Experiment 10

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

Theory:

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

Streaming is all about handling data in real-time. Like live dashboards, fraud alerts, or anything that reacts instantly. Batch processing waits until a pile of data is ready, then processes it all at once. It has higher delay but good for structured non-urgent tasks.

Batch Analysis:

- Collects data over time, then processes it in one go.
- More delay between input and result.
- Good for stuff like monthly reports or end-of-day summaries.
- Examples: Hadoop, Hive.

Streamed Analysis:

- Processes data on the fly, as it's generated.
- Near-zero delay. Great for immediate reactions.
- Used in live systems like stock trading, chat apps, etc.
- Tools: Kafka, Spark Streaming.

Key takeaway:
Batch = delayed + grouped.
Stream = real-time + continuous.

2. How does data streaming take place in Apache Spark?

Apache Spark handles streaming via Spark Streaming, which is built to manage live data efficiently and reliably across clusters.

1. **Getting the data in:**
   Spark can pull in data from Kafka, Flume, socket connections, or files (like logs dumped into HDFS or S3). This data flows in non-stop.
2. **Micro-batching:**
   Instead of processing every single data point one-by-one, Spark chunks the stream into tiny batches. This approach gives it the power of batch engines with near-real-time speed.
3. **Processing:**
   Spark applies your usual transformations (like map, filter, etc.) to each micro-batch. The logic stays mostly the same as traditional Spark programs.
4. **Sending the output:**
   After processing, Spark sends the results to wherever they're needed like databases, dashboards, storage, or even another pipeline.
5. **Resilience:**
   Spark handles crashes by checkpointing and replication. If something breaks, it can pick up where it left off without losing data.

Conclusion: This experiment showed how Apache Spark can handle both batch and streaming data. Spark Streaming, with its micro-batch architecture, makes real-time analytics scalable, reliable, and efficient. It's a solid option for apps that need to react instantly to data, like fraud detection, alerting systems, or live user behavior tracking.