Assignment - 1

Sec-A

① Big Data integration refers to the process of combining data from different sources into a unified view. It involves extracting data from various structured, semi-structured and unstructured sources, transforming it into a common format, and loading it into a central system for analysis.

② Value in Big data refers to the insights and actionable information that can be extracted from large datasets. The value comes from analyzing data to uncover trends, patterns and correlations that can drive business decisions, improve efficiency and create competitive advantages.

③ Data explosion refers to the rapid and exponential growth in the volume of data being generated from various sources like social media, sensors, devices and business processes. This increases in data requires advanced storage, processing and analysis techniques to extract meaningful insights.

④ A NOSQL model refers to a type of database that is designed to handle large volumes of unstructured & semi-structured data. Unlike traditional relation databases, NOSQL databases use flexible data models such as key-value pairs, document based column based or graph-based models. They are optimized for scalability and fast data access.

Sec-B

1. HDFS (Hadoop Distributed File System) is a distributed storage system used to store large volumes of data across a cluster of machines. It follows a master-slave architecture where:

- Namenode - Acts as the master and manages the metadata and file system structure.
- DataNode - Acts as the slave and stores the actual data.

HDFS splits large files into smaller blocks and distributes them across the cluster. It ensures fault tolerance by replicating the data across multiple nodes. HDFS supports high-throughput data access and parallel processing making it deals for handling bigdata workloads.

2. - Volume: managing large amounts of data generated at high speed.
- Variety - Handling data in different formats (Structured, Semi-structured, unstructured).
- Velocity - Processing data at a high rate of incoming flow.
- Scalability - Designing a system that can scale horizontally as data volume increases.
- Veracity: Ensuring the accuracy and consistency of data.

- Security: Protecting sensitive data from unauthorized access and breaches.

SEC-C

① Processing in BigData infrastructure refers to the methods and frameworks used to analyze and extract insights from large datasets. It includes:

- Batch Processing: Processing large datasets in bulk (using Hadoop MapReduce).
- Real time Processing: Processing data as it is generated (using Apache Storm or Spark Streaming)
- Stream Processing: Handling continuous data streams to provide real-time insights efficient Processing enables organizations to uncover trends, patterns and actionable insights quickly.

② A relational Database management System (RDBMS) is a database system that organizes data into tables (relations) with rows and columns. key concept include.

- Tables - Store data in a structured format.
- Primary key - A unique identifier for each rows.
- Foreign key - A reference to a primary key in another table to establish relationships.
- SQL (Structured Query Language).
  - Used to query and manipulate data.

Applications of RDBMS include banking, Customer relationship management (CRM)

and inventory management, where structured data and complex queries are involved.