Unit -1

Data Warehousing

* **Datawarehousing Definations:**

(1) A process of transforming data into information and making it available to users in a timely manner to make a difference.

[Forrester Reasearch] 1996

(2) A single complete and consistent store of data obtained from variety of different sourses made available to end users in a what they can understand and use in a business context.

[Barry Devlin]

(3) A data warehousei is subject oriented, integrated, time-varying, non-volatile collection of data that is used primarily in organizational decision making.

[Bill Inmon]

(4) **subject oriented**: Data that gives information about a particular subject instead of about company's ongoing operations.

(5) **Integrated**: Data that is gather into the data warehouse from variety of sources and merged into a coherent whole.

(6) **Time variant**: All data in the data warehouse is identified with a particular time period.

(7) **Non-volatile**: Data is stable in data warehouse. more data is added but old data is never removed this enables consistent picture of business.

- **ODS (operational Data store):**
  (1) centralized repository for real time and integrated operational-data.
  (2) focuses on current, granular data for operational reporting and synchronization.
  (3) provides a consistent and up-to-date data-source for decision making.

- **Data Marts:**
  (1) subjets of Data warehouse or data repository.
  (2) Designed for specific groups or department.
  (3) optimized for performance and tailored to meet particular analytical needs.
  (4) can be independent (standlone) or dependent (sourced from a central Data warehouse)

- **Operational system:**
  (1) A system that is used to run a business in real time, based on current data, also called a system of record.

- ODS | Data Warehouse:

ODS	Data Warehouse
(1) stores current operational data	(1) stores historic and current data
(2) perform numerous quick and simple queires on small amount of data.	(2) perform complex queries on large amount of data.
(3) small amount of transactional data	(3) large amounts of static data.
(4) Day to Day Decisions, current operational results, Tactical reporting	(4) Long-term Decisions, Strategic reporting, Trend decisions.
(5) Near to Normal form	(5) star schema.
(6) frequency: Twice daily, daily, weekly,	(6) frequency: weekly, monthly. quaterly

- OLTP (online Transaction processing):
  (1) it is a database system designed for managing and processing day-to-day opertional transactions in real time.
  (2) it focuses on handling large volume of short, atomic transactions, such as inserting, updating, or deleting records. and prioritizes fast response times and data integrity.
  (3) OLTP database typically use a normalized schema to minimize redundancy.

* OLAP (online Analytical processing):

(1) OLAP is database system designed for complex data analysis, reporting, and decision making support.

(2) It specializes in read-heavy operations involving data aggregation, multidimensional analysis, and trend analysis.

(3) OLAP databases often use denormalized or star/snowflake schemas to optime query performance and can handle large volumes of historical data for business intelligence and reporting purposes.

OLAP	OLTP	Data Warehouse
(1) complex data analysis and reporting	(1) managing dayday ops. transactions	(1) centralized data repository for reporting and analysis.
(2) Read heavy, complex queries	(2) Write heavy frequent transactions	(2) mix of data loading and complex queries.
(3) Denormalized or star/snowflake	(3) Normalized	(3) star, snowflake and data vault models.
(4) involves aggregation and trend analysis	(4) involves simple record retrieval, inserts, updates, and deletes.	(4) supports complex queries, data transformation, and reporting.
(5) prioritises analytical capabilities over low-latency processing	(5) prioritises fast response times for real-time transactions	(5) provides reasonable query performance or analytical purposes
(6) handles large volumes of historical data	(6) smaller volume focused on current transaction	(6) very large volume of historical data and entire wide data
(7) eg. Business intelligence, data mining	(7) e-commerce websites, CRM systems.	(7) business analysis, reporting and trend forecasting.

* Data Warehouse Architecture:

(1) Data warehouse architecture refers to the design and structure of data warehousing system, including its components and how they interact to store, and provide access to data for analytical and reporting purposes

(2) There are 3 types of architectures

(1) Tier 1
(2) Tier 2
(3) Tier 3

(3) Tier single data warehouse:

- In this architecture data is stored in a single database server
- suitable for small-scale data warehousing needs
- Limited scalility and performance.

