Alaa bin Othman
Nov 30 · 6 min read · Listen

The Implementation of MapReduce on STC Data

Contents

1. Introduction

1.1 Saudi vision 2030

All success stories start with a vision, and successful visions are based on strong pillars.

The Saudi vision 2030, revealed in 2016 by Crown Prince Mohammed bin Salman, is founded on three pillars: A Vibrant Society, a Thriving Economy, and an Ambitious Nation.

Saudi Vision 2030: Overview — Vision 2030

1.2 STC Company

STC or “Saudi Telecom Company” is one of the biggest telecommunication and digital services companies in Saudi Arabia and the Middle East. The company offers landline and fixed infrastructure, mobile, and data services. STC offers mobile, broadband, and cloud computing services.

1.3 STC and Saudi Vision 2030

Like any other company in Saudi, STC wants to take advantage of the opportunities that Saudi vision 2030 opens.

One of the Pillars of the Saudi 2030 vision is “a V**ibrant Society**”, STC is taking advantage of the 2030 vision by strengthening the economy of Saudi Arabia.

To support the “T**hriving Economy**” pillar, STC launched the Saudi Vision Cable project as one of its attempts to achieve vision 2030.

The Saudi Vision Cable spans 1,160,000 meters and it is fully owned by STC Group. The Saudi Vision cable is the first ever high-capacity submarine cable in the Red Sea region that will provide seamless connectivity up to 18Tbps/fiber pair with a total of 16 fiber pairs through four landings in Jeddah, Yanbu, Duba, and Haql.

STC Launches “Saudi Vision Cable”, the First high-capacity Submarine Cable in the Red Sea The official Saudi Press Agency (spa.gov.sa)

STC is doing that because the government plans to spend 13.3 billion USD in the telecommunication industry to achieve the objective of the Saudi 2030 vision.

The main goal of the “**Ambitious Nation**” pillar is to empower the Saudis and the private companies to take better steps and continue improving.

That is what STC did. It takes better steps, said the company, like introducing the revised CARE strategy and coming up with CARE 2.0.CARE 2.0 continues to improve STC employees’ skills and enriches client experiences. The table below shows the column’s name along with its description.

2. STC Dataset Description

The data set used in this project is **Uncommon handheld devices,** it has been taken from the official website of STC company. see this link: Open Dataset (stc.com.sa). It describes the uncommon handset devices usage by customers, in the span of 2 years and with specific customer demographics.

The STC dataset contains **714023 rows** and **20 columns**. The columns are about both the STC

2	MODEL_NAME	Name of devise model
3	BRAND_FULL_NAME	The full name of devise brand
4	BRAND_NAME	The device brand short name
5	VENDOR_NAME	Name of device vendor
6	OS_NAME	Device operating system name
7	DEVICE_TYPE	The type of the device
8	_2G_FLG	Version 2G type of the internet service
9	_3G_FLG	Version 3G type of the internet service
10	_4G_FLG	Version 4G type of the internet service
11	WIFI_FLG	Device WIFI availability
12	BLUETOOTH_FLG	Device Bluetooth availability
13	DUAL_SIM_FLG	Whether the device has 2 SIM or not
14	TOUCH_SCREEN_FLG	Whether the device has touch screen or not
15	GENDER_TYPE_CD	User gender
16	AGE_B	User age
17	NATIONALITY_CD	User nationality entered as alpha-3 country code
18	NATIONALITY_NAME	User nationality entered as country name
19	SAUDI_NON_SAUDI	User nationality is Saudi or not
20	DEVICE_COUNT	Count of device purchased

STC Dataset Description

2.1 The reason for choosing STC data and How it's related with the 2030 vision (problem statement)

We chose this dataset as it contains a large number of types that are not commonly used by users in Saudi Arabia and the percentage of their acquisition by users of different nationalities.

And the analysis of this data is very important for a company like STC, as it is one of the most important companies contributing to the **technical transformation**. As this helps the company and its parties concerned with the technical transformation in line with **Vision 2030** by knowing what can be provided and whether the current provider or the one under development is compatible with these types of devices.

