which accounted for much more meaningful patterns and these results are documented as part of APPENDIX-A. This has been accepted by the experts in SAP. To sum up it does make sense to further investigate regarding the possibility and feasibility to develop such kind of algorithms OR to tie up with external Data Mining vendors and integrate their products with the Work bench.

Bibliography

Monographs

[Berry and Linoff 1997]
Michael Berry and Gordon Linoff. Data Mining Techniques: For Marketing, Sales and Customer Support. New York: Wiley Computer Publishing, 1997.

[Delvin 1997]
B. Devlin: Data Warehouse from Architecture to Implementation, Addison-Wesley, 1997.

[FU-CH-2003]
Biao Fu, Henry Fu :A Step-to-Step Guide to SAP –Business Information Warehouse, Addison - Wesley,2003.

[Frawley 1992]
William Frawley, Gregory Piatetsky-Shapiro, Christopher Matheus. "Knowledge Discovery in Databases: An Overview." AI Magazine, Fall 1992, 213-228.

[Fayyad 1996]
Usama Fayyad et al. Advances in Knowledge Discovery and Data Mining. Cambridge: MIT Press, 1996.

[Groth 1998]
Robert Groth. Data Mining: A Hands-on Approach for Business Professionals. Upper Saddle River, New Jersey: Prentice Hall PTR, 1998.

[Han and Kamber 2001]
Jiawei Han and Michelle Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2001.

[Hand et al. 2004]
David Hand, Heikki Mannila, and Padhraic Smyth. Principles of data mining. MIT press, Cambridge, 2004.

[Inmon 1999]
Inmon, W.H.: SAP and Data Warehousing. Kiva Productions, 1999

[Kimball 1996]
Kimball, R.: The Data Warehouse Toolkit. Second Edition, John Wiley, 1996

[McDonald et al 2003]
Kevin McDonald, Andreas Wilmsmeier, David C. Dixon, W.H.Inmon: Mastering SAP Business Information Warehouse, Wiley Publishing Inc., 2003.

[Moxon 1996]
Bruce Moxon "Defining Data Mining, The Hows and Whys of Data Mining, and How It Differs From Other Analytical Techniques" Online Addition of DBMS Data Warehouse Supplement, August 1996.

[Müller and Lernke 2003]

Johann-Adolf Müller and Frank Lemke. Self-Organising Data Mining: Extracting Knowledge From Data. Victoria, British Columbia, Canada: Trafford Publishing, 2003.

[Rud 2001]
Olivia Rud. Data Mining Cookbook: Modelling Data for Marketing, Risk, and Customer Relationship Management. New York: Wiley Computer Publishing, 2001.

[SPSS 2004]
SPSS. Clementine 7.0 User's guide, 2004.

[Witten and Frank 2000]
Ian Witten and Frank Eibe. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. San Francisco: Morgan Kaufmann Publishers, 2000.

[Zaki and Ho 2000]
Mohammed Zaki and Ching-Tien Ho. Large-Scale Parallel Data Mining. Berlin: Springer, 2000.

Internet sources

Dictionaries

[Techtarget 2005]
"Data," TechTarget.
http://searchstorage.techtarget.com/sDefinition/0,,sid5_gci211894,00.html(12.01.2005).

[Princeton 2005]
"Data," Princeton.
http://www.cogsci.princeton.edu/cgi-bin/webwn2.0?stage=1&word=data (30.11.2004).

[Witnessminer 2005]
"KDD," Witnessminer.
http://www.witnessminer.com/kdd_definition.htm (06.01.2005).

Articles and other internet resources

[Albion 2005]
ALBION RESEARCH LTD. "Market Basket Analysis".
http://www.albionresearch.com/data_mining/market_basket.htm (15.03.2005)

[B.Inmon-2005]
http://www.billinmon.com//library/articles/ (10.03.2005)

[Bao 2005]
Ho Tu Bao. "Knowledge engineering: Knowledge discovery and data mining techniques and practice".
http://www.netnam.vn/unescocourse/knowlegde/knowlegd.htm (25.01.05)

[Chapple 2005]

Mike Chappel. “Data Mining: An Introduction”.
http://databases.about.com/library/weekly/aa100700a.htm (20.01.05)

[CRISP, 2005]
CRISP (Cross Industry Standard Process for Data Mining).
http://www.crisp-dm.org/Process/index.htm (06.01.2005)

[Dastani 2005]
Parsis Dastani “Data Mining – An Introduction“.
http://www.data-mining.com/miningmining.htm (25.01.05)

[FILETEK 2005]
FILETEK, The Future of Data Warehousing: Alternative Storage by Bill Inmon
http://www.filetek.com/papers/Inmon/inmon.htm (10.03.2005)

[Goransson 2005]
Olof Goransson. “Market Basket Analysis”.
http://www.megaputer.com/products/pa/algorithms/ba.php3 (15.12.2004)

[GSE&IS 2005]
GSE&IS. “Applied Categorical & Nonnormal Data Analysis--Multinomial Logistic Regression Models”.
http://www.gseis.ucla.edu/courses/ed231c/notes3/mlogit1.html (20.12.2004)

[IDUG 2005]
IDUG, Data Warehouse Administration
http://www.idug.org/idug/member/journal/mar98/faceoff.html (15.03.2005)

[ITTOOLBOX2005]
ITTOOLBOX, Dormant Data
http://businessintelligence.ittoolbox.com/documents/document.asp?i=2236 (15.03.2005)

[Kumar and Joshi 2004]
Vipin Kumar and Mahesh Joshi. “Tutorial on High Performance Data Mining”.
http://www-users.cs.umn.edu/ (06.01.2005)

[Lidal and Dingsoyr 2005]
Endre Lidal and Torgeir Dingsoyr. “An Evaluation of Data Mining Methods and Tools”.
http://www.idi.ntnu.no/~dingsoyr/project/report.html#SECTION00710000000000000000
(31.12.2003)

[Palace 2005]
Bill Palace. “Data Mining”.
http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/index.htm
(12.01.2005)

[SU 2003]
http://www.datawarehousing.com/whatis.asp (10.03.2005)

[SAP-2003]

http://help.sap.com/bestpractices/industries/businessintelligence/v131/documentation/DataWarehousing_tec_EN.pdf (11.04.2005)

[SAPPRE 2004]
SAPNET, "Analysis Process Designer "
https://websmp202.sap-ag.de/~form/sapnet?_SHORTKEY=01100035870000161446& (15.03.2005)

[**SAPHELP 2005-1]**
http://help.sap.com/saphelp_bw30b/helpdata/en/e3/e60138fede083de10000009b38f8cf/frameset.htm (10.03.2005)

SAPNET 2005-1]
http://service.sap.com/~form/sapnet?_SHORTKEY=01100035870000471520&

[SAPNET 2005-2]
http://service.sap.com/~form/sapnet?_SHORTKEY=01100035870000453136&

[SAPNET 2005-3]
http://service.sap.com/~form/sapnet?_SHORTKEY=01100035870000471520&

[STATSOFT 2005]
STATSOFT. "Cluster Analysis".
http://www.statsoftinc.com/textbook/stcluan.html (15.03.2005)

[SAPHELP 2005-1]
SAP HELP PORTAL, Technical content.
http://help.sap.com/saphelp_nw04/helpdata/en/e3/e60138fede083de10000009b38f8cf/frameset.htm

[SAPHELP 2005-2]
SAP HELP PORTAL, Technical content.
http://help.sap.com/saphelp_nw04/helpdata/en/e3/e60138fede083de10000009b38f8cf/frameset.htm

[SAPHELP 2005-3]
SAP HELP PORTAL, BW statistics.
http://help.sap.com/saphelp_nw04/helpdata/en/f2/e81c3b85e6e939e10000000a11402f/content.htm

