Data engineering - Assignment 1
- Mitushi Vishwakarma

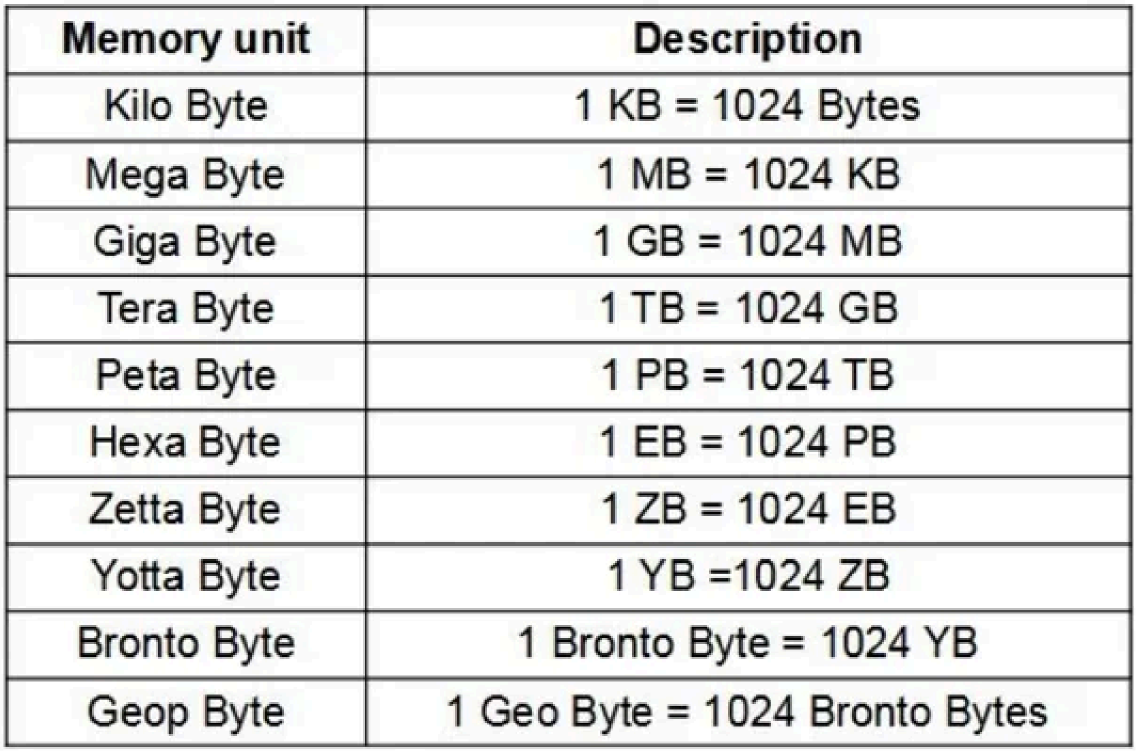
Question 1 : What is data engineering ?

Ans : Data engineering involves the tasks of designing, building and maintaining data infrastructures. Data infrastructures include databases, big data repositories, pipelines for moving data between data systems. Data engineers performs following tasks :

- Develop and optimize data systems
- Make data available for analysis
- Selecting right databases, cloud architecture

Question 2 : What is data and the different ways to measure it ?

Ans : Data is raw information generated on the daily basis by various sources. Sources can include texts, images, videos, clickstreams, conversations, social media platforms, live streams, IOT devices, etc. The measurement units of data in bytes is shown below.



Memory unit	Description
Kilo Byte	1 KB = 1024 Bytes
Mega Byte	1 MB = 1024 KB
Giga Byte	1 GB = 1024 MB
Tera Byte	1 TB = 1024 GB
Peta Byte	1 PB = 1024 TB
Hexa Byte	1 EB = 1024 PB
Zetta Byte	1 ZB = 1024 EB
Yotta Byte	1 YB =1024 ZB
Bronto Byte	1 Bronto Byte = 1024 YB
Geop Byte	1 Geo Byte = 1024 Bronto Bytes

Question 3 : What are different types of data ?

Ans : There are mainly 3 different types of data :

1. Raw : Raw data, also referred to as sourced, eggy, or primary data, are the data gathered from a source. Unprocessed data in format used on source e.g JSON. No schema applied.
2. Processed : Raw data with schema applied. Stored in event tables/destinations in pipelines.
3. Cooked : Processed data that has been summarized.Cooked data are raw data that has been processed.

Question 4 : What is Big Data and its properties ?

Ans : Big data refers to extremely large and diverse collections of structured, unstructured, and semi-structured data that continues to grow exponentially over time.

The properties of big data are :

1. Volume - How much data you have
2. Velocity - How fast data is getting to you
3. Variety - How different your data is
4. Veracity - How reliable your data is

Question 5 : What is Data Warehouse and its features ?

