CHAPTER - 28

Framework for Development of Meteorological Data Warehouse

N.Showri Raju, G.R.Maruthi Sankar and R.Nagarjuna Kumar
Central Research Institute for Dryland Agriculture, ICAR, Santoshnagar, Hyderabad-500059, Andhra Pradesh.
Scientist (Computer Applications in Agriculture); 2Principal Scientist (Agricultural Statistics)
E-mail address : nsraju@crida.in

ABSTRACT

The details of a frame work for development of data warehouse are discussed in this paper. The concepts of data ware housing, its data ware house process, its architecture, components, data sources, data transformation, apart from free and shareware and commercial software available are described. The details of power of Data Warehouse are also described with a procedure for a better reporting of data and analysed results. The data Warehouse should be made to deliver clear indications on how the research or business enterprise is performing. There is a need to plot out the expected users for the Data Warehouse in the enterprise so that they will have the appropriate reports in a format which is quickly understandable.

INTRODUCTION

The concept of data warehousing is not hard to understand. The notion is to create a permanent storage space for the tera and pita bytes of data needed to support weather reporting, analysis, and other Business intelligence functions. This paper designs and implements the meteorological data warehouse as well as the meteorological data report based on Business intelligence tools. The purpose is to apply data warehouse technology in the meteorological research area. Using On-Line Analytical Processing (OLAP) and the multidimensional report, we could get the beneficial data. The system generates meteorological data report, which can publish the report to the browser. This research is

beneficial for meteorological phenomenal studies. On the surface, it may seem not effective to store data in more than one place. The advantages, however more than justify the effort and cost of doing this exercise. Based on the meteorological information on different weather parameters, background and ideology of data warehouse construction, this article analyzes multiple meteorological data sources, designs the meteorological data warehouse architecture, target data model and ETL process. We have applied this architecture to provide a dynamic and flexible weather data warehouse that provides a wide variety of weather data from different sources to different weather-based applications. The weather data warehouse is very much required to take effective climate change mitigation strategies, weather aberrations and also Produce Reports for Long Term Trend Analysis, aggregations of weather changes and other aspects.

What is Data Warehousing?

A data warehouse is a collection of data in support of management decision-making process that is subject-oriented, integrated, time-variant, and nonvolatile. The data warehouse is focused on the concept (for example, sales) rather than the process (for example, issuing invoices). It contains all the relevant information on a concept gathered from multiple processing systems. This information is collected and stored at regular intervals and is relatively stable.

A data warehouse is an integrated store of information collected from other systems that becomes the foundation for decision support and data analysis. Although there are many types of data warehouses, based on different design methodologies and philosophical approaches, they all have these common traits.

- Information is organized around the major subjects of the enterprise reflecting a data-driven design.
- Raw data is gathered from the nonintegrated operational and legacy applications, cleansed, and then summarized and presented in a way that makes sense to end users.
- Based on the feedback from end users and discoveries in the data warehouse, the data warehouse architecture will change over time, reflecting the iterative nature of the process.

The data warehousing process is inherently complex and, as a result, is costly and time-consuming. Over the past several years, Microsoft company has been working within the software industry to create a data warehousing platform that consists of both component technology and leading products that can be used to lower the costs and improve the effectiveness of data warehouse

creation, administration, and usage. Microsoft also has been developing a number of products and facilities, such as Microsoft SQL Server version 7.0, that are well suited to the data warehousing process. Coupled with third-party products, that can be integrated using the Microsoft Data Warehousing Framework, customers have a large selection of interoperable, best-of-breed products from which to choose for their data warehousing needs.

The SQL Server 7.0 offers broad functionality in support of the data warehousing process. In conjunction with the Data Warehousing Framework, Microsoft plans to deliver a platform for data warehousing that helps reduce costs and complexity, and improves effectiveness of data warehousing efforts.