[SAPHELP 2005-4]
SAP HELP PORTAL, Administrator workbench
http://help.sap.com/saphelp_nw04/helpdata/en/a8/6b023b6069d22ee10000000a11402f/content.htm

[SAPDOCS 2005]
SAPNET, "Data Mining and APD in SAP BW 3.5"
http://service.sap.com/~form/sapnet?_SHORTKEY=01100035870000585703& (10.03.2005)

[**Thearling 2005**]
Kurt Thearling. "An Introduction to Data Mining: Discovering hidden value in your data warehouse".
http://databases.about.com/gi/dynamic/offsite.htm?site=http%3A%2F%2Fwww.thearling.com%2Ftext%2Fdmwhite%2Fdmwhite.htm (12.01.2005)

[UCLA 2005]
UCLA ACADEMIC TECHNOLOGY SERVICES. "Multinomial Logistic Regression, Contrived Examples".

http://www.ats.ucla.edu/stat/stata/code/odds_ratio_mlogit.htm (25.12.2004)

[Whitehead 2005]
John Whitehead. “An Introduction to Logistic Regression”.
http://personal.ecu.edu/whiteheadj/data/logit/ (23.12.2004)

[Wilson 2005]
Bill Wilson. “Induction of Decision Trees”.
http://www.cse.unsw.edu.au/~billw/cs9414/notes/ml/06prop/id3/id3.html (15.03.2005)

[W.H.Inmon 1999]
http://www.billinmon.com//library/articles/dwdef.asp

[Zadok and Stolfo 2005]
Erez Zadok, Salvatore Stolfo. “Data Mining Methods for Detection of New Malicious Executable”.
http://www1.cs.columbia.edu/ (30.12.2004)

[BW310 2005]
Coursematerial, BW310 Data Warehousing, SAP AG-2005.

[BW305 2005]
Course material, BW305 BI Warehouse - Reporting and Analysis, SAP AG-2005.

[SAP NET- Course Material BW310]
BW310-Data Warehousing 2003

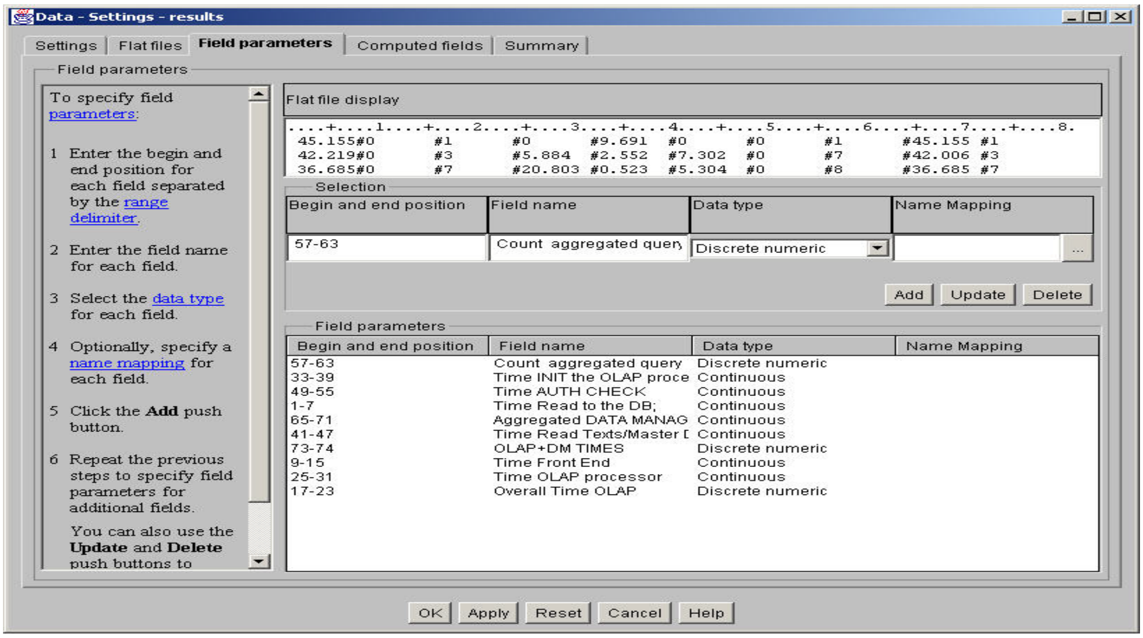
[Lesley 2004]
Clem Lesley: A Presentation on Data Mining with SAP BW 3.5. SAPNET.

[TABW30 2003]
Business Information Warehouse - Extraction and Special Topics, TABW30, Section 2, Unit 3, SAP AG, 2005

APPENDIX-A

IBM Intelligent Miner Cluster Analysis

1. **Data Selection -** The data extracted in to a flat file and loaded in to the intelligent miner



To specify field parameters:

1 Enter the begin and end position for each field separated by the range delimiter.

2 Enter the field name for each field.

3 Select the data type for each field.

4 Optionally, specify a name mapping for each field.

5 Click the **Add** push button.

6 Repeat the previous steps to specify field parameters for additional fields.

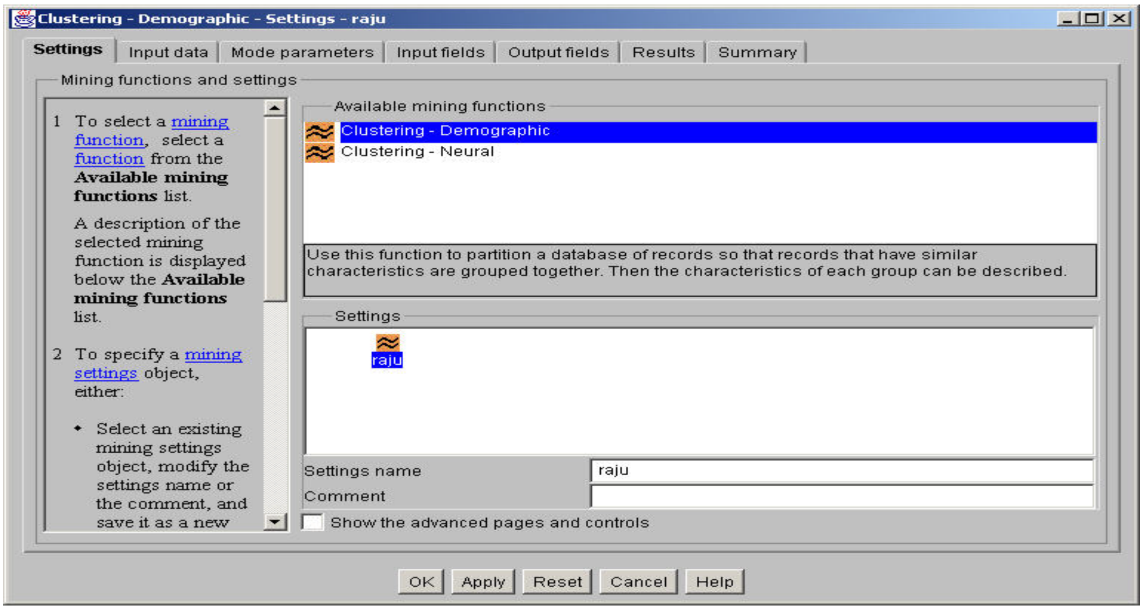
You can also use the **Update** and **Delete** push buttons to

```
....+....1....+....2....+....3....+....4....+....5....+....6....+....7....+....8.
 45.155#0      #1      #0      #9.691  #0      #0      #1      #45.155 #1
 42.219#0      #3      #5.884  #2.552  #7.302  #0      #7      #42.006 #3
 36.685#0      #7      #20.803 #0.523  #5.304  #0      #8      #36.685 #7
```

Begin and end position	Field name	Data type	Name Mapping
57-63	Count aggregated quer)	Discrete numeric	

Begin and end position	Field name	Data type	Name Mapping
57-63	Count aggregated query	Discrete numeric	
33-39	Time INIT the OLAP proce	Continuous	
49-55	Time AUTH CHECK	Continuous	
1-7	Time Read to the DB;	Continuous	
65-71	Aggregated DATA MANAG	Continuous	
41-47	Time Read Texts/Master [	Continuous	
73-74	OLAP+DM TIMES	Discrete numeric	
9-15	Time Front End	Continuous	
25-31	Time OLAP processor	Continuous	
17-23	Overall Time OLAP	Discrete numeric	

2. **Model Selection**



1 To select a mining function, select a function from the **Available mining functions** list.

A description of the selected mining function is displayed below the **Available mining functions** list.

