Week 5

1. What distributed graph processing framework operates on top of Spark?
    a. MLlib
    b. GraphX
    c. Spark streaming
    d. ALL

a. **Incorrect -MLlib**:

   This is Spark's machine learning library, not specifically for graph processing.

b. **Correct- GraphX**

   It is a distributed graph processing framework that operates on top of Apache Spark. It allows users to process and analyze graph data at scale by leveraging Spark's core functionality for distributed data processing.

c. **Incorrect -Spark Streaming**:

   This is used for processing real-time data streams, not for graph processing.

d. **ALL**:

   Incorrect, as only GraphX is specifically designed for graph processing on Spark.

2.Which of the following frameworks is best suited for fast, in-memory data processing and supports advanced analytics such as machine learning and graph processing?

a) Apache Hadoop MapReduce
b) Apache Flink
c) Apache Storm
d) Apache Spark

a) Apache Hadoop MapReduce

- **Incorrect**: Apache Hadoop MapReduce is primarily designed for batch processing and relies on disk-based storage, making it slower compared to in-memory processing. It is not specifically designed for advanced analytics like machine learning or graph processing.

b) Apache Flink

- **Incorrect**: Apache Flink is excellent for real-time stream processing and supports complex event processing. However, it is primarily optimized for stream processing rather than providing a comprehensive suite for batch processing, machine learning, and graph analytics as Spark does.

c) Apache Storm

- **Incorrect**: Apache Storm focuses on real-time stream processing with low latency. While it excels in handling real-time data, it does not offer the same breadth of support for in-memory processing, machine learning, and graph processing as Spark.

d)**Apache Spark**

- **correct**:**Apache Spark** is designed for fast, in-memory data processing. It provides advanced analytics capabilities, including machine learning (through MLlib) and graph processing (through GraphX). Its in-memory computation greatly speeds up processing tasks compared to disk-based systems. Spark's flexibility and performance make it ideal for handling complex analytics and iterative algorithms efficiently.

3. A financial institution needs to analyze historical stock market data to predict market trends and make investment decisions. Which Big Data processing framework is best suited for this scenario?

a. Apache Spark
b. Apache Storm
c. Hadoop MapReduce
d. Apache Flume

a) Correct- Apache Spark

Explanation: Apache Spark is well-suited for analyzing historical data due to its fast in-memory processing capabilities. It supports a variety of data analytics tasks and is ideal for predictive analytics and machine learning.

b) **Incorrect- Apache Storm**: This is for real-time stream processing, which is less suited for historical data analysis.

c) **Incorrect- Hadoop MapReduce**: While it can handle large-scale data processing, it is slower than Spark for iterative algorithms used in predictive modeling.

d) **Incorrect- Apache Flume**: It is primarily used for data ingestion, not analysis.

4.A telecommunications company needs to process real-time call logs from millions of subscribers to detect network anomalies. Which combination of Big Data tools would be appropriate for this use case?

a. Apache Hadoop and Apache Pig
b. Apache Kafka and Apache HBase
c. Apache Spark and Apache Hive
d. Apache Storm and Apache pig

a. **Incorrect**- **Apache Hadoop and Apache Pig**: Hadoop is designed for batch processing, and Pig is used for ETL tasks in batch mode. This combination is not suited for real-time processing.

b. **Correct**- **Apache Kafka and Apache HBase**

- **Apache Kafka**: Kafka is a distributed event streaming platform that excels at handling real-time data streams. It can ingest high volumes of log data efficiently and serve as a buffer for processing.
- **Apache HBase**: HBase is a NoSQL database that provides real-time read/write access to large datasets. It complements Kafka by offering a scalable and distributed storage solution where processed data can be stored and queried quickly.

c. **Incorrect**- **Apache Spark and Apache Hive**: While Spark can handle real-time processing (with Structured Streaming) and is powerful for analytics, Hive is more oriented towards batch processing and querying rather than real-time analytics.

d. **Incorrec**t- **Apache Storm and Apache Pig**: Storm is good for real-time stream processing, but Pig is used for batch processing and ETL tasks. Combining Storm with Pig does not align with the need for real-time analytics and storage.

5. Do many people use Kafka as a substitute for which type of solution?

a. log aggregation
b. Compaction

c. Collection
d. all of the mentioned

a. **Correct**- **Log aggregation-** Apache Kafka is commonly used for log aggregation. It efficiently collects, processes, and stores log data from various sources in a distributed manner.

b. **Incorrect -Compaction**: Kafka supports log compaction, but it's not a substitute for log aggregation.

c. **Incorrect -Collection**: Kafka is used for collecting logs, but the term 'substitute' more directly refers to log aggregation.

d. **Incorrect -All of the mentioned**: Not all options are directly applicable.

