

BIG DATA - TERM PROJECT

Implementing Map-Reduce Tools

Students:

Osman Araz, 16011020
Muhammet Çeneli, 15011025

Instructor:

Mehmet Aktaş

Delivery Date: 14.05.2020

1. About the System:

In this project, we designed and implemented a big data processing application for statistical analysis of pain pills in the USA and some other countries. We took a step further and implemented our project as generic as possible that is suitable for any appropriate dataset.

2. Technical Challenges:

The most challenging part was installing and setting the Hadoop software. We struggled with it almost a week. We overcame it with lots of readings and tutorials. Implementing the system and GUI was relatively simpler.

3. System Details:

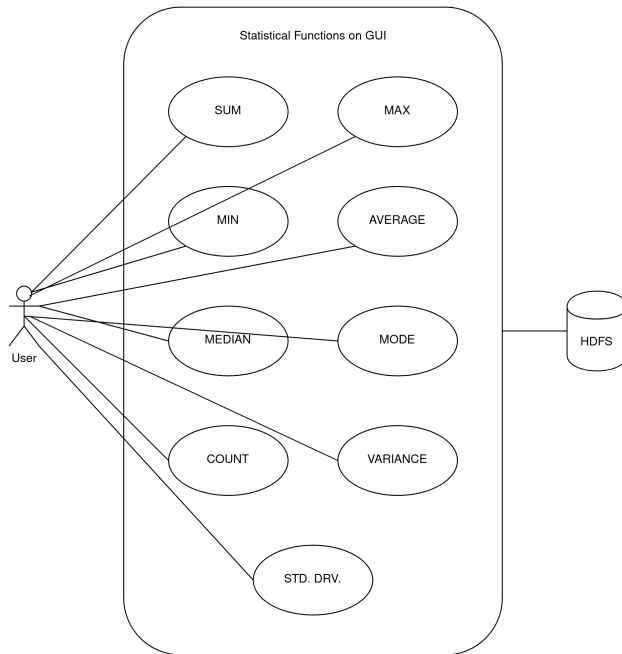
We created a software that allows people to perform some statistical analyzes on their own datasets. These statistical functions are as follows:

- Sum
- Maximum
- Minumum
- Average
- Median
- Mode
- Count
- Variance
- Standard Derivation

Our software is highly flexible and generic. People can select the feature in the dataset that will be performed for statistical analysis. It is also possible to combine features for creating the input key.

We used Java for Map-Reduce functions. For GUI, we used Qt with C++. We connect these with command line arguments.

4. Use Case Scenario:



5. Performance Evaluation:

Elapsed times of a statistical analysis processing (SUM function) for some dataset sizes are given below:

Dataset Size (# of rows)	Elapsed Time (seconds)
10	20.21
10.000	20.51
100.000	23.13

6. Comments:

We learned a lot about network communication and big data techniques in this project.

7. Screenshot:

Hadoop Statistical Tools

Hadoop path:

/home/hadoop/hadoop313

Dataset path:

/home/hadoop/Documents/samples.csv

Target column:

3

... depends on columns:

1, 2

Statistical function:

☐ Sum

☐ Median

☐ Mode

☐ Average

☐ Maximum

☐ Minimum

☐ Count

☒ Variance

☐ Std. Dev.

START HADOOP

PROCESSES COMPLETED. RESULTS:

ABBEVILLE, HYDROCODONE	VAR: 0.2666666666666667
ABERDEEN, OXYCODONE	VAR: 5.733333333333334
ABINGDON, HYDROCODONE	VAR: 5.866666666666667
ABINGDON, OXYCODONE	VAR: 1.2753623188405798
ABSECON, OXYCODONE	VAR: 18.666666666666668
ACCOKEEK, HYDROCODONE	VAR: 0.19230769230769232
ACCOKEEK, OXYCODONE	VAR: 22.77441269841271
ACCORD, OXYCODONE	VAR: 0.8904761904761902
ADAMS, HYDROCODONE	VAR: 1.6885057471264373
ADAMSVILLE, HYDROCODONE	VAR: 8.0
ADDISON, HYDROCODONE	VAR: 0.11111111111111115
ADDISON, OXYCODONE	VAR: 0.0
ADEL, HYDROCODONE	VAR: 0.3333333333333333
ADELPHI, HYDROCODONE	VAR: 0.0

YTÜ - Big Data Term Project, May 2020