Building a Data Warehouse

Shreyas Gopal
CSE B

Introduction - What is a Data Warehouse

A central repository of historical information a company can analyse and use to gain valuable insight or Business Intelligence.

Data is extracted on a periodic basis from source systems, which are applications such as ERP systems that contain important company info. Data from these systems is moved to a dedicated server that contains a data warehouse. When it is moved it is cleaned, formatted, validated, reorganized, summarized, and supplemented with data from many other sources.

Why Build a Data Warehouse?

- Key Source of Business Intelligence (BI)
- Data Analysis and Mining for trends and patterns
- Security - Limit access to sensitive information in a warehouse.
- Ad-hoc Reporting and End-User Report Generation

Introduction - Building the Warehouse

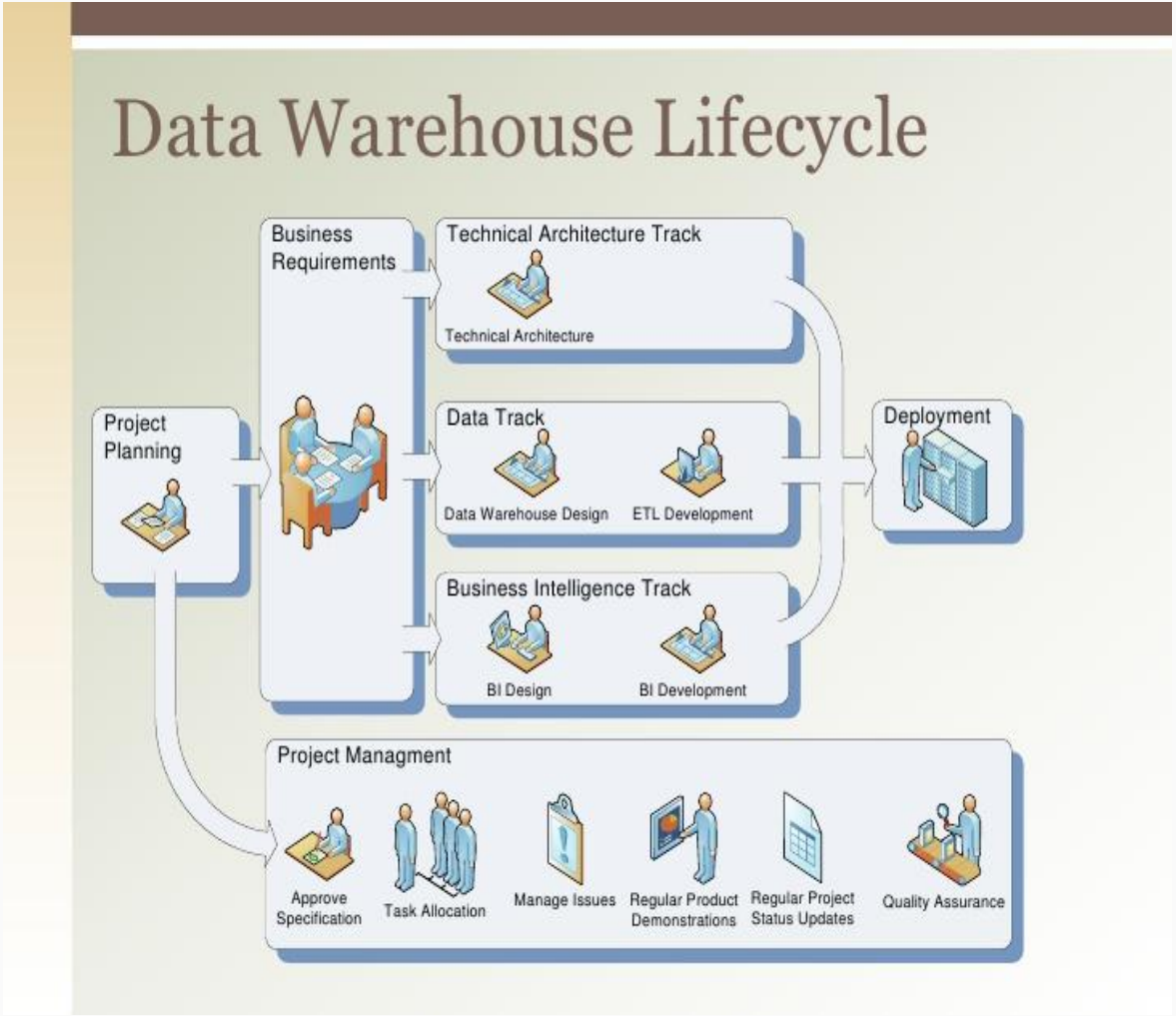
- Building a data warehouse is a continuous iterative process evolving with the organization.