(4) Two-tier Data Warehouse:

- seperate data storage (data warehouse server) from client applications.
- provides better scalability and performance compared to the single-tier architecture.
- still relatively simple and suitable for smaller data warehousing environments.

(P) Three-tier Datawarehouse:

*This is the most common data warehouse arch.

- consists of three tiers
  (1) Data source layer
  (2) Data warehouse layer
  (3) Client layer

- Data source layer: where data is extracted from various source systems, such as databases, applications, and external sources.

- Data warehouse layer: where data is transformed, cleaned, and stored in the data warehouse

- Client layer: where end-users access data for reporting and analysis through various tools and applications.

* ETL (Extract, Transform Load)

(1) Extract (E):

- The first step in ETL is extracting data from various source systems, such as data bases, files, applications, or external sources.
- Data is gathered, often in raw or unprocessed form and transferred to a staging area for processing

(2) Transform (T):

- Transformation is where data is cleaned, standardized and transformed to fit the data warehouse schema and business requirements.
- Tasks may include data cleansing, data enrichment, data validation, and aggregation.
- This step ensures data quality, consistency, and compatibility with the data warehouse structure.

(3) Load (L):

- The final step involves loading the transformed data into the data warehouse.
- Data is loaded into appropriate tables or data marts within the data warehouse structure.
- Depending on the architecture, this could be a full load (reloading all data) or incremented load (only new or changed data)

* Data warehouse design approaches :

(1) Bill Inmon - Top-Down approch

(2) Ralph Kimball - Bottom-up approach

(1) Top-Down approach :

(1) The Top down approach begins with the overall enterprise view and focuses on creating a comprehensive, centralized data warehouse that serve the entire organization.

(2) Characteristics :

(1) Enterprise wide-perspective : it starts with a high level understanding of the organization's data needs and business goal.

(2) Single central Repository : it aims to build a single all-encompassing data warehouse that stores data from various source systems.

(3) strategic and long term : it is often a strategic, long term initiative that involves significant planning and investment

(4) Normalized Data : Data is stored in normalized form.

(3) Benefits : (1) Provides consistent view of Data across organization

(2) Reduces data Redundancy, improve integrity

(3) supports complex cross-function analytics.

(4) challenges : (1) Requires substantial time, resources, and upfront planning

(2) may face resistance from business unit due to longer time for delivery.

(3) complex to implement.

(2) Bottom-up approach:

(1) Bottom up approach begins with specific business units or departments creating their data marts or data warehouses to address their immediate needs.

(2) Business unit centric: it starts with the needs of individual business unit or department.

(3) incremental building: Data marts are built Immediately to address specific business requirements.

(4) Faster implementation: Business units can see results more quickly, as they don't have to wait for centralized data warehouse to be fully developed.

(5) star schema or Denormalized: Data marts often use start schema or denormalized structures for faster query performance.

(6) Benefits: ① Rapid response specific business needs
② Allows busines units to have more control over their data.
③ easir to implement and show value quickly.

(7) challenges: ① may lead to data silos and duplication of efforts.
② integration and consistency challenges may arise when trying to unify data marts.
③ potential for lack of enterprise-wide data governance and standards.

* **Dimensional Modelling:**
  - modelling techniques determine how data is structured within the data warehouse to support efficient querying and reporting.

(1) **star schema:**
  - **structure:** In a star schema, the data warehouse is organised into two main types of tables fact tables and dimension tables.
  - **fact table:** contains the quantative data (measures) and foreign keys that link to dimension tables.
  - Dimension tables: stor descriptive or categorical information (attributes) related to the data in in fact table.
  - **Relationship:** fact table are the center of the schema connected directly to dimension tables. This creates star like structure.
  - **simplicity:** starschemas are easy to understand, query and maintain, making them idea for most data warehouse scenarios.
  - **performance:** queries involving star schema typically perform well because joins are straight forward and aggregation is efficient

(2) **snowflake schema:**
  - **structure:** The snowflake schema is an extension of the star schema where dimension tables are normalized into sub-dimensions. This results in more complex, heierarchical structure.
  - **normalization:** Dimension tables are devided into multiple related tables to reduce redundancy and improve data integrity.