2 To specify a mining settings object, either:

- Select an existing mining settings object, modify the settings name or the comment, and save it as a new

Available mining functions:

- Clustering - Demographic
- Clustering - Neural

Use this function to partition a database of records so that records that have similar characteristics are grouped together. Then the characteristics of each group can be described.

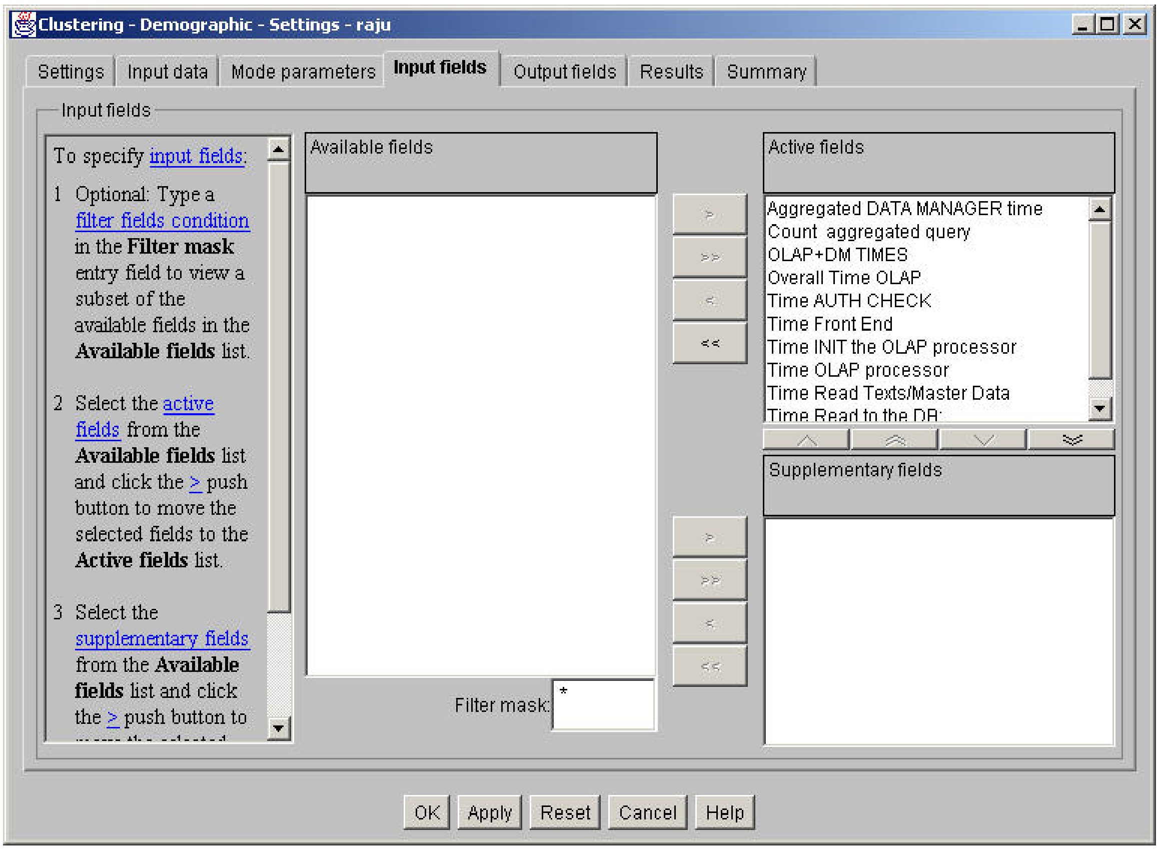
Settings: raju

Settings name: raju

Comment:

☐ Show the advanced pages and controls

3. **Attributes for analysis**



To specify input fields:

1 Optional: Type a filter fields condition in the **Filter mask** entry field to view a subset of the available fields in the **Available fields** list.

2 Select the active fields from the **Available fields** list and click the > push button to move the selected fields to the **Active fields** list.

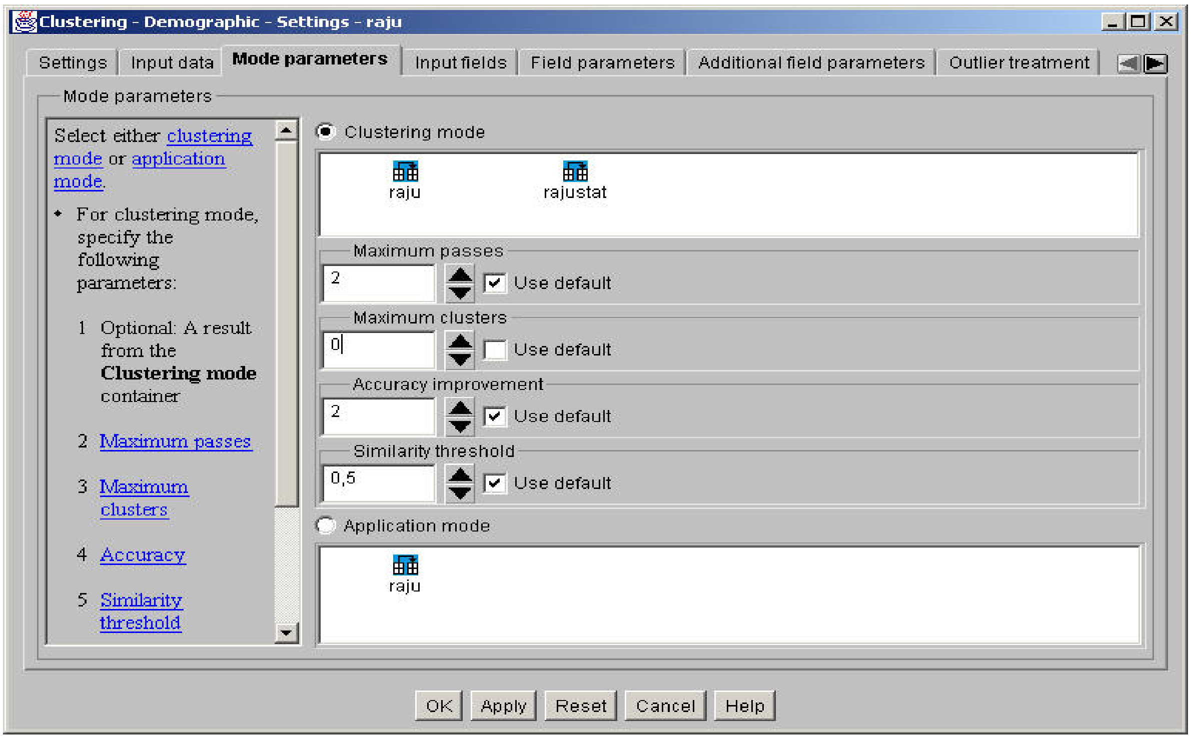
3 Select the supplementary fields from the **Available fields** list and click the > push button to

Filter mask: *

Active fields:

Aggregated DATA MANAGER time
Count aggregated query
OLAP+DM TIMES
Overall Time OLAP
Time AUTH CHECK
Time Front End
Time INIT the OLAP processor
Time OLAP processor
Time Read Texts/Master Data
Time Read to the DB:

4. Modal Parameters



Select either clustering mode or application mode.

