## **Big Data Programming**

## **Exam 2**

1. Under What circumstances is Map reduce better Performing than Apache Spark? Please give a practical example.

Ans: In terms of speed and handle size of data Spark is always powerful but efficient but, in some cases, Map Reduce is better option in terms of cost management where time is not a variant. Mar Reduce uses batch processing. It is a cost-effective solution for processing large amount of archived data. So, when it is not required real time processing map reduce can is suitable as it can be used in less memory and less memory means less cost in terms of hardware use. You need to process updated data say every day and data should be ready next day for reviews or analysis. So, data can be scheduled to be processed in batch during night hours. For example, a daily expenditure summery of a big company. Say it gets updated in every morning. Then all the transaction can be processed for all the branches over night processed through spark and can be summarized or analyzed next morning. In this case we will require less memory (RAM) less cost. It will not a be a better choice to buy expensive machine for Spark.

1. Explain RDD and how it differs from Data Frame and Data Set?

Ans: Apache Spark provides 3 type of Data model API’s.

RDD is one of them. resilient distributed dataset (RDD), which is a collection of elements partitioned across the nodes of the cluster that can be operated on in parallel. It is an Immutable distributed collection of objects. It is data set, so it has properties of Data Set API that is another API of Spark. It allows users to write parallel computations, using a set of high-level operators, without having to worry about work distribution and fault tolerance. [Stack overflow]. In a case of we lose some partition of RDD , we can replay the transformation on that partition in lineage to achieve the same computation, rather than doing data replication across multiple nodes. This characteristic is the biggest benefit of RDD because it saves a lot of efforts in data management and replication and thus achieves faster computations. It also consists lazy evaluation that also saves memory space when the RDD is initialized but not used.

**Data Frame:** A DataFrame is a distributed collection of data organized into named columns. It is conceptually equivalent to a table in a relational database or a R/Python Dataframe. Along with Dataframe, Spark also introduced catalyst optimizer, which leverages advanced programming features to build an extensible query optimizer. [Apache] So, a DataFrame has additional metadata due to its tabular format, which allows Spark to run certain optimizations on the finalized query. So, it is formatted in detail. It has a query optimization that also helps in some cases.

**Data Set:** Dataset API is an extension to DataFrames that provides a type-safe, object-oriented programming interface. It is a strongly-typed, immutable collection of objects that are mapped to a relational schema. It basically overcomes the drawbacks of Data frames and it also consist some great features of RDD. It differs from RDD by consisting data frame entities like more meta data and collection of row objects.

1. Provide an example type of problem you would use SparkML and outline a pipeline to solve it.

Ans:

One really useful example of SparkML to teach the machine to detect spam texts. It requires the machine to consistently compare the features and hash them into vectors. By iteration the system can be used to add more feature that would get the prediction percentage of the email to be spam higher. It will be more accurate.

* The Pipeline API can be used to detect. The dataset examples are Documents, Email messages, or other content received from external systems that can contain spam content.

The first to Step for SparkML is Data Ingestion and Data Cleansing.

* **Data Ingestion**: We’ll load the datasets (text files) for the content that has the spam data as well as the data that doesn’t contain any spam (called ham data).
* **Data Cleansing**: In our sample application, we don’t perform any specific data cleaning. We just aggregate all the data into a single DataFrame object.

We will use 60% train data and 40% test data. Having less trained data might not give us optimize prediction but for sake of speed we will do this now. Let’s see what solution if might provide us with.

The four Steps for Pipeline This problem are:

* Tokenizer
* HashingTF
* IDF
* LR

We load the data and convert them into data frames. We use methods to convert the dataset to RDD to data frame. We then split the data fram (both spam data and no spam data files) Then we create the four components of the pipleline, Tokenizer, HashingTF, IDF. The last step is to create a regression model in this case it would be Logistic Regression model by using the train Data.

1. Use of Facebook with different Big Data FrameWork and Tools.

Ans:

* + **Map Reduce:** Map Reduce uses batch processing so tasks that does not require instant analyzation is can be done using map reduce by Facebook. For example, Facebook posts, liked pages continuous searched topic, trending subjects this data set can be analyzed by Facebook with the help of Map Reduce. The petabytes of data sets of user info can be analyzed with lots of parrelelize jobs and by using job tracker it can be combined. Although map reduce takes lot of time compare to Spark, but It can be and was used by Facebook until it started using Corona instead of map reduce to delay the timing. Still it can be used as a daily basis to summarize task that is not required instant result.
  + **SparkML:** SparkML is will mostly be used to suggest adds and pages that the user has interest. Using the likes and post of the user the data can be analyzed and predicted and by SparkML it can be trained to suggest adds or pages that is related to user interest. Using SparkML algorithm during data analysis data can be imported ingested, cleaned and then pipelined to train to machine.
  + **Hive:** Hive has a SQL querry optimization. Hive can be used daily basis in Facebook. It can be used to do process the transactions of Advertisement account. As it is a language for real-time queries and row-level updates transactions can be done instantly. It is designed for Online analytical processing (OLAP). So individual Facebook advertisement accounts can be handled as it has some SQL familiarities. But in reality Hive is used for almost everything in Facebook on top of Corona Hive is used to analyze peta bytes of data every day.