SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A++'
Grade Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER APPLICATIONS

19CAE730 – Fundamentals of NOSQL database System
II YEAR III SEM

UNIT III – Data warehousing schemas

What is a Data Warehouse?

A **Data Warehouse** is a collection of data gathered and organized so that it can easily by analyzed, extracted, synthesized, and otherwise be used for the purposes of further understanding the data.

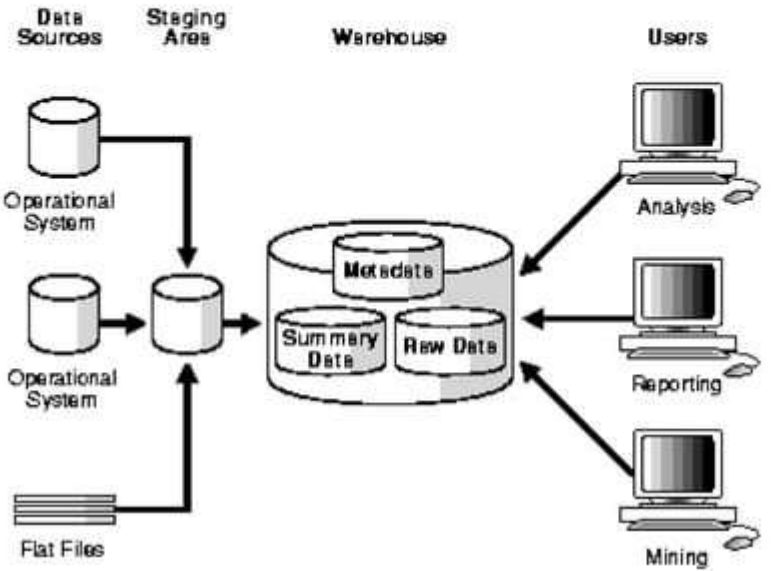
A **Data warehouse** is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process

THE ARCHITECTURE

A typical Data warehouse consists of

- Source
- Staging Area
- Data warehouse/ Data Mart
- End User

Data Warehouse Architecture (with a Staging Area)



General Stages of Data Warehouse

- Off line Operational Database
  - Data warehouses are developed by copying the data off an operational system to another server where the processing load of reporting against the copied data does not impact the operational system's performance.
- Off line Data Warehouse
  - Data warehouses are updated from data in the operational systems on a regular basis and the data warehouse data is stored in a data structure designed to facilitate reporting.

General Stages of Data Warehouse

- Real Time Data Warehouse
  - Data warehouses at this stage are updated every time an operational system performs a transaction
- Integrated Data Warehouse
  - Data warehouses at this stage are updated every time an operational system performs a transaction. The data warehouses then generate transactions that are passed back into the operational systems.