- For clustering mode, specify the following parameters:

1 Optional: A result from the **Clustering mode** container

2 Maximum passes

3 Maximum clusters

4 Accuracy

5 Similarity threshold

● Clustering mode

Maximum passes: 2 ☑ Use default

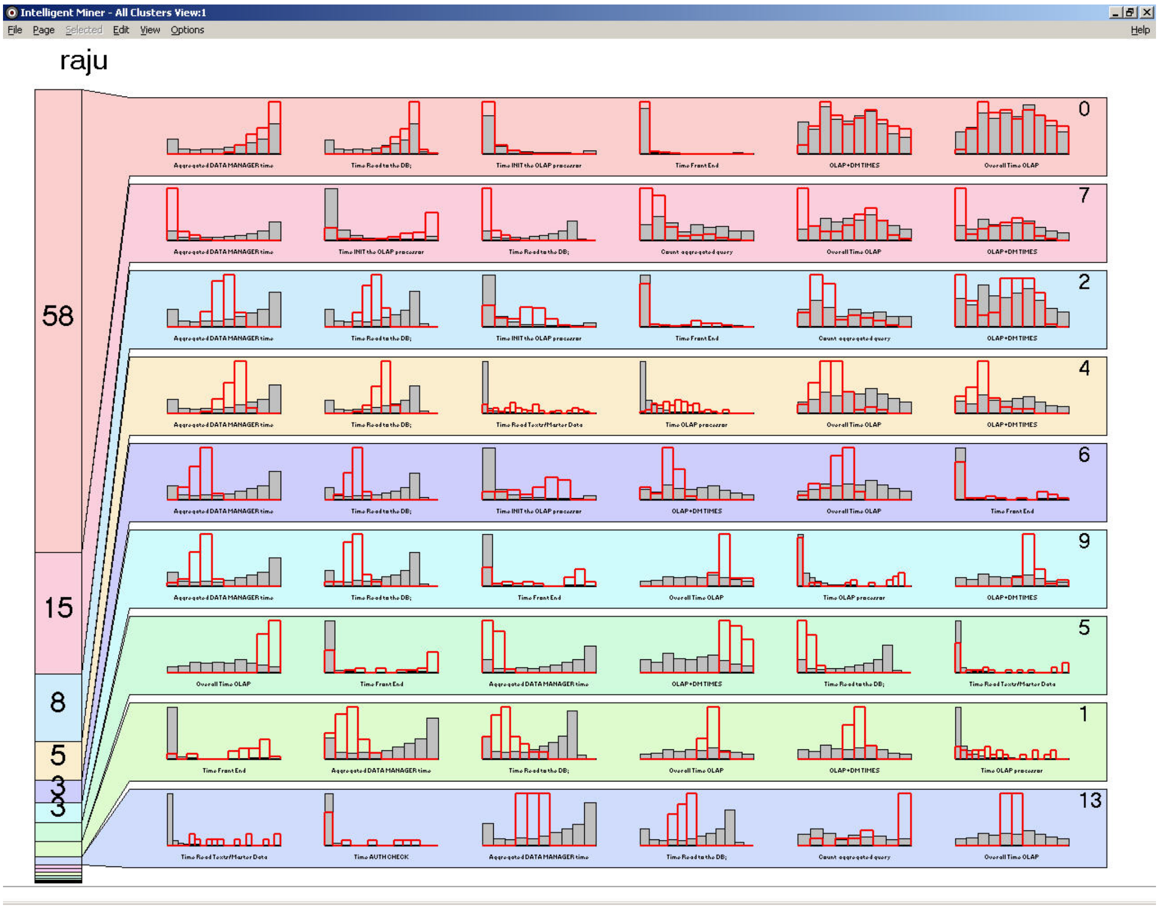
Maximum clusters: 0 □ Use default

Accuracy improvement: 2 ☑ Use default

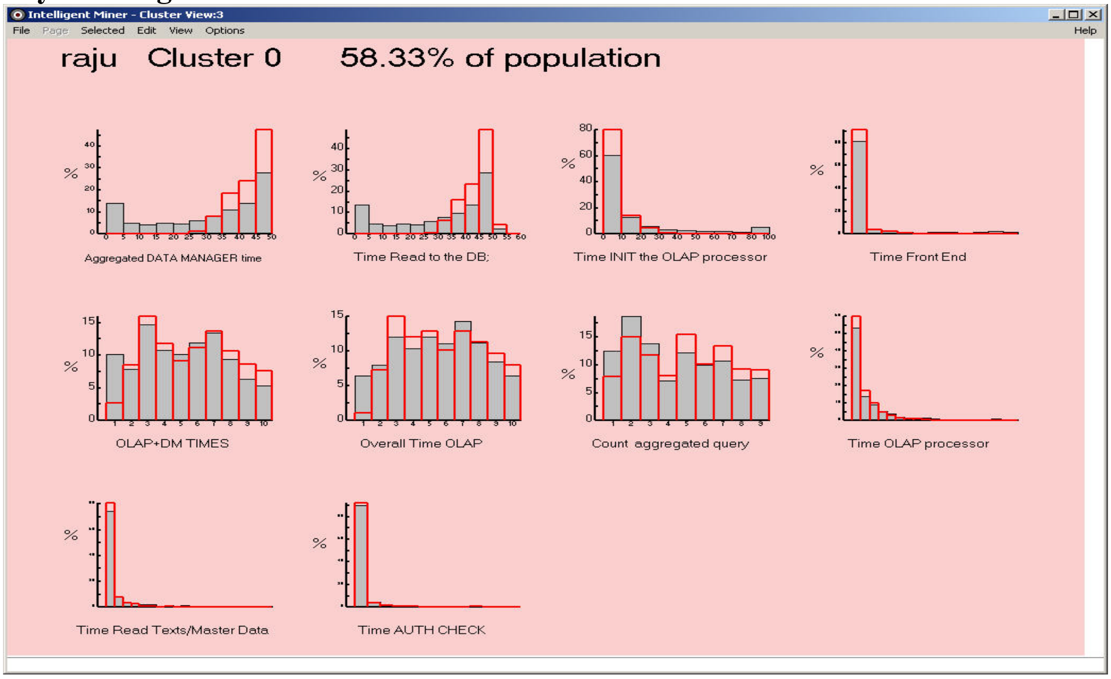
Similarity threshold: 0,5 ☑ Use default

○ Application mode

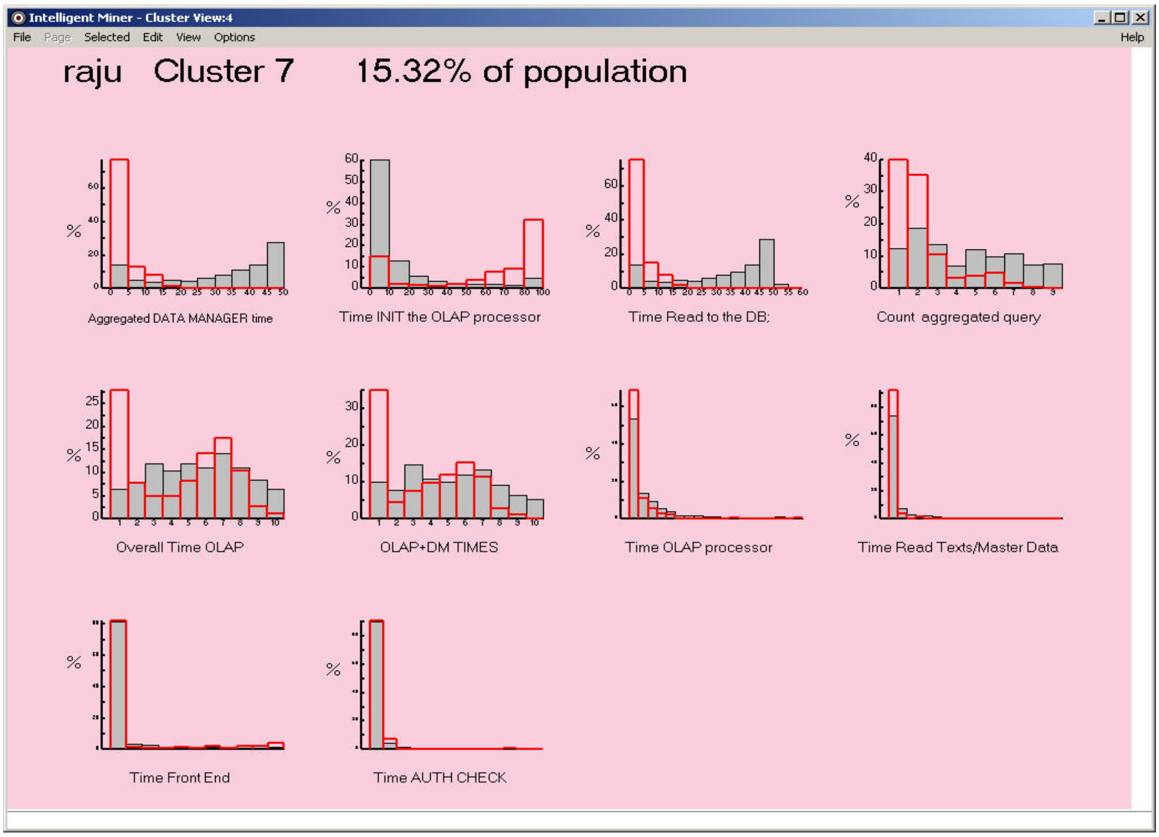
5. View of all clusters



