IDRBT's Working Paper No. 4

An Approach to Establish Data Warehouse For Banks in India

- P. Radha Krishna

Abstract

Operational data is the source for all business intelligence applications, but the data is typically not in the correct format to support the decision making process in a business. Further, nowadays, banks are storing more information than ever before. Decision makers must have the right information at the right time to help them make more informed and intelligent decisions. The data in the operational database represents current transactions, however the decisions are based on a different time frame; that is there is no time component. On the other hand, data in operational databases are stored with a functional or process orientation, what really decision-makers would like to have is subject orientation of data, which facilitates multiple views for data and decision making. ***Data Warehousing*** *and* ***Data Mining*** *are the right solution that makes the above possible.*

One of the main goal of implementing a data warehouse is to turn the wealth of corporate data into information that can be used in the daily decision making process. Dr. Vasudevan Committee on Technology Upgradation in the banking sector recommended that all banks should put in place their Data Warehouse strategy by January 1,2001. In this paper, we highlighted the need of data warehousing and data mining for banks and provided an approach to build the warehouse suitable to banks.

1. Introduction

The banking industry is becoming increasingly dependent on Information Technology to retain its competitiveness and adapt with the ever-evolving business environment. The industry which is essentially becoming a service industry of a higher order, has to rely on technology to keep abreast with global economy that technology has thrown open.

As far as Indian Banking scenario is concerned the Government has stipulated that banks have to computerize their operations at the earliest. Institute for Development and Research in Banking Technology (IDRBT) has undertaken the mammoth task of

---

Acknowledgement

The author would like to thank Dr. V. P. Gulati, Director, IDRBT for the encouragement and continuous support given during the present work. The author would also thank Ms. S. Subhashini and Mrs. Sangeeta Sam, Research Associates, IDRBT for rendering assistance in the preparation of this working paper.

May, 2000

networking all Indian banks under the Indian Financial Network (**INFINET**). The prospect of INFINET encompassing the Indian banking scenario gives mammoth volume of data to be handled by banks. This data could be of different forms and on different platforms

Day after Day Mountains of data is produced directly as a result of banking activities, and as a by-product of various transactions. A vast amount of information is about their customers. Yet, most of these data remains locked within archival systems that must be coupled with operational systems to generate information necessary to support strategic decision-making.

A variety of approaches for computer-aided decision-making systems have appeared over time under different terms like Management Information Systems (MIS), Executive Information Systems (EIS), and Decision Support Systems (DSS).

The term Management Information System is not new to the banking sector. Since the early 80s, banks have been using the Management Information Systems to the process of generation various reports which are used for analysis at the Corporate/Head offices for their decision making for own use as well as for conveyance to authorities in charge of regulation. Often, these reports are generated through computers and can be generated at any point of time. However, the usage of the terms **Data Warehousing** and **Data Mining** are relatively new. These terms have gained significance with the growing sophistication of the technology and the need for predictive analysis with What-if simulations.

This paper is organized as follows: Section 2 describes the need of Data Warehouse for banks. Section 3 lists the Data Warehouse & Data Mining Applications in Banking Systems. Section 4 introduces the Related Technical Concepts. Section 5 gives Recommendations for Data Warehousing and Data Mining applications in Banks given by Dr. Vasudevan Committee on "Technological Upgradation in Banks". Section 6 provides An approach to build a Data Warehouse for a Bank. Section 7,8 and 9 presents Case Studies on Data Warehouse and Data Mining. Section 10 and 11 illustrates Return on Data Warehouse Investments and Management values respectively. Section 12 lists Selected Data Warehouse Technology Vendors. Finally, section 13 concludes this paper.

2. Need of Data Warehousing and Data Mining for Banks

The development of management support systems is characterized by the cyclic up and down of buzzwords. Model based decision support and executive information systems were always restricted by the lack of consisted data. Now-a-days data warehouse tries to cover this gap by providing actual and decision relevant information to allow the control of critical success factors. A data warehouse integrates large amounts of enterprise data from multiple and independent data sources consisting of operational databases into a common repository for querying and analyzing. Data warehousing will gain critical importance in the presence of data mining and generating several types of

analytical reports which are usually not available in the original transaction processing systems.

Banking being an information intensive industry, building a Management Information System is a gigantic a task. It is more so for the public sector banks, which have a wide network of bank, branches spread all over the country. It becomes more difficult due to prevalence of varying degrees of computerization. At present, banks generate MIS reports largely from periodic paper reports/statements submitted by the branches and regional/zonal offices. Except for a few banks, which have been using technology in a big way, MIS reports are available with a substantial time tag. Reports so generated have also a high margin of error due to data entry being done at various levels and likelihood of varying interpretations at different levels.

Though computerization of bank branches has been going at a good pace, MIS requirements have not been fully addressed to. It is on account of the fact that most of the Total Branch Computerization (TBC) software packages are transaction processing oriented. In most banks large databases are in operation for normal daily transactions. In most cases, these operational databases have not been designed to store historical data or to respond to queries but simply to support all the applications for day-to-day transactions.

The present information systems evolved from the legacy of the old. They exist as collection of separate islands of information that have developed as response to certain operational needs. They have not been designed to meet the information requirement on real time basis of decision-makers cutting across departments. Due to contingent nature of the developmental process, the hardware and software platforms that have been used in these operational information systems lack compatibility. As a result whenever decision-makers information requirements have to be met by pulling out data from various operational databases, special efforts are needed to be made for collating these data. Another important consequence of the disparate nature of the existing system is the lack of subject orientation in the system. This in turn reduces the utility of the system to the decision makers. Another major shortcoming of the present system is its inability to provide consistent data for different variables for a reasonably long duration. This apart, the most critical deficiency of the present information system is the lack of information about the availability of data. In this connection, an application of *Data Warehousing* along with *Online Analytical Processing* (OLAP) and *Data Mining* techniques appears to be the appropriate solution. Further data warehouses provide a central repository for large amounts of diverse and valuable information.

3. Data Warehouse & Data Mining Applications in Banking Systems

The Warehouse infrastructure can support a wide range of applications and reports to meet exact business needs. Some of the applications for banking system are below:

Risk management

In Banking, the most important of data warehousing is building Risk Management Systems. Risk Management System will identify the risks associated with a given set of assets. This means understanding the way in which the market is likely to move in the future, based on past performance. The key measure here is volatility, but there are many others, and this is highly complex and technical area. One of the applications of Risk management is **Asset and Liability management** (ALM), and the entire system can be implemented with Data Warehousing.

Campaign Analysis

Accurately targeting customers in campaigns and promotions and analyzing their responses to promotion episodes are the keys to enabling the transition from mass marketing to mass customization. Most organizations launch many different kinds of promotional campaigns for many different products using many different media. This application enhances the organization's understanding of the entire process from selecting customers to be targeted to analyzing how they responded. Campaign Analysis allows you to measure the responsiveness to campaigns by households and by individual customers. It provides the ability to measure the effectiveness of individual campaigns and different media and offers the ability to conduct cost-benefit analyses of campaigns.

Customer Profile Analysis

Customer Profiling allows organizations to distinguish, in the mass of customers, the many microsegments that make up the whole. Increasingly, customer segmentation is forming an essential element of marketing strategy as markets become more fragmented especially where customer segments exhibit distinct and different characteristics. The profiling and segmentation of customers facilitate the building of genuine customer relationships in an era of one-to-one marketing.

Profiling customer behavior aims at extracting patterns of their activities from transactional data, and using these patterns to provide for service provisioning, trend analysis, abnormal behavior discovery, etc. It has becoming increasingly important in a variety of application domains such as fraud detection, commercial promotion etc.

Loyalty Analysis

One of the keys to profitability in any enterprise is customer loyalty. Yet so few organizations measure customer loyalty in a structured way or seek to understand the underlying causes of customer attrition. The Loyalty Analysis application allows you to measure customer loyalty from different viewpoints such as duration of relationship; range of services and products consumed; and the demographic, psycho-graphic and geographic influences on customer attrition. By itself, the Loyalty Analysis application

measures and monitors customer loyalty and facilitates the development of customer retention programs. The loyalty of customers can be assessed in the context of their value, their contact history, the segments they belong to and the transaction events that may influence their loyalty.

Customer Care Analysis

Customers interact with organizations in many ways using different touch points to initiate inquiries, provide feedback or make suggestions. This information provides valuable insight into the behavior of customers and the track records of the organizations servicing customers. The likely level of satisfaction or dissatisfaction of a customer can be determined by their customer contact history. Analyzing customer contacts is an essential ingredient in maintaining and nurturing customer relationships and preserving the loyalty of customers into the future.

