International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 1 Issue 8, October - 2012

# Data warehousing-Implementation, Architecture and Future needs

Gaurav Nama

Gyan chand verma

Garima Ojha(Assistant professor)

## Abstract

*Data warehouses was coined by bill inmon in early 1990,since then data warehouses have been at the forefront of information technology as a way for organisation for effectively use of digital information. A data warehouse is a database use for data analysis and reporting.it is a central repository of data which is created by integrating data from multiple source system. A data ware house stores current and also historical data and can be used for making report for senior management.*

*Data in data ware house are uploaded from uploaded from operational system. The data may pass through an operational data store for some additional operations before they are used in data warehouse for reporting*



## 1. Introduction

The definition of data warehouses focuses on data storage. It is a special database containing large stocks of enterprise data, operational data and its related metadata used by an enterprise or by a multinational company .The data stored in data warehouse can be used to make further strategy and analyse the business by top level of enterprise

The warehouse is a subject-oriented, integrated, time –varient and nonvolatile collection of data. The ware house is a business intelligence tool, tool to extract data, transform and load data into database and to manage it. Thus by its huge benefits a warehouse has now become a much needed and essential part of a big enterprise. Data warehouse maintains its functions in three layers

1. Staging
2. Integration
3. Access

Raw data is stored in staging layer for use by developers for analysis and support. The second layer which is called as integration layer used to integrate data and to provide a level of abstraction from users and the access layer is for getting data out for users.

The main thing about data warehouse is that its scope is the whole organisation and the data warehouses are subdivided into data marts. These data marts stores subsets of data from a ware house. A data warehouse supports cross functional decision support system (DSS) as to manage the business as it provides detail, consistent, normalised, business data for any further manipulation.

To build a warehouse a collection of data from transaction systems was analysed, cleaned and restructured. The data then summarised, and arranged in a format to support and reporting.

The warehouse is refreshed time to time periodically using transaction system as this is the source of data.

So data is taken from transaction system and then converted into warehouse format

Data in warehouse is termed as business data, the three currency features for deciding its currency level are:

1. Current data- It is the view of current data
2. Point –in-time – It is view of data at a particular moment
3. Periodic data- It is the representation of data by periods

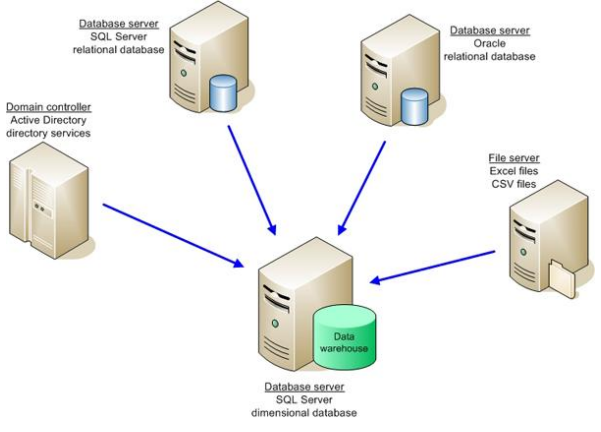


Fig 1. Data warehouse having database server.

## Advantages of data warehouses

1. Integrating different data from multiple sources.
2. Performing some new types of analysis.
3. Reducing historical data access cost.
4. Standerdizing data across the organisation
5. Improving time for analysis
6. Sharing data and allow other to access data
7. Reducing development burden on top management level.
8. Information is stored safely in data warehouse even if source system data is purged over time
9. Warehouses work with conjunction with and can enhance the value of operational business applications,(CRM) Customer Relationship management.
10. Data warehouses can facilitate decision support system (DSS) applications such as trend reports.
11. A data warehouses enhances quality and consistency-It convert data from numerous source system into a common format so there confidency in decisions, since all departments of an enterprise work together with each other so they want data in a common format.
12. A data warehouses high ROI- companies that have their data warehouse have generated more revenue and saved more money and time than companies that haven't invested in data warehouses.
13. There is no limit of extraction of data and access of data.



## Applications of data warehouse:

1. in decision support
2. in trend analysis
3. in any financial forecasting
4. in churn prediction for telecom subscribers and credit card users
5. in analysis of fraud in insurance
6. in analysis of card record
7. in logistic management
8In inventory management
9. Agriculture

## Implementation of data warehouses and its design:

It consists of following steps to be followed strictly and step by step:

**1. Problem definition analysis**
**2. Requirements**
**3. Prototyping and design making.**
**4. Review**
**5. Deployment and training**
**6. Operation**
**7. Enhancement**
**8. Help desk**

## Problem definition analysis:

This is the important and must overlooked step in which all major problems are fined with their solutions. Neglecting this step lead to a disaster in data warehouse. Some of these problems are:

**Do I need a warehouse or not?**

This is to be answer very seriously. It is very simple for large companies but smalls ones have to think on this as it is very costly to make a data warehouse so they can try an alternative if warehouse is not much needed.