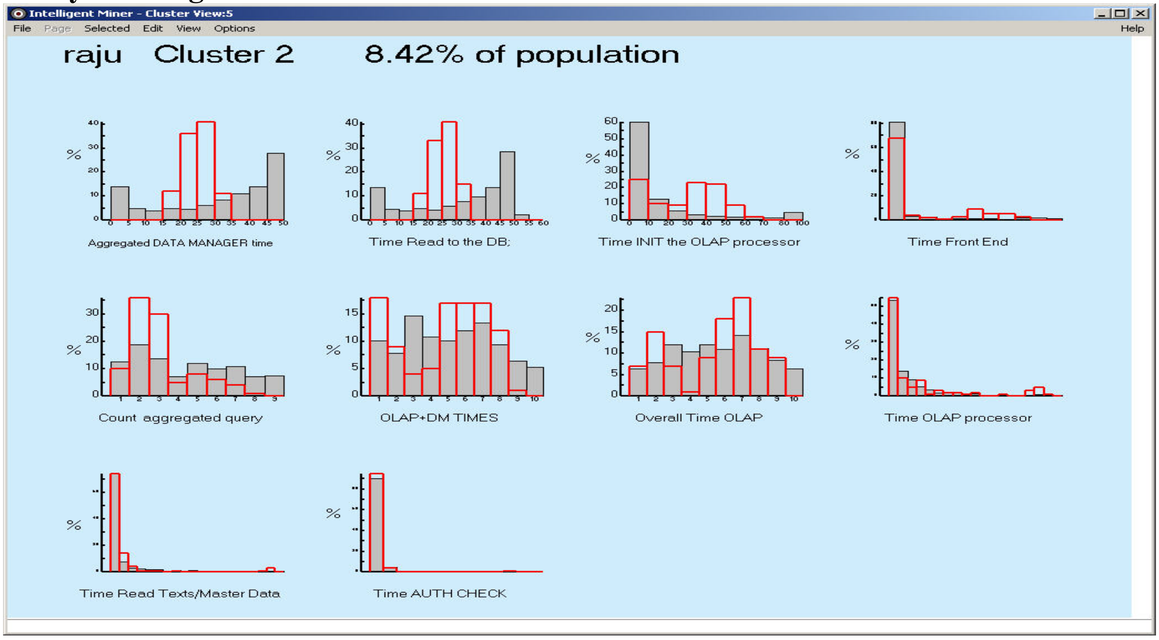
6. Analysis of segments



7. Analysis of segments



8. Analysis of segments



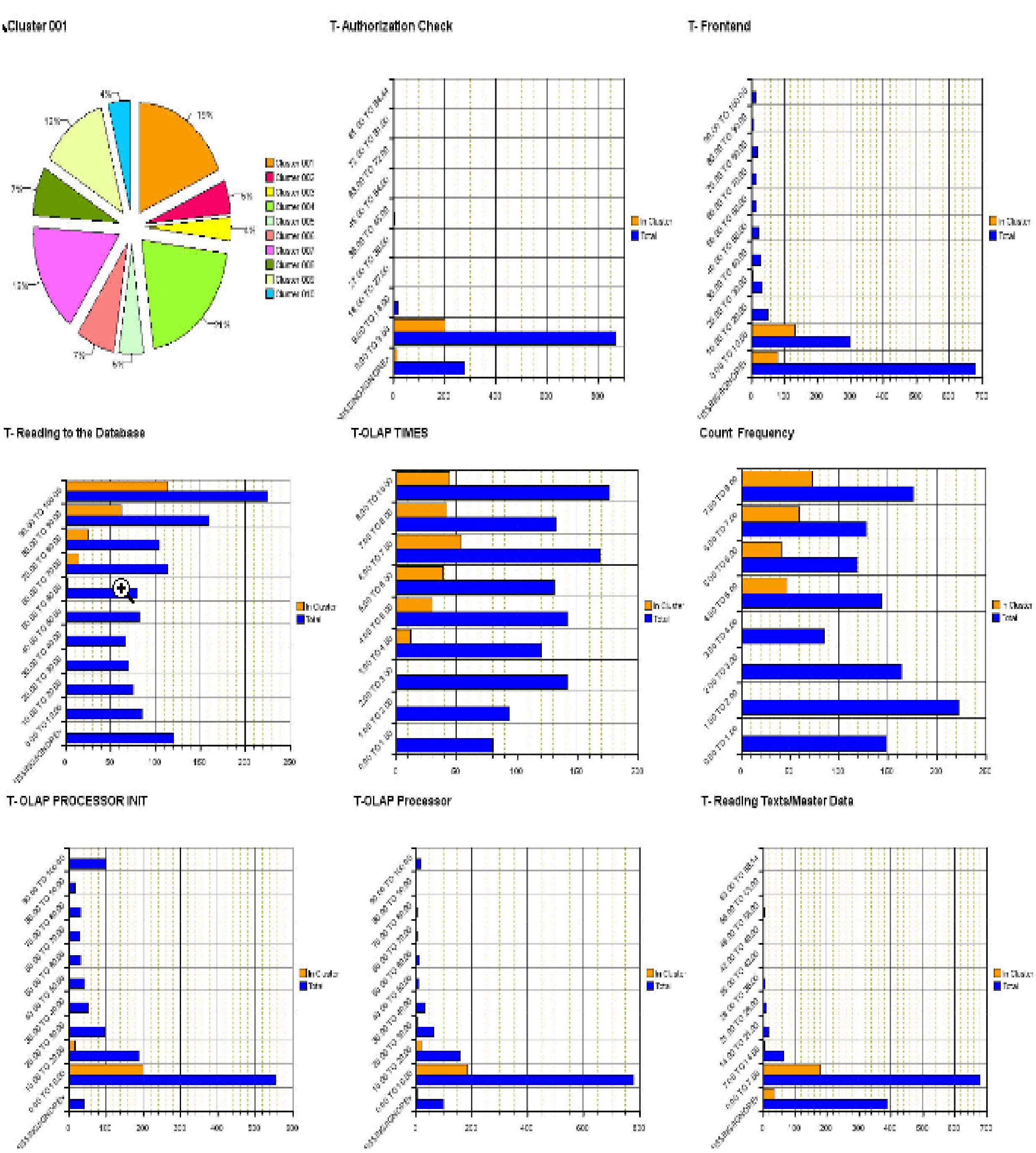
To summarize, the results clearly shows that this demographic algorithm takes into consideration data sets more precisely and the number of clusters are justified by the algorithm based on the data distribution. A big cluster with 58% data set which consists of similar data, but all the remaining clusters shows valuable patterns and this has been judged by the experts at SAP.