(3) Fact constellation (Galaxy schema):

- Structure: Fact constellation represents a data warehousing scenario where multiple fact tables share dimension tables: its like a collection of star schemas.
- Multiple stars: In this approach, several fact tables exist, each associated with its set of dimension tables.
- Complexity: Fact constellations are complex to design and query compared to star schema. They are typically used when different business areas require their own fact tables but share some dimensions.
- Flexibility: provides flexibility for accomodating diverse needs across an organization.
- Challenges: Maintainance and query complexity can be challenges in fact constellation schemas due to the presence of multiple fact tables

* OLAP (online Analytical processing) operations:

(1) Roll up
(2) Drill down
(3) slice
(4) dice
(5) Drill through
(6) Drill across

(1) **Roll up:**

(1) Roll up also known as drill-up or aggregation, is an OLAP operation that involves moving from Lower level, more detailed data to higher level summaries.

(2) Rolling up monthly sales data to quaterly to yearly totals.

(3) Roll-up is used to view data at different levels of granularity, allowing users to see the big picture and identify trends.

(2) **Drill Down:**

(1) Drill down or Roll down, is the reverse of roll-up it involves moving from higher level summaries to more detailed data.

(2) example: Drilling down from yearly sales to quaterly or monthly sales.

(3) Drill down helps users explore detailed data to investigate anomalies, identify causes, or gain deeper understanding of specific aspects of the data.

(3) Slice:

- Slice is an OLAP operation where users selects a single dimension value to view a 2D cross-section of the data cube.
- Slicing by product category to see sales data for a specific category across time
- slicing allows user to focus on specific aspect of the data and isolate it for analysis, providing a simplified view.

(4) Dice:

- Dicing is similar to slicing but involves selecting specific values from multiple dimensions to view subset of data
- Dicing to view sales data for a particular product category in specific region during particular time frame.
- Dicing allows users to create custom views of data by selecting specific dimension values, facilitating more detailed analysis.

(5) Drill through:

- Drill through enables users to access underlying detailed data records from summarized data.

(6) Drill across:

- Drill across is an OLAP operation that involves navigating between different dimension hierarchies to explore data relationships.

* OLAP models:

(1) MOLAP (multiDimensional OLAP)
(2) ROLAP (Relational OLAP)
(3) HOLAP (Hybrid OLAP)

(1) MOLAP:

- Data is stored in multidimensional cubes, enabling fast query performance due to pre-aggregation and optimized indexing.
- ideal for complex analytical tasks and interactive exploration of multidimensional data. Provides quick and efficient responses to use queries.
- eg. IBM cognos, oracle Essbase.
- suitable for smaller to medium-sized datasets requiring rapid analysis and dedicated OLAP environment.

(2) ~~HOLAP~~ : HOLAP:

- combination of MOLAP and ROLAP, detailed data stored in relational tables, aggregated data stored in multiDimensional cubes.
- Balances the performance benefits of MOLAP with flexibility of ROLAP; offers compromise between speed and versatility.
- eg. Microsoft SQL Serve analysis services with HOLAP storage mode, IBM DB2 OLAP.
- suitable for versatile analytical environments where both performance and flexibility are crucial. ideal for organizations with diverse analytical needs.

(3) ROLAP:

- ~~Store~~ Data stored in relational databases, making it flexible and compatible with existing databases. leverages standard SQL for querying.
- offers scalability and compatibility with large datasets and existing infrastructure. provides flexibility in modelling and supports ad-hoc querying.
- eg. SAP BW, Micro-strategy.
- suitable for large data sets, integration with relational database and environment where querying is preffered.