Types of Data Warehouse

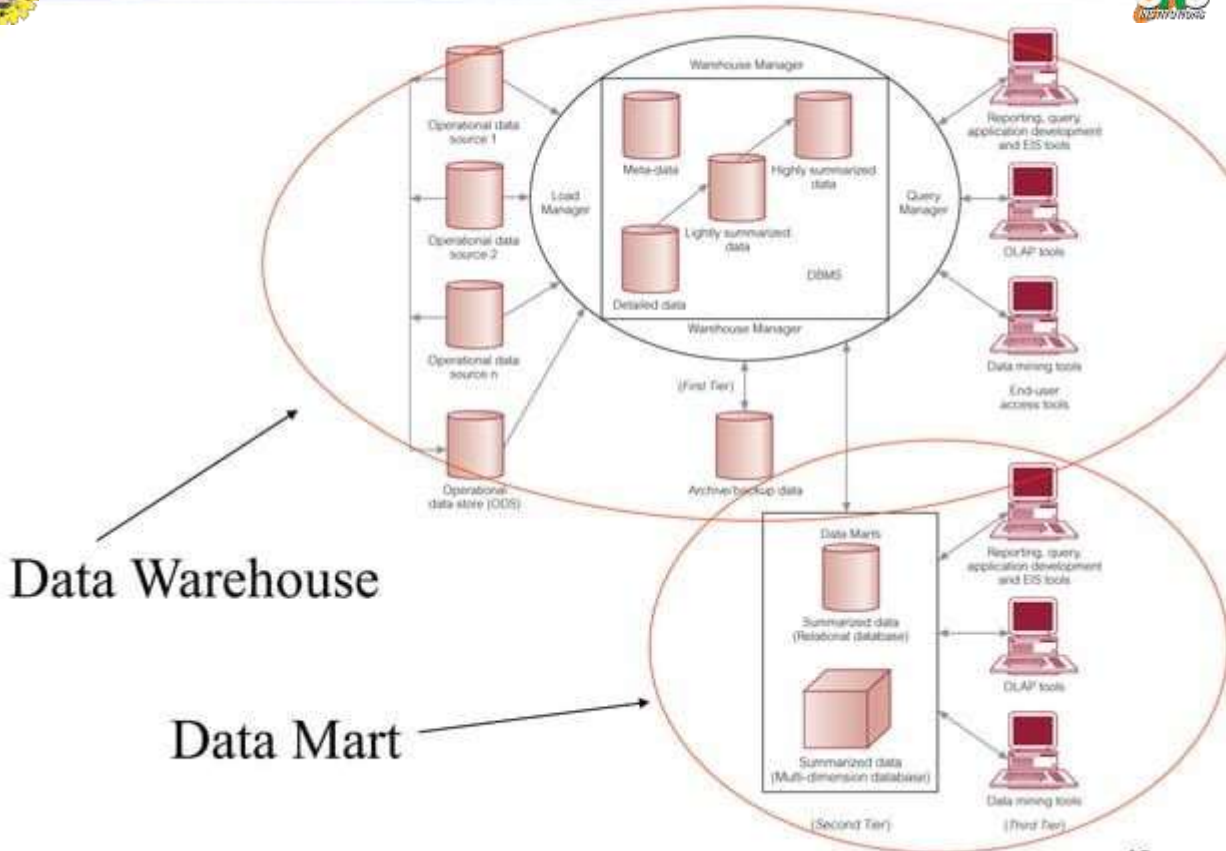
- Enterprise Data Warehouse
  - provide a control Data Base for decision support through out the enterprise.
- Operational data store
  - has a broad enterprise under scope but unlike a real enterprise DW. Data is refreshed in rare real time and used for routine business activity.
- Data Mart
  - is a sub part of Data Warehouse. It support a particular reason or it is design for particular lines of business.

Data Mart

- A data mart is a smaller version of a data warehouse
  - Usually containing data related to a single unit of an organization
- Usually data mart focuses on the requirements of only one department or business function of a company
- Data mart can be a useful first step to a full-scale data warehouse

Why creating Data Mart?

- To provide data for users in a single department
- To improve end-user response time due to the reduction in the volume of data
- Building a data mart is simpler compared with establishing a corporate data warehouse
- The cost of implementing data marts is normally less than implementing a data warehouse
- Users of a data mart are more clearly defined and can be more easily targeted


Warehouse Manager
Operational data source 1
Operational data source 2
Operational data source n
Operational data store (ODS)
Load Manager
Meta-data
Highly summarized data
Lightly summarized data
Detailed data
DBMS
Warehouse Manager
Query Manager
Reporting, query, application development and EIS tools
OLAP tools
Data mining tools
End-user access tools
(First Tier)
Archive/backup data
Data Warehouse
Data Marts
Summarized data (Relational database)
Summarized data (Multi-dimension database)
Reporting, query, application development and EIS tools
OLAP tools
Data mining tools
(Second Tier)
(Third Tier)
Data Mart

Benefits of data warehousing

- A data warehouse provides a common data model for all data of interest regardless of the data's source
- Prior to loading data into the data warehouse, inconsistencies are identified and resolved.
- Information in the data warehouse is under the control of data warehouse users so that, even if the source system data is purged over time, the information in the warehouse can be stored safely for extended periods of time.