Business Performance Analysis

The Business Performance Analysis application for banking exploits the industry-specific, transaction-level data in a typical retail banking enterprise. Analyzing a bank's business performance requires an understanding of customer behavior, including their usage patterns of the different services the bank offers. It facilitates analysis of product performance, branch and ATM activity and utilization. It also provides the information needed to formulate effective up-sell and cross-sell strategies. It provides the business intelligence information most needed by sales and marketing executives, as well as strategic planners in banks everywhere. The atomic data stored in this key portion of the warehouse becomes the engine that powers the entire solution and, when combined with the related applications, will revolutionize the way a bank manages and satisfies its customers.

Sales Analysis

The Sales Analysis application allows analysis of sales from a variety of viewpoints such as sales by channel, outlet or organizational unit; sales by product, product category or group; and sales by region and by season. This application offers organizations an integrated perspective on sales results and enables sales managers to understand the underlying trends and patterns in their sales data.

Profitability Analysis

In any organization, it is essential to understand profitability in order to determine pricing, award discounts, allocate resources or develop strategy. But profitability is a many-faceted concept and can be considered in the context of an organization, a channel, a product, a product category, a brand, a customer or a customer segment. And most organizations will also wish to measure gross profits, net profits and margins. In the retail

banking industry the measurement of customer profitability is key to managing the business effectively. Customer profitability is influenced by a range of factors that includes the volume and type of business conducted, the range of products purchased and the utilization by the customer of automated transaction facilities. The capability to measure and analyze profitability from many different viewpoints and by many different dimensions is the objective of this application.

4. Related Technical Concepts :

A. What is a Data Warehouse?

Data warehouse can be formally defined *as 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 come from different sources

For want of better description, in general, a data warehouse is a collection of data copied from other systems and assembled into one place. Once assembled it is made available to end users, who can use it to support a plethora of different kinds of business decision support and information collection activities.

W.H.Inmon (1993), in his landmark work Building the data Warehouse, offers the following definition of a data warehouse: "A data warehouse is a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision making process".

- **Subject Oriented:** The Data warehouse data is arranged and optimized to provide answers to questions coming from diverse functional areas within an enterprise. Therefore, the Data Warehouse contains data organized and summarized by topics such as finance, marketing etc. For each of these topics the Data Warehouse contains specific subjects of interest – customers, departments, regions, and so on.

- **Integrated:** The Data Warehouse is centralized consolidated database that integrates data derived from the entire organization. Thus the data warehouse consolidates data from multiple and diverse source with diverse formats. Data integration implies a well-organized effort to define and standardize all data elements. This integration effort can be time consuming but, once accomplished, it provides a unified view of the overall organization situation.

- **Time Variant:** In contrast to the operational data that focus on current transactions, the Warehouse Data represent the flow of data through time. In short, the Data Warehouse contains data that reflect what happened last week or last year. The Data Warehouse can even contain projected data generated through statistical and other models. It is also time-variant in the sense that once data are

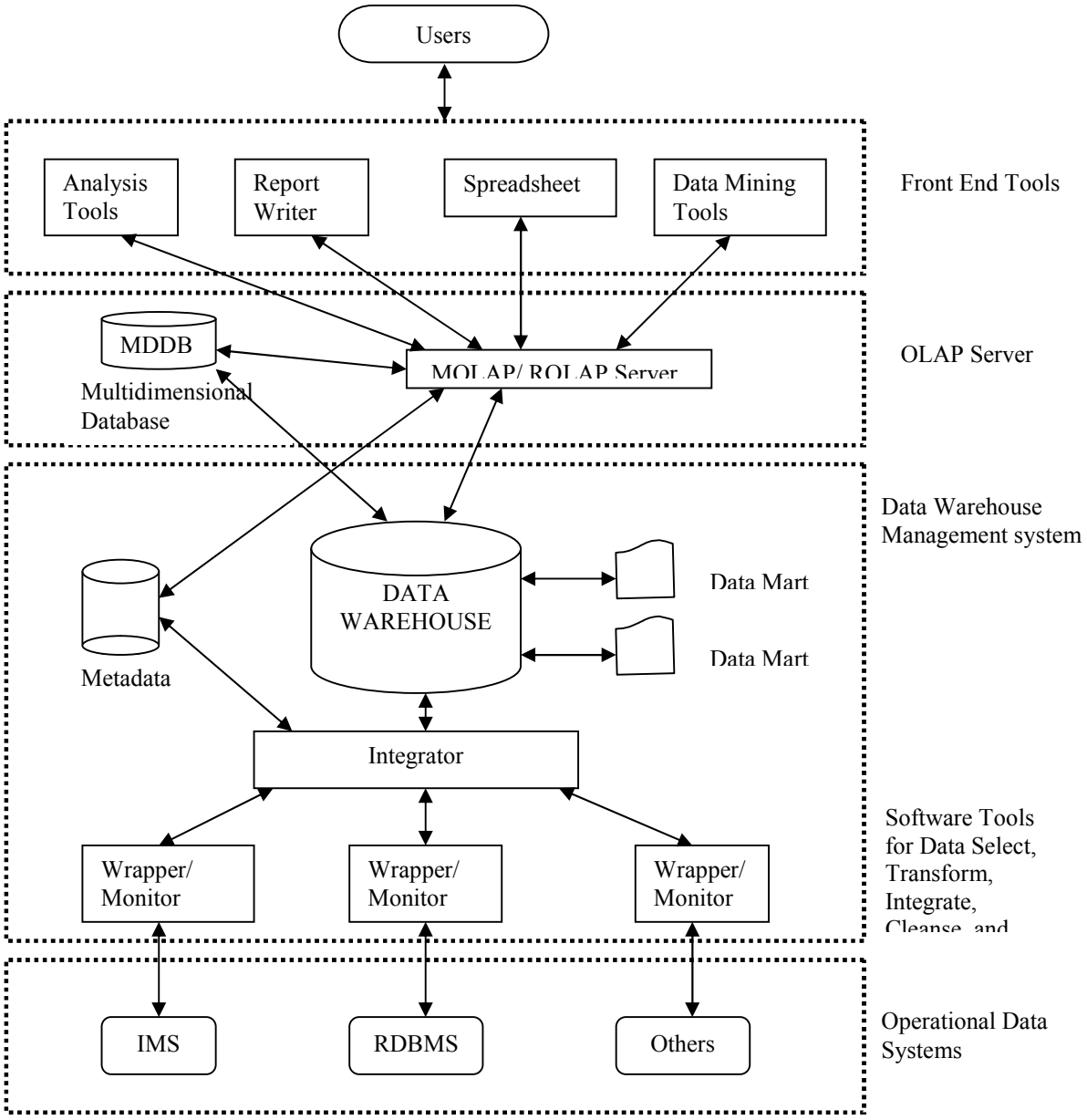
periodically uploaded to the Data Warehouse, all time-dependent aggregations are recomputed.

- **Non-volatile:** Data in a data warehouse is used only for queries. Once data enter the Data Warehouse, they never removed. Because the data in the Warehouse represent the enterprise's entire history, the operational data, representing the near-term history, are always added to it. Because data are never deleted and new data are always added, the Data Warehouse is always growing, That's why the DSS DBMS must be able to support multigigabytes and even multiterabyte size databases and multiprocessor hardware.

B. Data Warehouse Components

Data Warehouses have a distinct structure. There are different levels of summarization and detail that describe the Data Warehouse. The different components of the Data Warehouse are Current Detail Data, Older Detailed Data, Summarized Data and Meta Data. Fig. 1 shows a typical enterprise-wide data warehouse architecture.

Fig.1. Data Warehouse Architecture



i. *The Current Detailed Data*

The heart of a Data Warehouse is its *current detail*, where the bulk of data resides. Current detail comes directly from operational systems and may be stored as raw data or as aggregations of raw data. Current detail, organized by subject area, represents the entire enterprise, rather than a given application.