APPENDIX-B

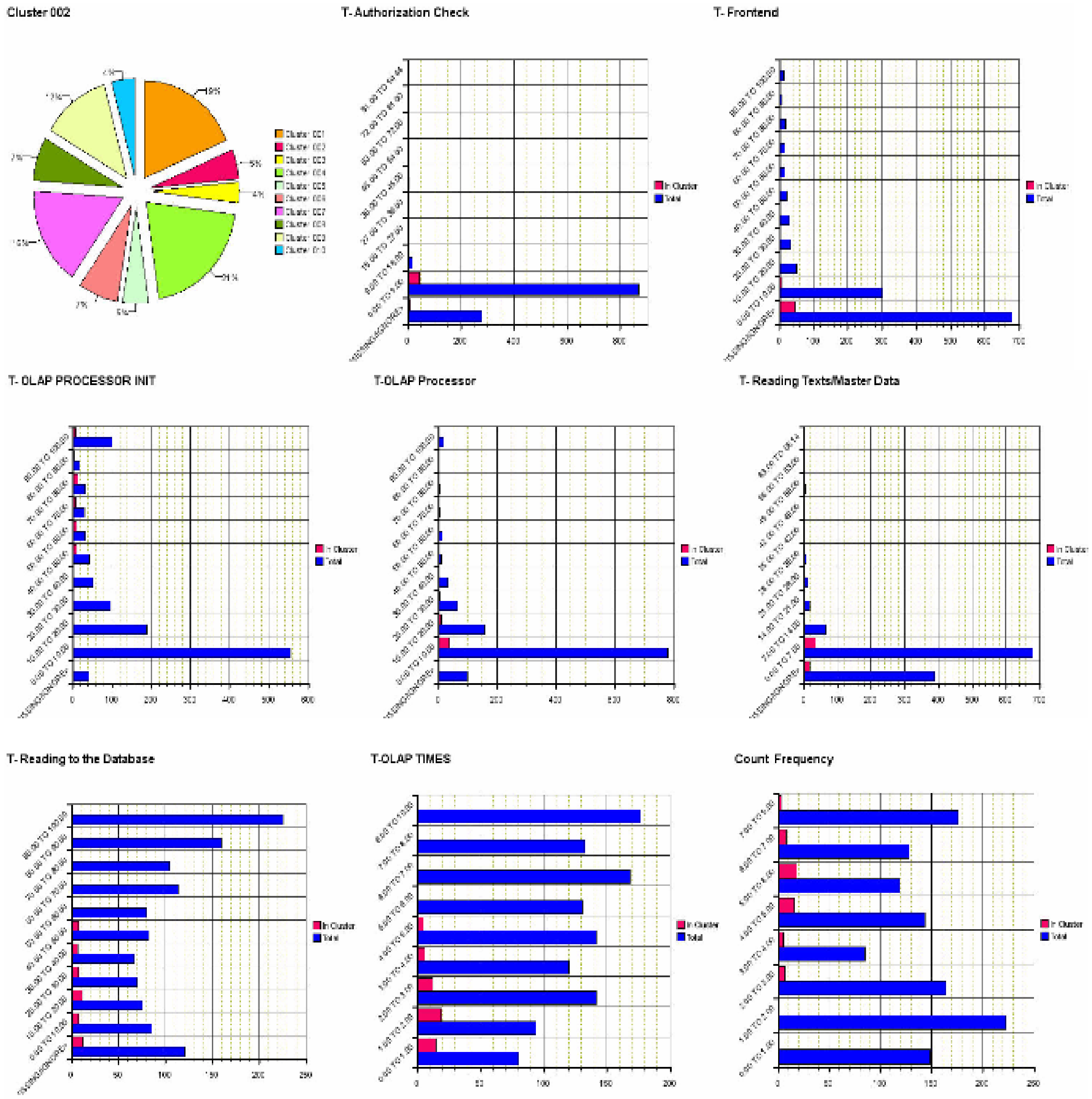
Results of SAP Cluster Analysis

As a sample, the following screens represent the five clusters and the values distribution details.

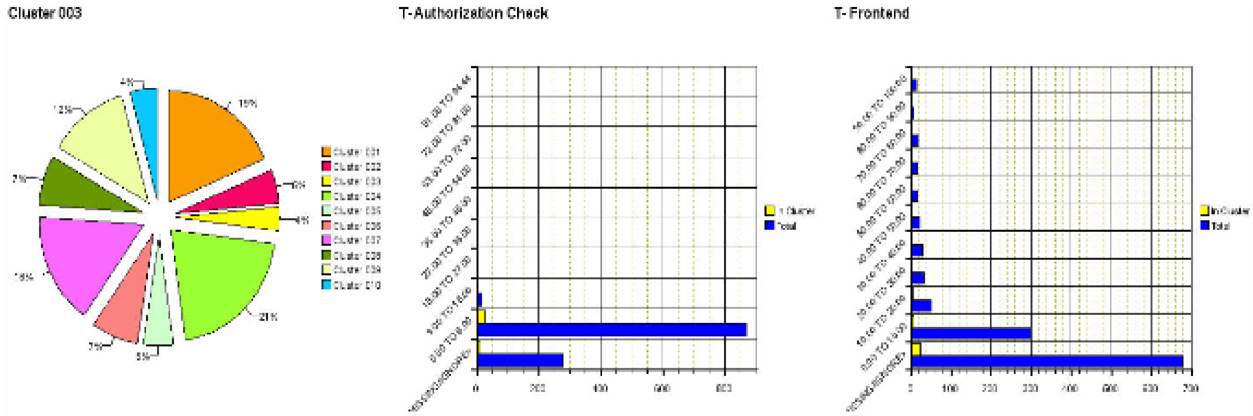
1. **Cluster 1**



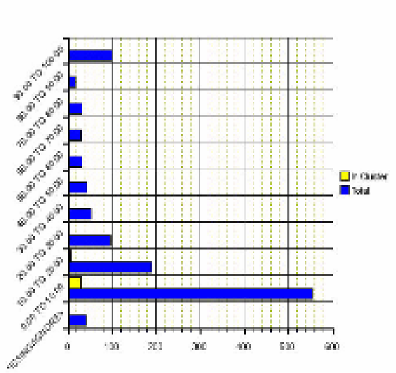
2. **Cluster 2**


Cluster 002
19%
5%
4%
21%
5%
7%
15%
7%
12%
4%
Cluster 001
Cluster 002
Cluster 003
Cluster 004
Cluster 005
Cluster 006
Cluster 007
Cluster 008
Cluster 009
Cluster 010
T- Authorization Check
In Cluster
Total
0
200
400
600
800
T- Frontend
In Cluster
Total
0
100
200
300
400
500
600
700
T- OLAP PROCESSOR INIT
In Cluster
Total
0
100
200
300
400
500
600
T-OLAP Processor
In Cluster
Total
0
200
400
600
800
T- Reading Texts/Master Data
In Cluster
Total
0
100
200
300
400
500
600
700
T- Reading to the Database
In Cluster
Total
0
50
100
150
200
250
T-OLAP TIMES
In Cluster
Total
0
50
100
150
200
Count Frequency
In Cluster
Total
0
50
100
150
200
250