Consists of Two Stages:

1. Conceptual Model construction in accordance with User Demands (Data Warehouse Design).
2. Data Acquisition and Data Access Specification.

Development Cycle

- Must take into consideration the users' requirements concerning reporting and analysing.
- Otherwise it will become a "data jail", and authorized individuals cannot access necessary data.



Development Cycle - Continued

Ideal Steps:

1. One or more business processes are selected.
2. Granularity of individual processes is established.
3. Fact Table dimensions are established.
4. Necessary measures for the fact table are collected.
5. Identify Equivalent entities in different operational systems.

Development Cycle - Continued

There are two approaches to building a data warehouse:

- Top-Down Approach: It is a systemic method which minimizes the integration problems, but is expensive, of long standing and has a low flexibility.
- Bottom-Up Approach: It is a flexible method that allows the organization to go further with lower costs, to build independent data marts and to evaluate the advantages of the new system as they go along.
    - Experimental methodology and difficult to gain results proportional to efforts in early stages.

Development Cycle - Continued

Observing this development process from a Software Engineering perspective we can make use of the advantages of both the above approaches and use one of the following:

1. Waterfall Approach; which requires a structured and systematic analysis at each step, before going forward;
2. Spiral (iterative) Approach; which allows fast generation of more and more developed functional systems.

Development Cycle - Conclusion

The most adequate method for developing a data warehouse is the iterative one. In this approach, more iteration is made, a new version resulting at every iteration. The business subjects are approached one after the other. The method provides a scalable architecture and answers the informational demands of the whole organization. It also allows an efficient management of the users' requirements and reduces the possible risks.

Building the Data Warehouse - Stages

The following are the stages of building a warehouse as I perceive them:

1. Development of a Feasibility Study
2. Business Line Analysis
3. Data Warehouse Architecture Design
4. Selection of the Technological Solution
5. Planning the Project Iterations
6. Detail Designing,
7. Data Warehouse Testing and Implementation,
8. Deployment and Roll-out.

1. Development of Feasibility Study

- Strategic analysis, including the evaluation of organization business lines.
- Feasibility study is presented to some important managers within the organization.
- The roles and responsibilities for all the people involved in the project must also be established in this stage.

2. Business Line Analysis

- Main Purpose is business understanding and business requirement identification.
- Following goals must be achieved:
  - Achieving a global view on organization’s activity and users’ requirements,
  - Establishing the data warehouse scope,
  - Identifying the business directions and purposes,
  - Establishing the priorities of users’ requirements,
  - Establishing the necessary data for solving the requirements,
  - Defining the process for data warehouse iterative population and for data validation according to the business requirements.

2. Business Line Analysis - Continued

- Preliminary Data-Source Audit must be done; to identify necessary data.
- To achieve this purpose, some analyses regarding the following are to be done:
  - the current technological architecture of the organization: computing equipments, operation systems, database management systems, networks, development, tools for communication and data access etc.;
  - the relations between the different systems within the organization and their level of integration;
  - the available documentation, its accuracy and up-to-dateness;
  - data quality and possible extracting tools.

3. Data Warehouse Architecture Design

The architecture is the **logical and physical foundation** the data warehouse is built on.

First, the data warehouse logical architecture is defined. This is the configuration of the required data collections:

- a central repository storing the data of the entire organization,
- an optional operational data store,
- one or more data marts,
- one or more metadata repositories.

3. Data Warehouse Architecture Design - Continued

Once the logical configuration is defined the following architectures need to be designed:

- Data Architecture
- Application Architecture
- Technical Architecture
- Support Architectures needed for data warehouse implementation.

