Getting Started with Stream Processing with Spark Streaming

* You're tracking trending news stories over the last 2 days with your streaming application and you receive a new batch of updates every hour. News received a day ago is only half as important as the news received in the last hour. How would you specify this forgetfulness in your streaming k-means algorithm?

Ans: **Half-life of 24 batches**

* What is checkpointing in streaming Spark applications?

Ans: **Periodically saving data so state can be reconstructed from this intermediate context**

* What are the basic components of any Spark application?

Ans: **Driver, Executor and Receiver**

* Which of these operations require real-time processing?

Ans: **Requests on sale days to an e-commerce site**

* How is stream processing in Spark an extension of its batch processing framework?

Ans: **Stream processing operates on DStreams, which are made up of sequences of RDDs. Spark processes these individual RDDs in a stream just as it would during batch processing.**

* When are transformations considered stateful?

Ans**:** **When they operate across multiple RDDs rather than a single RDD**

* You receive log messages at the rate of two messages per second and your batch interval is five seconds. If the sliding window size is 20 seconds, how many RDDs are included in the window at any point in time?

Ans: **4**

* In a sliding window operation, let's say the summary function you want to apply across a window is a multiplicative operation represented by this pseudocode:

multiple(a, b) {

return a \* b

}

What is the pseudocode for the inverse function of this in the sliding window operation?

Ans: **divide(a, b) {**

**return a/b**

**}**

* What attributes can represent a student in school?

Ans: **Age, grade, test scores**

* What is the major difference in applying k-means clustering on batch vs streaming data?

Ans**: Determining how relevant older data is compared with newer data in the stream**

* What is the relationship between a DStream and an RDD?

Ans: **A DStream is made up of a sequence of RDDs, where every RDD contains entities which are received within a batch interval**.

* Under what conditions would you use the updateStateByKey()function? Ans: **Summarizing across an entire DStream where each RDD in the stream is a pair in RDD.**
* You're tracking trending news stories over the last 24 hours with your streaming application and you receive news updates every hour. What kind of decay factor will you use to perform streaming k-means clustering?

Ans**: 0 < decayFactor < 1**

* Which of these are examples of stream processing?

Ans: **Analyzing log messages to see whether all pages of the e-commerce site are up and running**

* Why are RDDs, the basic Spark programming abstraction, considered resilient?

Ans: **RDDs keep track of their lineage; they can be reconstructed from the source even when cluster nodes crash.**

* What are the two operations that can be performed on an RDD in Spark?

Ans**:** **A transformation to create a new RDD and an action to retrieve results from an RDD**

* Why does MapReduce NOT work well for streaming data?

Ans: **It operates on a huge number files stored in reliable storage, typically running jobs which can take hours or even days.**

* You receive log messages at the rate of two messages per second and your batch interval is five seconds.How many messages are included in one RDD in the DStream?

Ans**:10**