Binary Journal of Data Mining & Networking 5 (2015) 22-25
http://www.arjournals.org/index.php/bjdmn/index

Research Article

ISSN: 2229 -7170

A Roadmap: Designing and Construction of Data Warehouse

Dinesh Moriya [1] , Mrs. Gouri Gosawi[2]

*Corresponding author:

Dinesh Moriya

[1]Asst. Prof. M.L.B.Girls College Bhopal
Department of Computer Science
[2]Asst. Prof. Unique College Bhopal
Department of Computer Science

Abstract

Data warehousing is not about the tools. Rather, it is about creating a strategy to plan, design, and construct a data store capable of answering business questions. Good strategy is a process that is never really finished; A defined data warehouse development process provides a foundation for reliability and reduction of risk. This process is defined through methodology. Reliability is pivotal in reducing the costs of maintenance and support. The data warehouse development enjoys high visibility; many firms have concentrated on reducing these costs. Standardization and reuse of the development artifacts and the deliverables of the process can reduce the time and cost of the data warehouse's creation. In today's business world, data warehouses are increasingly being used to help companies make strategic business decisions. To understand how a warehouse can benefit you and what is required to manage a warehouse, you must first understand how a data warehouse is constructed and established.

Keywords: Data Mining,Assocation Mining,Data werehouseing

Introduction

A data warehouse is a collection of consistent, subject-oriented, integrated, time-variant, non-volatile data and processes on them, which are based on available information and enable people to make decisions and predictions about the future. Over the last years, data warehouses enjoy a lot of attention both from the industrial and the research community. The reason lies in their great importance: making predictions about the (near) future, has always been desirable for business companies. Defined a data warehouse as follows:

Subject-oriented, meaning that the data in the database is organized so that all the data elements relating to the same real-world event or object are linked together;

Time-variant, meaning that the changes to the data in the database are tracked and recorded so that reports can be produced showing changes over time;

Non-volatile, meaning that data in the database is never over-written or deleted, but retained for future reporting; and,

Integrated, meaning that the database contains data from most or all of an organization's operational applications, and that this data is made consistent [5].

Key Planning of Data Warehouse

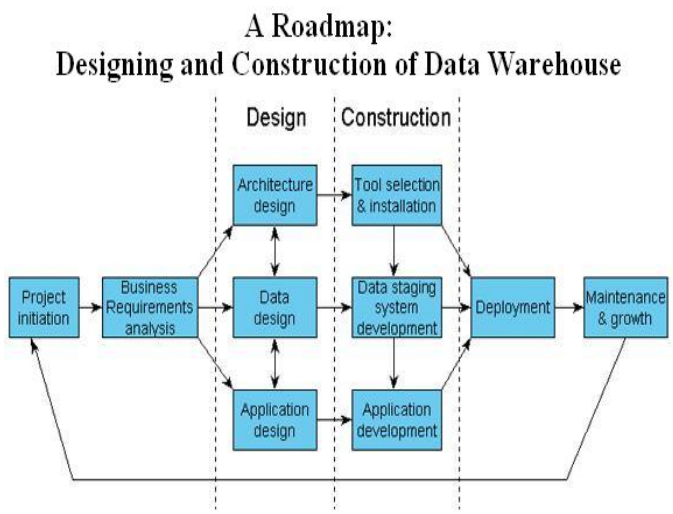
The key planning of data warehouse are identical to the steps for any other type of computer application. Users must be involved to determine the scope of the warehouse and what business requirements need to be met. After selecting a focus area, for example, analyzing the use of state government records over time, a data warehouse team of business users and information professionals compiles a list of different types of data that should go into the warehouse. After business requirements have been gathered and validated, data elements are organized into a conceptual data model. The conceptual model is used as a blueprint to develop a physical database design [3].

Data warehouses are computer based information systems that are home for "secondhand" data that originated from either another application or from an external system or source. Warehouses optimize database query and reporting tools because of their ability to analyze data, often from disparate databases and in interesting ways. In other words, data warehouses are read-only, integrated databases designed to answer comparative and "what if" questions.

Designing and Construction of Data Warehouse

The designing and construction of data warehouse can be summarized as follows:

Project initiation
Requirements analysis
Design (architecture, databases and applications)
Construction (selecting and installing tools, developing data feeds and building reports)
Deployment (release & training)
Maintenance



It is advisable to conduct a *pilot* exercise before embarking on a full-scale development effort. This will include most of the above steps, and provides an opportunity to:
understand new concepts and processes, and identify potential problems;
make more realistic plans and manage expectations;
evaluate alternative tools;
Demonstrate benefits and gain management commitment.
Testing should be an integral part of construction, not a separate step in the development process.

Project initiation

No data warehousing project should commence without:
a clear statement of business objectives and scope;
a sound business case, including measurable benefits;
an outline project plan, including estimated costs, timescales and resource requirements;
high level executive backing, including a commitment to provide the necessary resources;
A small team is usually set up to prepare and present a suitable project initiation document. This is normally a joint effort between business and IT managers. If the organization has limited data warehousing experience, it is useful to obtain external advice at this stage. If the project goes ahead, the project plan and business case should be reviewed at each stage. It is widely regarded as good practice to develop a data warehouse in small, manageable phases. Thus the analysis, design, construction and deployment steps will be repeated in cycles. It is generally a good tactic to provide something that is not already available during the first phase, as this will help to stimulate real interest. This could be new data or enhanced functionality. It is also better to start with something relatively easy, which the warehousing team can deliver whilst still learning the ropes.

