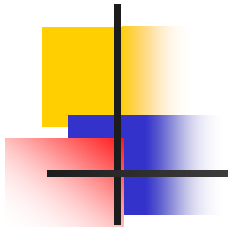
UNIT-III

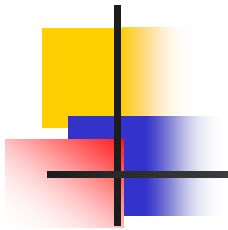
Design of DW

Project Planning, Management and Requirement (Data Warehouse Design)



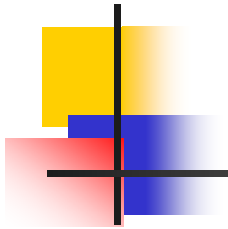
Introduction

- Data warehouses are **C**omputer **B**ased **I**nformation **S**ystem.
- That are home for **“second hand”** data that originated from either **another application** or from an **external system or source**.
- Warehouse **optimize database query and reporting tools** because of their **ability to analyze data from disparate databases and in interesting ways.**



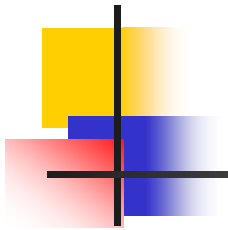
Cont…

- They are the way for managers and decision makers to extract information quickly and easily in order to answer questions about their business.
- In other words data warehouses are **read-only** and integrated database designed to answer comparative and “what if” questions.



Cont...

- Unlike operational databases that are set up to handle transactions and that are kept up to date as of the last transactions.
- Data warehouses are analytical, subject-oriented, and are structured to aggregate transactions as a snapshot in time.



Planning a Data Warehouse

- Data Acquisition/Collection
- Metadata
- Data Marts
- Trustworthiness/Security

Physical Structure of Data Warehouse

- There are three basic architectures of constructing data warehouse.
- Centralized
- Distributed
- Federated
- Tiered

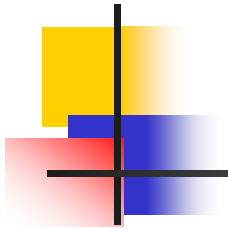
Conceptual Modeling of Data Warehouse

- There are three basic models:
- Star Schema
- Snowflake schema
- Fact constellation

Data Warehouse Design

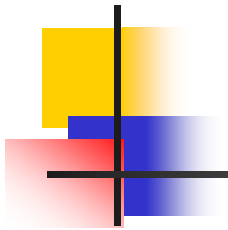
- For designing a data warehouse
- Starting with the **need** and **motivation,** the sequential steps involved for developing a data warehouse are.....

1. **Content in a data warehouse**
2. **Hardware**
3. **Networking infrastructure requirement for the data warehouse**
4. **The soft environment and various Tools**



Cont...

5. The issues involved in **multiprocessor architecture** and **I/O devices** have been discussed.
6. The **range of tools available** in the market with their **mutual compatibility** requirements and their features and capabilities.

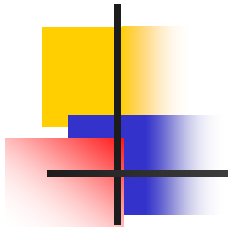


Design of Data Warehouse

- The job of designing and implementing a data warehouse is very challenging and difficult.
- There are so many questions while designing a data warehouse.
- Where to start?
- Which data should put first?
- Where is the data available?
- Which query should be answered?

Data Warehouse Design Process

- Top-down, bottom-up approaches or a combination of both
  - **Top-down:** Starts with overall design and planning (mature)
  - **Bottom-up:** Starts with experiments and prototypes (rapid)
- From software engineering point of view
  - Waterfall: structured and systematic analysis at each step before proceeding to the next
  - Spiral: rapid generation of increasingly functional systems, short turn around time, quick turn around
- Typical data warehouse design process
  - Choose a business process to model, e.g., orders, invoices, etc.
  - Choose the *grain* (*atomic level of data*) of the business process
  - Choose the dimensions that will apply to each fact table record
  - Choose the measure that will populate each fact table record

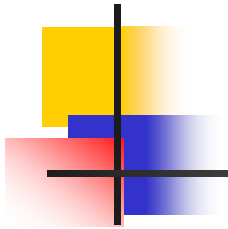


Cont…

- How would you bring down the scope of the project to something smaller and manageable?
- How it would be scalable to gradually upgrade to build a comprehensive data warehouse environment?
- **The recent trend is to build data mart before a large data warehouse is built.**
- People want something smaller, so as to get manageable**.**