Current detail is the lowest level of data granularity in the Data Warehouse. Every data entity in current detail is a snapshot, at a moment in time, representing the instance

when the data are accurate. Current detail is typically two to five years old. Current detail refreshment occurs as frequently as necessary to support enterprise requirements.

ii. *The Older Detailed*

The older data is data that is frequently accessed and is stored as a level of detail consistent with current detailed data. While not mandatory that it be stored on an alternate storage medium, because of the anticipated large volume of data coupled with infrequent access of the data, the storage medium for older data is usually removable storage such as automatic tape library.

iii. *Summarized* (or *Aggregated*) data

The data in data warehouse is stored in the summarized form (in contrast to transactional data in operational systems), so that long term management analysis of data can be performed reliably, repeatedly and quickly. The summarized data is generally classified as Lightly summarized data and Highly summarized data

Lightly summarized data are the hallmark of a Data Warehouse. All enterprise elements (department, region, function, etc.) do not have the same information requirements, so effective Data Warehouse design provides for customized, lightly summarized data for every enterprise element (see Data Mart, below). An enterprise element may have access to both detailed and summarized data, but there will be much less than the total stored in current detail.

Highly summarized data are primarily for enterprise executives. Highly summarized data can come from either the lightly summarized data used by enterprise elements or from current detail. Data volume at this level is much less than other levels and represents an eclectic collection supporting a wide variety of needs and interests. In addition to access to highly summarized data, executives also have the capability of accessing increasing levels of detail through a "drill down" process.

iv. *Meta Data*

In simple terms, Metadata is data about data. That is, the data contains characteristics and definition of the data in the warehouse, as well as metrics. Metadata allows for easier data access and analysis and provides a means of documentation in the event of an audit. This concept is so powerful that it can bring tremendous value to an organization. It describes the structure, content, keys, indexes etc of data. Metadata has several roles and used in a data warehouse environment. Metadata contains informational data about the creation, management and usage of the data warehouse. For business users, metadata can provide information about the data in the data warehouse (such as what it means, how to access it, and when it was last updated) as well as information about reports, spreadsheets and queries related to the data. For data

warehouse administrators, metadata is involved in all aspects of their job; Meta data defines, documents, and drives the processes of the data warehouse.

In a data warehouse environment, there are two types of metadata – *Technical* and *Business*. Technical metadata describes data elements, as they exist in the source systems, the data warehouse and data marts, and the interim data staging areas. For example, technical meta data could include the technology definitions for operational data in DB2, Oracle or SQL Server databases.

Technical metadata also includes specifications on how the data is extracted, transformed, cleansed, and aggregated at each stage and the schedules for the data warehouse processes that accomplish this. This meta data is used by data warehouse administrators, power users, and the tools that drive the processes of the data warehouse.

In contrast, business metadata is used by business users and by decision support tools. The information is related to the technical metadata, but the presentation is very different. Business metadata provides a subject-oriented view of data. It describes data objects like databases, tables, and columns, as well as informational objects like queries, charts and reports. The metadata also contains the business dimensions, hierarchies, and formulas needed by business users to simplify their query and data navigation, and support more in-depth analysis.

Like technical metadata, business metadata includes information about transformations, aggregations, and schedules. However, all business metadata is given in business terms rather than technology terms. Business metadata should provide business users all the information needed to understand, locate, and use the data in the data warehouse in a way that fits naturally with their data analysis tasks.

Because technical and business metadata are very distinct in content and use, separate metadata stores, sometimes called dictionaries or repositories, are often used with each being optimized for its particular audience. In these cases, it is key that there is an open architect interchange mechanism and flow of metadata, in both directions, between the technical and business repositories. This mechanism can also be used to import metadata about data and information objects from available sources such as relational databases, CASE tools, modeling tools, repositories, and decision support tools.

The ability to interchange metadata is key for enabling single entry of metadata and metadata synchronization. The repositories must be open, allowing extensibility of metadata objects and attributes and providing public interfaces to access and maintain metadata content. The use of technical and business repositories does not eliminate other meta stores. Some tools will continue to have their own metadata store for flexibility and performance. However, these repositories greatly simplify the tasks of data warehouse administrators and business users, and enhance their productivity.

C. Data Mart

Data Mart is a subset of the enterprise-wide data warehouse. Data Marts are highly focussed sets of information that are designed in the same way as Data warehouses, but are implemented to address the specific needs of a defined set users who share common characteristics. For example, the west region may only care about data that pertains to them and would like to work with that geographic subset. The marketing department might only care about customer, product, and sales data, and wants that subject subset. A group in the Marketing Department needs to do data analysis that is best supported by a dimensional structure, so a data mart with the data structure and summaries appropriate for multi-dimensional analysis is set up for them.

In some cases, the size of the enterprise might result in a very expensive and requires years to build. Therefore, some companies have started building the Data Marts first, so that they can benefit much sooner from Data warehouse. For reasons such as lower costs and risks, the data marting solution is gaining in popularity.

Data marts fall into two broad categories: 1.*Subset data marts* created from a parent data warehouse or parent data mart. 2.*Incremental data marts* used as independent information resources or as data warehouse building blocks. These categories reflect the two approaches to data warehousing that have evolved Top-Down (subset) and *Bottom-Up* (incremental).

In the classic ***top-down approach,*** an enterprise of galactic data warehouse is constructed and populated. Subset data marts are then created by taking portions of the enterprise data warehouse and creating information resources to serve specific user groups with homogenous characteristics or needs. The incremental ***bottom-up approach*** to data warehousing uses incremental data marts as the building blocks of the enterprise data warehouse. Individual incremental data marts are created and deployed. They are used to test and perfect the methodologies, processes, and tools used in the creation of the corporate information resources. As the data marts prove themselves to be valuable corporate resources, the organization can justify the time the expense associated with the enterprise data warehouse.

Each strategy has its own set of merits and demerits, and each should only be used where appropriate. Although each is viable in a suitable site, the misapplication of strategy can doom a project from start.

D. Data Warehouse design approaches

There are two popular design schemas: Star schema and Snowflake schema.

Star schema

Star schema contains a denormalized central fact table for the subject area and multiple dimension tables for descriptive information about the subject's dimensions. The fact table can contain many millions of rows. Commonly accessed information is often preaggregated and summarized to further improve performance. While the Star schema is primarily considered a tool for the database administrator to increase performance and simplify data warehouse design, it also represents data warehouse information in a way that makes better sense to end users. There are two disadvantages for Star schema:

- Denormalization schema may require too much disk storage
- Very large dimension tables can adversely affect performance, partially offsetting benefits gained through aggregation.

Snowflake schema

Dimension tables in the snowflake schema are normalized. To understand how disk storage can be efficiently utilized, consider the deposit database in which there are 200,000 transactions rolling upto 20 branches. In a star schema, the corresponding

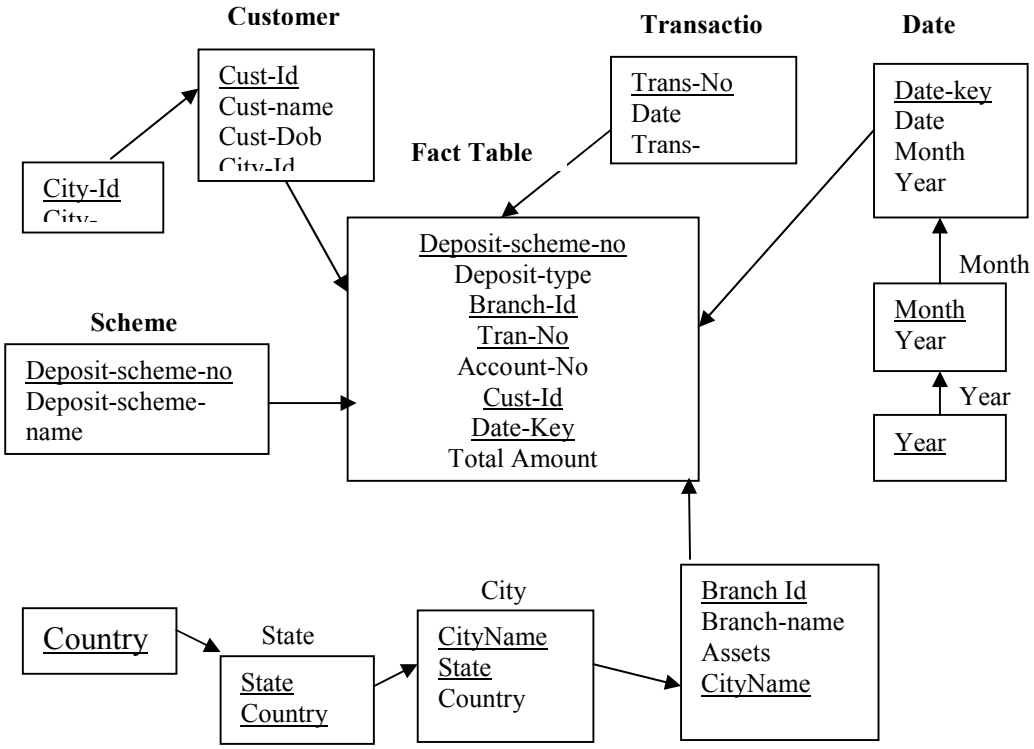


Fig.2 Sample Snowflake schema for Deposit

dimensional schema would have 200,000 rows and each row would store all of the relevant information for every level of hierarchy above or equal to its own level. In this case every kilobyte of attribute data elements costs 100 megabytes of disk space. Normalizing this dimension table avoids the additional disk storage. The main disadvantage of the snowflake (versus the star) is the relative complexity of the normalized snowflake data structure. In addition, the overall maintenance becomes more difficult to manage as the data model complex. Fig.2 shows a Snowflake schema for a sample Deposit DW with one fact table and dimension tables representing Customer, Branch, Scheme, Transaction, Time hierarchies.

