University of Puerto Rico- Mayaguez Campus

Department of Computer Science and Engineering

CIIC 5995/8995: Big Data Analytics
Final Exam

Name: **LEO RAMIRO SOLORZANO** Student ID: **502-149526**

Section:

Answer the following questions.

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

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key.

MAP PHASE takes an input pair and produces a set of intermediate key/value pairs. The MapReduce library groups together all intermediate values associated with the same intermediate key ***I*** and passes them to the Reduce function.

REDUCE PHASE accepts an intermediate key ***I*** and a set of values for that key. It merges together these values to form a possibly smaller set of values. Typically, just zero or one output value is produced per Reduce invocation.

MapReduce is useful in a wide range of applications, including distributed pattern-based searching, distributed sorting, web link-graph reversal, web access log stats, document clustering and machine learning.

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

An RDD is, essentially, the Spark representation of a set of data, spread across multiple machines, with APIs to let you act on it. An RDD could come from any datasource, e.g. text files, a database via JDBC, etc.

The formal definition is: RDDs are fault-tolerant, parallel data structures that let users explicitly persist intermediate results in memory, control their partitioning to optimize data placement, and manipulate them using a rich set of operators.

Spark is framework for performing general data analytics on distributed computing cluster like hadoop. It provides in memory computations for increase speed and data process over Mapreduce. It runs on top of existing hadoop cluster and access hadoop data store (HDFS), can also process structured data in Hive and Streaming data from HDFS, Flume, Kafka, Twitter.

Keep data in memory is a good idea because (1) Partial results stay in memory (2) Next task does not need to do I/O (3) Performance Gain**.**

(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.

Supervised learning is the machine learning task of inferring 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. A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

A training set of examples with the correct responses (targets) is provided and, based on this training set, the algorithm generalises to respond correctly to all possible inputs.

Applications: Sentiment Analysis, Information extraction, Handwriting recognition, Spam detection, Pattern recognition, Speech recognition, etc.

(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.

Unsupervised machine learning is the machine learning task of inferring a function to describe hidden structure from "unlabeled" data (a classification or categorization is not included in the observations). Since the examples given to the learner are unlabeled, there is no evaluation of the accuracy of the structure that is output by the relevant algorithm—which is one way of distinguishing unsupervised learning from supervised learning.

Correct responses are not provided, but instead the algorithm tries to identify similarities between the inputs so that inputs that have something in common are categorised together.

Application: Behavioral-based detection in network security has become a good application area for a combination of supervised- and unsupervised-machine learning. This is because the amount of data for a human security analyst to analyze is impossible to review to find patterns and anomalies.