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 refers to a technique in data processing where information is generated, transferred, and analyzed on a continuous basis in real-time or with minimal delay. It is particularly useful in scenarios that demand instant feedback or actions—such as live video streams, banking transactions, or IoT-based sensor data.

Rather than waiting for an entire dataset to be collected, streaming systems process data incrementally as it flows in, allowing for immediate analysis and quicker decision-making.

Feature	Batch Processing	Stream Processing
Data Processing	Processes a large volume of data at once.	Processes data as it arrives, record by record.
Latency	High latency, as processing happens after data collection.	Low latency, providing near real-time insights.
Throughput	Can handle vast amounts of data at once.	Optimized for real-time but might handle less data volume at a given time.
Use Case	Ideal for historical analysis or large-scale data transformations.	Best for real-time analytics, monitoring, and alerts.
Complexity	Relatively simpler to implement with predefined datasets.	More complex, requires handling continuous streams.
Data Scope	Operates on a finite set of data.	Operates on potentially infinite streams of data.
Error Handling	Errors can be identified and corrected before execution.	Requires real-time handling of errors and failures.
Resource Usage	Resource-intensive during processing, idle otherwise.	Continuous use of resources.
Cost	Cost-effective for large volumes of data.	More expensive due to continuous processing.

2) How data streaming takes place using Apache spark.

Ans:

Apache Spark includes a robust module known as **Spark Structured Streaming**, designed to process data streams in real-time. It enables seamless and continuous handling of incoming data, blending the ease of using SQL or DataFrame queries with Spark's ability to scale and recover from faults efficiently.

Core Components and Workflow

1. Data Input Sources

The streaming workflow starts with ingesting data from various sources. Spark continuously

monitors and reads data from real-time input channels, such as:

- Apache Kafka
- File systems (by tracking newly added files)
- Network sockets
- Amazon Kinesis
- Custom data sources

These sources deliver data in a live stream, which Spark captures and processes on the fly.

1. Streaming Data as a Table

Spark conceptualizes real-time data as a never-ending or *unbounded* table, where each incoming record is treated as a newly inserted row. This structure allows users to perform familiar operations—such as select, filter, groupBy, and complex SQL queries—on data that is constantly evolving.

2. Query Execution

Users define logical queries on streaming data, like calculating averages or tracking counts. Under the hood, Spark converts these logical instructions into an optimized physical execution plan. This layered planning ensures efficient and scalable processing, even with massive, real-time workloads.

3. Micro-Batch Processing

Instead of processing each data point as it arrives, Spark Structured Streaming groups incoming records into short-duration intervals called *micro-batches*. For instance, it might process new data every second. This batching strategy strikes a balance between real-time responsiveness and high-throughput performance, making it well-suited for both latency-sensitive and large-scale scenarios.

4. 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 streaming data processing represent two fundamental paradigms in the world of analytics. **Batch processing** deals with static datasets gathered over time and is excellent for deep analysis and complex transformations. In contrast, **streaming analytics** enables immediate processing of real-time data, which is essential for scenarios like fraud detection, live dashboards, or real-time alerts.

Modern data platforms often combine both strategies to form **hybrid architectures**—taking advantage of batch processing for in-depth insights and streaming for timely, reactive decision