What is Big Data?

Data that is being generated in massive volumes around the globe in varying forms. Large amounts of data that cannot be processed by traditional systems. Big data follows the concepts of 4 V's:

-Volume
-Veracity
-Velocity
-Variety

What is a Database?

- A database is a collection of data that is organized and stored in a way that allows for efficient retrieval, updating, and management of that data. A database can be thought of as an electronic filing cabinet that stores information.
- Databases can be used to store a large amount of structured data for a business or entity. It can be in the form of inventory, sales, employee data etc. These databases can be managed using software called Database Management Systems or DBMS which allow the creation, updation and maintenance of databases.
- Databases can be structured (SQL) or non-structured (NoSQL). SQL databases store structured data in the form of tables which contain rows and columns. NoSQL databases store unstructured data in various other forms.

What is a Data Warehouse?

- A large collection of cleaned, filtered and organized business data ready to be used for analytics and making business related decisions.
- Data is brought into the DW from a variety of different source systems. Raw data is brought into the DW through a process called ETL (can also be ELT).
- DWs can be on prem or cloud based. The future of DWs for many business is moving their DWs to the cloud as it comes with numerous benefits over on prem alternatives.
- Some characteristics that a DW should have:
    - Centralized and consistent location for data
    - Must be accessible and have high query performance
    - User friendly (easy for users to query and use data)
    - Should load data consistently and repeatedly
    - A DW could also have a built in reporting or dashboarding tool

What is a Data Lake?

A data lake is a large, centralized repository that allows organizations to store all their structured and unstructured data at any scale. Unlike traditional databases, data lakes store raw data without any predefined schema or organization, allowing for flexible analysis and processing of the data. It is mostly useful for AI and ML engineers since they can make use of the raw & and unfiltered data to create models and predictions. Because of its flexibility in terms of volume and type of data that can be stored, it creates a challenge when it comes to security and data governance.