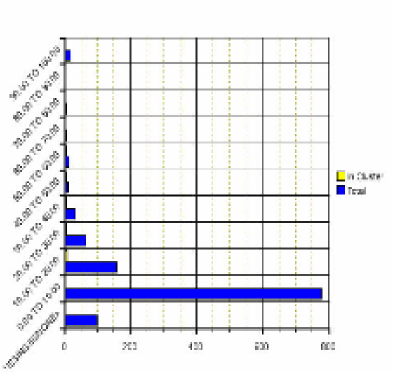
3. Cluster 3


Cluster 003
19%
6%
4%
21%
5%
7%
16%
7%
12%
4%
Cluster 001
Cluster 002
Cluster 003
Cluster 004
Cluster 005
Cluster 006
Cluster 007
Cluster 008
Cluster 009
Cluster 010
T- Authorization Check
In Cluster
Total
0
200
400
600
800
T- Frontend
In Cluster
Total
0
100
200
300
400
500
600
700

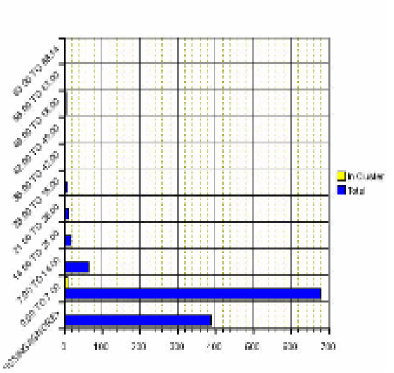
T- OLAP PROCESSOR INIT



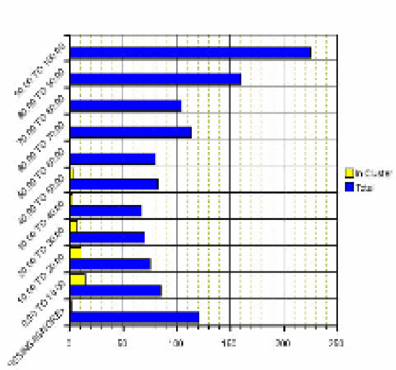
T-OLAP Processor



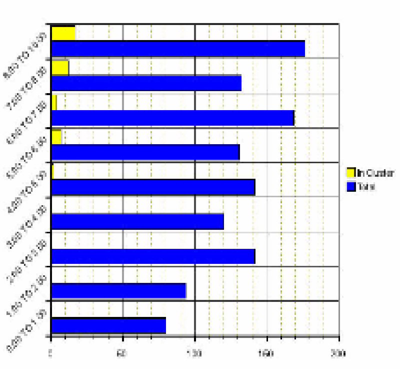
T- Reading Texts/Master Data



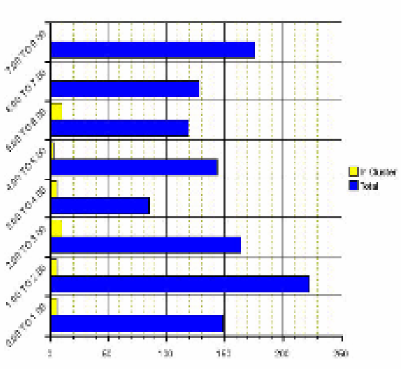
T- Reading to the Database



T-OLAP TIMES

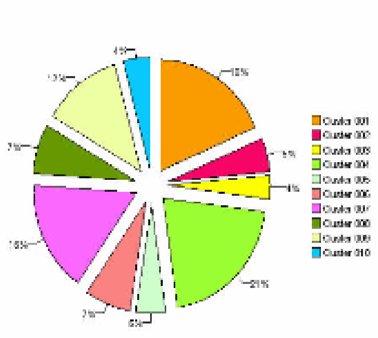


Count Frequency

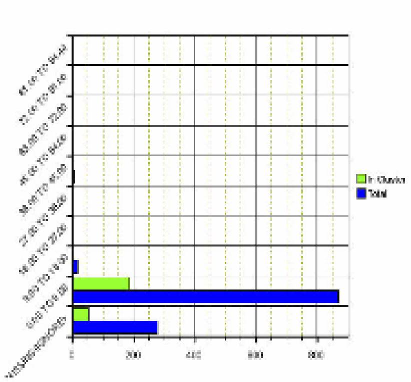


4. Cluster 4

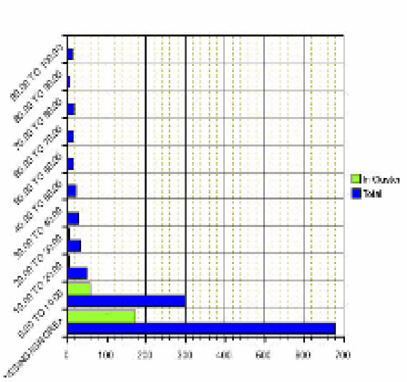
Cluster 004



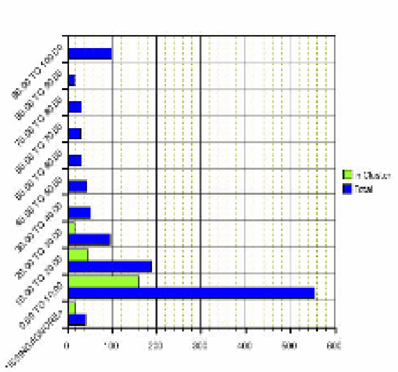
T- Authorization Check



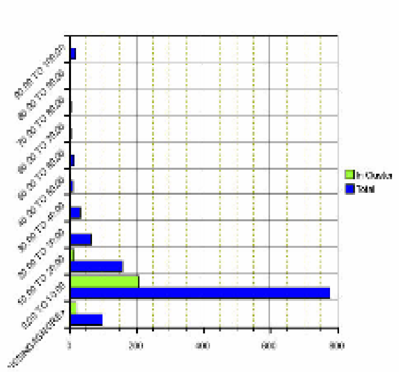
T- Frontend



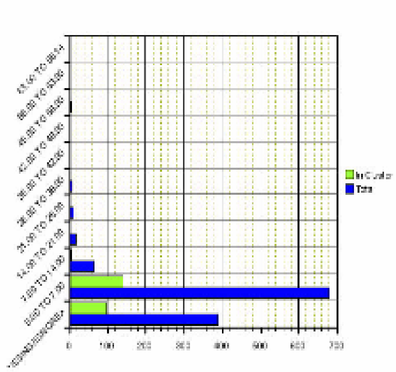
T- OLAP PROCESSOR INIT

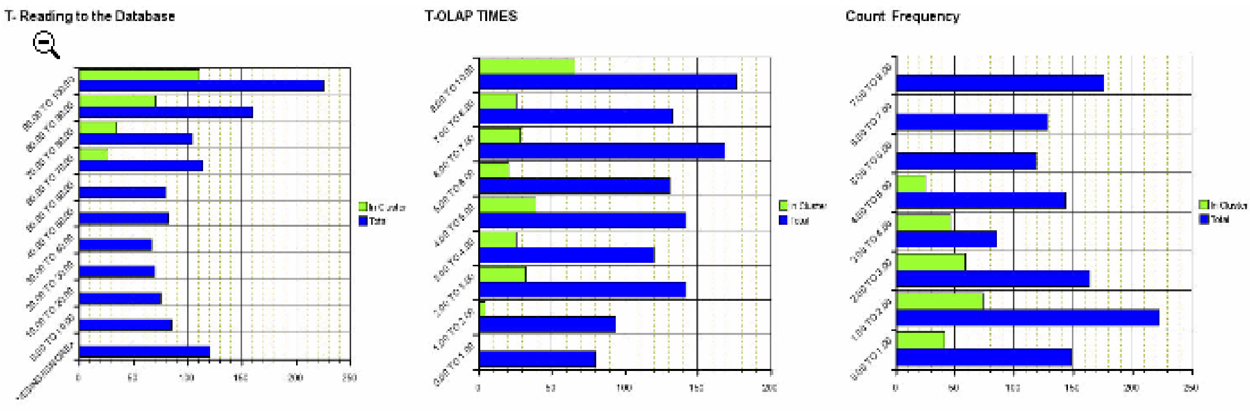


T-OLAP Processor

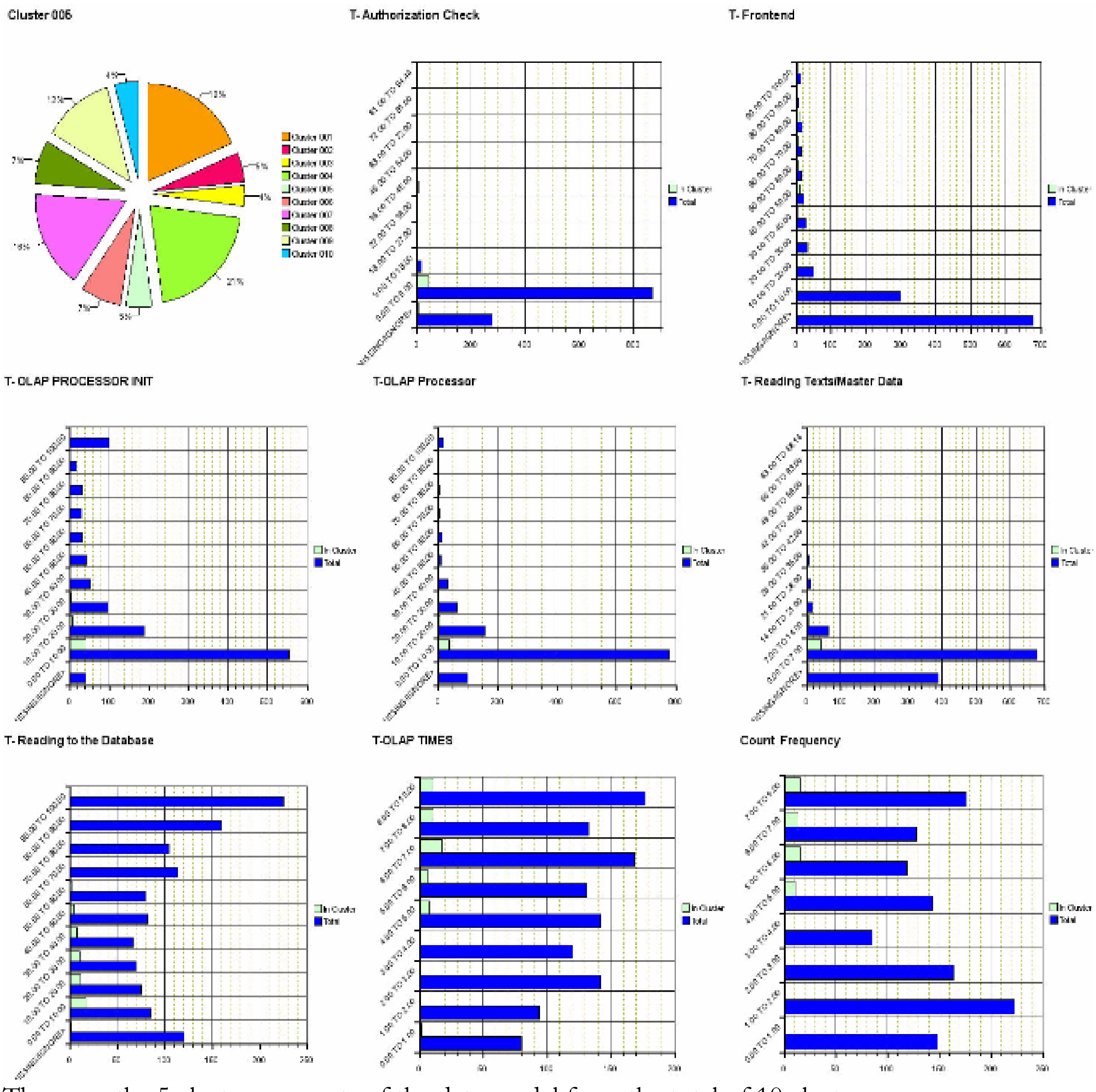


T- Reading Texts/Master Data





5. Cluster 5



These are the 5 cluster segments of the data model from the total of 10 clusters.

APPENDIX-C – Related Internet Links

General information's about Data Mining

http://www.the-data-mine.com
http://www.dmreview.com
http://www.datawarehousingonline.com/
http://datawarehouse.ittoolbox.com/
http://www.kdnuggets.com
http://itmanagement.earthweb.com/datbus/
http://www.thearling.com/index.htm#wps

Data mining software providers