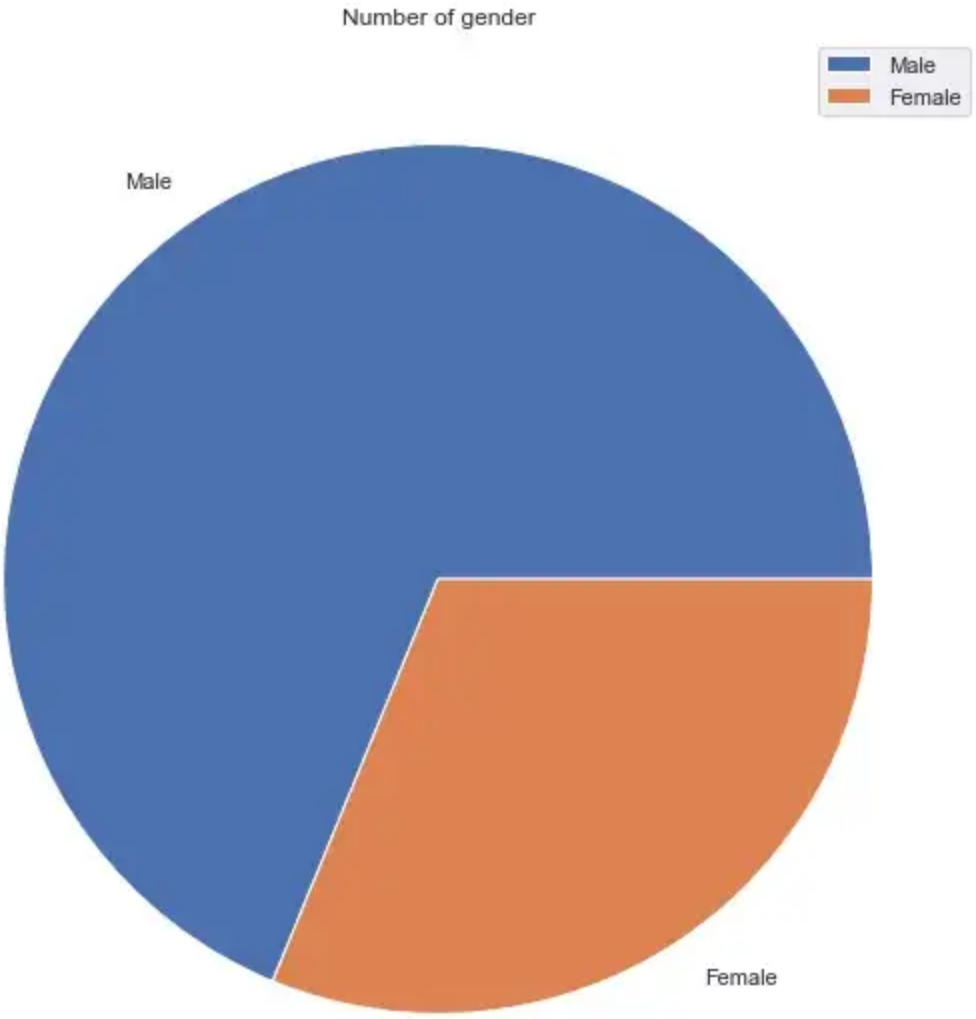
On the other hand, STC's sales performance is based on the quality of the devices used. Achieving the goals related to sales reflects the company's strength, stability of its indicators,

devices' performance by using MapReduce to count them for each performance category.