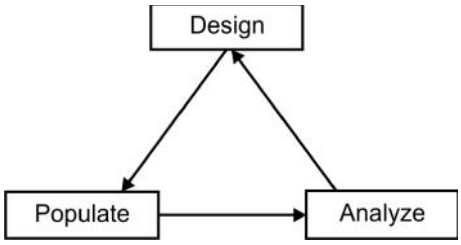
Data Warehousing Process

From the information technology perspective, data warehousing is aimed at the timely delivery of the right information to the right individuals in an organization. This is an ongoing process, not a one-time solution, and requires an approach different from that required in the development of transaction-oriented systems.

A data warehouse integrates the operational data by using consistent naming conventions, measurements, physical attributes, and semantics. The first step in the physical design of the data warehouse are determining which subject areas should be included and developing a set of agreed-upon definitions. This requires interviewing end users, analysts, and executives to understand and document the scope of the information requirements. The issues must be thoroughly understood before the logical process can be translated into a physical data warehouse.

Following the physical design, operational systems are put in place to populate the data warehouse. Because the operational systems and the data warehouse contain different representations of the data, populating the data warehouse requires transformations of the data: summarizing, translating, decoding, eliminating invalid data, and so on. These processes need to be automated so that they can be performed on an ongoing basis: extracting, transforming, and moving the source data as often as needed to meet the business requirements of the data warehouse.

Finally, the information is made available for browsing, analyzing, and reporting. Many tools assist in analysis, from simple report writers to advanced data miners. Ultimately, the analysis drives the final iterations of the data warehousing process, causing revisions in the design of the data warehouse to accommodate new information, improve system performance, or allow new types of analysis. With these changes, the process restarts and continues throughout the life of the data warehouse.



Data Warehousing Architecture

Many methodologies have been proposed to simplify the information technology efforts required to support the data warehousing process on an ongoing basis. This has led to debates about the best architecture for delivering data warehouses in organizations. Two basic types of data warehouse architecture exist: *enterprise* data warehouses and *data marts*.

Enterprise Architecture

The enterprise data warehouse contains enterprise-wide information integrated from multiple operational data sources for consolidated data analysis. Typically, it is composed of several subject areas, such as customers, products, and sales, and is used for both tactical and strategic decision making. The enterprise data warehouse contains both detailed point-in-time data and summarized information, and can range in size from 50 gigabytes (GB) to more than 1 terabyte (TB). The Enterprise data warehouses can be very expensive and time-consuming to build and manage. They are usually created from the top down by centralized information services organizations.

Data Mart Architecture

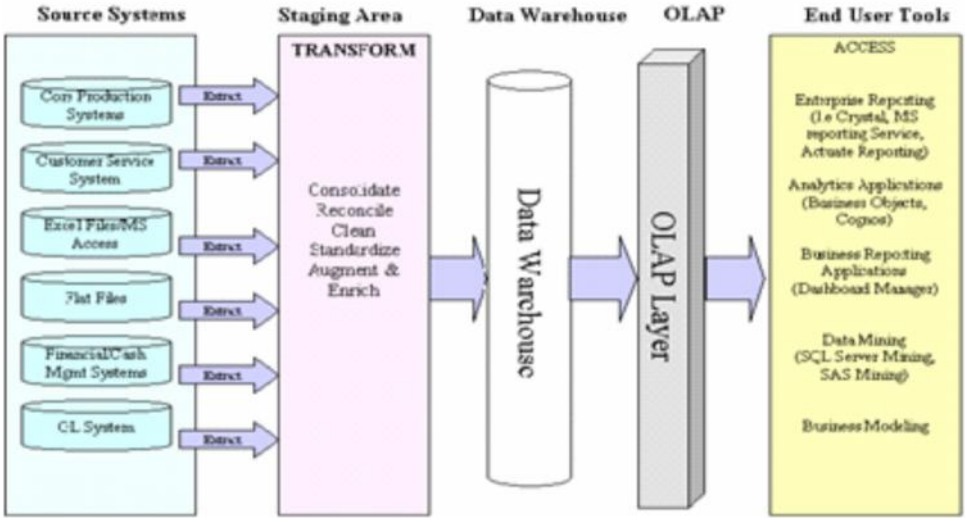
The data mart contains a subset of enterprise-wide data that is built for use by an individual department or division in an organization. Unlike the enterprise data warehouse, the data mart is usually built from the bottom up by departmental resources for a specific decision-support application or group of users. The data marts contain summarized and often detailed data about a subject area. The information in the data mart can be a subset of an enterprise data warehouse (dependent data mart) or can come directly from the operational data sources (independent data mart). The enterprise data warehouses and data marts are constructed and maintained through the same iterative process described earlier. Furthermore, both approaches share a similar set of technological components.

