الأكاديمية السعودية الرقمية
SAUDI DIGITAL ACADEMY

CODING DOJO

Analyzing Saudi Stock Exchange (Tadawul) with Hadoop

DaQuest Team

Today's Agenda

01 Introduction

02 Data Review

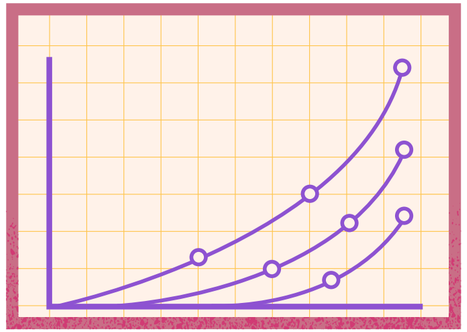
03 Data Preprocessing

04 Our Approach: Using Different MapReduce Functions

05 Results and What Can Be Improved

Introduction

- Saudi’s 2030 Vision and Tadawul
- Business problem



Tadawal Dataset:

- This is the data of Saudi stock market companies from 2001–12–31 until 2020–04–16.
- Collected from Saudi Stock Exchange (Tadawul) website.
- This dataset has 14 columns and 593819 rows.

Tadawal Dataset:

Each row in the database represents the price of a specific stock at a specific date:

(symbol (Integer	The symbol or the reference number of the company.
(name(String	Name of the company.
(trading_name (String	The trading name of the company
(sectoer (String	The sector in which the company operates.
(date (Date	The date of the stock price.
(open (Decimal	The opening price.
(high (Decimal	The highest price of the stock at that day.
(low (Decimal	The lowest price of the stock at that day.
(close (Decimal	The closing price.
(change (Decimal	The change in price from the last day.
(perc_Change (Decimal	The percentage of the change.
volume_traded (Decimal)	The volume of the trades for the day.
value_traded (Decimal)	The value of the trades for the day.
(no_trades (Decimal	The number of trades for the day.

Data preprocessing :

- Change the column Name
- Add year, month and day columns
- Drop unnecessary columns:

  1-Symbol
  2-Company Name
- Add the categorical target column

  We added a new column from the per_change column in the dataset, which is the Change_category (the classification target).
  We create the classes from the per_change as below:

  Good change: if the per_change > 0
  Bad Change: if the per_change < 0
  Stable: if the per_change = 0
- Handle the null values

Business Problem:

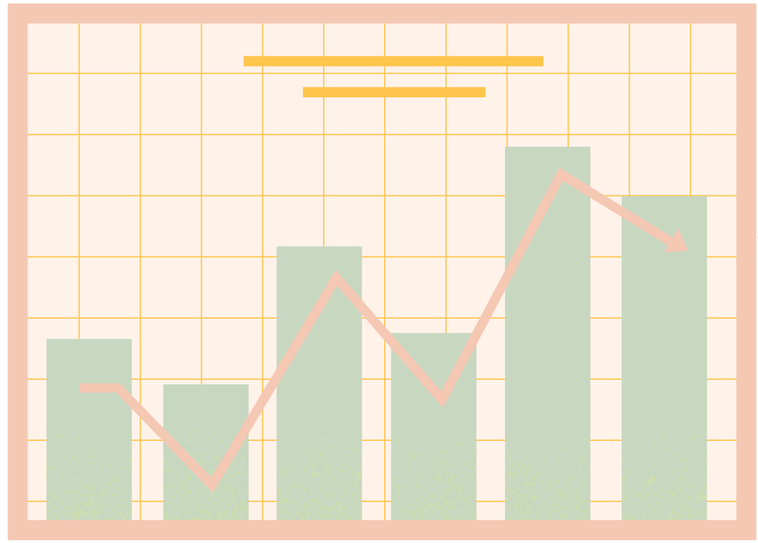
- Count change category for each sector.
- Count number of companies in each sector.
- Count the change categories for each year, month, and day.
- Count the number of shares in each year.

First Business Problem:

1-Count change category for each sector:

Our dataset contains a variety of sectors, it is critical to assess each one's performance using the change categories. We achieve this by grouping each sector's good, stable, and bad changes.