Benefits of data warehousing

- Data warehouses can work in conjunction with and, hence, enhance the value of operational business applications, notably customer relationship management (CRM) systems.
- Data warehouses facilitate decision support system applications such as trend reports, exception reports, and reports that show actual performance versus goals.
- Data warehouses provide retrieval of data without slowing down operational systems.

Disadvantages of data warehouses

- The data warehouse is usually not static. Maintenance costs are high.
- Data warehouses can get outdated relatively quickly.
- There is often a fine line between data warehouses and operational systems. Duplicate, expensive functionality may be developed. Or, functionality may be developed in the data warehouse that, in retrospect, should have been developed in the operational systems and vice versa..

Business intelligence

- **Business intelligence** (**BI**) refers to skills, technologies, applications and practices used to help a business acquire a better understanding of its commercial context. Business intelligence may also refer to the collected information itself.
- BI applications provide historical, current, and predictive views of business operations. Common functions of business intelligence applications are reporting, OLAP, analytics, data mining, business performance management, benchmarks, text mining, and predictive analytics.

Types of business intelligence tools

- Spreadsheets
- Reporting and querying software
- OLAP
- Digital Dashboards
- Data mining
- Process mining
- Business performance management

Data mining

- **Data mining** is the process of extracting hidden patterns from large amounts of data. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery.
- As more data is gathered, with the amount of data doubling every three years, data mining is becoming an increasingly important tool to transform this data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery.
- Data mining can be applied to data sets of any size.

Data integration

- Data integration is the process of combining data residing at different sources and providing the user with a unified view of these data .
- Data integration appears with increasing frequency as the volume and the need to share existing data explodes.
- It has been the focus of extensive theoretical work and numerous open problems remain to be solved. In management practice, data integration is frequently called Enterprise Information Integration.

OLAP & OLTP

OLAP	OLTP database
Designed for analysis of business measures by category and attributes.	Designed for real time business operations.
Optimized for bulk loads and large, complex, unpredictable queries that access many rows per table.	Optimized for a common set of transactions, usually adding or retrieving a single row at a time per table.
Loaded with consistent, valid data; requires no real time validation.	Optimized for validation of incoming data during transactions; uses validation data tables.
Supports few concurrent users relative to OLTP.	Supports thousands of concurrent users.

Analysis Services

A middle-tier server for OLAP and data mining; manages multi-dimensional cubes of data for analysis and provides rapid client access; allows you to create data mining models from both OLAP and relational data sources