E. On Line Analytical Processing (OLAP)

OLAP is a software technology to extract data out of transaction processing systems and turning it into information. OLAP provides fast, consistent, interactive access to a wide variety of possible views of information reflecting the real dimensionality of the enterprise as understood by the user.

OLAP's main purpose in an organization is to empower users with the ability to review historical data for the purpose of reporting, aggregating, trending, summarizing, averaging and graphing. It is used to summarize, consolidate, view, apply formulae to, and synthesize data according to multiple dimensions. Traditionally, a relational approach (relational OLAP) has been taken to build such systems. Relational databases are used to build and query these systems.

A data warehouse is a repository from which data are gathered from internal and external sources. The storing of data is designed in such a way that it meets the needs of analytical processing. This is the reason behind the explosive growth in data warehousing, as banks seek a way of providing staff with access to information across all delivery channels in a timely manner. For example to create and update customer behavior profiles, hundreds of millions of call records must be processed everyday. Therefore a scalable infrastructure to support filtering, mining and analyzing massive transaction data continuously is needed which ultimately supports such profiling. Such infrastructure provides maximum flexibility to drill down, drill up, drill across, pivoting along with analytical tools for extracting business intelligence from data. Also such an infrastructure can be developed with data warehousing and OLAP technology. The OLAP server can be used for analyzing patterns in multiple dimensions and at multiple levels, and comparing patterns similarity. From a performance point of view, it supports indexed catching, reduces database access dramatically and extends main memory based reasoning. From a functional point of view, it allows us to deliver powerful solutions for profiling, pattern generation, analysis and comparison, in a simple and flexible way.

F. DataMining

Banks use data warehousing and data mining methodology to build long-term relationships with their customers and because it helps banks to quickly and smoothly adapt to business changes. To make data useful, bank enterprises collect data from almost every platform and data format; clean and transform data into information that users will understand; and stores the information in an open and efficient data warehouse structure. Data analysis is the ability to look at the same information in a variety of ways. This is where *datamining* comes in. To explore the information stored in data warehouse structure, datamining, OLAP, query and reporting, statistical analysis, data visualization and application-development interfaces are included and are mostly client/server and web enabled.

Data mining *can be formally defined as the process of extracting hidden and interesting implicit information (or knowledge) from large databases and then using the knowledge to make crucial business decisions*

Datamining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases and visualization to address the issue of information extraction from large databases. It involves running advanced queries on a database and building models to answer critical business questions without consuming time or assembling fragmented databases for analysis. It also refers to analytical process and summary on operational data.

The crux of data mining is that the information mined should have been previously unknown, should be valid, should be capable of being transformed into business advantage aiding decision making. Data mining forms one of the basis KDD (knowledge discovery in databases). The essence of data mining can be stated as the discovery of information without a previously formulate hypothesis.

5. Recommendations by Dr. Vasudevan Committe on "Technological Upgradation in Banks "

The following are the recommendations to establish a Data Warehouse for banks:

- The Standing committee on Legal Issues relating to Electronic Banking recognizes the need for data warehouses both at the individual bank level and at industry level. The argument that it is too early for such technology in India does not hold good for the banking industry which is primarily an industry dealing with facts and figures. For implementing various regulatory guidelines including the latest one on ALM, a robust MIS, founded on data warehousing and data mining, at individual bank level is essential. The structure, configuration and design of the data warehouse may, however, differ from bank to bank.

- It is not necessary to wait for all bank branches to be computerized for setting up of data warehouse. Neither is it necessary for all branches to have the same TBC software package. Data warehouse can be established even across multiple computer platforms as long as the transaction details are made available to the data warehouses in standardized formats. Therefore, banks should standardize the data formats and start supplying the data on a continuous basis from the branches, which have already been computerized. It is expected that the computerized branches themselves would provide the critical data for a data warehouse to go live. ***The committee recommends that all banks should put in place their data warehouse strategy by January 1, 2001***. The banks with a large number of computerized branches may start their pilot projects by warehousing certain categories of data (if not all the transactions) by April 1, 2001. Some illustrative application areas are:
  Investment Analysis, Credit Analysis, Customer Base Analysis and Defaulters Analysis.

- For building data bases at the individual customer level within a bank or at the Industry level, it may be advisable to follow a unique identification number for all bank customers. A Task Force may be set up by IBA to explore feasible methodology for working out a unique identification system.

- While building the industry level data warehouse, legal questions relating to confidentiality of information may arise. The Standing committee on Legal Issues relating to Electronic Banking may examine this issue. However, for the data collected under the regulatory provisions, the Reserve Bank of India could establish a Data warehouse on Banking and Finance. The Department of Banking Supervision, the Department of Banking Operations and Development and the Exchange Control Department of the Reserve Bank of India have already been receiving large amount of data. The Department of Statistical Analysis & Computer Services and the Department of Economic Analysis & Policy have also been receiving various statistical returns. Data so collected do not have any legal sensitivity and can well be used for data warehousing and data mining.

- The Indian Banks' Association may initiate the process of building another Industry Level Data Warehouse, based on agreements to be signed by the participating banks on sharing of data. This data warehouse may mask the customer information, but it should be based on individual customer information so that the participating institutions can derive the benefits of business segmentation analysis and trend forecasting on various banking operations.

The implication of adopting such technology in a bank would be as under:

- All transactions captured at the branch level would get consolidated at a central location. Such a central location could be called the data Warehouse of the concerned bank. For this to happen, one of the requirements would be to establish connectivity between the branches on the one hand and data Warehouse platform on the other.

- For banks with large number of branches, it may not be desirable to consolidate the transaction details at one place only. It can be decentralized by locating the services on regional basis. The regional Data marts as developed can provide mutual back up and could be linked to the central Data Warehousing server so that for the purpose of MIS at the bank level, data can be accessed from all the regional data marts.

- By way of data mining techniques, data available at various computer systems can be accessed and by a combination of techniques like classification, clustering, segmentation, association rules, sequencing, decision tree, various ALM reports such as Statement of Structural Liquidity, Statement of Interest Rate Sensitivity etc. or accounting reports like Balance Sheet and profit & Loss Account can be generated instantaneously for any desired period/date.

Significant cost benefits, time savings, productivity gains and process re-engineering opportunities are associated with the use of data warehouse for information processing. Data can easily be accessed and analyzed without time consuming manipulation and processing. Decisions can be made more quickly and with confidence that the data are both **time-relevant** and **accurate**. Integrated information can be also kept in categories that are meaningful to profitable operation.

Trends can be analyzed and predicted with the availability of historical data and the data warehouse assures that every one is using the same data at the same level of extraction, which eliminates conflicting analytical results and arguments over the source and quality of data used for analysis. In short, data warehouse enables information processing to be done in credible and efficient manner.

6. Design and Implementation of a Data Warehouse

The creation of the Data Warehouse for an enterprise poses a number of design and technological challenges. The design of such a database must be flexible enough to allow for integration of any new data source that may be emerge due to function as a decision support system, for a heterogeneous group of users having very complex and diverse data requirements. Given this broad requirement, the architecture of the Data Warehouse needs to be worked out and a roadmap using a software engineering approach for iterative and modular construction of the system has to be laid down.

The actual design and implementation of a data warehouse is a long and complex process, consisting of the following basic activities:

A. Requirement analysis and specification.
B. Data Warehouse design
C. Data Warehouse Implementation
D. Data Warehouse Deployment
E. Data Warehouse Review
F. Data Warehouse Maintenance and Administration.