3. Data Warehouse Architecture Design - Continued

- **The Data Architecture:** Has the purpose to organize the data sources and collections and to define the quality and management standards, both for data and metadata.
- **The Application Architecture:** Presents the software components that provide the implementation of the business functionality within the data warehouse

3. Data Warehouse Architecture Design - Continued

- **The Technical Architecture:** Provides the proper infrastructure for data and application architectures.
- **The Support Architecture:** Includes tools for backup/recovery, archiving, performance monitoring, as well as the organizational functions necessary for the technological investment management.

4. Selection of the Technological Solution

The purpose of this stage is to identify the possible tools for implementing data and application architecture and for providing technical and support architecture functions.

At the same time software components are selected: operation systems, databases management systems, development and analysis tools etc

If the data warehouse will have data marts too, besides the relational technology, a client-server architecture is necessary to allow data accessing and multidimensional analysis.

4. Selection of the Technological Solution - Continued

According to their function, these tools can be classified into the following categories:

- data extracting and transforming tools
- data cleaning, data loading and refreshing tools
- data access, security providing tools
- version control and configuration management tools
- database management data backup and recovery, disaster recovery tools
- performance monitoring, data modeling, metadata management tools

5. Planning the Project Iterations

In this stage, the identified business and technical requirements are refined for leading to the proper detail level for data warehouse development and implementation.

Bottom-Line:

A lot of process analysis and review of previous iterations is done to refine the next set of iterations. Documents are rewritten and verified and efficiency is compared to the amount of progress to examine actual work done.

6. Detail Designing

In this stage:

- The data warehouse physical model (database schema) is finalized,
- Metadata is defined,
- Data source list is updated to include all the information necessary for the implementation of that subject.

The data warehouse physical model must respond to the users' informational demands.

7. Data Warehouse Testing and Implementation

Once the planning and design stages are completed, the current iteration for data warehouse implementation may start.

Referring to the specified aspects of the installed database management system, the data warehouse logical and physical design are finalized:

- the physical design of the fact and dimension tables are finalized;
- the most proper indexes are established, taking into consideration the estimated size of the data warehouse and the queries supposed to be performed;
- a decision about table partitioning is taken, knowing that it’s easier to manage a partitioned data warehouse, but its performances are lower.

8. Deployment and Roll-Out

In the operation stage, besides the data warehouse employment by the final users, its maintenance and development are provided too. The IT specialists have to do several specific activities for achieving this purpose:

1. Refreshing Data
2. Statistical Calculations
3. Continuous evaluation of DB Size
4. Database Disaster Recovery

8. Deployment and Roll-Out - Continued

- **Periodical refreshing of the data warehouse.** From time to time, the new changes made in the operational systems have to be loaded into the data warehouse, so that its users could have the most recent information at their disposal. Note: Usually, this process is performed when the operational systems are not in use and it consists in extracting, cleaning, transforming and loading data into the data warehouse.
- **Computing certain statistical indicators to pursue the data warehouse evolution, performance and maintenance.** Some of these indicators are number of queries performed in a certain day, average response time, number of users per day, frequency of using data warehouse subjects, average duration of a work session.

8. Deployment and Roll-Out - Continued

- **Evaluation of the database size.** The data warehouse size increases at every loading operation and can cause serious troubles. There are several techniques that can be used for reducing the negative consequences the increase of the data warehouse volume over certain limits:
  - aggregating the detailed data, and then archiving and deleting it from the data warehouse;
  - limiting the storing interval at a certain period of time, and then archiving and deleting the data;
  - deleting the unused data, which can be identified on the basis of statistical indicators regarding the data warehouse.

8. Deployment and Roll-Out - Continued

- **Database disaster recovery.** The information within the data warehouse have a strategic importance for the managers of the organization. That is why a special attention has to be given to disaster recovery procedures.

Conclusion

From the above 26 slides, it is obvious that the process of data warehouse development and building is a very complex one.

A data warehouse cannot be developed if its goals are not clear and well understood. But at the same time the User Needs and Requirements must also be managed.

Having in mind that, a data warehouse must offer support in the decision making process, only the understanding of all these aspects guarantees the success of a long-term project.

References

- Building a Data Warehouse, Manole Velicanu, http://revistaie.ase.ro/content/42/velicanu.pdf

- Why You Need a Data Warehouse, http://www.jamesserra.com/archive/2013/07/why-you-need-a-data-warehouse/

Thank You