6. Which of the following features of Resilient Distributed Datasets (RDDs) in Apache Spark contributes to their fault tolerance?

a. DAG (Directed Acyclic Graph)
b. In-memory computation
c. Lazy evaluation
d. Lineage information

a. **Incorrect - DAG (Directed Acyclic Graph)**: While the DAG is essential for understanding how Spark schedules tasks, it is the lineage information specifically that ensures fault tolerance.

b. **Incorrect - In-memory computation**: This improves performance but does not directly contribute to fault tolerance.

c. **Incorrect -Lazy evaluation**: This optimizes execution and resource usage but does not specifically address fault tolerance.

d. **correct -Lineage information**: RDDs maintain lineage information, which is the history of transformations applied to the data. This lineage information allows Spark to recompute lost data in case of failures, ensuring fault tolerance.

7.Point out the correct statement.

a. Hadoop do need specialized hardware to process the data
b. Hadoop allows live stream processing of real-time data

c. In the Hadoop mapreduce programming framework output files are divided into lines or records
d. None of the mentioned

a. Hadoop do need specialized hardware to process the data

- **Incorrect**: Hadoop is designed to run on commodity hardware, meaning it does not require specialized or high-end hardware. It is intended to scale across many inexpensive, standard machines.

b. Hadoop allows live stream processing of real-time data

- **Incorrect**: Traditional Hadoop, specifically the MapReduce framework, is designed for batch processing and does not natively support real-time stream processing. For real-time processing, frameworks like Apache Storm or Apache Flink are more appropriate.

c. In the Hadoop MapReduce programming framework output files are divided into lines or records

- **Correct**: In Hadoop MapReduce, the output is indeed processed and written as lines or records, where each record represents a unit of data processed by the framework. This is how data is commonly handled in MapReduce jobs.

d. None of the mentioned

- **Incorrect**: The third statement is accurate regarding how Hadoop MapReduce processes and outputs data.

8. Which of the following statements about Apache Pig is true?

a. Pig Latin scripts are compiled into HiveQL for execution.
b. Pig is primarily used for real-time stream processing.
c. Pig Latin provides a procedural data flow language for ETL tasks.
d. Pig uses a schema-on-write approach for data storage.

a. **Incorrect**: **Pig Latin scripts are compiled into HiveQL for execution**:

- Pig Latin is compiled into a series of MapReduce jobs.

b. **Incorrect**: **Pig is primarily used for real-time stream processing**:

- Pig is used for batch processing.

c. **correct - Pig Latin provides a procedural data flow language for ETL tasks.** Pig Latin is a scripting language used in Apache Pig for expressing data transformation tasks in a procedural manner, making it suitable for ETL (Extract, Transform, Load) processes.

d. **Incorrect**: **Pig uses a schema-on-write approach for data storage**:

- Pig uses schema-on-read.

9. An educational institution wants to analyze student performance data stored in HDFS and generate personalized learning recommendations. Which Hadoop ecosystem components should be used?

a. Apache HBase for storing student data and Apache Pig for processing.
b. Apache Kafka for data streaming and Apache Storm for real-time analytics.
c. Hadoop MapReduce for batch processing and Apache Hive for querying.
d. Apache Spark for data processing and Apache Hadoop for storage.

a. **Incorrect - Apache HBase for storing student data and Apache Pig for processing**: While HBase is a NoSQL database suitable for real-time read/write access, Pig is used for ETL tasks and batch processing, which may not be as efficient as Spark for complex analytics and recommendations.

b. **Incorrect - Apache Kafka for data streaming and Apache Storm for real-time analytics**: Kafka is used for streaming data, and Storm is for real-time analytics. This combination is more suited for real-time data processing rather than batch analytics and recommendation generation.

c. **Incorrect - Hadoop MapReduce for batch processing and Apache Hive for querying**: While MapReduce and Hive are both part of the Hadoop ecosystem, MapReduce is less efficient for iterative processing compared to Spark. Hive is used for querying but is more oriented towards batch processing rather than real-time analytics and personalized recommendations.

d. Correct - Apache Spark for data processing and Apache Hadoop for storage.

- **Apache Spark for data processing**: Spark is a powerful and versatile data processing engine that supports complex analytics, machine learning, and iterative algorithms. It is well-suited for analyzing large datasets and generating recommendations. Spark's in-memory computation capabilities provide high performance for such tasks.

- **Apache Hadoop for storage**: Hadoop's HDFS (Hadoop Distributed File System) is a scalable and reliable storage system designed for storing large volumes of data across a distributed cluster. It is ideal for storing the large datasets of student performance data.

10.A company is analyzing customer behavior across multiple channels (web, mobile app, social media) to personalize marketing campaigns. Which technology is best suited to handle this type of data processing?

a. Hadoop MapReduce
b. Apache Kafka
c. Apache Spark
d. Apache Hive

a. **Incorrect - Hadoop MapReduce**: While it can handle large-scale data, it is less efficient for iterative and real-time analytics compared to Spark.

b. **Incorrect - Apache Kafka**: Primarily used for message streaming, not data processing.

c. correct - Apache Spark - Apache Spark is highly suitable for analyzing customer behavior across various channels due to its fast processing capabilities and support for complex analytics and machine learning.

d. **Incorrect - Apache Hive:** Used for querying and not for complex data processing and analytics.