A. Requirement analysis and specification

The requirements phase addresses the high-level needs of the entire warehouse environment. By maintaining a high-level perspective, the project can avoid the "analysis paralysis" that often plagues waterfall approaches, keeping the total time spent on this phase to a minimum.

Both business and technical (including the infrastructure) requirements are gathered during this phase. Interviews, workshops, and analysis of existing documents and systems may be used to gather and confirm the necessary facts. The resulting requirements document is reviewed by all affected parties. That document includes identification of business objectives as well as a technical feasibility analysis. Recommendations may include the number of project builds and the priority of each build.

B. Data Warehouse Design

The design phase focuses on one project build at a time. By narrowing the scope before drilling to this level of detail, the warehouse can continue to evolve rapidly in an iterative manner. Activities for this phase include:

- Detailed analysis and requirements for the selected build.
- Detailed design for the data model.
- Detailed specification of the process model for extraction, transformation, and loading.
- Selection of Software and Hardware
- Creation of the application model or selection of exploitation tools.
- Design of additional aspects such as the security and metadata models.

Both the project manager and the warehouse architect can expect to be involved full-time in this phase. Additional resources, such as business content experts, the data administrator, warehouse administrator, construction manager, and IT personnel are consulted during this phase. Business and technical representatives provide reviews of the outlined design.

Detailed analysis and requirements for the selected build

The requirements document is used as input to this phase, which produces a detailed design document for the selected build. In subsequent builds, the design document for previous builds may be used as input to ensure that the work incorporates previous decisions.

Detailed design for the data model

The content and structure of the data Warehouse are reflected in its data model. The data model is the template that describes how information will be organized within the integrated warehouse framework. It identifies major subjects and relationships of the model, including keys, attributes and attribute groupings. A designer should always remember that decision support queries, because of their broad scope and analytical intensity, require data models to be optimized to improve query performance.

According to Ralph Kimball, a design methodology for Data Model is the "Nine-Step method":

1. Choosing the subject matter
2. Deciding what a fact table represents
3. Identifying and conforming the dimensions
4. Choosing the facts
5. Storing pre-calculations in the fact table
6. Rounding out the dimension table
7. Choose the duration of the database
8. The need to track slowly changing dimensions
9. Deciding the query properties and the query models.

The design should resolve the inconsistencies in data formats, semantics and usage across multiple operational systems and define procedures for aggregating, reconciling and summarizing data to make it more relevant and useful for users.

Detailed specification of the process model for extraction, transformation, and loading

Even if the data are cleaned perfectly in the legacy environment, there is always a need for recleansing data as it passes the integration and transformation process. This Process includes

- The encoding of data according to a single scheme
- The conversion of data using a common formula
- The standardization of data into a common data structure
- The structuring of data into a common structure
- The interpretation of data according to a common definition

- The summarization of data into a common level of granularity
- The structuring of keys based on a common definition of the basis of the key
- The indexing of data according to a common key
- The movement of data from many operating systems into a common operating system
- The movement of data from one hardware architecture to another hardware architecture.

Data transformation may involve the decoding and translating of field values, addition of a time attribute (if missing in source data), summarization of data, or the calculation of derived values.

Selection of Software and Hardware

The technical platform must be examined in terms of scalability, load and query performance, maturity, portability, and cost. Large data warehouses, loading, complex query performance, and response time are all factors that may drive technical platform decisions. Estimates of data volumes are likely to be underestimated because of historical needs and redundant warehouse tables. As detailed user requirements are developed, data volumes, query frequency, and geographical distribution patterns should be confirmed.

- Hardware- decisions must be made regarding the data server hardware platform: mainframe, SMP(Symmetric Multiprocessor), or MPP (Massively parallel Processor). Storage requirements, I/O bandwidth, and number of processors must all be considered.
- Operating System- Choice of operating system, such as MVS, UNIX or NT, is closely tied to hardware, and will have an impact on the availability of tools available for the selected environment.
- DBMS- The choices in this area are among standard function relational DBMS, special purpose relational DBMS optimized for queries, or proprietary multidimensional DBMS. Database replication services to address data distribution in a client/server environment and other mechanisms for moving data may be required.
- Application Servers/Workstations- A two-tier or three-tier client server topology must be selected based on user distribution, workload, and application development tools. User workstations may need to be upgraded and included a standard workstation configuration.

Creation of the application model or selection of exploitation tools

This activity provides support for a standard set of end-user tools, including maintenance and enhancement of structured decision support applications. Support for

the technology infrastructure, including platforms, networks and workstations are necessary, but might not be unique to the data warehouse, and should follow technical services and "HELP" desk practices.

Design of additional aspects such as the security and meta data models

This set of activities addresses the need to document, reuse, and communicate the meaning of data and its system component relationships. This step creates confidence in the data because the data's definition, origin and subsequent derivations are stated and available for inquiry. A metadata repository should be populated with keys and attributes, business data description, physical data structures, source structures, mapping and transformation rules, frequency, derivation, summarization algorithms etc. Meta data generated from multiple data warehousing tools must be coordinated.

C. Warehouse Implementation:

During the implementation phase, implementation teams code and populate the warehouse with data and develop the applications for end-user analysis and reporting. Business users and the IT organization rigorously test the warehouse and applications to verify that all acceptance criteria are met.

One important aspect of this phase is preparation for the rollout of the production system. The data warehouse consultant and project manager's review the process for creating, updating, and maintaining the warehouse with the warehouse administrator. A maintenance document is prepared that records this information for future reference. Individuals from the business units and IT organization are involved in the testing of exploitation applications as an efficient means of knowledge transfer.

C. Deployment

The deployment phase is the rollout of the data warehouse and end-user applications to end-users and IT staff. Making sure that users are well trained and that applications and data are readily accessible helps promote widespread acceptance of the entire project. The faster business users gain some benefit from the warehouse project, the more likely they are to support further development or enhancement efforts.

E. Review

Three executions of the review phase are conducted with each build:

- Following the construction phase to access the implementation process and learn from successes and setbacks.
- 3-6 months later to review the deployment phase and ensure that the transition to

support has gone smoothly and that users have access to the warehouse.

- 18-24 months after initial construction to measure any tangible benefits, calculate ROI, and ensure that the warehouse environment is continuing to meet the business community requirements.

F. Maintenance and Administration

Once implemented, the warehouse requires on-going maintenance. It is essential that attention be given during the construction process to this on-going element of the warehouse life cycle.

A survey of data warehousing projects has revealed that once development begins, a kind of lethargy sets in. Further development of the warehouse slows and finally ceases as the organization confronts the realities of long-term maintenance. Development team members are often not well suited to carry out these activities, or their energies are tied up in providing support rather than addressing the development of the next builds.

Operations support staff must be appropriately trained and actively consulted during the building of the warehouse. Taking this step helps ensure a smooth transition to the operations staff without requiring development resources to provide such on-going support. Maintenance activities include:

- Addition of new users.
- Delivery of user training.
- Addition of queries, pre-joined tables and views, aggregate data.
- Continual data quality management.
- Warehouse systems management, including data storage monitoring and management.
- Provision of minor enhancements to support slight change requirements.

7. Case Study 1: An approach to build a Data Warehouse for a Bank

Building a well-designed and successful and complete Data Warehouse from Legacy systems for banks will take at least 2 to 3 years. The simplest way to achieve goal with minimum time period is start by building 'data marts'. This is a natural and understandable way to break the total solution into bite-sized pieces. All that is a required is to spend a few weeks before designing the first data mart to step back and get a perspective on where the organization is going and what the succession of data marts will be over the next few years. By interviewing managers, end-users and legacy system database administrator, we probably can make a good list of all the data marts that may be required over, say, the next four years.

Building Data Warehouse/ Datamarts can be according to hierarchical organization of banks. Initially DW/DM can be implemented for regional offices, once it is successfully

complete, warehouse will be rolled out to zonal offices and to the head office. On the other way, initially the focus can be in support of the particular area say, marketing functions, the breadth and depth of data being captured will allow bank to deploy data marts in support of other areas such as risk management, compliance reporting, profitability etc. There can be as many different data marts running off the main warehouse, covering many different areas of bank's business.

Armed with list of prospective data marts, we only really need focus on the central concept: Conforming Dimensions. By making sure that we confirm our data mart dimensions as we build successive data marts, we guarantee that the separate data marts will eventually snap together into a single, useful, integrated Data warehouse.

