211720118011

Kafka and spark streaming

Homework 9

Q1. What is Apache Spark Streaming?

Apache Spark Streaming is a scalable fault tolerant streaming processing system that supports both batch and streaming workloads. Spark streaming a extension of a core sprk single execution engine and unified programming and to lead some unique benefits over other traditional streaming system.

Q2. Describe how Spark Streaming processes data?

Apache spark streaming receives input data streams in live and divides the data into batches spark streaming provides a high level abstraction called discretized stream or dstream which represent a continuous stream of data.

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 and stream generated by transforming the input stream.

Q4. What is a Streaming Context object?

A Streaming Context extends object implement 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.

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])
- saveAsTextFviiles(prefix, [suffix])