Data Engineering Batch 1 **Day 1: 17/01/2024**

Name: Pradip Bochare

Statement: Data Warehousing concepts

- Data warehousing:
  - The process of creating data warehouses to store a large amount of data is named Data Warehousing.
  - Data Warehousing helps to improve the speed and efficiency of accessing different data sets and makes it easier for company decision-makers to obtain insights that will help the business and promoting marketing tactics

- Data Warehouse:
  - Subject oriented, integrated, time variant, non-volatile collection of data in support of management's system.
  - A Data Warehouse is a collection of software tools that facilitates analysis of a large set of business data used to help an organization make decisions.

- Need of Data Warehousing:
  - Data Warehousing is a progressively essential tool for business intelligence. It allows organizations to make quality business decisions.
  - The data warehouse benefits by improving data analytics, it also helps to gain considerable revenue and the strength to compete more strategically in the market.

- Features of Data Warehouse:
  - Subject-oriented
  - Integrated
  - Time-variant
  - Non volatile

- Subject oriented:
  - Data are organized according to the subject instead of application.
  - It mainly focuses on modelling and analysis of data for decision makers, not on daily operations or transaction processing.

- Integrated:
  - Constructed by integrating multiple, heterogeneous data sources like relational databases, flat files, on-line transaction records.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources.

- Time-variant:
  - The time horizon for the data warehouse is significantly longer than that of operational systems. i.e. provide information from a historical perspective (e.g., past 5-10 years).

- Non-volatile:
  - No updates are allowed. Once the data entered into the data warehouse, they are never removed.
  - The data in warehouse represent company's history, the operational data representing near term history are always added to it.

Decision Support System (DSS):

- In a typical business environment with an increasing competitive market, we cannot ponder over decisions for too long.
- Hence the business needs managers with quick decision making.
- Therefore, to succeed in business today, any company needs information systems that can support the diverse information and decision-making needs.

DSS architectural styles:

- OLTP (Online Transaction Processing)
  used by traditional operational systems (RDBMS).
- OLAP (Online Analytical Processing)
  used by Data Warehouse.

OLTP (Online Transaction Processing):

- **OLTP** is a methodology to provide end users with access to large amounts of data
- It works in an intuitive and rapid manner to assist with deductions based on investigative reasoning.
- **OLTP** refers to a class of systems that facilitate and manage transaction-oriented applications, typically for data entry and retrieval transaction processing.

❖ Benefits of OLTP:

1. Simplicity & Efficiency:
   Reduced paper trails and the faster and more accurate forecasts for revenues and expenses are both examples of how OLTP makes things simpler for businesses.
2. OLTP systems maintain data integrity and they also provide fast query processing in environments having multiple access.

❖ Pitfalls of OLTP:

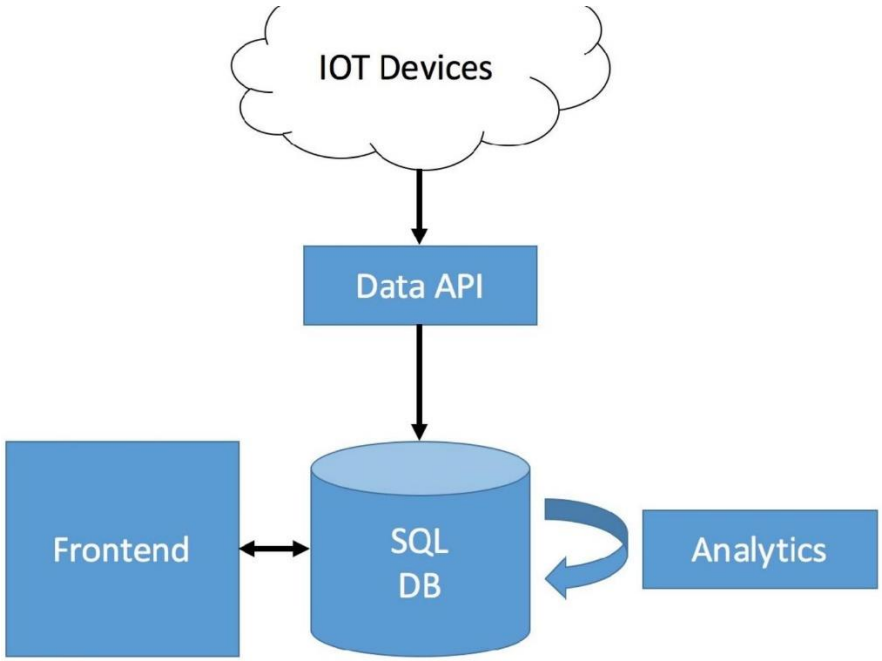
1. OLTP requires instant update.
2. The data what we get from OLTP is not suitable for data analysis.
3. To perform one simple transaction even with the normalized structure, we need to query multiple tables by using joins.

Data Engineering

Data engineering is the process of designing and building systems that let people collect and analyze raw data from multiple sources and formats. These systems empower people to find practical applications of the data, which businesses can use to thrive.

ETL (Extract, Transform, Load)

Extract, transform, and load (ETL) is the process of combining data from multiple sources into a large, central repository called a data warehouse. ETL uses a set of business rules to clean and organize raw data and prepare it for storage, data analytics, and machine learning



Data Classification:

Raw Data

- Unprocessed data in format used on source e.g JSON
- No schema applied

Processed data

- Raw data with schema applied
- Stored in event tables/destinations in pipelines

Cooked data

- Processed data that has been summarized.

Big Data Properties:

- **Volume**
  How much data you have

- **Velocity**
  How fast data is getting to you

- **Variety**
  How different your data is

- **Veracity**
  How reliable your data is