Name-Vedant Dhoke Div-D15C Roll No-9 A.Y.-2024-25

Experiment 10:

Aim:- To perform Batch and Streamed Data Analysis using Apache Spark.

Theory:

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

Streaming:
Streaming refers to the continuous flow of data generated in real-time from various sources like sensors, logs, social media, etc. In data streaming, data is processed as it arrives, enabling real-time analytics.

Batch Data:

- Batch data is collected over a period of time and then processed in chunks or batches.
- Examples include daily sales reports, monthly transaction records, etc.
- It is processed using tools like Apache Spark, Hadoop, etc.

Stream Data:

- Stream data is continuously generated and processed in real-time.
- Examples include live tweets, sensor data, website activity logs, etc.
- Requires tools that support real-time processing like Apache Spark Streaming, Apache Flink, etc.

2. How Data Streaming Takes Place Using Apache Spark.
Apache Spark enables real-time data processing through its component called Structured Streaming. It allows users to treat streaming data similarly to static data

by using high-level DataFrame APIs. Internally, Spark continuously reads new data from sources like Kafka, sockets, or files, and processes it incrementally.

Data streaming in Spark works by dividing the incoming stream into small batches, which are processed sequentially. Each batch is treated as a micro-batch that goes through transformations and actions similar to batch processing. Spark then updates the output, which can be directed to various destinations like the console, file systems, or databases. This model combines the benefits of batch processing with near real-time performance, offering a powerful framework for stream analytics.

Conclusion:

Apache Spark provides a unified platform for both batch and real-time data analysis. Batch processing is ideal for analyzing large datasets collected over time, while streaming is essential for scenarios where immediate insights are needed. With the help of Spark Structured Streaming, users can build scalable and efficient data pipelines that support real-time decision-making and analytics.