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 real-time as it arrives.
Batch Data is collected over a period and then processed together as a group. It is suitable for high-volume, less time-sensitive operations.
Stream Data is generated continuously (like sensor data, logs, transactions) and requires real-time or near-real-time processing.
Batch Processing is ideal for complex analytics and operations that don't require immediate results.
Stream Processing enables instant insights and actions based on incoming data, making it ideal for live dashboards, fraud detection, etc.

2. How Data Streaming Takes Place Using Apache Spark.

Apache Spark uses Spark Streaming or Structured Streaming for processing real-time data streams.
Data from sources like Kafka, Flume, or socket connections can be ingested continuously using the Spark Streaming API.
In Spark Streaming, the data stream is divided into small time-based batches called micro-batches.
Each micro-batch is then processed using standard Spark transformations and actions.
With Structured Streaming, Spark processes data using the same DataFrame and SQL APIs, treating streaming data as a continuously growing table.
The output can be stored or pushed to dashboards, databases, or file systems for further use.

Conclusion:

Apache Spark efficiently handles both batch and stream processing through its unified engine. While batch processing is suitable for historical data analysis, stream processing enables real-time analytics on continuously incoming data. Spark's Structured Streaming model simplifies real-time data handling using familiar APIs and ensures fault-tolerant and scalable performance. Through this experiment, the practical understanding of how Apache Spark processes both static and dynamic data was achieved, highlighting its significance in real-world big data applications.