211720104037

MODULE 9

Kafka and Spark Streaming

1. What is Apache Spark Streaming?

Apache Spark Streaming is **a scalable faulttolerant streaming processing system that natively supports both batch and streaming workloads**.

2. Describe how spark streaming processes data?

Spark Streaming **receives live input data streams and divides the data into batches, which are then processed by the Spark engine to generate the final stream of results in batches**. Spark Streaming provides a high-level abstraction called discretized stream or DStream, which represents a continuous stream of data.

211720104037

3. What are D-Streams?

Discretized Stream or DStream is **the basic abstraction provided by Spark Streaming**.

It represents a continuous stream of data, either the input data stream received from source, or the processed data stream generated by transforming the input stream.

4. What is a streaming context object?

Public class StreamingContext extends Object implements Logging.

Main entry point for Spark Streaming functionality. It provides methods used to create DStream s from various input sources. It can be either created by providing a Spark master URL and an appName, or from a org. apache.

5. What are some of the common transformations on DStreams supported by Spark Streaming?

DStreams support many of the transformations available on normal Spark RDD's. Some of the common ones are as follows. Return a new DStream by passing each element of the source DStream through a function func.

...

UpdateStateByKey Operation

- Scala.
- Java.
- Python.

6. What are the output operations that can be performed on DStreams?

Some of the output operations are **print(), save()** etc.. The save operation takes directory to save file into and an optional suffix. The print() takes in the first 10 elements from each batch of the DStream and prints the result.