Assignment 3

Abhijeet Sharma, Afan Ahmad Khan

Friday, February 6, 2016

INTRODUCTION

The program takes a hadoop directory path as input containing multiple gzipped csv files and creates a plot displaying the average ticket prices of all airlines per month(restricted for airlines active in 2015).
The program also benchmarks and compares the cost of computing (A) mean and (B) median price, and (C) fast median for (i) single threaded Java, (ii) multi-threaded Java, (iii) pseudo-distributed MR, and (iv) distributed MR.
The program evaluates the performance of the following configurations i-A, i-B, ii-A, ii-B, iii-A, iii-B, iv-A, iv-B, iv-C.

Fine print:
1. This is a Group assignment of two students.
2. We Provide code that can run in single-threading, multi-threading, pseudo-distributed mode as well as on EMR.
3. We have produced a graph that plots and compares the cost of computing (A) mean and (B) median price and (C) fast median for (i) single threaded Java, (ii) multi-threaded Java, (iii) pseudo-distributed MR, and (iv) distributed MR.
4. We have included a script that executes everything and produces the graph. For example, if you use the Unix make command, you have two targets pseudo and cloud such that typing make pseudo will create a HDFS file system, start hadoop, run your job, get the output, and produce the graph. Typing "make cloud-mean"" will run the code on EMR for mean calculation.
5. We have only output airlines with flights in 2015.
6. This one page report documents our implementation and describes our results. The report is automatically constructed as part of running the project to include the plot.
7. We have submitted a tar.gz file which unpacks into a directory name "Sharma_Khan_A3". The directory contains a README file that explains how to build and run our code.

SYSTEM SPECIFICATION:

1. Java 1.7.0_79
2. Ubuntu 14.04 64-bit
3. 8GB RAM
4. Pandoc (https://github.com/jgm/pandoc/releases/tag/1.16.0.2)
4. R Packages:
4.1 Rcpp
4.2 R.utils
4.3 ggplot2
4.4 rmarkdown
4.5 plyr

Install the required R Packages

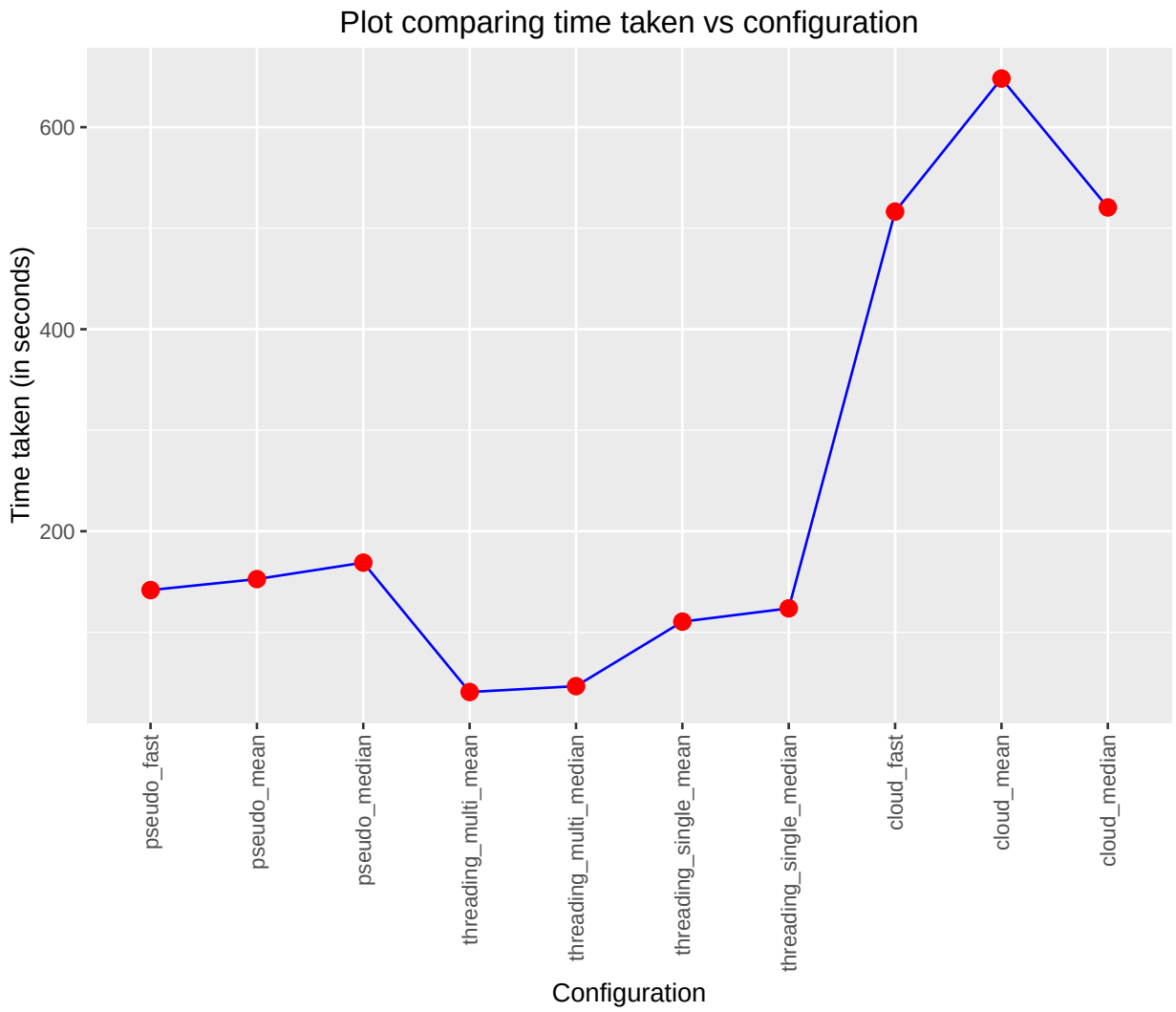
```
install.packages("Rcpp")
install.packages("R.utils")
install.packages("rmarkdown")
```

```
install.packages("ggplot2")
install.packages("plyr")
```

The R Markdown document can be run individually with the below command.

```
Rscript -e "rmarkdown::render('Assignment3_Report.Rmd')"
```

Plot



Implementation

1. We have implemented a Single Thread Program, a Multi Threaded Program, A Pseudo-Distributed map reduce system and a AWS EMR implementation of a map reduce system.
2. In the Map Reduce system, The Mapper calls the map method with "Carrier Code" as Key and a record of a csv file as Value.
3. The record is checked for sanity tests and Intermediate Key, Value Pairs are sent to the Reducer
4. The Key is the carrier code of the record. The Intermediate value is a Text object containing the

price, Month and Year of a single record.
5. The reducer calls the reduce method passing an iterable of values. The reducer calculates the mean/median/fast median price of a carrier per Month and outputs to a specific output file.
6. The R script reads the multiple time files and plots a line graph comparing the time performance(in seconds)
for all the different configurations.
7. Plot is drawn with ggplot2 package.
8. Make file is created to automate all the above steps.