**Big Data**

A single smartphone generates around 40 Exabytes of Data every month. This data is too big and complex for normal ways of processing it, for example using your own CPU to process this data.

Big Data is mainly characterized by three things,

* **Volume:** The amount of data that is being generated. For example, hospitals generate around 2314 Exabytes of data annually.
* **Velocity:** The speed by which data is generated. Or in other words, number of transactions done.
* **Variety:** Data can be generated in different ways, mainly divided into structured, semi-structured, unstructured. Structured would be in Excel format (Tables etc.), semi-structured can be JSON format data, unstructured would be images, videos, audio etc.

To store and process this data, different techniques are used such as Hadoop, Spark etc. These techniques distribute this data into smaller chunks and store them accordingly. And then, to process this data, parallel processing is done (MapReduce etc.) to make use of the data effectively and efficiently.

There are many applications of Big Data. For example, Big Data has revolutionized the advertising market. It was used in US Elections and still is being used (through the help of targeted marketing).

**Database**

It is a place where your transactional data is stored. Transactional data is the one which is used for CRUD operations i.e., Create, Read, Upload and Delete. These databases can be of different types, such as Relational Databases, NoSQL and they are also further divided into different types (for example, NoSQL Database can be Key-Value Database, Document Database, Graph Database etc.)

Database is usually highly normalized, to maintain consistency and reduce duplication of data. Due to highly normalized structure of Database, transactions can be performed at a very high speed in these databases. However, it is not suitable for analytical purposes (due to high number of joins required to get data suitable for Analysis, which slows down the queries, data of different regions/departments may be stored in different databases and etc.)

**Data Warehouse**

Data Warehouse is mainly used to store analytical data. Analytical data is usually **aggregated data** (to the granularity needed). Data Warehouse is not used for transactional purposes and instead, maintains historical data (and this data is updated after a scheduled amount of time)

This historical data is aggregated from all the sources of Data sources (with the help of ETL Process, i.e., Extract Transform Load. Extract is used to ingest data from databases, Transform is used to transform the transactional data into aggregated data as needed for analysis, and Load is used to load the transformed data into the Data Warehouse)

Data Warehouse is also very useful in capturing and analyzing *historical changes* in the data. For example, a person phone number was changed so we can analyze that in one year of period, on average how many people changes their phone number (this is not possible in Databases because it only keeps the updated record. Data Warehousing, using a good schema can track these slowly changing dimensions).

Data Warehouse mostly uses Star Schema or Snowflake Schema because these contains less number of relations and hence faster query execution on large amount of data. Furthermore, Data Warehouses are mostly Columnar, that is data is stored in columns rather than rows (as in OLTP Databases). This is also so as to keep the query performance good.

**Data Lake**

It is usually the storage area previous to Data Warehouse. While Data Warehouse only always support structured data in a pre-defined schema (Star/Snowflake Schema), Data Lake is much more flexible in this. It allows storage of any kind of data (structured, semi-structured, unstructured). Due to this, Data Lake is able to store much larger volume of data at a much lower cost and complexity (as no schema is present). This data can then be used for analysis (however will require complex queries to be written so as to extract the meaningful data).

Flexibility of this data lake can also be a disadvantage, if the data lake is not properly governed. It can lead to disorganized data which would be very difficult to use. Therefore, it is important to have clear policies and good governance practices in order for a data lake to be useful.