4. Exploratory Data Analysis (EDA)

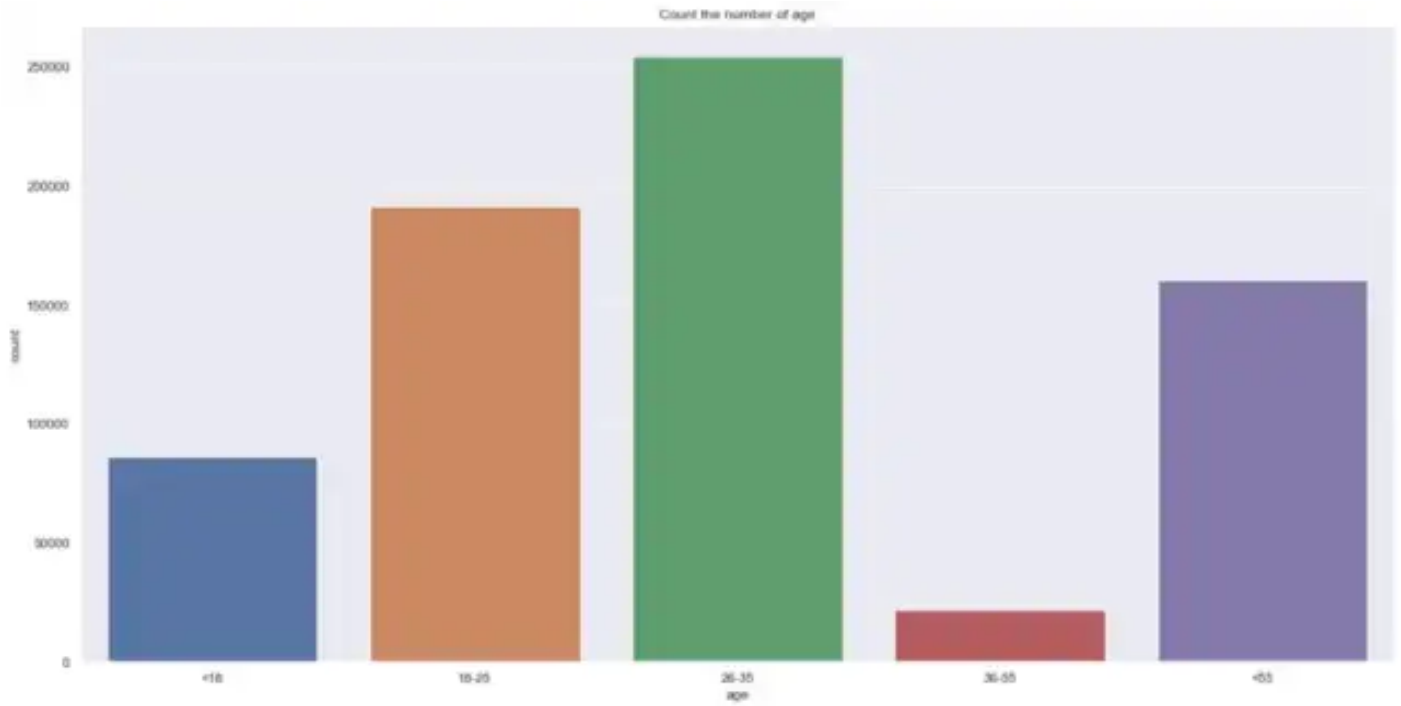
(1) Users gender count

As we can see from the chart, majority of our customers' gender is male. About 33% of the customers are female.



