Kafka and Spark Streaming

Homework

Sindhu.V

202001139

3rd year CSE C

1. What is Apache Spark Streaming?

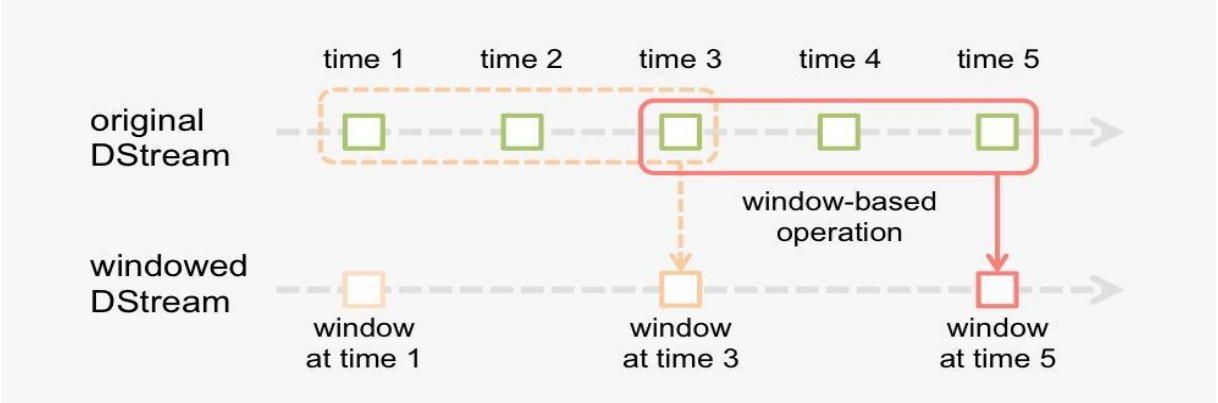
Apache Spark Streaming is a scalable fault-tolerant streaming processing system that natively supports both batch and streaming workloads.

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.

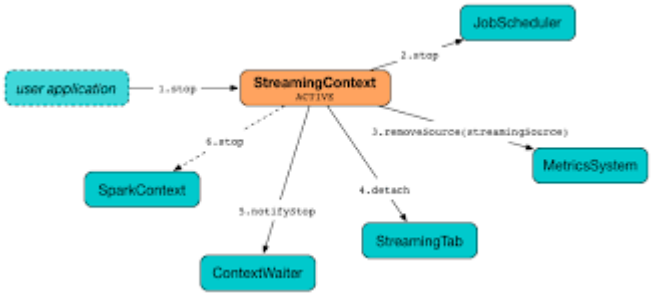
3. What are DStreams?

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 streaming Context object?

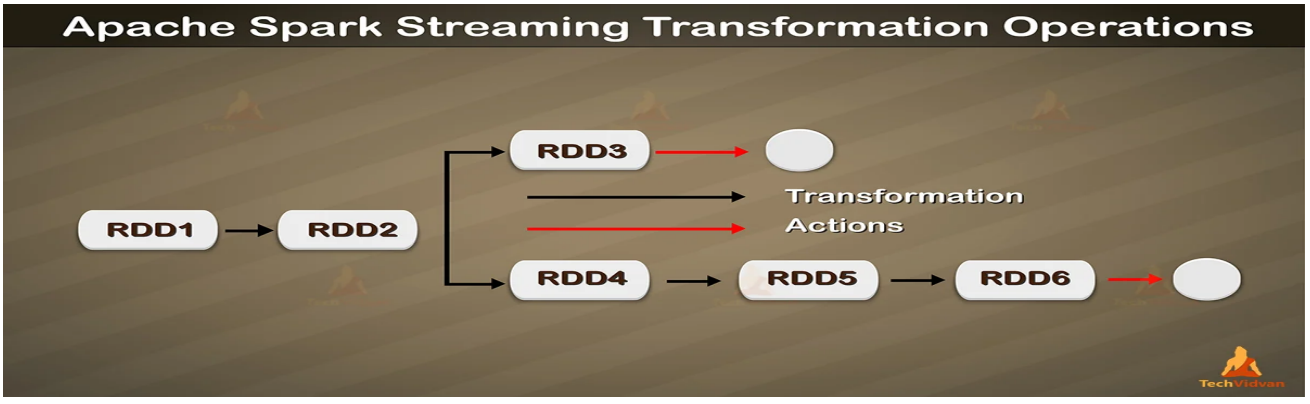
Public class Streaming Context extends Object implements Logging. Main entry point for Spark Streaming functionality. It provides methods used to create DStream s from various input sources. It can be either created by providing a Spark master URL and an appName, or from a org.



5. What are some of the common transformations on DStreams

supported by Spark Streaming?

Map function in Spark passes each element of the source DStream through a function and returns a new DStream.



FlatMap function in Spark is similar to Spark map function, but in flatmap, input item can be mapped to 0 or more output items.

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

Spark DStream also support two types of Operations: Transformations and output Operations-

i. Transformation

There are two types of transformation in DStream:

a. Stateless Transformations

The processing of each batch has no dependency on the data of previous batches. Stateless transformations are simple RDD transformations. It applies on every batch meaning every RDD in a DStream. It includes common RDD transformations like map(), filter(), reduceByKey() etc.

Although these functions seem like applying to the whole stream, each DStream is a collection of many RDDs (batches).

Stateless transformations are capable of combining data from many DStreams within each time step. For example, key/value DStreams have the same join-related transformations as RDDs— cogroup(), join(), leftOuterJoin() etc.

Output Operation

Once we get the data after transformation, on that data output operation are performed in Spark Streaming. After the debugging of

our program, using output operation we can only save our output. 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**.**