Ans : Data Warehouse (DW) is a Subject oriented, integrated, time variant, non-volatile collection of data in support of management's system.
Data warehouse is a relational database used for data reporting and analysis based on OLAP. Data is sourced from operational systems. Operational systems access operational databases.

1. **Subject-oriented** – A data warehouse is always a subject oriented as it delivers information about a theme instead of organization's current operations.
2. **Integration** - A data warehouse is built by integrating data from various sources of data such that a mainframe and a relational

database. In addition, it must have reliable naming conventions, format and codes.

3. **Time-Variant –** In this data is maintained via different intervals of time such as weekly, monthly, or annually etc.
4. **Non-Volatile –** As the name defines the data resided in data warehouse is permanent. It also means that data is not erased or deleted when new data is inserted.

Question 6 : What is OLTP(Online Transaction Processing) ?

Ans : OLTP or Online Transaction Processing is a type of data processing that consists of executing a number of transactions occurring concurrently—online banking, shopping, order entry, or sending text messages. OLTP is a methodology to provide end users with access to large amounts of data.

Queries on database is OLTP.

- 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.

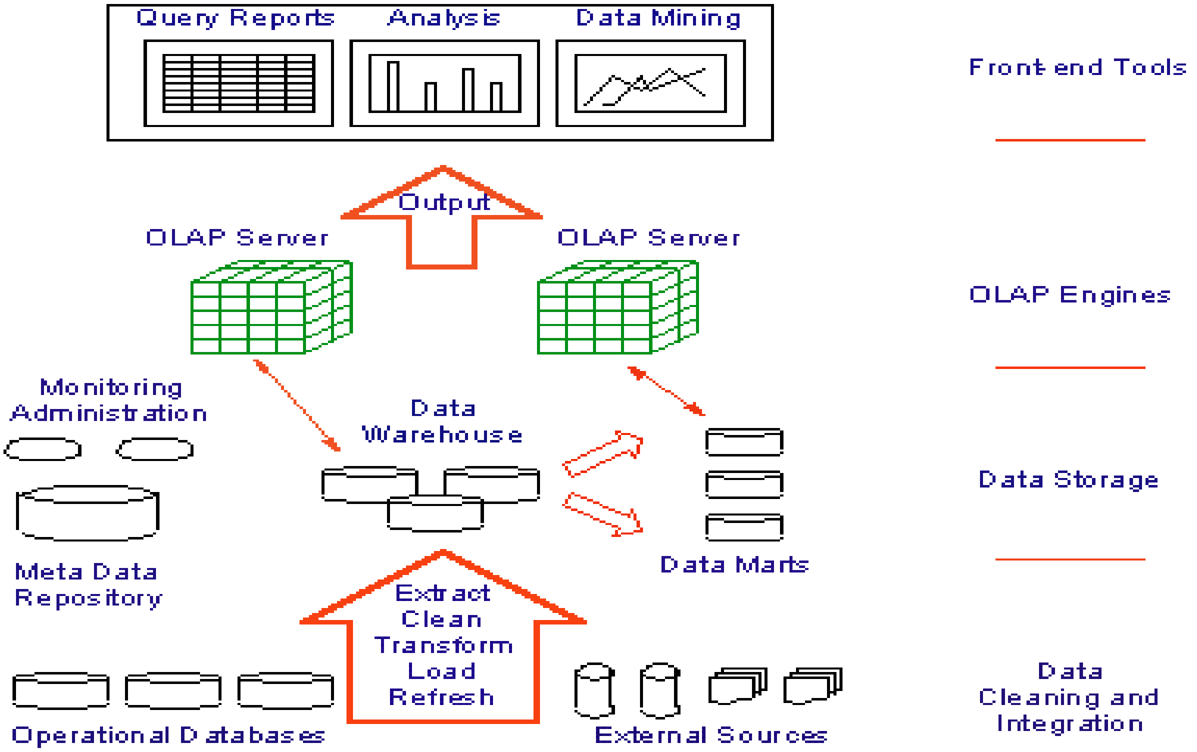
Question 7 : What is OLAP (Online Analysis Processing)?

Ans : Online analytical processing (OLAP) is software technology you can use to analyze business data from different points of view. Organizations collect and store data from multiple data sources, such as websites, applications, smart meters, and internal systems. It is an approach used to analyze data from multi dimensions.

Queries on data warehouse is OLAP.

- OLAP Server receives the data from data warehouse by which it represents the data in a user understandable format which actually supply analytical functionality for the DSS system.
- OLAP Server generally performs data analysis in two forms.

- ROLAP(Relational OLAP)
- MOLAP(Multi-dimensional OLAP)


Query Reports
Analysis
Data Mining
Front end Tools
Output
OLAP Server
OLAP Server
OLAP Engines
Monitoring
Administration
Data
Warehouse
Data Storage
Data Marts
Meta Data
Repository
Extract
Clean
Transform
Load
Refresh
Data
Cleaning and
Integration
Operational Databases
External Sources