Department of Applied Informatics
Big Data Analytics Academic Year
2022-2023

Second group assignment
Due date: Friday, 03 February 2023

You are given a dataset, which comes from Kaggle, of consumer transactions. Each line is for one transaction and contains product codes purchased. The dataset can be downloaded as follows:

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

1. Using the implementation of the FP-Growth algorithm of the MLLib library, which is part of Apache Spark, construct a program that will compute frequent tuples and take as a parameter at runtime a parameter N, where N is the minimum occurrence frequency of each product and will output a related file.

2. Thresholding the following transaction values: 5000, 10000, 50000,
Perform 3 runs and record: a. the minimum,
maximum and average time required for the calculation. b. You will do the above: i. For a worker ii. For two workers

3. Comment on the results of the above measurements.

The time of each execution can be seen in your browser at <ip-of-master>:8080

Instructions 1. The work is done exclusively in groups of 2 people.
2. You will deliver:
2.1. A code file and the command you use to run it (see also in the lesson example).
2.2. A report of up to 3 pages on
the steps you followed, the problems you encountered and how to solve them. 2.3. The commands and parameters you used in a .txt file.
2.4. A file with the resulting frequent datasets. 2.5. The output of the Spark run. 3. The assignment should be submitted
electronically via openeclass in the
form **of a single**
zip **file .**

References: 1. Chapter
6 from the book "Mining from Big Data Sets", also available
online at http://www.mmds.org/ (English)
2. Chapter 5.6 from the book "Introduction to Data Mining" by Ning Tan, Steinbach, Kumar.
3. "Data-Intensive Text Processing with MapReduce". Available online at https://lintool.github.io/MapReduceAlgorithms/
4. https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.fpm.FPGrowth.html#pyspark.ml.fpm.FPGrowth
5. https://spark.apache.org/docs/latest/ml-frequent-pattern-mining.html