Some of the questions that need to be answered before the beginning of the design process of the data warehouse include:

- Who are the end-users?
- What end-user tools will be used?
- What platforms are currently being used or contemplated for the future?
- Where id the data originally stored?
- When does the data warehouse need to be operational?

Present case study provides an approach to a data warehouse for deposit data of a typical nationalized bank which maintains more than 200 branches serving 800,000 customers located all over the country. As is common with banking institutions, this bank was operating on multiple platforms and systems, with disparate data scattered throughout the decentralized branches. Their biggest challenge was finding a better way of gathering and analyzing bank-wide enterprise information. All transactions captured at the branch level would get consolidate at a central location called DWH of the nationalized bank.

Data residing in the warehouse might be derived from a number of different sources. These sources are described below:

1. Transactional data: Data collected as a result of operational transactions. This include Loan data, Deposit data etc.

2. External data: Data residing outside of the bank business. External data consists of information taken from census reports, stock/exchange market reports etc.

3. Internal data: Enterprise specific data such as interest rates.

4. Legacy data: Archived data that is stored apart from operational data, as it was not needed for near-term analysis.

The present case study focuses on the deposit data that flow into the data warehouse daily. The data warehouse presents a multidimensional, logical view of the deposit data and is generally called a multidimensional deposit database. This deposit data warehouse has been implemented using a dimensional model or star schema design. Fig.3 shows a typical star schema for Deposit Warehouse.

Determining "Clean" data

Depending on the state of the data, this is often the most challenging phase in creating the data warehouse. For the most part, the deposit data was fairly clean. The only data problems were:

- Inconsistent use of state names: sometimes the same state would roll-up to different regions.

- Inconsistent use of scheme-id descriptions: the same scheme-id is found with different scheme names.

Both cases were handled the same way. When a data consistency was found, the record was not immediately loaded, but written to an error file. At the end of the load process, a report can be generated which highlights the discrepancy between source record and the warehouse record. The user can review the problem and generate a new version of the dimension record in the warehouse from the source record.

Identifying Meta data

The deposit data warehouse metadata consists of the following components for quick and flexible query analysis:

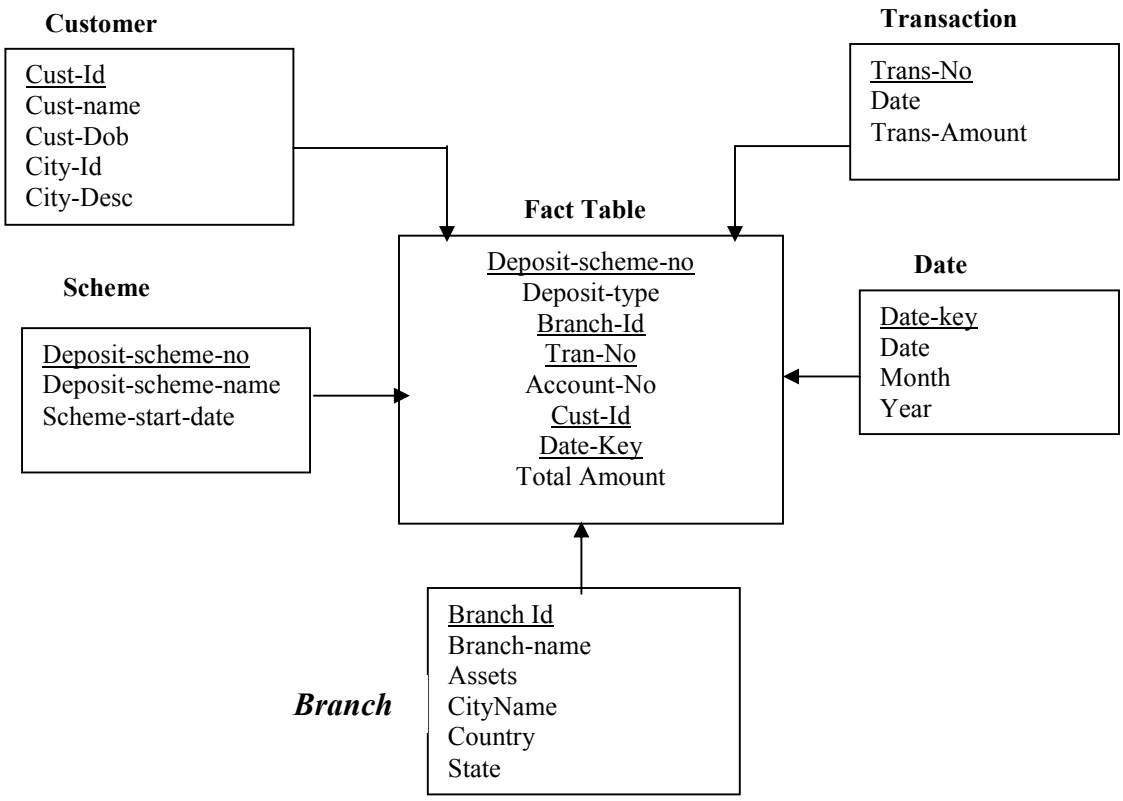
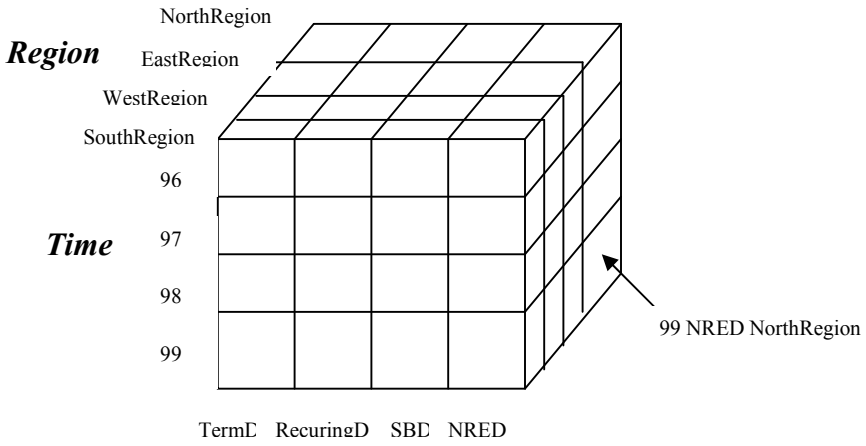


Fig. 3. A typical star schema of Deposit Data Warehouse

1. Tables, structures, views, integrity constraints, aliases, macros, ownership
2. Studies associated with each database, databases associated with each study
3. Data integration, conversion, summarization
4. Hierarchies, granularity
5. When data extracted from source to target database
6. Who, what and when data accesses; size of request and time required
7. Where data is stored and when it was collected
8. Identification and location of variable formats for tables and reports
9. Adverse event dictionary, data entry dictionary
10. Document control, protocol versions used in study, status of submission.

Building multidimensional deposit data cube

To interact with Deposit data warehouse, OLAP operations such as drill-down, roll-up, pivot and slice are used. A 3-D data cube of a Deposit data warehouse is shown below:



The above deposit data warehouse consists of three dimensions: Region, Time, Scheme and Deposit-amount as measure. A concept hierarchy is associated with each dimension as follow,

Region (*Region-name, State, District, Branch*)
Time (*Year, Month, Day, Hour*)
Scheme (*Scheme, Scheme-Id, Interest-rate, Maturity-period*)

Drill-down and roll-up operations can be performed along each dimension. For example one may start with high-level cube which consists of Region-name, Year and Scheme and drill-down to examine the total deposit amount by Scheme, by Region and by Month. Also the results of OLAP operations can be viewed in the tabular form. For example, distribution of deposit amount by region and by year is as follows:

	1996	1997	1998	1999
NorthernRegion	17,00,000	15,00,000	20,00,000	25,00,000
EasternRegion	17,50,000	16,00,000	15,00,000	20,00,000
WesternRegion	25,00,000	17,00,000	25,00,000	16,00,000
SouthernRegion	20,00,000	17,00,000	17,00,000	15,00,000

Further focussing on a particular year, say 1996, for different schemes, distribution of deposit amount is as follow:

		1996		
	Term	Recurring	SavingsBank	NonResidentialExternal
	Deposits	Deposits	Deposits	Deposits
NorthernRegion	500,000	450,000	400,000	350,000
EasternRegion	450,000	400,000	500,000	400,000
WesternRegion	800,000	600,000	500,000	600,000
SouthernRegion	600,000	500,000	400,000	500,000

