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 real-time data. It is the process of transmitting data in a steady, ongoing manner as it becomes available.
    - **Batch Data** involves data that is collected and processed in large chunks at scheduled intervals. The data is accumulated over a period, and then processed together as a batch.
    - **Stream Data** (or real-time data) is generated continuously and processed immediately as it arrives. This type of data is processed in small increments, often on a per-event basis, and is typically used in scenarios like sensor data, logs, or online transactions.

2. **How data streaming takes place using Apache Spark?**

    - Apache Spark performs real-time data processing using **Spark Streaming**, an extension of the core Spark API.
    - Data is ingested in small chunks called **micro-batches**. These micro-batches are processed as streams in near real-time.
    - Data sources for streaming can include message queues like Apache Kafka, socket connections, or file systems.
    - Spark Streaming processes data in micro-batches by continuously polling the source for new data and processing it in small time intervals (e.g., seconds).
    - Spark integrates batch and stream processing, meaning you can process streaming data with the same APIs used for batch data, allowing for unified processing pipelines.

Conclusion:

Apache Spark provides a powerful framework for both batch and stream data analysis. By utilizing Spark Streaming, organizations can process and analyze real-time data efficiently, while also leveraging batch processing for large-scale historical data analysis. The integration of these two processing models allows Spark to cater to a wide range of data processing needs, from time-sensitive streaming applications to periodic batch processing, making it a versatile tool for modern data engineering tasks.