Data Warehousing Components

A data warehouse always consists of a number of components, including:

- Operational data sources.
- Design/development tools.
- Data extraction and transformation tools.
- Database management system (DBMS).
- Data access and analysis tools.
- System management tools.

Several years ago, Microsoft recognized the need for a set of technologies that would integrate these components. This led to the creation of the Microsoft Data Warehousing Framework, a roadmap not only for the development of Microsoft products such as SQL Server 7.0, but also for the technologies necessary to integrate products from other vendors.



Many companies would build data warehouses to predict developments in a particular field so that based on the data warehouse can be obtained by new policies that benefit. In building a data warehouse, there are four steps that must be considered, the fourth thing is the identification of sources of data, extraction, transformation, and loading data into the database.

1. Identify data sources

The first step before developing data warehouses, things to note is to identify existing data sources, for example in building a library data warehouse,

there are two types of data sources that need to be considered, namely internal and external data. The internal data is data that already exists in the library, for example a database collection of books, while the external data source is data that does not exist in library databases, external data can be combined to create the appropriate variable.

The data clustering is one technique used in the identification of data sources. Basically clustering of data is a process for classifying a set of data without a class attribute that has been defined previously, based on the principle of conceptual clustering is to maximize and minimize intra-class similarity. For example, a set of objects can be a set with clustering to make classes and then became a set of rules that can be derived based on a certain classification.

The analysis with clustering, it is helpful to build partitions of a large number of objects with based on the principle of "divide and conquer" which decomposition of a large-scale system, into smaller components, to simplify the design process and implementation. Because of the variety of data types and different purposes of the data warehouse, it is unrealistic to expect that data warehouse can capable of handling all types of data. The data warehouse system should be constructed specifically for special data types such as in a relational database, transaction database, spatial database, multimedia database and so forth.

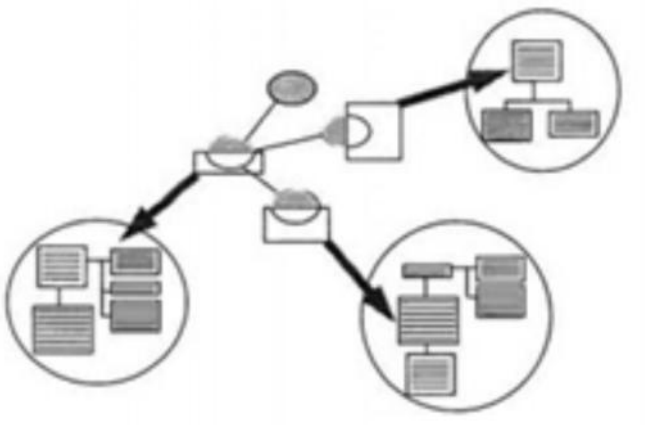
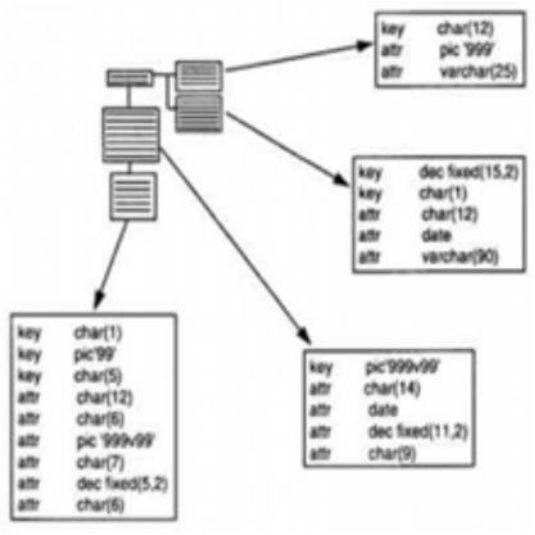


