University of Puerto Rico
Mayagüez Campus

Department of Computer Science and Engineering

CIIC 5995/8995: Big Data Analytics
Final Exam

Name: Andres R Hernandez TOTAL: /100

Student ID: 802-14-3188

Section: 116 **June 29, 2017**

Answer the following questions.

I) **(25 pts) Problem I: MapReduce**: Briefly explain the concept of MapReduce, and provide an example on what problems you can solve with it.

MapReduce works on key-value pairs. The pairs go through two phases defined by the user, the map and the reduce, to get to their final state. Also, several map reduce programs can be used one after the other to achieve a final goal. In the map phase, the key value pairs get transformed to an intermediate result that the reducers then can combine into a final result. If the MapReduce is running on HDFS, it has several advantages since it can run several mappers and reducers on parallel making the running time of a program more efficient.

MapReduce is good for when you can do you analysis in one pass, when the data used is so big that it won't fit in memory and to do batch processing. A good example of this would be batch processing of a Website’s logs to find patterns and analyze what people look for the most or when they are looking at the website.

II) **(25 pts)Problem II: Spark RDD**: Briefly explain the concept of the Spark RDDs and contrast their performance with that of MapReduce.

Spark RDDs or Resilient Distributed Datasets are read-only datasets that Spark uses to store and transform its data. The computations on RDDs are made using the Directed Acyclic Graph (DAG) model, which means that results are pipelined along an operator pipeline so there is no result materialization in between. This leads to performance gains. Spark also uses Lazy Evaluation to evaluate transformations and actions of RDDs. This means that nothing will be computed until Spark is forced to produce an output of a transformation by using an Action. Transformations of RDDs are how you morph an RDD into a new RDD. This has closure like a relational model. Actions of RDDs are when you compute a value of the RDD. By default, an RDD is kept in memory for a short time. If swapped from RAM, it will be recomputed. You can force the persistence of the RDD in memory by calling cache() or persistence() method. This will ensure that future operations see data from memory and makes the programs run faster. RDDs can also be cached with disk support or, if necessary, copied to disk and read when needed. This is faster than recomputing is most cases.

The performance of Spark RDDs versus MapReduce is a lot better because it does not have to write to disk as often. But both models have their advantages:

Spark advantages:
-It runs in memory on the cluster, so it does not have to go to disk as often.
-Since it runs in memory, it is faster for doing several iterative computations.
-It is interactive.
-It has streaming and sql capabilities.

MapReduce advantages:
-Good when you can do you analysis in one pass.
-Good when the data used is so big that it won't fit in memory.
-Good to do batch processing.

III) **(25 pts)**<u>**Problem III: Supervised Learning**</u>: Briefly explain the concept of Supervised Learning, and provide an example on what problems you can solve with it.

In supervised learning, algorithms learns to make predictions from a set of examples. The task is to infer a function from labeled training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object and a desired output value. Most implementations utilize regression to delimit the behavior of the machine. It is often used for classification and neural networks.

A possible example of a problem that can be solved with this concept is the sentiment analysis of social media. If you have a large data set of manually classified and analyzed data, you can input that into a Supervised Machine Learning algorithm to determine the sentiment in a post about any particular topic. This is very useful and has many possible applications.

IV) **(25 pts)**<u>**Problem IV: Unsupervised Learning**</u>: Briefly explain the concept of Unsupervised Learning, and provide an example on what problems you can solve with it.

In unsupervised learning, algorithms learns the structure of the data by itself. The task is inferring a function to describe the hidden structure from unlabeled data. Since the examples given to the machine are unlabeled, there is no evaluation of the accuracy of the structure that is output by the algorithm. Most implementations utilize clustering to delimit the behavior of the machine. It is often used for principal component analysis and outlier detection.

A possible example of a problem that can be solved with this concept would be the placement of antennas in a wireless network. A clustering algorithm can be used to determine the most optimal placement of antennas so that every building can be reached by the network.