**What data I need to store?**

This is an another question which to be answer before making of warehouses in which it is decided that what kind of data we have to store in data warehouse and house much data.

**Take advice from professionals.**

Advice should be taken from any professionals or from watching the status of any other company to decide that if you really needed a warehouse or to focus on an alternative.

## Requirements:

It is very crucial for any success of a project, there is numbers of requirements analysis and the one method to work is as follows:

1. State the problem you are willing to solve.
2. Clearly identify all data sources.
3. Clearly identify the users of completed system.
4. Calculate the specific budget which includes –time, money and personnel.
5. Ask from users what they want from system to do And their requirements.
6. Ask the management persons to specify their success criteria.
7. Separate all the requirements from "desirements" and focus on to design requirements,
8. Group and bubble up all the requirements.
9. Generate a priority wise requirements table listing the all requirement, where it came from, the and

priority. Keep this table at high-level. A table with a dozen requirements can be easier to manage than one with hundreds.
10. Make a detailed development schedule including hardware, software, personnel, documentation, and reviews. Include all outsourcing requirements and long lead-time items.
11. Get a sign-off of all the requirements, resource allocation, and schedule from top management before you go further.

## Prototyping:

There two methods for prototyping:

**1. Rapid prototyping:** For small to medium projects.
**2. Structured prototyping:** For large to complex projects
.
**Rapid prototyping:** There are five methods for this methodology:

1. Make a small team of database programmers, hardware technicians, designers, documentation and a decision support specialist, and a single manager.
2. Make small "focus group" having users (both novice and experienced) and managers. These are the people who will provide the feedback necessary to build the prototyping cycle.
3. Generate a user's manual and then user interface first.
4. Use general tools specifically designed for rapid prototyping. Stay away from C, C++, and SQL or from any other language. Instead of it use the visual development tools included with the database.
5. Remember that prototype is NOT the final application. Prototypes are just meant to be copied into production models. Once if the prototypes are successful, then begin the development process using development tools, such as C, Java, SQL, etc.

**Structured prototyping:**

It is used when project has 10 or more people or when multiple companies are performing the development process; we required a more structured approach
This is a more disciplined approach to the warehouse development process.
In this method we require more document, quality control is more critical and reviews are also increases.

## Development and documentation:

When the requirements analysis is well underway, prototyping is working and focus group are becoming happy, then its time is to begin the development process

Coordinating hardware and software purchases and server and the hardware installation, software and the database development, documentation, guides, manuals, reviews and testing become a job for full time The key to handle all this is to maintain is a good written schedule that every can view.

Working with vendors can be a frustrating experience. Hardware incompabilities, data incompatibilities, software bugs, late delivery are more the norm than the exception. Outsourcing can help you but you should continually involve to insuring success

## Test and review:

This test and review process should take place throughout the development process including prototyping, designing, deployement, enhancement etc. This work is to be given to a single person specially to a trained engineer. Testing is time consuming, tedious work, preparing reviews can take much longer time but finally it lead to success and protect from various problems like budget misconceptions etc.

## Deployment and training:

After ending development, quality assurance, documentation now it's time to start put them together. This can be much time taken process.

Training sessions should be run concurrently with installation work to save time. Each user should trained with good training which makes him to gain knowledge and each time when a user is promoted the user must be trained and each time the enhancement is made training sessions should be scheduled.

## Operation:

There are two or more than two servers in the data warehouses. Work such as backups, bug fixes, software updates, maintenance of hardware, media services, security patches, account maintenance and some other similar tasks should be performed regularly.

Maintenance and operation performed of such services requires an trained staff, as it is not the responsibility of users to take care of it so if any one providing these services then they need an on-site support from either an agency which is outsourced or from a staff from data warehouse

The present trend is to outsource these important services. Almost all the companies are outsourcing the entire data in data warehouse access it via internet or any other private network. This outsourcing result into a substantial savings. The outsourcing agency which works for us should be well working regularly and available when you need they and security requirements should be discussed before them to protect your data.

