ASSIGNMENT-3

MODULE-3(GROUP TASK)

BIG DATA PROCESS MAPPING

Big Data Process Mapping is the **structured flow of steps** used to collect, process, store, analyze, and visualize large volumes of data (Big Data). It helps organizations understand how data moves from its source to meaningful insights.

Big Data is characterized by the **5 V ’s**:

- Volume (large amount of data)
- Velocity (high speed of data generation)
- Variety (different data types)
- Veracity (data quality)
- Value (useful insights)

Data Generation / Data Sources

Data is generated from multiple sources:

- Social media
- IoT devices
- Sensors
- Websites
- Mobile apps
- Business transactions

Data types:

- Structured (databases)
- Semi-structured (JSON, XML)
- Unstructured (images, videos, text)

Data Collection (Ingestion)

Data is collected and transferred into the system.

Tools used:

- Apache Kafka
- Apache Flume
- APIs
- Log collectors

Two types:

- Batch processing (large data at intervals)
- Real-time streaming

Data Storage

Big data is stored in distributed systems.

Common technologies:

- Hadoop Distributed File System (HDFS)
- NoSQL databases
- Cloud storage

Storage must be:

- Scalable
- Fault-tolerant
- Distributed

Data Processing

Data is cleaned and transformed.

Two processing methods:

- Batch Processing (e.g., Hadoop MapReduce)

- Real-time Processing (e.g., Apache Spark)

Activities:

- Data cleaning
- Filtering
- Aggregation
- Transformation

Data Analysis

Advanced analytics is performed:

- Machine Learning
- Data Mining
- Statistical analysis
- Predictive analytics

Goal: Extract meaningful patterns and insights.

Data Visualization & Reporting

Insights are presented using:

- Dashboards
- Charts
- Reports
- BI tools (Power BI, Tableau)

This helps decision-makers understand trends.

Example: E-commerce company

- Collects customer browsing data
- Stores it in cloud storage
- Processes purchase patterns