University of Puerto Rico
Mayaguez Campus
Department of Computer Science and Engineering
CIIC 5995/8995: Big Data Analytics
Final Exam

Name: Luis G. Rivera	TOTAL: /100
Student ID: 802-14-6612	
Section: 116	June 29, 2017

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 consists in two phases for processing data, Map and Reduce. In the Map phase, the data gets sorted in key-value pairs set by the user and then sent to the reduce phase. In the reduce phase the data gets filtered as the user specified in the algorithm to get the desired results. It's a highly parallelizable framework when using with HDFS, when using both, the map and reduce phases can be parallelized using all the HDFS nodes. This framework is a good option for batch process a huge amount of data such that won't be able to fit on memory. A good example is processing tweets posted in a month to get the most trending hashtag.

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

The Spark RDD represents an immutable, partitioned collection of elements that can be operated on in parallel. It's a great option to perform in-memory computations on large clusters in a fault-tolerant manner. It's lazy evaluated, in other words, it's data won't be available or transformed until it's required and an action triggers the execution. Also you can store the data in a persistent memory or disk.

Spark and MapReduce have their advantages and disadvantages, it all depends of the use case.

Spark performance advantages:

- It runs in memory on the cluster, making it fast when processing data
- It can be used for streaming data and has sql features.

MapReduce performance advantages:

- It's great when analyses are made in one pass
- It's good when you analyze data that doesn't fit in memory and for batch processing.

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

With supervised learning algorithms, the user trains the machine in order to make predictions based on the training data given to the machine. Each example in the training data is interpreted as an input-output pair where the second ones is the desired output value. Regression is frequently used on these implementations. A good example that can be solved with supervised learning is a sentiment analysis of certain data as twitter posts. According to what you want to analyze on those posts you give the training data to the machine.

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

Using unsupervised learning the machine trains itself. It will be able to find the structure or relationships between different inputs. There are several unsupervised learning techniques like clustering, which will create different cluster of inputs and will be able to put any new input in appropriate cluster, anomaly detection, Hebbian Learning and learning Latent variable models such as Expectation–maximization algorithm, Method of moments (mean, covariance) and Blind signal separation techniques like Principal component analysis, Independent component analysis, Non-negative matrix factorization, and Singular value decomposition. A good example that can be solved with unsupervised machine learning is implementing a product recommendation system, where it recommends you certain products based on your purchase history.