Name: Nayaab Jindani D15C Rollno. 20

Experiment No: 10

Aim:

To perform Batch and Streamed Data Analysis using Apache Spark.

Theory:

1. What is Streaming? Explain Batch and Stream Data.

Streaming refers to the continuous flow of data that is processed in real-time or near realtime. It involves analyzing data as it arrives, allowing immediate insights and reactions.

Batch Data:

- Batch data processing deals with large volumes of static data that are collected over a period and then processed together.
- It is suitable for applications where immediate results are not required.
- Example: Processing daily sales data at the end of the day.

Stream Data:

- Stream data processing handles continuous, unbounded data arriving in real-time or micro-batches.
- It is ideal for applications needing real-time analytics, like fraud detection, live dashboards, etc.
- Example: Processing logs from a web server or transactions from a banking system as they occur.

2. How Data Streaming Takes Place Using Apache Spark?

Apache Spark Streaming is an extension of the Spark Core API that enables scalable, highthroughput, fault-tolerant processing of live data streams.

How It Works:

- Spark Streaming ingests data in mini-batches from various sources such as Kafka, Flume, HDFS, or socket connections.
- Each mini-batch is treated as an RDD (Resilient Distributed Dataset) and processed using Spark's core operations.
- Once processed, the results can be stored in databases, file systems, or dashboards.

Key Components:

- **DStreams (Discretized Streams):** The basic abstraction in Spark Streaming. Internally, a DStream is a sequence of RDDs.
- **Data Sources:** Real-time data can come from Kafka, socket, files, etc.
- **Window Operations:** Perform computations over sliding windows of data (e.g., last 10 minutes).
- **Transformations:** Just like batch RDDs, DStreams can use map, filter, reduce, etc.

Use Cases:

- Real-time fraud detection.
- Social media sentiment analysis.
- Log processing.
- Monitoring systems and alerts.

Conclusion:

Apache Spark provides powerful capabilities for both batch and stream data processing, making it a unified framework suitable for a wide range of big data applications. While batch processing is ideal for historical data analysis and scheduled jobs, streaming is essential for real-time insights and event-driven applications. Spark Streaming bridges the gap between these two paradigms by providing a consistent and scalable platform for analyzing both static and live data, enabling organizations to react to information as it happens.