BLM4120 Big Data Processing
Term Project

19011618 Umut Sümer
17011080 Engin Deniz Çağlar

Introduction:

In this project we implemented a big data processing application for statistical analysis of “Used Car Dataset” obtained from Kaggle. We analyzed the price, exterior color and the brand of the cars.

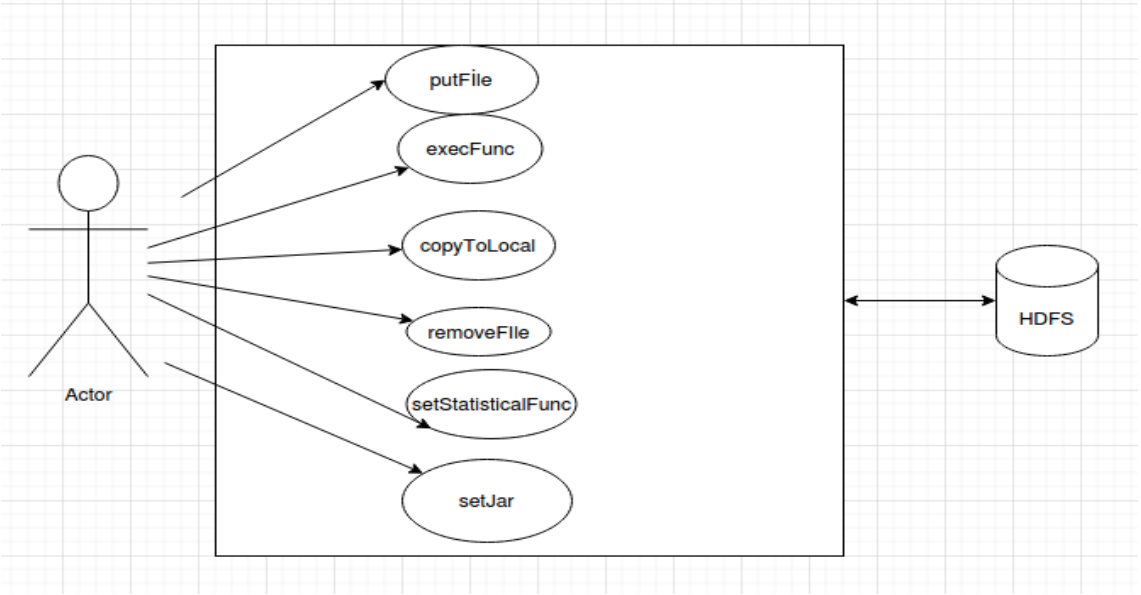
Motivation:

We were curious about the car prices and their correlations with other specifications. Therefore we found a large dataset which contains the price, car brand, city, color etc. Analyzing such a large database would be tough with traditional ways but using a distributed file system made it easy.

Implementation Environment:

Hadoop 3.2.1
IntelliJ
Ubuntu 16.04
Java Swing

Use-Case Diagram:



Technical Challenges:

Setting up the Multi Node Hadoop Environment was a challenge by itself. Some tutorials from blogs helped us overcome this. Since we are not familiar with hadoop file system, we faced some problems while trying file operations. Some of the issues we had are shown below:

Problem 1)

```
hadoopuser@hadoop-master:~$ start-dfs.sh
Starting namenodes on [hadoop-master]
pdsh@hadoop-master: hadoop-master: rcmd: socket: Permission denied
Starting datanodes
pdsh@hadoop-master: hadoop-slave2: rcmd: socket: Permission denied
pdsh@hadoop-master: hadoop-slave1: rcmd: socket: Permission denied
Starting secondary namenodes [hadoop-master]
pdsh@hadoop-master: hadoop-master: rcmd: socket: Permission denied
```

Solution 1)

Problem 2)

We formatted the hadoop file system twice. After formatting the file system second time, datanodes started to crash after every execution.

Solution 2)

We removed the files in data folders. For master node, namenode folders are removed. For slave nodes, datanode folders are removed.

Explanation:

We created an application which uses statistical functions on a large dataset. The statistical functions are :

1-Range : Range between prices according to the cities that cars sold.

2-Mode : Most used car color.

3-Average : Average prices for brands.

4-Standard Deviation : Standard deviation of prices according to brands.

5-Sum : Total number of cars of all brands.

Experience and Discussion:

We experienced the multi-node cluster system. With the help of using distributed computers, we processed the data very fast against one node system. Using Ubuntu for this project, helped us to understand how to use bash properly. Now we can use Hadoop Distributed File System when we work with large-scaled data.