Kafka and Spark Streaming

Homework 9

Q1. What is Apache Spark Streaming?

Apache Spark Streaming is a scalable fault tolerant streaming processing system that natively supports both batch and streaming workloads. Spark streaming provides a highlevel abstraction called discretized stream (DStream) which represents a continuous stream of data.

Q2. Describe how Spark Streaming processes data?

Apache spark streaming receives input data streams in live and divides the data into batches which are then processed by Spark engine to provide the final stream of results in batches.

Q3. What are DStreams?

Discretized Streams (DStreams) is the basic abstraction provided by Spark Streaming which 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. DStreams can be produced by performing high- level operations on existing DStreams or by using input data streams from sources like Kafka and Kinesis.

Q4. What is a StreamingContext object?

A StreamingContext object is created which is the main entry point of all Spark streaming functionality is used to initialize a Spark streaming program. A StreamingContext object can be created from a SparkConf object and can also be created from an existing SparkContext object. There is only one StreamingContext can be active in a JVM at the same time.

Q5. What are some of the common transformations on DStreams supported by Spark Streaming?

>map(function)
>flatMap(function)
>filter(function)
>repartition(numPartitions)
>union(otherStream)
>count()
>countByValue()
>reduce(function)
>reduceByKey(function, [numTasks])

Q6. What are the output operations that can be performed on DStreams?

>print()
>save()
>foreachRDD(func)
>saveAsTextFiles(prefix, [suffix])
>saveAsHadoopFiles(prefix, [suffix])
>saveAsTextFiles(prefix, [suffix])