UNIT – I

1. Big Data and Data Science, Big Data Architecture:

Big Data and Data Science:

- Big Data refers to large and complex datasets that are difficult to process and analyze using traditional data processing methods. It involves high-volume, high-velocity, and high-variety data. Big Data solutions often utilize technologies like distributed computing, NoSQL databases, and data parallelism to handle the immense scale of data.

- Data Science is an interdisciplinary field that combines scientific methods, algorithms, processes, and systems to extract knowledge and insights from structured and unstructured data. It involves data cleaning, data preparation, data analysis, and machine learning techniques to make data-driven decisions and predictions.

- **Characteristics of big data**

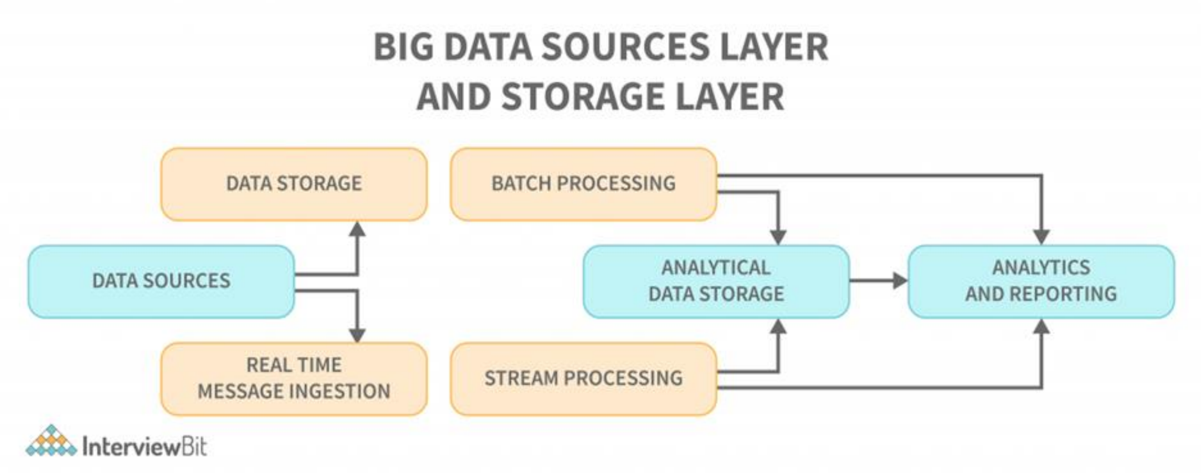
  - **Volume.** Big data is enormous, far surpassing the capabilities of normal data storage and processing methods. The volume of data determines if it can be categorized as big data.
  - **Variety.** Large data sets are not limited to a single kind of data-instead, they consist of various kinds of data. Big data consists of different kinds of data, from tabular databases to images and audio data regardless of data structure.
  - **Velocity.** The speed at which data is generated. In Big Data, new data is constantly generated and added to the data sets frequently. This is highly prevalent when dealing with continuously evolving data such as social media, IoT devices, and monitoring services.

Data Science	Big Data
Data Science is an area.	Big Data is a technique to collect, maintain and process huge information.
It is about the collection, processing, analyzing, and utilizing of data in various operations. It is more conceptual.	It is about extracting vital and valuable information from a huge amount of data.
It is a field of study just like Computer Science, Applied Statistics, or Applied Mathematics.	It is a technique for tracking and discovering trends in complex data sets.
The goal is to build data-dominant products for a venture.	The goal is to make data more vital and usable i.e. by extracting only important information from the huge data within existing traditional aspects.

Data Science	Big Data
Tools mainly used in Data Science include SAS, R, Python, etc	Tools mostly used in Big Data include Hadoop, Spark, Flink, etc.
It is a superset of Big Data as data science consists of Data scrapping, cleaning, visualization, statistics, and many more techniques.	It is a sub-set of Data Science as mining activities which is in a pipeline of Data science.
It is mainly used for scientific purposes.	It is mainly used for business purposes and customer satisfaction.
It broadly focuses on the science of the data.	It is more involved with the processes of handling voluminous data.

2. **Big data Architecture:**

Big Data Architecture in data analytics is the design and arrangement of various components, technologies, and processes that allow organizations to effectively store, process, and analyze large and complex datasets (Big Data). The architecture is crucial in enabling data analysts and data scientists to derive meaningful insights and valuable information from massive volumes of data. Below are some key components and concepts in Big Data Architecture:



1. Data Sources: Big Data often originates from multiple sources, including databases, sensors, social media, web logs, and more. These sources continuously generate large amounts of data that need to be collected and ingested into the Big Data system.

2. Data Ingestion: In this phase, data from various sources is collected, validated, and loaded into the Big Data platform. Data ingestion tools and processes ensure the smooth flow of data into the storage layer.

3. Data Storage: Big Data requires distributed storage systems capable of handling enormous volumes of data across multiple nodes. Commonly used storage solutions include Hadoop Distributed File System (HDFS), cloud-based object storage (e.g., Amazon S3), and NoSQL databases (e.g., Apache Cassandra, MongoDB).

4. Data Processing: To analyze Big Data effectively, parallel processing frameworks are used. Apache Hadoop and Apache Spark are popular choices for distributed data processing, allowing data analysts to perform complex computations across the entire dataset.

5. Data Transformation: Data may need to undergo cleansing, transformation, and enrichment before analysis. This step involves converting raw data into a structured and usable format suitable for analytics.

6. Data Querying: Big Data architectures often provide query interfaces to access and retrieve relevant subsets of data efficiently. Technologies like Apache Hive and Apache Pig enable querying using SQL-like languages and data flow programming, respectively.

7. Data Analytics: Data analysts use various analytical techniques, including descriptive, diagnostic, predictive, and prescriptive analytics, to gain insights and make data-driven decisions.

8. Data Visualization: Visualizations are essential for presenting complex Big Data analysis results in a more understandable and intuitive manner. Tools like Tableau, Power BI, or custom web-based dashboards are used for this purpose.

9. Scalability and Fault Tolerance: Big Data architectures are designed to scale horizontally by adding more nodes to handle growing data volumes. They also incorporate fault tolerance mechanisms to ensure data integrity and reliability.

10. Security and Data Privacy: Due to the sensitive nature of the data involved, robust security measures are implemented to protect against unauthorized access and data breaches.