Chapter 3:

Data Warehouse

What is Data Warehouse?

- **Data warehouse** is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process.
- **Data warehousing** is the process of constructing and using data warehouses.
- **Data warehouse** allows "**knowledge workers**" (such as **managers**, **analysts**, and **executives**) to use the warehouse to quickly and conveniently obtain an overview of the data and to make **sound decision** based on information in the warehouse

The goal of a data warehouse

- The **goal** of a data warehouse is **to support decision making with data.**

Data Warehouse: Subject-Oriented

- Organized around major subjects, such as customer, product, sales.
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.
- Constructed by integrating multiple, ***heterogeneous data sources***

Cont…

Data Warehouse: Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems.
  - Operational database: current value data.
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key operational data may or may not contain “time element”.

Cont…

Data Warehouse: Non-Volatile

- A physically separate store of data transformed from the operational environment.
- Operational update of data does not occur in the data warehouse environment.
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing:
    - **initial loading of data** and **access of data**

Data Warehouse vs. Operational DBMS

- **DBMS - tuned for OLTP (Online Transactional Processing):**
  - access methods, indexing, concurrency control, recovery mechanism are desirable
- **Warehouse - tuned for OLAP:**
  - Complex OLAP queries, multidimensional view, consolidation are desirable.
- Indexing, concurrency control, recovery mechanism are not desirable in warehouse

Cont…

- OLTP and OLAP differs in
  - User and system orientation
  - Data contents they operate
  - Database design used
  - View
  - Data Access patterns

Design of a Data Warehouse

- The basic steps involved in the design process of data warehouse mainly involves business analysis framework which give clear understanding of what can a business analyst gain from having a data warehouse?
- Some of the gains may include:
- Provide a competitive advantage by presenting relevant information
- Enhance business productivity as it enables to quickly and efficiently gather information that accurately describe the organization
- Facilitate customer relationship management by providing consistent view of customers and items across all lines of business, all departments and all markets

Cont…

- As data warehouse can be seen from various views.
- In data warehousing literature, an n dimensional (n-D) cube is called **a base cuboid.**
- Base cuboid shows some information about every attribute at most refined granularity.
- The top most 0-D cuboid, which holds the highest-level of summarization, is called **the apex cuboid**

Concept Hierarchy

- **Dimensions** are organized into concept hierarchies. **A concept hierarchy** defines a sequence of mappings from a set of low-level concepts to higher-level and more general concepts. As shown in the concept hierarchy, each level refers to values of some type. The type of hierarchy define ordering which can be partial ordering or total ordering.

Cont…

- Location dimension can be seen as a total ordering
- ***continent → country → Region → Zone → city → kifle ketema → kebele***
- Time dimension shows partial ordering
- ***second → minute → hour → day → {month →quarter, week} →year***

Typical OLAP Operations

- In multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple level of abstraction defined by concept hierarchies. This organization provides users with flexibility to view data from different perspectives.
- Different OLAP data cube operations exist to materialize these views:
- Roll up (drill-up)
- Drill down (roll down)
- Slice and dice
- Pivot (rotate)

Cont…

Roll up (drill-up)

- Data is summarized with increasing generalization (for example, weekly to quarterly to annually).
- For example: from cities to countries, form second to minute.

Drill down (roll down)

reverse of roll-up: from higher level summary to lower level summary or detailed data, or introducing new dimensions. For example: from region to town, from year to month.

Slice

- performs selection on one dimension of a given cube resulting in a sub-cube (say time = Q1 from different dimensions like time, location and item).

Cont…

Dice

- performs defines a sub cube by performing a selection on two or more dimension.
- Dice for(location=”Ambo” or “ Wolliso” and Time =”Q1” or “Q2”, Item=”Mobile”, or “Compuer”)
- From (location (Ambo, Wolliso, Ginch, Guder, Bako), Time=(quarters(Q1,Q2,Q3,Q4)),Item(mobile, computer, stabilizer, divider, cable)).

Pivot (rotate)

- *reorient the cube, visualization, 3D to series of 2D planes.*

Data Warehouse Design Process

- Data warehouse design process consists of 4 steps
- Choosing a **business process** to model
- Choosing **the dimensions** that will apply to each fact table record
- Choosing the **grain (atomic level of data)** of the business process that will be represented in the fact table
- Choosing the **measure** that will populate each fact table record

Three Data Warehouse Models

- there are three data warehouse models described as

Enterprise warehouse,

- collects all information about subjects that span the entire organization (*customers*, *products*, *sales*, *assets*, *personnel*).

Data Mart,

- Its scope is confined to specific, selected groups

Virtual warehouse

- a set of views over operational databases

Characteristics of Data Warehouse

- distinctive characteristics of Data warehouses are,
- Multidimensional conceptual view
- Generic dimensionality
- Unlimited dimensions and aggregation levels
- Unrestricted cross-dimensional operations
- Dynamic sparse matrix handling
- Client-server architecture

Cont…

- Multiuser support
- Accessibility
- Transparency
- Intuitive data manipulation
- Consistent reporting performance
- Flexible reporting

Difficulties of Implementing Data warehouse

- Difficulties or challenging points of Data warehouse are,
- Project management
- The administration of a data warehouse
- quality control of data.