Experiment 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 a method of processing data that is generated and consumed in real time or near real time. This is especially useful when immediate feedback or actions are required—like monitoring sensor inputs, processing transactions, or managing live feeds.

Unlike traditional methods that wait until all data is collected, streaming systems process information as soon as it's available. This allows for quicker decision-making and timely insights.

2. **How data streaming takes place using Apache spark.**
   Apache Spark handles real-time data using its feature called Structured Streaming. This enables developers to process continuous flows of data with ease using familiar SQL or DataFrame APIs. It is designed for scalability and reliability.
   How it works:

1. **Input Source**
   The streaming starts with data being read continuously from sources such as:

- Apache Kafka
- File directories (watching for new files)
- Network sockets
- Amazon Kinesis
- Other custom data providers

Spark receives this live data stream for processing.

2. **Data as an Unbounded Table**
   In Spark, incoming streaming data is viewed as a growing table where each new entry adds a new row. Standard operations like filtering, selecting, or grouping can be applied, just like with regular tables.
3. **Query Definition**
   Users define operations on the data stream (like counting values or calculating stats). Spark translates this into a logical plan, then converts it to a physical execution plan for efficient performance.
4. **Micro-Batch Mechanism**
   Instead of processing each incoming piece of data one by one, Spark gathers data for small time windows (e.g., every second), and processes them in groups. This “micro-batching” keeps it responsive but still efficient.

5. **Output Destination**
   After the data is processed, results are stored or displayed via:

- Console (for quick checks)

- Kafka
- Databases
- Filesystems

The output can be configured in different ways:

- **Append**: Adds only new results
- **Update**: Updates only changed results
- **Complete**: Outputs the full result every time

Spark also provides built-in fault tolerance to recover from failures automatically.

Conclusion:
Batch and stream analysis serve different purposes in data processing. Batch processing handles large datasets collected over time and is suitable for deep, detailed analysis. Stream processing deals with real-time data, allowing immediate insights and actions. While batch ensures thoroughness and precision, streaming offers speed and responsiveness. Using both together enables systems to be both intelligent and fast, handling a wide range of data needs effectively.