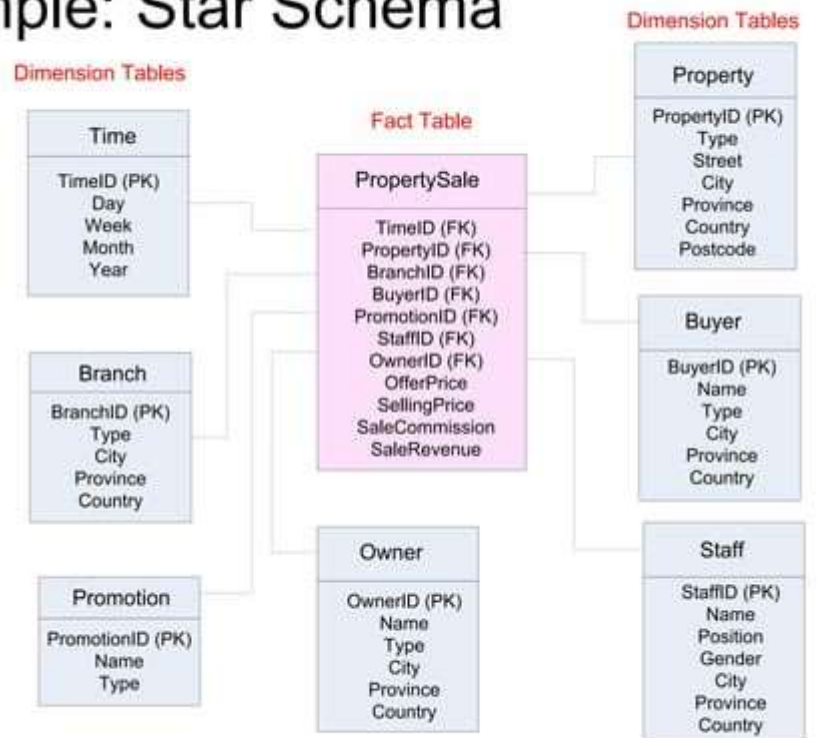
Data Warehousing Schemas

- Star Schemas
- Snowflake Schemas

Star Schema

- The **star schema** is perhaps the simplest data warehouse schema. It is called a star schema because the entity-relationship diagram of this schema resembles a star, with points radiating from a central table. The center of the star consists of a large fact table and the points of the star are the dimension tables

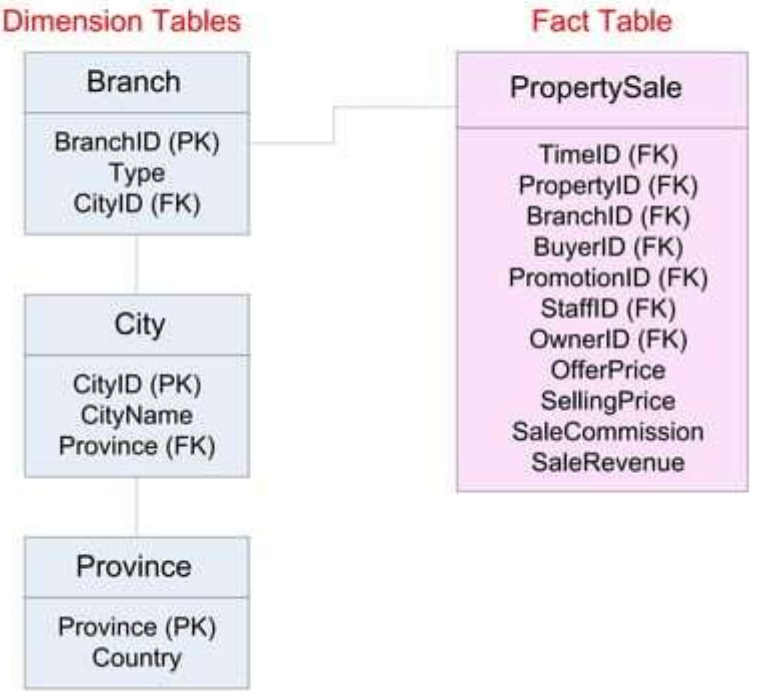
Example: Star Schema



Snowflake Schema

- Snowflake schema is a variant of the star schema where dimension tables contain normalized data
  - e.g. ‘city’ and ‘province’ can be splitted as separated tables to normalize dimension tables
  - ‘Starflake’ schema is a hybrid structure that contains a mixture of star (denormalized) and snowflake (normalized) schemas
    - This allows dimension tables to be present in both forms for different query requirements

Example: Snowflake Schema



Important aspects of Star Schema & Snow Flake Schema

- In a star schema every dimension will have a primary key.
- In a star schema, a dimension table will not have any parent table.
- Whereas in a snow flake schema, a dimension table will have one or more parent tables.
- Hierarchies for the dimensions are stored in the dimensional table itself in star schema.
- Whereas hierarchies are broken into separate tables in snow flake schema. These hierarchies helps to drill down the data from topmost hierarchies to the lowermost hierarchies.

Dimension Modeling

- Dimensional modeling is a technique for conceptualizing and visualizing data models as a set of measures that are described by common aspects of the business.
- It is especially useful for summarizing and rearranging the data and presenting views of the data to support data analysis.
- Dimensional modeling focuses on numeric data, such as values, counts, weights, balances, and occurrences.