Cont…

- The recent trend is to build **data marts** before a real large data warehouse.
- A data warehouse can be built either on a **bottom-up or a top-down approach.**
  - **Top-down: Starts with overall design and planning (mature)**
  - **Bottom-up: Starts with experiments and prototypes (rapid)**
- **Top-Down:** We can design a global data warehouse for an entire organization and split it up in to individual **data marts** or **sub data warehouses** dedicated for individual departments.
- **Bottom-up:** Alternatively individuals **data marts** can be built for each department and finally they all get integrated in to a **central data warehouse.**



Cont…

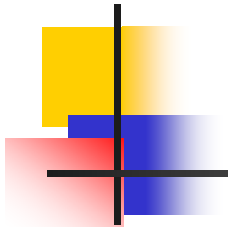
- The **bottom-up** approach is more realistic but the integration of individual data mart should be made easier with advanced planning and preparation.
- In building a data warehouse, the organization has to make arrangements for information flow from various Internal Information Systems and databases as well as from internal sources of information.

Cont...

- This requires close involvement of the users in identifying the information requirements and identifying sources of the same from both internal and external data sources and information system in time.

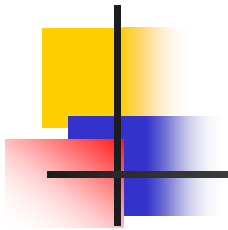
Requirement
(Important steps to design a Data Warehouse)

1. Choose a subject matter.(one at a time)
2. Decide what the fact table represents
3. Identify and confirm the dimensions
4. Choose the facts
5. Store pre-calculations in the fact table
6. Define the dimensions and table
7. Decide the duration of the database and periodicity of updation
8. Track slowly the changing dimensions
9. Decide the query properties and the query methods



Cont…

- All the above steps are required before the data warehouse is implemented.
- The final step 10 is to implement a simple data warehouse or data mart.
- **First only the data marts are identified, designed and implemented.**
- **Secondly the data warehouse will come out gradually.**



Implementation

- A data warehouse can not be purchased and installed.
- Its implementation requires the integration of many products.

Followings are the steps of the Data Warehouse Implementation

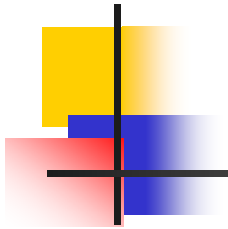
Step 1. Collect and analyze business requirements

Step 2. Create a data model and physical design for the data warehouse after deciding the appropriate hardware platform.

Step 3. Define the data sources.

Step 4. Choose the DBMS and software platform for data warehouse.

Step 5. Extract the data from operational data sources, transform it , clean-up and load in to the data warehouse model or data mart.



Cont…

- **Step 6.** Choose database access and reporting tools.
- **Step 7.** Choose database connectivity software.
- **Step 8.** Choose data analysis (OLAP) and presentation software (client GUI).
- **Step 9.** Keep refreshing the data warehouse periodically

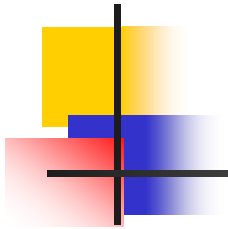
Detailed of step-1 to step-9

- **Step-1:**
- Choosing the subject matter is the most crucial decision.
- This will emerge after the user interaction and interviews.
- The most cost effective, highly demanded subject should be taken first.
- The subject should be such that it can answer the user's business questions, with the data source being available.

Major Issues in Data Warehouse

- There are **three major issues** that will be faced in data warehouse development.

1. Heterogeneity of data source requiring substantial efforts in data conversion.(data source to data conversion)
2. Maintaining timeless and high-quality levels of data integrity, reliability and authenticity.
3. Further data may be quit old and historical, while old data of past is essential for a data warehouse, but it will be relevant and useful in the data warehouse form.



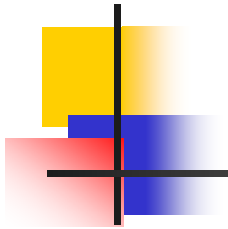
Cont...

3. Another issue is the tendency of the data warehouse to grow very large.

   So discrete decisions should be made by the designer of the data warehouse in limiting the size of the warehouse

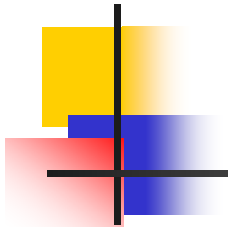
Implementing a Data Warehouse

- **Designing and rolling out a DW is a complex process consisting of the following activities:**
  - Define the architecture, do capacity planning and select the storage servers, database and OLAP servers (ROLAP vs MOLAP) and tools,
  - Integrate the servers, storage and client tools,
  - Design the warehouse schema and views,
  - Define the physical warehouse organizations, data placement, partitioning and access method,



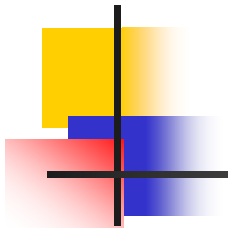
Cont...