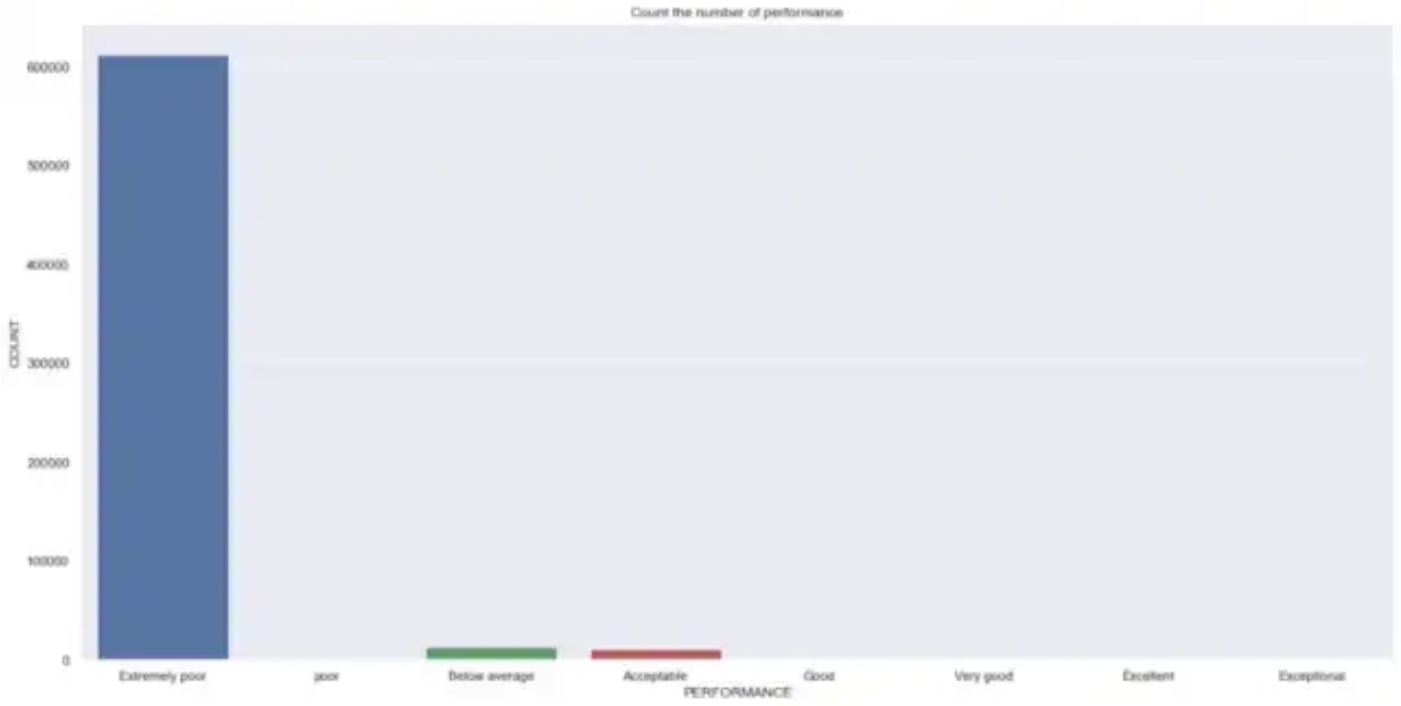
(2) Users age range count

As we can see from the chart, customers under the age of 18 are less than 10000. Customers between the ages of 18–25 and between 26–35 are more inclined to purchase a device whether its for personal use or for work. Customers older between 36–55 are usually content with the



(3) Count the number of performances

This graph shows the categories of the performance columns, we have several categories, from the graph we can see that most of our data falls under the extremely poor category. And there is a few in below average and acceptable categories and there is not much in other categories.



finding unique values for all features

```
stc_d['AGE_B'].unique()
```

step 2

dropping unnecessary values

```
stc_d.drop(stc_d[(stc_d['NATIONALITY_NAME']=='18-25') | (stc_d['NATIONALITY_NAME']== '<18  ') |
                 (stc_d['NATIONALITY_NAME']== 'N') | (stc_d['NATIONALITY_NAME']== 'NA   ')].index, axis=0,inplace=True)
```

```
stc_d.drop(stc_d[stc_d['DEVICE_COUNT'] > 15000 ].index ,axis=0, inplace = True) # outliers
```

step 3

removing unnecessary columns

```
stc_d.drop(columns="BRAND_FULL_NAME",axis=1,inplace=True)
```

step 4

removing duplicated columns after encoding

```
stc_d.drop(columns="_3G_FLG",axis=1,inplace=True)
stc_d.drop(columns="_2G_FLG",axis=1,inplace=True)
stc_d.drop(columns="_4G_FLG",axis=1,inplace=True)
stc_d.drop(columns="WIFI_FLG",axis=1,inplace=True)
stc_d.drop(columns="MODEL_NAME",axis=1,inplace=True)
stc_d.drop(columns="OS_NAME",axis=1,inplace=True)
stc_d.drop(columns="VENDOR_NAME",axis=1,inplace=True)
stc_d.drop(columns="BRAND_NAME",axis=1,inplace=True)
stc_d.drop(columns="DEVICE_TYPE",axis=1,inplace=True)
stc_d.drop(columns="SAUDI_NON_SAUDI",axis=1,inplace=True)
stc_d.drop(columns="NATIONALITY_NAME",axis=1,inplace=True)
stc_d.drop(columns="AGE_B",axis=1,inplace=True)
stc_d.drop(columns="GENDER_TYPE_CD",axis=1,inplace=True)
stc_d.drop(columns="DUAL_SIM_FLG",axis=1,inplace=True)
stc_d.drop(columns="TOUCH_SCREEN_FLG",axis=1,inplace=True)
stc_d.drop(columns="BLUETOOTH_FLG",axis=1,inplace=True)
```

```
stc_d["CAL_DT"]=pd.to_datetime(stc_d["CAL_DT"])
```

```
stc_d["CAL_DT"]=pd.to_datetime(stc_d["CAL_DT"]).dt.strftime('%Y')
```

```
stc_d["DEVICE_COUNT"]=stc_d["DEVICE_COUNT"].astype(str).astype(int)
```