Fig. 1 : Identifying data from various data sources

2. Data Extraction

At this stage takes the longest time because this step is a step we want to retrieve data from various sources. The issues of interest in this phase is duplication data and inconsistencies data. The data collected in the transaction process is often placed at different locations. So it takes the ability of the system

to collect data quickly. If the data is stored in the regional branch, the data must be uploaded to a centralized server. This can be done daily, weekly, or monthly depending on the amount of data, security and cost. Data can be summarized first before sending them to a central storage place. For example, a hardware store may send data which showed that 10 rolls of cable have been sold on this day by employee number 10 compared to the transaction detail data transmission.



Figu. 2 : Add the Attribute

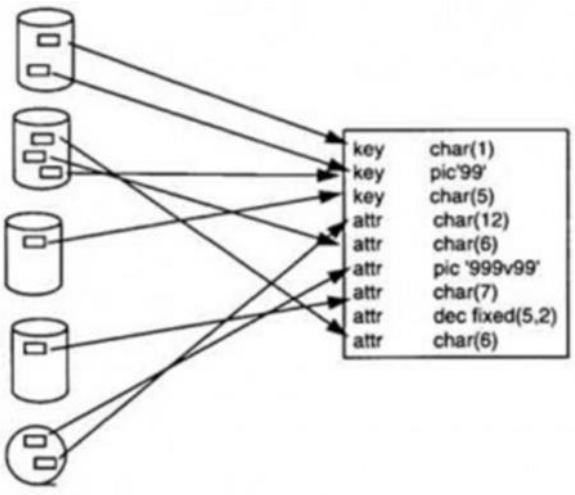


Fig. 3 : Defining System Record

3. Data Transformation

The next stage after extracting data from various sources is a transformation. Transformation of data associated with the conversion process from data source to data destination. This process consists of 2 steps: mapping data from all data elements by using transformation rule 1-to-many and many-to-1 generation of code that creates a real transformation program.

Transformation is needed to ensure consistency of data, data transformation can be done when the extraction of data or when entering data into the data warehouse. This integration will be a complex problem when the data is very large. The data transforming assuming that the data has been stored in a single repository. In the final step, the data has been extracted from multiple databases into a single database. The type of summarizing in this step is similar with summarizing during the extraction phase. Some companies choose to summarize the data in a single repository. The aggregate functions are frequently used, among others summarizations, averages, minimum, maximum, and count. The important in the transformation of data includes the following.

Data Cleansing

The cleaning process performed on the data already collected in order to remove the incorrect records, attributes-attributes standardize, rationalize the structure of the data, and controlling the lost data. The data are inconsistent and many errors make the results inaccurate data warehouse. It is very important to make the data consistent. Data cleansing can also help companies to consolidate records. This is very useful when a company has a lot of records for a customer. Any records or customer files have the same customer number, but different information in each file.

Standard Forms

Furthermore, after experiencing the process of data cleansing, data transferred into the standard form. The standard form is a form of data to be accessed by the algorithm, a data warehouse. The standard form is usually in the form of a spreadsheet. The form works well because of spreadsheet rows is represent the cases and columns represent the features.