Multidimensional Data Model (MDM)

- Focused on a collection of numeric measures.
- Each measure depends on a set of dimensions
- Data are presented as multidimensional array

Unified Dimensional Model (UDM)

- The role of a Unified Dimensional Model (UDM) is to provide a bridge between the user and the data sources.
- A UDM is constructed over one or more physical data sources, and then the end user issues queries against the UDM using one of a variety of client tools.
- Advantages of creating UDM
  - More readily understood model of the data
  - Isolation from heterogeneous backend data sources
  - Improved performance for summary type queries.

Basic Concepts

- Dimensional modeling has several basic concepts:

  · Facts
  · Dimensions
  · Measures (variables)

Fact Table

- Contain numeric measures of the business
- Contains facts and connected to dimensions
- two types of columns
  - facts or measures
  - foreign keys to dimension tables
- May contain date-stamped data
- A fact table might contain either detail level facts or facts that have been aggregated

Steps in designing Fact Table

- Identify a business process for analysis(like sales).
- Identify measures or facts (sales dollar).
- Identify dimensions for facts(product dimension, location dimension, time dimension, organization dimension).
- List the columns that describe each dimension.(region name, branch name, region name).
- Determine the lowest level of summary in a fact table(sales dollar).

Types of Facts (Measures)

- **Additive** - Measures that can be added across all dimensions.
- **Semi Additive** - Measures that can be added across few dimensions and not with others.
- **Non Additive** - Measures that cannot be added across all dimensions.

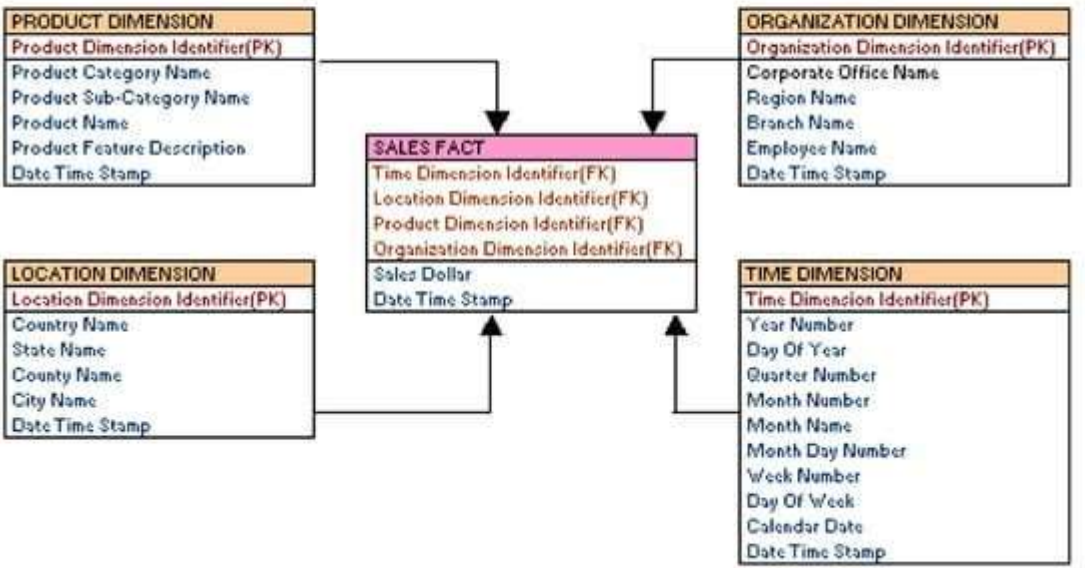
Dimension Tables

- A dimension is a structure, often composed of one or more hierarchies, that categorizes data. Dimensional attributes help to describe the dimensional value. They are normally descriptive, textual values. Several distinct dimensions, combined with facts, enable you to answer business questions. Commonly used dimensions are customers, products, and time.