step 2

Using label encoder on the columns

```
le = preprocessing.LabelEncoder()
stc_d["2G_FLG"]=le.fit_transform(stc_d["_2G_FLG"])
stc_d["3G_FLG"]=le.fit_transform(stc_d["_3G_FLG"])
stc_d["4G_FLG"]=le.fit_transform(stc_d["_4G_FLG"])
stc_d["WIFI"]=le.fit_transform(stc_d["WIFI_FLG"])
stc_d["BLUETOOTH"]=le.fit_transform(stc_d["BLUETOOTH_FLG"])
stc_d["TOUCH_SCREEN"]=le.fit_transform(stc_d["TOUCH_SCREEN_FLG"])
stc_d["DUAL_SIM"]=le.fit_transform(stc_d["DUAL_SIM_FLG"])
stc_d["GENDER"]=le.fit_transform(stc_d["GENDER_TYPE_CD"])
stc_d["MODEL"]=le.fit_transform(stc_d["MODEL_NAME"])
stc_d["BRAND"]=le.fit_transform(stc_d["BRAND_NAME"])
stc_d["VENDOR"]=le.fit_transform(stc_d["VENDOR_NAME"])
stc_d["OS"]=le.fit_transform(stc_d["OS_NAME"])
stc_d["DEVICE"]=le.fit_transform(stc_d["DEVICE_TYPE"])
stc_d["AGE"]=le.fit_transform(stc_d["AGE_B"])
stc_d["NATIONALITY"]=le.fit_transform(stc_d["NATIONALITY_NAME"])
stc_d["SAUDI"]=le.fit_transform(stc_d["SAUDI_NON_SAUDI"])
```

7. MapReduce

Definition

> *MapReduce is a programming paradigm that enables massive scalability across hundreds or thousands of servers in a Hadoop cluster. As the processing component, MapReduce is the heart of Apache Hadoop.*

steps for MapReduce

First : We used the STC cleaned data as an input to the MapReduce function.

Second : We saved the performance column as a text file, which include several categories of the

```
np.savetxt(r'C:/Users/DELL/OneDrive/Desktop/Lina/Big data & AI bootcamp/stcPerformance.txt', stc_d.PERFORMANCE, fmt='%s')
```

Third : We changed the file extension from .txt to .CSV

```
from mrjob.job import MRJob
from mrjob.step import MRStep

class PerformanceBreakdown(MRJob):
    def steps(self):
        return [
            MRStep(mapper=self.mapper_get_performance,
                   reducer=self.reducer_count_performance)
        ]

    def mapper_get_performance(self, _, line):
        Performance = line.split('\n')
        yield Performance, 1

    def reducer_count_performance(self, key, values):
        yield key, sum(values)

if __name__ == '__main__':
    PerformanceBreakdown.run()
```

MRJob function

Fourth : We used the MRJob function. Which includes the mapper and the reducer functions . The mapper function maps the “Performance” with its count . And the reducer function sums the values for each “Performance” category .

Output

```
Total input paths to process : 1
number of splits:2
Submitting tokens for job: job_1669074792817_0008
Submitted application application_1669074792817_0008
The url to track the job: http://sandbox-hdp.hortonworks.com:8088/proxy/application_1669074792817_0008/
Running job: job_1669074792817_0008
Job job_1669074792817_0008 running in uber mode : false
 map 0% reduce 0%
 map 1% reduce 0%
 map 3% reduce 0%
 map 12% reduce 0%
 map 20% reduce 0%
 map 27% reduce 0%
 map 36% reduce 0%
 map 45% reduce 0%
 map 53% reduce 0%
 map 62% reduce 0%
 map 67% reduce 0%
 map 100% reduce 0%
 map 100% reduce 68%
 map 100% reduce 100%
Job job_1669074792817_0008 completed successfully
Output directory: hdfs:///user/root/tmp/mrjob/1661211802__RatingsBreakdown.root.20221122.214416.005229/output
ounters: 49
```

```
job output is in hdfs:///user/root/tmp/mrjob/1661211802__RatingsBreakdown.root.20221122.214416.005229/output
Streaming final output from hdfs:///user/root/tmp/mrjob/1661211802__RatingsBreakdown.root.20221122.214416.005229/output...
["Acceptable"]  10522
["Below average"]       13206
["Excellent"]   84
["Exceptional"] 26
["Extremely poor"]      612394
["Good"]        970
["Poor"]        76113
["Very good"]   687
```

result of MapReduce

8. Conclusion

based on the charts STC will understand their customer types “gender and age” and will be able to provide better products to suit them. furthermore, looking at the results of using MapReduce we were able to determine the amount of STC devices performance for each category, this will help the company evaluating the performance of their products which will increase their sales in the future.