Data Reduction and Feature

After the data is in standard spreadsheet form, efforts should be made to reduce the number of features. There are several reasons to reduce the number of features in a spreadsheet. A bank may have hundreds of features when they wanted to predict credit risk. This means that companies have data in very

large quantities. Working with a large of data can make prediction algorithm decreases performance.

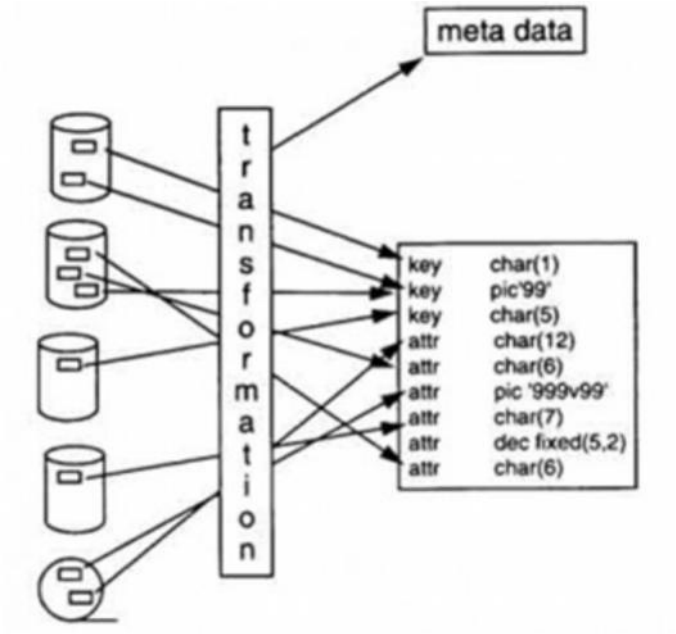


Fig. 4 : The process of transformation

Loading data into the data warehouse

After the extraction, transformation and data cleansing performed, the data loaded into the data warehouse. The categorized two types of loading data is loading data containing the operational database and loading data into the data warehouse from changes that occur of the operational database.

Some types of information can be produced from the formation of a data warehouse, the information may be based on association rules, characteristic rules, classification rules, discriminate rules, clustering, sequential pattern, and deviation analysis. Furthermore, there is another classification according to level of abstraction of the information obtained, among others, generalized knowledge, primitive-level knowledge and multiple levels of knowledge. A flexible data warehouse system can gather information at various levels of abstraction. Various Free and Shareware and commercial data warehousing tools are as follows.

