Name : Sandeshkumar.M.Yadav **Div : D15C** **Roll : 61**

Experiment 10

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

Theory:

1) What is streaming. Explain batch and stream data.

Ans:

Streaming is a data processing method where data is continuously generated, transmitted, and processed in real-time or near real-time. This approach is used when immediate insights or actions are required, such as in live video feeds, financial transactions, or sensor data in IoT systems.

Instead of waiting for a complete dataset, streaming systems handle data as it arrives, enabling timely processing and responses.

Aspect	Batch Data	Stream Data
Definition	Data is collected, stored, and processed in large chunks or batches.	Data is continuously generated and processed in real-time.
Example	Processing monthly sales reports.	Monitoring live stock market prices.
Latency	High (processing happens after data collection).	Low (near real-time processing).
Storage	Data is stored first, then processed.	Processed immediately as it arrives.
Use Cases	Data warehousing, analytics reports, backups.	Fraud detection, live analytics, IoT monitoring.
Tools	Hadoop, Apache Spark (batch mode).	Apache Kafka, Apache Flink, Spark Streaming.

2) How data streaming takes place using Apache spark.

Ans:

Apache Spark provides a powerful framework called Spark Structured Streaming to handle real-time data streams. It allows for continuous processing of data as it arrives, combining the simplicity of SQL/DataFrame operations with the scalability and fault-tolerance of Spark.

Key Components and Process

1. Data Source (Input)

The streaming process begins with a data source. Spark reads data continuously from streaming sources like:

- Apache Kafka
- File systems (monitoring new files in a directory)
- Sockets
- Amazon Kinesis
- Other custom sources

These sources send data in real-time, which Spark ingests as a stream.

2. Streaming Data as a Table

Spark treats streaming data as an unbounded table. Each new data item is like a new row being added to this table. Onecan perform operations like select, filter, groupBy, and even SQL queries on this streaming table.

3. Query Execution

The user defines a query on the streaming data (e.g., count words, calculate averages).
Internally, Spark builds a logical plan and then optimizes it into a physical plan for execution.

4. Micro-Batch Processing

Spark Structured Streaming processes data in micro-batches.
Instead of processing each event individually, it collects data for a short interval (e.g., every second) and processes it together.
This approach balances real-time performance with processing efficiency.

5. Output Sink

After processing, the results are written to an output sink, such as:

- Console (for testing/debugging)
- Kafka
- Databases
- File systems

You can choose different output modes:

- Append: Only new rows are written.
- Update: Only updated rows are written.
- Complete: The entire result table is written.
- Fault Tolerance

Conclusion:

Batch and streamed data analysis are two core approaches in data processing. **Batch analysis** processes large volumes of data collected over time, ideal for historical insights and complex computations. **Streamed analysis**, on the other hand, processes data in real-time as it arrives, enabling immediate decision-making. While batch is suited for accuracy and completeness, streaming excels in speed and responsiveness. Together, they offer a powerful hybrid approach for modern data-driven systems.