Distribution of deposit amount of a particular region, say southern region for the Year 1996 at a lower level of abstraction is as follows:

		1996		
	Term	Recurring	SavingsBank	NonResidentialExternal
	Deposits	Deposits	Deposits	Deposits
NorthernRegion	500,000	450,000	400,000	350,000
EasternRegion	450,000	400,000	500,000	400,000
WesternRegion	800,000	600,000	500,000	600,000
SouthernRegion	600,000	500,000	400,000	500,000
*AndhraPradesh	350,000	300,000	150,000	250,000
*Tamilnadu	250,000	200,000	250,000	250,000

Drilling to further lower level of abstraction, distribution of deposit amount of a particular location for the Year 1996 is as follows:

		1996		
	Term	Recurring	SavingsBank	NonResidentialExternal
	Deposits	Deposits	Deposits	Deposits
NorthernRegion	500,000	450,000	400,000	350,000
EasternRegion	450,000	400,000	500,000	400,000
WesternRegion	800,000	600,000	500,000	600,000
SouthernRegion	600,000	500,000	400,000	500,000
*AndhraPradesh	350,000	300,000	150,000	250,000
**Krishna	100,000	100,000	50,000	100,000
**Hyderabad	250,000	200,000	100,000	150,000
*Tamilnadu	250,000	200,000	250,000	250,000

The above case study pertains to only the deposit information and on a whole data warehouse for a bank covers several such areas.

8. CaseStudy 2: *Pattern discovery and Data mining of loan data*

A proper decision support system should exist to overcome the difficulties arise during the loan operation. For example, loan officer must be able to identify potential credit risks during the loan approval cycle to minimize loan defaults. Bankers are finding Datamining as suitable approach for information mining to enable banks in data analysis and decision support.

Establishing Mining Goals

The objective of this case study to identify the interesting patterns related to customer profile and asset type in the loan data of a nationalized bank using data mining techniques.

Data collection and selection for Analysis

In this application, a profile of around 300 customers were taken from nationalized bank with details pertaining to occupation, residential address, gross and net income, sanctioned loan amount, outstanding amount, and months pending. The data was obtained as flat files. With the available data, relations pertaining to age, netsalary, monthspending, sanctioned loan amount were attempted for customer segmentation to analyze and predict behavioral patterns.

Data preparation for Analysis

After selecting the desired database tables and identifying the data to be mined, the data to be preprocessed for further analysis. The original data was transformed into MSAccess database format through which DBMiner a data mining tool, developed by the data Mining Research Group led by Dr. Jiawei Han, can take input using ODBC connectivity. Also the attribute "assettype" using months pending information. Customers with their loan considered as Non-Performance Asset (NPA) form risk group.

Loans are classified as under:

AssetType	MonthsOverDue
Standard	0 to 1
OverDue	2 to 6
SubStandard(NPA)	7 to 24
Doubtful	>=25

Datamining

In this case study, the selected loan data of a nationalized bank is mined using association rule and classification techniques. For the present study a tool has been developed in java that allows users to interactively mine association rules from input Data. DBMiner tool has been used for applying classification techniques. Association rules are presented in the form "in 80% cases, if an individuals NetSalary is between 0~5000, and Age is between 40~50 then, that loan is considered as Non-Performance Asset (NPA)".

Classification analyzes a set of training data (i.e., a set of objects whose class label is known) and constructs a model for each class based on the features in the data. A set of

classification rules generated by such a classification process can be used for better understanding of future data. In this application, classification process is used to classify loans based on the features in the data and help predict the kind of loan based on particulars of individuals.

Evaluating the mining results

The results obtained by mining the data are quite outstanding and provide a deep insight into the model constructed using the available data. Mining tools used in this application explores numerous hidden information that help from achieving goals. Characteristics of customers with high probability of default when granted loan were identified and thus customers who were not falling in that risk group were promoted.

Following are the observations inferred from the association rules and classification rules:

Rules : NetSalary and Assettype

- *In **15%** cases, if an individuals NetSalary is between 0~5000, that loan is considered as **NPA**.*
- *In **6%** cases, if an individuals NetSalary is between 5~10000, that loan is considered as **NPA**.*

Rules: NetSalary, SanctionedLoanAmount and AssetTtype

- *In **19%** cases, if an individuals NetSalary is between 0~5000 and SanctionedLoanAmt is between 0~100,000, that loan is considered as **NPA**.*

Rules: NetSalary, Age and Asset type

- *In **9%** cases, if an individuals NetSalary is between 0~5000 and Age is between 20~40, that loan is considered as **NPA**.*
- *In **19%** cases, if an individuals NetSalary is between 0~5000 and Age is between 40~60, that loan is considered as **NPA**.*

Rules: NetSalary, Servicetype and Asset type

- *In **14%** cases, if an individuals NetSalary is between 0~5000 and Service (OTH), that loan is considered as **NPA**.*
- *In **16%** cases, if an individuals NetSalary is between 0~5000 and Service (SER), that loan is considered as **NPA**.*
- *In **23%** cases, if an individuals NetSalary is between 15000~20000 and Service (SER), that loan is considered as **NPA**.*

Since the doubtful assets does not present in the sample data, we could not actually obtained any interesting pattern for doubtful assets. The data mining on complete data may result in other interesting patterns.

9. Case Study 3 : *Data mining for the analysis of credit card transactions*

With the advent of new technologies, people are increasingly showing their inclination towards electronic means of payment. Credit card margins continue to be squeezed by a combination of high charge-off and rising account acquisition costs. Record levels of delinquencies, personal bankruptcies and resulting charge-off coexists in a saturated market where offers are quickly becoming commodities. In this environment, accurate risk prediction is of utmost important.

In order to remain competitive, credit card issuers are turning to data mining to uncover information from their massive databases. This application deals with credit card from a nationalized bank. Mining the credit card data not only discover the customer segments but also helps extracting additional hidden information that may guide the bank in developing models tailored for specific business goals, such as to detect fraud or accurately target customers.

Identifying Objectives

Credit card issuers are engaged in improving response rates while also identifying the best candidates in terms of profitability and risk. Issuers that most accurately match customer risk profile and behavioral attributes with differentiable products will seize a competitive advantage.

The drivers for data mining the card transactions are:

- Determine whom to solicit as client, possibly with pre-approved credit limits.
- Customer retention – find and analyze which customer characteristics will help to offer services that will keep the best credit card customers for the long term.
- Customer attrition – discover which customers are most likely to leave for lower-interest-rate cards.
- *Fraud detection – find purchase patterns and trends to detect fraudulent behavior at the time of credit card purchases.*
- Payment or default analysis – identify specific patterns that will help predict when and why cardholders default on their monthly payments.
- Market segmentation – correctly segment cardholders into groups for promotional and evaluation purposes.

Data mining makes the above possible by organizing the bank's credit card holders into related groups and then examining the past credit history, the purchasing profile, the

payment profile of each group, merchant details, etc. It uncovers vital knowledge hidden in the database so that the issuers can improve marketing of card products and related services, retain and attract good customers, increase market share, reduce cost, and increase return on investment.

Data Collection and Preparation

A profile of 200 customers were taken from Nationalized bank with details pertaining to occupation, residential addresses, gross and net incomes, outstanding amount and months pending. Transactions pertaining to last 3 years were also obtained for the customer profile given.

The flat files were initially imported into MS Access and queries done, to unearth possible information. In the present study, it was assumed that the difference between the income figures would indicate savings. Subsequent to preliminary analysis, they were imported into the application session of Pattern Recognition Workbench (PRW), *a powerful datamining tool developed by Unica Technologies Inc., Lincoln*.

Mining using Pattern Recognition

Pattern Recognition involves the discovery and characterization of patterns in the data. A pattern is an arrangement or an ordering in which some organization of underlying structure can be said to exist. Patterns in data are identified using measurable features or attributes that have been extracted from the data. The customers are divided into clusters on similarities. This helps in assessing the risk involved in the credit card transactions and identifying loyal customers & defaulters. PRW system is used to perform data mining tasks on the credit card data and to extract hidden information that can help identify risk segments. The PRW system combines machine learning, neural networks and statistical algorithms for building general pattern recognition and discovers various kinds of knowledge using multi-algorithmic from large relational databases.