Advanced Software Applications http://www.asacorp.com/
AIS Visual http://www.visualmine.com/
Alice http://www.alice-soft.com
Angoss http://www.angoss.com/
Assoc http://www.asoc.de
Attar Software http://www.attar.com/
Bissantz & Company http://www.bissantz.de/
Business Objects http://www.businessobjects.com/
Cogit http://www.cogit.com/
Cognos http://www.cognos.com/
Data Distilleries http://www.ddi.nl/
DataMind http://www.datamindcorp.com/
DataMiner http://www.dminer.com/
Datasage http://www.datasage.com/
Dialogis http://www.dialogis.de
Dimension 5 http://www.dimension5.sk/
HNC http://www.hnc.com/
human IT http://www.humanit.de/
Hyperparallel, http://www.hyperparallel.com/
IBM http://www.ibm.com/
Information Discovery http://www.datamining.com/
Integral Solutions http://www.isl.co.uk/
Magnify http://www.magnify.com/
Management Intelligenter Technologien http://www.mitgmbh.de/
MarketMiner http://www.marketminer.com/
Mathsoft http://www.mathsoft.com/
NeoVista http://www.neovista.com
Oracle http://www.oracle.com/
Prudential Systems http://www.prudsys.de/
Quadstone http://www.quadstone.com/
Rulequest http://www.rulequest.com
SAP AG http://www.sap.com/index.epx
Salford Systems http://www.salford-systems.com/
SAS http://www.sas.com/
SGI http://www.sgi.com/software/mineset/
SLP InfoWare http://www.slp-infoware.com
SPSS http://www.spss.com/datamine/
Syllogic http://www.syllogic.nl

Tandem http://www.tandem.com/
Thinking Machines http://www.think.com/
Torrent http://www.torrent.com/
TriVida http://www.trivida.com/
Unica http://www.unica-usa.com/
Wizsoft http://www.wizsoft.com/