## Enhancement:

As your users of data warehouses become more sophisticated they will want more and more capabilities such as---

-More successful data warehouse.
-Faster requirements etc.
If you completed their requirements they will again sing your praises so you should have to construct your warehouse scalable flexible at all phases to improve your data warehouse time to time regularly when needed by users.

## Help desk:

This is very important thing I am explaining here as because good manuals and good training is not sufficient for the effective use of your data warehouse. An available, knowledgeable and responsible help desk is required for overall success of your warehouse .as users will always find other users for a well-designed system and problems will may occur. We can't imagine a warehouse without a help desk as without it, it becomes dated and under-utilized

In my opinion there should be a help desk for every warehouse with capabilities of following:

- Telephone
- Fax
- E-mail

Help desk ensures the continued success of a data warehouse make your company in competition with other.

## Architecture of data warehouses:

Data warehouses architecture varies depending upon species of an organisation these are of 3 types:

1. Data warehouse architecture (basic).
2. Data warehouse architecture (with staging area).
3. Data warehouse architecture (with a staging area and data marts).

## Data warehouse architecture (basic):

In these there are 3 main things
-data sources (operational system and flat files).
-warehouse (metadata, summary data, and raw data)
-users (analysis, reporting and mining)
In these architecture end uses directly access data derived from several sources system through data warehouse
A metadata used in warehouse is used for taking business data to data warehouse
Summary data is used for pre-compute long operations in advance.

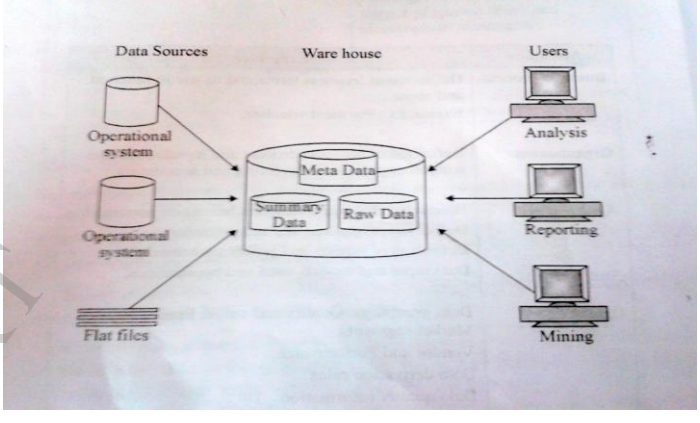


Fig 2. Data warehouse architecture (basic).

## Data warehouse architecture (with staging architecture):

It is used to clean and process your operational data before putting it into the warehouse; this is done programmatically in a staging area according to this architecture.
Almost all warehouses use a staging with this warehouse. A staging area simplifies buildings summaries and general warehouse management
So this architecture has mainly 4 things
-data sources (operational data and flat files).
-staging area (where data sources go before the warehouse).
Warehouse (metadata, summary data, and raw data).
-users (analysis, reporting, mining).
The main advantage of this architecture from previous one is that data entered in warehouse in this architecture is more clean and understandable for users.

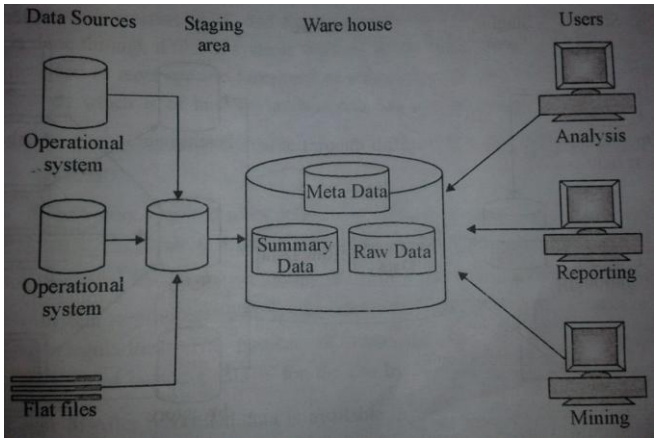


Fig 3. Data warehouse architecture (with staging area).

