School of Computer Science

COMP47470

Project 2
Using a Hadoop Cluster to Crunch Data From New York City Cab System

Teaching Assistant:	Leandro Almeida
Coordinator:	Dr Anthony Ventresque
Date:	Thursday 30th March, 2017
Total Number of Pages:	2

General Instructions

- In this project, you will create a Hadoop/HDFS cluster in a Cloud environment (Microsoft Azure), write program(s) and run jobs to process public available data.
- The first task is to set up a Hadoop cluster running Azure VMs. The cluster needs to have at least several (3+) nodes, including a master node.
- You will use the Yellow Cab data, from New York City Taxi And Limousine Commission dataset, available at http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml.
- You will need to load the years 2009 to 2016 of the dataset in your cluster.
- You will then need to answer the following questions using the dataset and mapreduce programs:
  - What is the average number of passengers per trip in general, per month, and per day of the week?
  - What is the average trip distance in general, per month, and per day of the week?
  - What is the most used payment type? Create an ordered list for payment types.
  - Create a graph showing the average number of passengers over the day (per hour). Create a version for work days and another for weekend days.
  - Create a graph showing the average trip distance over the day (per hour). Create a version for work days and another for weekend days.
  - Create a graph showing the average number of passengers over the day (per hour). Create a version for work days and another for weekend days.
- You are encouraged to collaborate with your peers on this project, but all written work must be your own. In particular we expect you to be able to explain every aspect of your solution if asked.
- We ask you to hand in an archive (zip or tar.gz) of your solution: code/scripts and a 5-10 page pdf report of your work (no need to include code in it).
- We also ask you to give us the url of a VM that would give access to your configured environment.
- The report should include the following sections:
  1. a short introduction
  2. a requirement section that answers the question *what* is the system supposed to do
  3. a description of the software running on top of mapreduce that answers the proposed questions .
  4. a series of sections that describe the different challenges you faced and your solutions. For instance, take one of the script, describe the difficulty you faced and your solution. These sections can be short – the objective here is to show how you crafted the solutions with the tools you have learnt so far.

5. a short conclusion

- The project is worth **10% of the total grade** for this module. The breakdown of marks for the project will be as follows:
  - Create the cluster in Azure VMs: 15%
  - Inserting the data from NYC Cab System in the cluster: 10%
  - Creating and executing the mapreduce jobs answering the proposed questions: 50%
  - Report: 25%
- **Due date: 21/04/2017**