Dimension Tables

- Contain textual information that represents attributes of the business
- Contain relatively static data
- Are joined to fact table through a foreign key reference
- Are usually smaller than fact tables

Examples for Dimensions and Facts



MEASURES

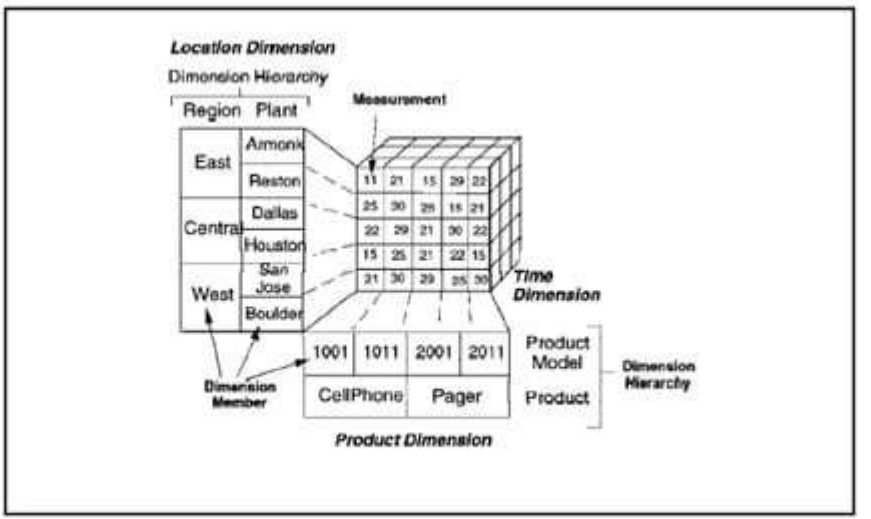
- A measure is a numeric attribute of a fact, representing the performance or behavior of the business relative to the dimensions. The actual numbers are called as variables.
- A measure is determined by combinations of the members of the dimensions and is located on facts.

Examples of Measures:

- Quantity Sold
- Unit Price
- Amount Sold
- Profit

CUBE

Cubes contain a set of data that is usually constructed from a subset of a data warehouse and is organized and summarized into a multidimensional structure defined by a set of dimensions and measures.



Advantages of SSAS Cubes

- SSAS is fast even on a large volume of data
- SSAS calculated measures are fast execution-wise and easy reusable
- They are defined centrally in the SSAS database, and the reports pick and choose the calculated measures they want.

ROLAP: Relational OLAP

- Special schema design: *star, snowflake*
- Special indexes: bitmap, multi-table join
- Special tuning: maximize query throughput
- Proven technology (relational model, DBMS), tend to outperform specialized MDDB especially on large data sets
- Products
  - IBM DB2, Oracle, Sybase IQ, RedBrick, Informix

ROLAP: Relational OLAP

- *Advantages:*
  - It can handle large amounts of data, ROLAP itself places no limitation on data amount

- *Disadvantages:*
  - Performance can be slow. Because each ROLAP report is essentially a SQL query (or multiple SQL queries) in the relational database, the query time can be long if the underlying data size is large .
  - It is difficult to perform complex calculations .

MOLAP: Multi Dimensional OLAP

- MDDB: a special-purpose data model
- Facts stored in multi-dimensional arrays
- Dimensions used to index array
- Sometimes on top of relational DB
- Products
  - Pilot, Arbor Essbase, Gentia

MOLAP: Multi Dimensional OLAP

- *Advantages*
  - Excellent performance
  - The storage is not in the relational database, but in proprietary formats.
  - MOLAP cubes are built for fast data retrieval, and are optimal for slicing and dicing operations.

- *Disadvantages:*
  - It is limited in the amount of data it can handle. Because all calculations are performed when the cube is built, it is not possible to include a large amount of data in the cube itself.
  - It requires an additional investment in human and capital resources are needed.

HOLAP: Hybrid OLAP

- Hybrid OLAP (HOLAP) refers to technologies that combine MOLAP and ROLAP.
  - *Advantages*
    - For summary-type information, HOLAP leverages cube technology for *faster performance*.
    - When detail information is needed, HOLAP can "drill through" from the cube into the underlying relational data.

Any
questions?