## Data warehouse architecture (with staging area and data marts):

In this warehouse architecture you can customize your warehouse for different groups within your organisation

You can do this by adding data marts in architecture, which are systems designed for particular line of business such as a purchasing, sales, and inventories are separated, as all the users have different work, so they just directly connect to their related data marts which stores subset of data from entire warehouse.
So this architecture mainly consists of 5 main things-
-data sources (operational systems and flat files).
-staging area (where data sources go before the warehouse).
-warehouse (metadata, raw data).
-data marts (like purchasing, sales, inventory).
-users (analysis, reporting, and mining)

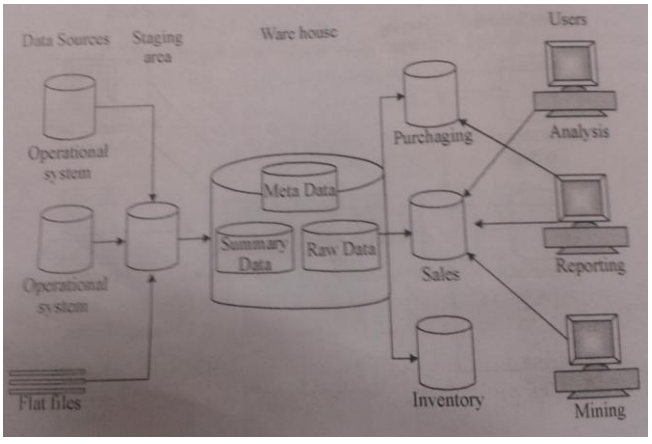


fig 4. Data warehouse architecture (with staging area and data marts)

## Disadvantages of data warehouses:

1. Data warehouses are not optimal for environment for some unstructured data.
2. As because data must be extracted, transformed and loaded into data warehouses there us an element of latency in data of data warehouses
3. Implementation of data warehouses take much time.
4. Data warehouses can have high costs of making, so companies with low budget can't able to have its own warehouses.
5. Data warehouses are can update much quickly. So there is a cost of delivering sub optional information to the organisation.
6. There is affine between warehouses and operational system. So duplicate, expensive functionality may be developed.
   Else functionality may be developed in warehouses that should have been developed in operational system
7. If any small companies make their own warehouse ,then it can make its business harder to deal with

## Future needs from data warehouses:

In this paragraph I am sharing my ideas to improve the data warehouse.

1. As we already know that it takes much costs to design a data warehouse for a company so many low and middle level company can't able to make their own data warehouse ,this is a major reason that many companies are falling in business competition

According to me there can be a solution to this answer that there can be a private firm with enough budget to make a big data warehouse in which many low and middle class company can store their summary and historical data for analysis and to survive in business competition .

These small companies can pay rent to store their business data to that private firm rather than to make whole personal data warehouse which is not in its budget.

So i think that this could be a good step for the wellness of small budget companies.

2. As I already discus earlier in the disadvantages of data warehouses that I take much time to make a warehouse, so new methods should to be developed to shorten the implementation time for data warehouses.

3. At this present time the storage of data warehouse is disk storage, since from the past twenty years. But the future of the storage of data warehouse is a storage media known as alternative storage instead of disk storage.
There can be two forms of alternative storage for warehouse

   - Near line storage

   -Secondary storage

   Near line storage is a siloed tape storage where siloed catridges of tape storage are managed robotically. This method has proven it a mature technology.

   Secondary storage is a type of disk storage whose disk is slower, less expensive than high performance disk storage.

   There are lot of reason for choosing alternative storage as storage for warehouse.

   The one reason is that queries that operate on warehouse data need long stream of data and also data is stored sequentially. It is unlike a job stream for online processing where a constant demand for different units of data from different parts of disk device, in warehouse processing that occurs is fundamentally different. Near line and secondary fits nicely for this model of job stream.

   The another big reason for having an alternative storage is that in a warehouse there is a lot of data including summary and historical data is stored far more than online data, so an alternative storage (near line or secondary storage) has ability to store far more data so alternative storage is the real future of warehouse.

   The greatest reason for choosing an alternative storage is that in this user can choose lowest level of granularity desired for warehouse. If high performance disk is used as only medium for storing data in warehouse, then the designer of warehouse ends up being restricted as how much data can be placed in data warehouse.
   But in alternative storage designer can store data at lowest level of detail that can exists.
   So this is another reason which confirms alternative storage as storage for data warehouse.

   In order to make alternative storage used at optimal level two types of software are required.

   First one is needed is that of activity monitor. This activity monitor sits between warehouse and users of warehouse and collects info about activity that is occurring inside warehouse.

   Second software is needed for environment for warehouse that operates on alternative storage is software that is called a cross media storage manager. Its work is to manage the traffic between the activities used storage and alternative storage.

   According to the rule the activity manager is first used to determine how much data is used to store in alternative storage then decision is taken to place data in alternative storage, cross media manager and storage is purchased and stored.

## References:

[1] www.wikipedia.com

[2] www.horsburgh.com

[3] www.ventechsolutions.com

[4] Management information system by Richa Sharma and Antima Saxena