Requirements analysis

Establishing a broad view of the business' requirements should always be the first step. The understanding gained will guide everything that follows, and the details can be filled in for each phase in turn [1].
Collecting requirements typically involves 4 principal activities:
Interviewing a number of potential users to find out what they do, the information they need and how they analyze it in order to make decisions. It is often helpful to analyze some of the reports they currently use.
Interviewing information systems specialists to find out what data are available in potential source systems, and how they are organized.
Analyzing the requirements to establish those that are feasible given available data.
Running facilitated workshops that bring representative users and IT staff together to build consensus about what is needed, what is feasible and where to start.

Design

The goal of the design process is to define the warehouse components that will need to be built. The architecture, data and application designs are all inter-related, and are normally produced in parallel.

Data design

This step determines the structure of the primary data stores used in the warehouse environment, based on the outcome of the requirements analysis. It is best to produce a broad outline quickly, and then break the detailed design into phases, each of which usually progresses from logical to physical:
The *logical* design determines what data are stored in the main data warehouse and any associated functional data marts. There are a number of data modeling techniques that can be used to help.
Once the logical design is established, the next step is to define the *physical* characteristics of individual data stores (including aggregates) and any associated indexes required to optimize performance.
The data design is critical to further progress, in that it defines the target for the data feeds and provides the source data for all reporting and analysis applications.

Construction

Warehouse components are usually developed iteratively and in parallel. That said, the most efficient sequence to begin construction is probably as follows:

Tool selection & installation

Selecting tools is best carried out as part of a pilot exercise, using a sample of real data. This allows the development team to assess how well competing tools handle problems specific to their

organization and to test system performance before committing to purchase.
The most important choices are the:
ETL tool
Database(s) for the warehouse (usually relational) and marts (often multi-dimensional)
Reporting and analysis tools
Clearly these need to be compatible, and it is worth checking reference sites to make sure they work well together.
It pays to define standards and configure the development, testing and production environments as soon as tools are installed, rather than waiting until development is well underway. Most vendors are willing to provide assistance with these steps, and this is normally well worth the investment [2].
Create a provisional set of aggregates;
Automate all regular procedures;
Document the whole process.
Substantial time should be allowed to resolve any issues that arise, establish appropriate data cleansing procedures (preferably within the source systems environment) and to validate all data before they are released for live use.

Deployment

It is too often assumed that the first version of a data warehouse can be rolled out in a matter of weeks, simply by showing all the users how to use the new reporting tools.
In practice, training needs to cover not just the basic use of the tools, but also the data that have been made available, and, more significantly perhaps, the new business processes or different ways of working that are intended. This training usually works best if delivered on a one-to-one basis.
As well as training, planning for deployment needs to cover:
Installing and configuring desktop PCs - any hardware upgrades or amendments to the 'standard build' need to be organized well in advance;
Implementing appropriate security measures - to control access to applications and data;
Setting up a support organization to deal with questions about the tools, the applications and the data. However thoroughly the data were checked and documented prior to publication, users are likely to spot anomalies requiring investigation and to need assistance interpreting the results they obtain from the warehouse and reconciling these with existing reports;
Providing more advanced tool training later, when users are ready, and assisting potential power users to develop their first few reports.
If the first users find errors and inconsistencies in the data, don't feel comfortable with the tool or can't be bothered to learn how to use it properly, or won't accept new procedures and responsibilities, all the time spent building the warehouse may ultimately be wasted. The following guidelines will help to reduce these risks:
Do not start deployment until the data are ready (available and validated) and the tools and update procedures have been tested;
Use a small, representative group to try out the finished system before rolling out, including users with a range of abilities and attitudes;
Do not grant system access to users until they have been trained.

Maintenance

A data warehouse is not like an OLTP system: development is never finished, but follows an iterative cycle (analyze – build – deploy). Also, once live, a warehousing environment requires substantial effort to keep running. Thus the development team should not anticipate handing over and moving on to other projects, but to spend half of their time on support and maintenance.
The most important activities are:
Monitoring the realization of expected benefits;
Providing ongoing support to users;
Training new staff;
Assisting with the identification and cleansing of dirty data;
Maintaining both feeds & meta-data as source systems change over time;
Tuning the warehouse for maximum performance (this includes managing indexes and aggregates according to actual usage);
Purging dormant data;
Recording successes and using these to continuously market the warehouse.

Conclusion

Data Warehousing is not a new phenomenon. All large organizations already have data warehouses, but they are just not managing them. Over the next few years, the growth of data warehousing is
going to be enormous with new products and technologies coming out frequently. In order to get the most out of this period, it is going to be important that data warehouse planners and developers have a clear idea of what they are looking for and then choose strategies and methods that will provide them with performance today and flexibility for tomorrow.

References

[1]. http://www.cis.drexel.edu/faculty/song/dolap/dolap99/paper/dolap99_Nectl.pdf

[2]. http://www.sas.com/offices/europe/slovakia/press/newsletters/SNLapril2008/warehouse_architercture_principles.pdf

[3]. http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-28/paper6.pdf

[4]. http://www.ablongman.com/samplechapter/0130813060.pdf

[5]. http://www.mnhs.org/preserve/records/datawarehouses.html