8. Post-Experiments Exercise

A. Extended Theory: (Soft Copy)

1) Different types of big data analytics tools

Big Data Analytics tools are broadly categorized into ***Batch Processing, Real-Time (Stream) Processing, and Interactive Analysis***. Some popular tools include:

A. Batch Processing Tools:

These tools handle large volumes of static data stored over time and process it in chunks (batches).

- **Apache Hadoop:** A batch processing framework that uses MapReduce for large-scale data processing.
- **Apache Hive:** Data warehouse infrastructure built on Hadoop for providing data summarization and query.
- **Apache Pig:** A high-level platform for creating MapReduce programs used with Hadoop

B. Real-Time / Stream Processing Tools:

These tools process data in real-time or near real-time as it flows into the system.

- **Apache Spark:** A fast, general-purpose cluster-computing system for both batch and stream processing.
- **Apache Storm:** A distributed real-time computation system for processing data streams.
- **Apache Flink:** Framework for stateful computations over unbounded and bounded data streams.

C. Data Ingestion and Messaging Tools:

These tools handle the transfer of data between different systems.

- **Apache Kafka:** A distributed event streaming platform, mainly used for building real-time data pipelines.
- **Apache NiFi:** Manages the flow of data between systems with real-time control.

D. Machine Learning and Graph Tools:

Used for data mining, predictive modeling, and graph computations.

- **Spark MLlib:** A machine learning library built on top of Apache Spark.
- **GraphX:** Spark's API for graphs and graph-parallel computation.

---

2) Apache spark and spark Framework

Apache Spark is a fast, general-purpose cluster computing framework designed for big data processing. It supports both batch and real-time stream processing and offers an advanced DAG (Directed Acyclic Graph) execution engine.

Core Features:

In-memory computing: Keeps data in memory between operations, reducing disk I/O and improving performance.

Distributed processing: Can handle huge volumes of data across clusters.

Ease of use: Supports APIs in Python, Java, Scala, and R.

Rich ecosystem: Includes libraries for SQL, machine learning, graph processing, and streaming.

Spark Ecosystem Components: Spark Core: The base engine for the overall Spark platform. Spark SQL: Used for working with structured data using SQL and DataFrame APIs. Spark Streaming: For real-time data processing using mini-batch processing. MLlib: A machine learning library that includes tools for classification, regression, clustering, etc. GraphX: Provides APIs for graph processing and computation.	Use Cases: Real-time fraud detection in banking Log processing and monitoring Machine learning model training at scale Big data ETL (Extract, Transform, Load) pipelines