Quiz 4

Name: Eric Solis   
Date: 07/22/2020

1. What is windowing in spark streaming?

Windowing is using a time/index frame to read data. For a data stream, we can set a window that would serve as a buffer for the incoming data, and that frame would be the only piece of data visible for our algorithm. The windowing could be sliding or not, this depends on our application and on whether we want data overlapping while processing.

1. What are broadcast variables and accumulators and a use case of each?
2. Whats the difference between a transformation and an action?

As spark uses lazy programming, processing is not performed until the very last moment. In this case, transformations are data processing techniques that modify our data, however, this actions are not performed until an action is called. Actions are simple functions that start the processiong of the previously stated transformations. Examples of actions are show(), count(), collect().

1. what’s the difference between Spark streaming and Structured streaming?

Spark streaming take RDDs as an input, that is unorganized text, while Structured streaming assign an schema to the incoming data. Structured streaming turns the lectures into a dataframe where the columns are the features of the data and the rows are the new lectures.

1. what’s the difference between Kafka vs Flume vs Nifi and when will you use which tool?

Nifi is an ingestion too with a GUI that allows to create data pipelines. Flume is a data ingestion tool for non structured and semi-structured data. And kafka is a distributed, fault-tolerant messaging broker. The difference is that kafka is a distributed tool that uses a client-consumer architecture to broadcast messages. By creating a topic from incoming data, clients can subscribe to these topics a read information from them improving the many-to-many communication.

I would use nifi for small data pipelines where the number of producers, clients is reduced and the information is not critical since it is not fault-tolerant.

Flume for a streaming application or to move files from one place to another. Since the configuration is made in a conf file, the applications are limited.

Kafka for a robust implementation of a messaging system or an application with critical data. Due to its fault-tolerance and its flexibility in data sources and sinks, it can be used in many different cases.

1. What is data skewness and how can we fix it?

There are two types of transformations in spark, the ones that does not affect the data partitions and the one that modify the data partitioning. The later is known as shuffles and it make data going from one partition o another like a sort operation.

This type of movements might result in having partitions with more data making difficult to parallel process the data decreasing the performance of the processing. Some techniques like merge sort or broadcasting are used to deal with data skewness.

1. Spark architecture?

The main elements of the spark architecture are the master node, the cluster manager and the worker nodes.

The master runs a piece of code (Scala, python, java) in a spark context, creates a DAG for the pipeline and the node manager assigns resources to every task in the job. Send the tasks to the worker nodes and they strart processing the data.

1. What is a Kafka Producer, Consumer, Broker?

Kafka producer is the element of the kafka architecture that produces or introduces data to the system.

Consumer is the one that ingests the data from kafka and gives feedback of its reception.

The broker is the kafka instance that keeps track of the incoming data, stores it in the partitions, assign an offset to the data and reads the feedback from the consumer.

1. What is the role of zookeeper in Kafka?

As kafka is a distributed system that does not follow a master-slave architecture, zookeeper act as a cluster manager and it allow the nodes to communicate and to distribute the data partitions among the cluster nodes and racks optimizing the fault-tolerance.

Zookeeper coordinates the tasks of the distributed system, like selecting the controller node in the cluster, the membership of the nodes to the cluster and the topics.

1. What are two types of offsets in Kafka and why do we need offsets?

The first one is the offset that keeps track of the incoming data. It assigns an index to the lectures to make it easy to access to a particular set of registers.

The second type is the committed offset and it is also a number that keeps track of the last element of data processed by our consumer. In this case, whenever the consumer request for more data, the broker knows exactly which is the following element that should be sent to the consumer.