ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

Department of Applied Informatics
Big Data Analysis
Academic year 2022-2023

First group work
Delivery date: Wednesday 7 December 2022

You are given a data set, which comes from Kaggle and concerns consumer transactions. Each row relates to one transaction and contains product codes that were purchased. The dataset can be downloaded as follows:

wget snf-803830.vm.okeanos.grnet.gr/data.zip

Your goal is to build a MapReduce program that identifies subsets of products that are frequently purchased, i.e., that appear in at least N transactions, where N is given as a parameter at runtime.

Solve the above problem with Hadoop & MapReduce in distributed configuration, for a) one node, b) two nodes. Set the following transaction values as thresholds: 5000, 10000, 50000

Your solution may consist of more than one MapReduce cycle.

Example of execution for subsets of products purchased at least 2 times:

Input file:
1 2 3 4
1 2 3
2 2
1

Exit:
[1, 2, 3] 2
[1, 2] 2
[1, 3] 2
[1] 3
[2, 3] 2
[2] 3
[3] 2

For each of the above configurations, perform 5 runs and after subtracting the highest and lowest values, record the average of the 3 remaining runs for each of the following times:
Elapsed Time, Average Map Time, Average Reduce Time.

Instructions

1 **The work is carried out exclusively in groups of 2 students. If you cannot find a partner, please use this link.**
2 The programming language is Java.

3 You will deliver:

3.1 Your code.

3.2 A report that will include:

3.2.1 The algorithm you used to solve the problem.

3.2.2 Any limitations and considerations of your solution. 3.2.3 Additional packages, if you used them.

3.2.4 The system parameters, if modified.

3.2.5 Graphs and tables with the times you measured. 3.2.6 Commentary on the times you observed.

3.3 The output files with the result of your program for each number of reducers you used.

3.4 The .jar you will create.

4 The assignment will be delivered **electronically exclusively via openeclass** in the form of **a** zip **file.**

5 To measure the times, you will use the job history server. It starts with the command:

```
~/hadoop/sbin/mr-jobhistory-daemon.sh start historyserver
```

And it terminates if we put a stop at the start. You can find it at the following address for your machine:

<MASTER_NAME>:19888

6 To run on a node, you can use the following script (without the optional parameters) to terminate the appropriate service:

```
~/hadoop/sbin/yarn-daemon.sh stop
```

Indicative References:

1 Chapters 2.1-2.3 from the book "Mining from Big Data Sets", also available online at http://www.mmds.org/

2 "Data-Intensive Text Processing with MapReduce." Available online at https://lintool.github.io/MapReduceAlgorithms/

3 Notes of the course