**Introduction to Cluster Analysis**

**Setup, Compilation, Execution and Output Guidelines**

**Setup and Pre-requisites:**

Hadoop Setup and Configuration- Minimum 2.6

Environment built on Linux

Input dataset all in the Hadoop home directory i.e. /usr/local/Hadoop along with the working code provided in the tar.gz folder

Latex and Pandocs

Our Makefile consist of two rules, one for the program to execute on the Pseudo Distributed System and one for the EMR

(make-emr and make-aws)

For EMR, we need to have the AWS CLI installation and configuration done

All the deliverables are to be placed in the hadoop home so that we don’t face any path conflict

We have used Latex for pdf Report generation (MyReport.pdf)

We have used MarkDown for HTML report generation (MyReport.html)

The R script used for generating the Graph from the output of the Hadoop reducer is TimeAnalysis.R

**Compilation and Execution:**

Pseudo Distributed System-

We are providing a jar file containing all compiled classes or one can compile the same using below commands

javac -cp /usr/local/hadoop/share/hadoop/mapreduce/lib/hadoop-annotations-2.6.3.jar:/usr/local/hadoop/share/hadoop/common/hadoop-common-2.6.3.jar:/usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-client-core-2.6.3.jar:/usr/local/hadoop/opencsv-2.2.jar  <filename>.java

jar cvf <filename>.jar <filename>\*.class

run the rule – make pseudo // This will start the Hadoop system, run the job and provide the output at <Hadoop-home>/output

EMR-

For the cluster setup and code to run on EMR, follow the rule

make emr

This will create a cluster, run the jar file at the mentioned location and terminates the cluster on the execution of the given job.

>>> We have 7 java files as a part of this submission :::

Cluster\_Analysis\_Average.java (run in pseudo distributed and AWS)

Cluster\_Analysis\_Median.java (run in pseudo distributed and AWS)

Cluster\_Analysis\_Fast\_Median.java (run in pseudo distributed and AWS)

Threaded\_Average.java

Threaded\_Median.java

Threaded\_Fast\_Median.java

Sequential.java (This is the sequential file implementation that outputs all the required outputs as per requirement)

>>> 1 R files ::

TimeAnalysis.R

>>> 1 sbt build file

Results ->

The timings for all the above files are calculated while running the dataset for A2. The detailed analysis is written out in the report with this submission. The time calculated is read by TimeAnalysis.R and the time graph (Time.png). This graph is integrated in the report called TimeAnalysis.pdf There is also a second graph called MyGraph.png which is the graph plotted for average ticket price.

**References:**

[https://docs.google.com/document/d/18HkaZ0tSAH8xjDOZjPULBfAJw5H\_vj33QSLemseRM2o/edit?pref=2&pli=1#](https://docs.google.com/document/d/18HkaZ0tSAH8xjDOZjPULBfAJw5H_vj33QSLemseRM2o/edit?pref=2&pli=1)

http://stackoverflow.com/

<https://developer.yahoo.com/hadoop/tutorial/module2.html>

<https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SingleCluster.html>

---------------------------------------END--------------------------------------------------