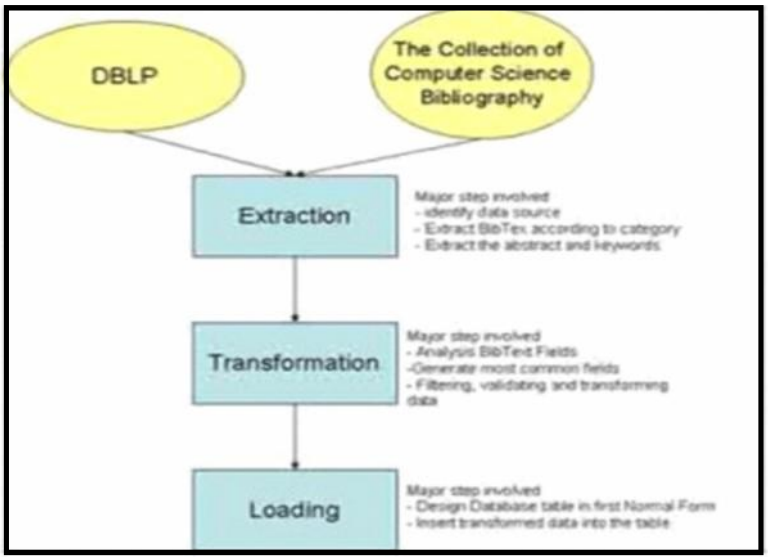


Fig. 5 : The framework in building a data warehouse

Free and Shareware

- ***AlphaMiner,*** open source data mining platform that offers various data mining model building and data cleansing functionality.
- ***Gnome Data Mining Tools***, including apriori, decision trees, and Bayes classifiers.
- ***IBM Intelligent Miner,*** University scholars can now receive free copies of DB2 UDB and Intelligent Miner for educational or research purposes.
- ***KNIME,*** extensible open source data mining platform implementing the data pipelining paradigm (based on eclipse).
- ***Machine Learning in Java (MLJ),*** an open-source suite of Java tools for research in machine learning.
- ***Orange,*** C++ components for data mining, includes pre-processing, modelling and data exploration techniques.
- ***RapidMiner,*** a leading open-source system for knowledge discovery and data mining.
- ***Weka,*** collection of machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost any platform.

Commercial

- ***Clementine from SPSS,*** leading visual rapid modeling environment for data mining. Now includes Clementine Server.
- ***Data Applied,*** offers a comprehensive suite of web-based data mining techniques, an XML web API, and rich data visualizations.
- ***IBM Intelligent Miner Data Mining Suite,*** now fully integrated into the IBM InfoSphere Warehouse software; includes Data and Text mining tools (based on UIMA).
- ***KXEN (Knowledge eXtraction ENgines),*** providing Vapnik SVM (Support Vector Machines) tools, including data preparation, segmentation, time series, and SVM classifiers.
- ***Microsoft SQL Server 2008,*** empowers informed decisions with predictive analysis through intuitive data mining, seamlessly integrated within the Microsoft BI platform, and extensible into any application.
- ***Oracle Data Mining (ODM),*** provides GUI, PL/SQL-interface, and Java-interface to Attribute Importance, Bayes Classification, Association Rules, Clustering, SVM, and more.
- ***Salford Systems Data Mining Suite,*** CART Decision Trees, MARS predictive modeling, automated regression, TreeNet classification and regression, data access, preparation, cleaning and reporting modules, RandomForests predictive modeling, clustering and anomaly detection.
- ***SAS Enterprise Miner,*** an integrated suite which provides a user-friendly GUI front-end to the SEMMA (Sample, Explore, Modify, Model, Assess) process.
- ***SPSS*** featuring Clementine, SPSS and other data mining tools.
- ***Statistica Data Miner,*** a comprehensive, integrated statistical data analysis, graphics, data base management, and application development system.
- ***Teradata Warehouse Miner and Teradata Analytics***, providing analytic services for in-place mining on a Teradata DBMS.
- ***XLMiner,*** Data Mining Add-In For Excel.

CONCLUSION

The power Data Warehouse consumers are business analysts and managers. Data Warehouses are meant to deliver clear indications on how the business is performing. Plot out the expected users for the Data Warehouse in the enterprise so that they will have the appropriate reports in a format which is quickly understandable. Always remember that data has to be presented attractively and in a format business managers will feel comfortable. Text files with lines of numbers will not suffice.

REFERENCES

Boehnlein Michael and Ulbrich-vom Ende Achim (1999). Deriving Initial Data Warehouse Structures from the Conceptual Data Models of the Underlying Operational Information Systems: DOLAP 99 Kansas City Mo USA.

Inmon., and William, H. (2000). Building the Data Warehouse: Getting Started.

Jarke, M., Jeusfeld, M., Quix, C., and Vassiliadis, P. (1999). Architecture and quality in data warehouses: An extended repository approach.

Rizzi Stefano., Lechtenborger Jens., Abell Alberto., and Trujillo Juan. (2006). Research in Data Warehouse Modeling and Design: Dead or Alive?: Conference held on 10th November, Arlington, Virginia, USA.

Rawat Laxmi (1998). Changing facets of weather and climate in Doon Valley, FRI.

Sharma Gajendra (2008). Data mining, data warehousing and OLAP (Second Edition), Katson Books.

Wah., and Teh Ying. (2007). Building Data Warehouse. Malaysia: Department of Information Science University Malaya.

❑❑❑