211720104133

SATHYANATHAN H

1.What is Apache Spark Streaming?

Apache Spark streaming is nothing but an extension of core Spark API that is responsible for fault-tolerant, high throughput, scalable processing of live streams. Spark streaming takes live data streams as input and provides as output batches by dividing them. These streams are then processed by the Spark engine and the final stream results in batches.

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.

3.What are DStreams?

Discretized Streams (DStreams)

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 StreamingContext 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?

Some of the common transformations on DStreams supported by Spark Streaming are:

- map(func)
- filter()
- flatMap()
- union()
- intersection() • distinct()
- groupByKey()
- reduceByKey(func)
- aggregateByKey(func)
- sortByKey()

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.