Analysis of results

This study on the typical data set produced some interesting results. The following is the brief summary of important findings/observations in the work done so far:

- The individuals with lower income group were found to be customers of more value as well as less defaulting in nature.
- Some customers were found consistently fail to pay the credit card outstanding balance even the amount is small; if the data on payments dates and amounts is incorporated into the data set, analysis can show a consistent pattern of payment behavior for customers.
- Among other factors noticed was that some data like the occupation listed, did not match with the income or spending pattern of the customer, calling attention to capturing and appraising customer credit worthiness

- Customers are spending more amount than their average savings in a month.
- In most of the cases, the outstanding amount is very high that the actual utilization amount.

The above results are tentative and indicative but not of absolute nature. However, the sample of 200 customers is insufficient to have concrete results as the results turn out to be symmetrical after restoration of missing values and noisy data.

10. Return on Data Warehouse Investment.

The costs of Data Warehouse projects are usually high. The cost involved in Data Ware House project: Hardware cost, Software cost, Personnel cost, Ongoing Operational cost, etc. It is important to have clear understanding of their real benefit, and of how to realize this benefit at a cost that is acceptable to the banking industry.

With data Warehouse, banks can keep costs from spiraling out of control or at least ensure that additional cost are focussed on the right business challenges.

Cost escalate quickly in 3 ways:

1. The overlong requirements study with numbers of untrained consultants wandering around enterprise interviewing managers followed by the over long information requirements program.
2. The data auditing tasks turn up problems with critical information that require changes to the underlying operational systems; and
3. High-value applications influence other managers to request new applications giving rise to increased requirements for processing power and programming talent, hence money. Experienced managers control costs by avoiding long requirements studies, but they spend money necessary to make sure that data are consistently accurate.

Data Warehouse makes money by increasing the yield of marketing and sales program, by uncovering and anticipating fraudulent activities, eliminating redundant application development and enhancing quality and timeliness of decision while dramatically increasing productivity.

For a bank active in trading, market risk is part of doing business. The risk of loss stems from changes in interest rates, foreign exchange, equity and other factors in a volatile worldwide market. As a result, risk management is crucial to any bank's stability and success. By using a sophisticated data warehouse to aid in day-to-day evaluation of risk such as operational, credit, and market factors managers at a bank can obtain the up-to-the-minute information they need to make critical decisions. The risk management reports that are necessary for the 7.30A.M. conference calls can be generated by those who need and use the information, with rapid turnaround time.

How can the cost be justified? Given the high costs, it is difficult to justify a data-warehousing project in terms of short-term benefits. As a point solution to a specific management information need, a data warehouse will often struggle to justify the associated investment. It is, as a long-term delivery mechanism for ongoing management information needs that data warehousing reaps significant benefits. But how can this be achieved?

There is scope for economies of scale when planning data warehousing projects; if focus were to be placed initially on the 20% of source systems which supplied 80% of the data to decision support systems, then an initial project which simply warehouses "useful" data from these systems would clearly yield cost benefits of future MIS projects requiring that data. Rather than targeting a specific business process or function, benefits should be aimed at the wider audience for decision support. Such a project would form an invaluable foundation for an evolving data warehouse environment.

Once this initial project is complete, emphasis can be placed on the growth of the warehouse as a global resource for unspecified future decision support needs, rather than as a solution to specific requirements at a particular time. In subsequent phases of the warehouse development, new data which is likely to play a major role in future decision support needs should be carefully selected, extracted and cleaned. It can then be stored along side the existing data in warehouse, hence maximizing its information potential. As new information needs emerge, the cost of meeting them will be diminished due to the elimination of the need to perform much of the costly extraction, cleaning and integration functions usually associated with such systems.

Over time, this environment will grow to offer a permanent and invaluable repository of integrated, enterprise-wide data for management information. This in turn will lead to massively reduced time and cost to deliver new decision support offerings, and hence to true cost justification. Finally, a data Warehouse results both tangible and intangible benefits to the industry.

11. Management Values

The banking industry in India can be a lot more disciplined about its operations. The primary input and output for banks remains money, The next important source of capital is the intellectual capital. How far the upcoming technologies can add to employee and work satisfaction, besides the organization becoming more geared to implement its job scrupulously remains to be seen. The INFINET and subsequently the application of technologies like data warehousing and data mining should be able to contribute positively to the management values.

Uncovering new revenues by repackaging information in corporate data store is a tantalizing idea. The applications of data warehouse and data mining technologies can be the best way to do it. A bank-wide data warehouse can form a better way of gathering

and analyzing customer transactions and individual profitability to determine which services could be better marketed, and where, which means, more customized services for customers and increased profitability for the banks. Data warehousing solution can be used to turn business data into a business advantage and meet aggressive ROI targets and thus contribute positively to the management value. However, it calls for new legislature concerning transparency on the part of the management in divulging information it considers an asset. The more useful a technology is, the greater should be the management's endeavor in implementing its fair share of work and hence greater should be the management value.

12. Selected Data Ware House Technology Vendors

A. Data Warehouse Suites

Given the complexity of data warehouse projects, many database vendors also provide suites of related technologies needed to build a data warehouse. These vendors offer commonly used databases for data warehouse implementations and other tools:

- Aredent Software (http://www.ardentsoftware.com).
- IBM DB2 Universal Database (http://www.software.ibm.com/is/swservers/index.html) With Oracle, a leading database vendor.
- Informix (http://www.informix.com/informix/solutions/dw/dwprod.html).
- NCR Teradata (http://www3.ncr.com/product/teradata/product.html).
- Oracle 8i (http://www.oracle.com/tools/InternetBITools.html)
- SAS (http://www.sas.com/software/data_warehouse).
- Sybase (http://www.sybase.com/bid).
- Showcase Corp. (http://www.showcase.com).
- MS SQL Server 7.0(http://www.microsoft.com/india/sql).

B. Data Marts Tools

The leading data mart vendors include the following

- Fiserv (http://www.fisver.com/resources).
- Informatica (http://www..informatica.com/products/pmart_tech.html).
- Information Builders (http://www.ibi.com/products/smartmart/overview.html).
- Sagent (http://www.sagent.com/products/index.html).
- Silvon Software (http://www.silvon.com/prod_serv/products.htm).

C. Data Warehouse Access Tools

Some leading vendors of data warehouse access tools, including query and reporting, OLAP, and data mining applications, include these companies:

- AlphaBlox (http://www.alphablox.com).
- Brio Technologies (http://www.brio.com).
- Business Objects (http://www.businessobjects.com).
- Cognos (http://www..cognos.com).
- Hyperion Solutions (http://www.hyperion.com/solutions.cfm).
- Hummungbird Communications (http://www.hummungbird.com).
- Microstrategy (http://www.strategy.com).
- Seagate Software (http://www.seagatesoftware.com).
- Sterling Software (http://www.infoadvan.com).
- IBM Intelligent Miner (http://www.software.ibm.com.iminer).
- DBMiner (http://www.dbminer.com).
- Pattern Recognition Workbench (unica-usa.com).

13. Conclusions

Climacteric competition and rising loan delinquency rates are seeing more banks exploring ways to use their data assets to gain a competitive advantage. This paper analyses how, in practice, data warehouse applications fits in with various different business problems at banking sector and also demonstrates how the bank-wide enterprise data warehouse can be implemented to provide atomic level information on all banking transactions, customers and all products for use in decision-support systems. The possibility of setting up a data warehouse, seems more remote when compared to the setting up of data marts, which can later be integrated into a bank-wide enterprise data warehouse. The integrated data store can be used to uncover a huge potential loss of revenue, which can be averted and which will further guide how to approach pricing and service grouping well into the future.

IDRBT in its constant endeavor to bring to the fore technological needs of the Indian banking industry and deliver the right solutions have embarked on a data warehousing project which is envisioned to provide the banking industry with the fore said benefits. The prospects that INFINET bring to the Indian banking scenario can only be enriched by the application of technologies like Data warehousing and data mining. Finally Data warehousing and Data mining tools are essential components in banking sector.

References

1. Harry S. Singh, “Data Warehousing – concepts, Technologies, Implementations, and Management “, Prentice Hall PTR, New Jersey.
2. Douglas Hackney, “Understanding and Implementing successful DataMarts, Addison-Wesley Developers Press”.
3. Report of the Dr. Vasudevan Committee on Technology Upgradation in the Banking Sector.
4. http://www.dw-institute.com

5. http://www.datawarehouse.org

6. http://www.datawarehouse.dci.com

7. http://www.webopedia.internet.com/TERM/d/data_warehouse.html

8. http://www.db.stanford.edu/warehousing

9. http://www.datamation.com/dataw