- Connect the sources using gateways, ODBC drivers,
- Design and implement scripts for data extraction, cleaning, transformation, load, and refresh.



Case study

- **Design a data warehouse for the following:**
  - World Bank
  - Government of Chhattisgarh
  - Hewlett-Packard
  - BSNL



Assignment-2

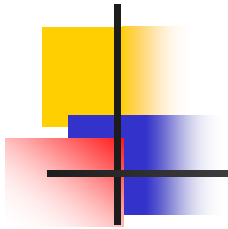
- Explain the data warehousing design using Oracle.

Data Warehousing design using Oracle

- Oracle 10g Data Warehousing enhancements include:
- an increase in the size limits of the database to support ultra-large databases of millions of terabytes in size and ultra-large files of terabytes in size.
- Improvements to Real Application Clusters (RAC) enable resources to be allocated automatically and means that operational data can be used immediately without the need to copy it to another database.

Cont...

- Enhancements to OLAP analytic, a data-mining GUI and a new SQL model allow query results to be treated as sets of multi-dimensional arrays on which complex inter-dependent operations - such as forecasting - can be run without the need to extract data to spreadsheets or perform complex joins and unions on the data.
- A new changed data capture facility based on Oracle Streams provides low or zero latency trickle feeds that combined with integrated extraction, transformation and loading (etl) enable real-time warehousing.



Cont...

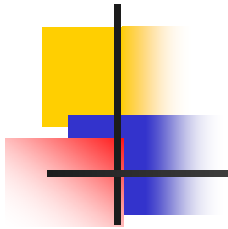
- The new features and enhancements for Oracle 10g are geared towards **grid computing** which is an extension of the **clustering features** (**R**eal **A**pplication **C**lusters) introduced with Oracle 9i.

Data warehouse design using Oracle

- A traditional data warehouse is all-about providing a **vehicle for reporting from summary and aggregate information (using de-normalized tables, summarized tables and materialized views).**
- Most data warehouse designers replicate the data warehouse summary data onto another instance to **avoid contention with the OLTP database**,
- This depends on the traffic on your system and the ability of your server to handle additional load (i.e. SMP processor capability).

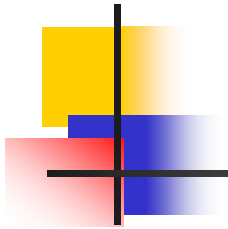
Cont...

- There are several data warehouse design factors:
- 1 - **Data Warehouse Query Performance** – A data warehouse pre-summarizes and pre-aggregates the OLTP data so that the queries can fetch the result sets with only a few data block touches.
- Make sure that your OLTP server has enough CPU resources to support Oracle parallel query, as you will need it to roll-up your summaries and aggregates.



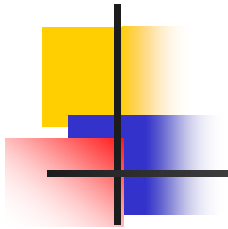
Cont…

- 2 - **Data warehouse schema design** –
- If your existing summary tables do not require joins into other OLTP tables, then you will not benefit from a star transformation approach.



Cont…

- Oracle 10g Data Warehousing is a guide to using the Data Warehouse features in the latest version of Oracle, Oracle Database 10g.
- Designed and implemented the code and by people with industry experience implementing warehouses using Oracle technology
- Oracle Database 10g software is best used for any application.

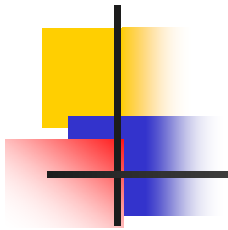


Cont...

- The new features of Oracle Database 10g and other Oracle products are used in the data warehouse.
- how to deploy the Oracle database and correctly use the new Oracle Database 10g features for your data warehouse.
- How to use tools such as Oracle to Discoverer and Reports to query the warehouse and generate reports that can be deployed over

Cont...

- the web and gain better insight into your business.
- This provides step by step instructions including screen captures to make it easier to design, build and optimize performance of the data warehouse or data mart.
- It must have reference for database developers, administrators and IT professionals who want to get to work now with all of the newest features of Oracle Database 10g.



IMPORTANT QUESTIONS

- What are the various steps to design a data warehouse?
- Give the various steps for data warehouse implementation.
- Explain the various issues faced in data warehouse development.
- Explain the concept of DW to the web and web to the DW.
- Write short notes on growth and maintenance of the data warehouse.