Big Data Mining and Applications - Term Project Overview

Lars Fikkers (108012050), Pascal Roose (108012051)

About

The term project was made with PySpark and executed/tested on Jupyter Notebook.

Check out our repository for all of the files including the output:
https://github.com/PascalRoose/bigdatamining/

Included in this directory is FinalReport.pdf which goes into depth about our approach findings and conclusion. Also in this directory is FinalPresentation.pdf

Detailed Responsibility

Member	Part
Pascal Roose	Data analyzation (statistics.py) Estimation by averages (averages.py)
Lars Fikkers	Implementing AI (decisiontree.py) Final Report

Environment setup

Host: Digital Ocean
Specs: (used multiple configurations)

- Ubuntu 18.04.3 (LTS) x64
- 2-32 vCPUs
- 4-64GB RAM / 25GB ROM
- Region Singapore(1)

Spark:

- Cluster mode
    - Master: http://104.248.150.122:8080/
    - Worker: http://104.248.150.122:8081/
- With Python (PySpark)
- With Jupyter Notebook as driver

Setup and installation

Requirements

```
sudo apt install python3
sudo apt install pip-python3
```

Apache Spark

Download, unpack and move

```
wget https://www-eu.apache.org/dist/spark/spark-2.4.4/spark-2.4.4-bin-hadoop2.7.tgz
tar xvf spark-2.4.4-bin-hadoop2.7.tgz
mv spark-2.4.4-bin-hadoop2.7 /usr/local/spark
```

Create then edit the configuration file using nano

```
mv /usr/local/spark/conf/
cp spark-env.sh.template spark-env.sh
sudo nano spark-env.sh
```

Add the following line: SPARK_MASTER_HOST = ‘your ip-address’

Exit using Ctrl+x, press enter to save

Start the master- and worker-node

```
cd /usr/local/spark/sbin/
./start-master.sh
./start-slave.sh spark://ip-address:port
```

PySpark

```
pip3 install pyspark
```

Jupyter Notebook

```
pip3 install jupyter
```

Connect Jupyter Notebook to Spark

Setup environment variables

```
export PATH=$PATH:/usr/local/spark/bin
export SPARK_HOME=/usr/local/spark
export PYSPARK_DRIVER_PYTHON=jupyter
export PYSPARK_DRIVER_PYTHON_OPTS='notebook'
```

Run Jupyter Notebook with PySpark

```
cd /path/to/homework/src
pyspark
```