our approach

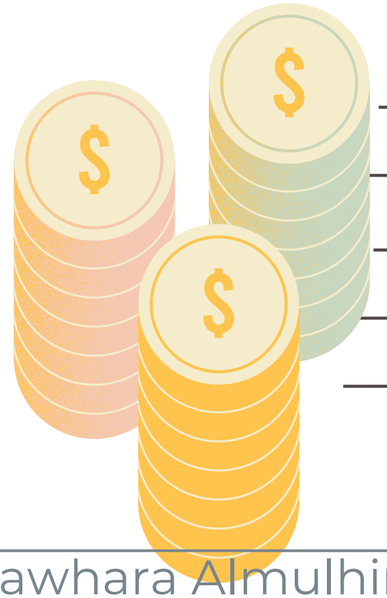


MapReduce Function:

```
from mrjob.job import MRJob # import the mrjob library
from mrjob.step import MRStep # import the mrStep library

class MRtarget(MRJob):
    def steps(self):
        return [
            MRStep(mapper=self.mapper_get_target,
                   reducer=self.reducer_count_target)
        ]

    # the map step: each line in the csv file is read as a key, value pair

    def mapper_get_target(self, _, target):
        yield target, 1


    # the reduce step: combine all tuples with the same key
    # then sum all the values of the tuple, which will give the total value

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

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

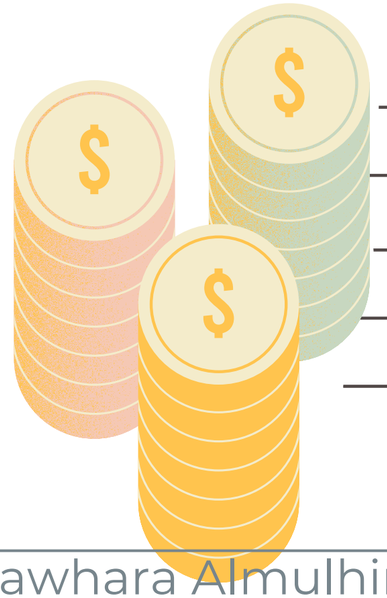
- Running Hadoop Cluster on a Sector_change data:

```
[maria_dev@sandbox-hdp ~]$ python project_MapReduce.py -r hadoop --hadoop-stream
ing-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar data._Sect
or_change.csv
```



Output:

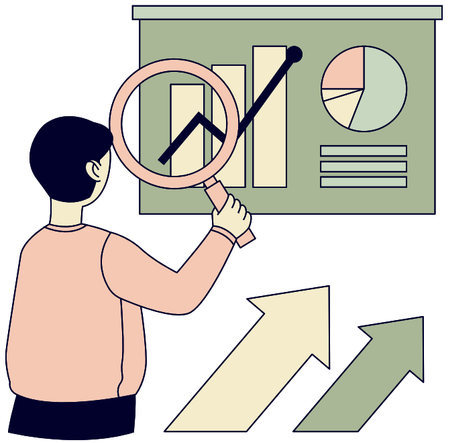
```
aria_dev.20221129.173214.064348/output...
"Communication Services,Bad Change"     10023
"Communication Services,Good Change"    9068
"Communication Services,Stable" 2257
"Consumer Discretionary,Bad Change"     24937
"Consumer Discretionary,Good Change"    23996
"Consumer Discretionary,Stable" 5923
"Consumer Staples,Bad Change"   28450
"Consumer Staples,Good Change"  27502
"Consumer Staples,Stable"       8734
"Energy,Bad Change"     7202
"Energy,Good Change"    6904
"Energy,Stable" 1625
"Financials,Bad Change" 67244
"Financials,Good Change"        63151
"Financials,Stable"     14455
"Health Care,Bad Change"        8259
"Health Care,Good Change"       8018
"Health Care,Stable"    1548
"IT,Bad Change" 394
"IT,Good Change"        360
"IT,Stable"     52
"Industrials,Bad Change"        28957
"Industrials,Good Change"       27568
"Industrials,Stable"    7412
"Materials,Bad Change"  63853
"Materials,Good Change" 61414
"Materials,Stable"      16277
"Real Estate,Bad Change"        19975
"Real Estate,Good Change"       18838
"Real Estate,Stable"    5650
"Utilities,Bad Change"  3804
"Utilities,Good Change" 3816
"Utilities,Stable"      1765
```



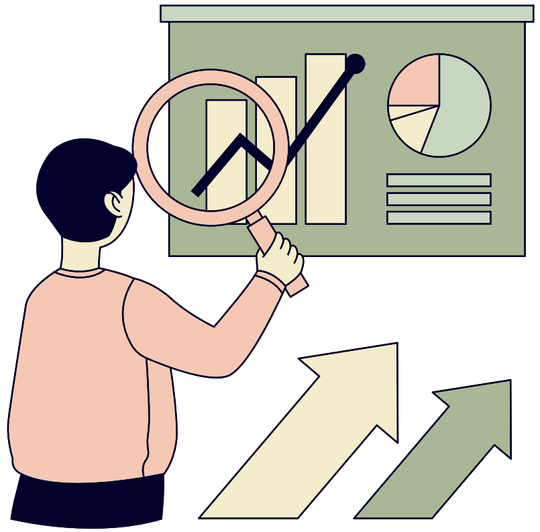
Second Business Problem:

2-Count number of companies in each sector:

We want to try another MapReduce problem that can help Tadawul to understand the Saudi Stock market better by knowing the number of trading companies in each sector to find the most popular sectors and the least popular ones.



Second Approach:

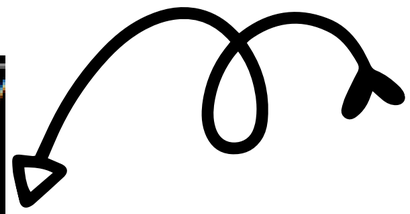


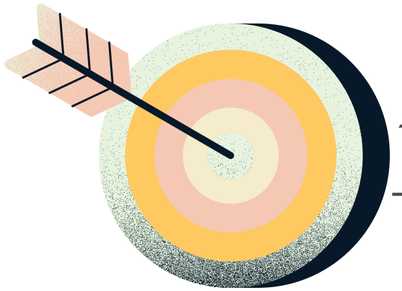
Running Hadoop Cluster on a Sector data:

```
[maria_dev@sandbox-hdp ~]$ python project_MapReduce.py -r hadoop --hadoop-stream
ing-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar sector_dat
a0.csv
```

Output:

```
Streaming final output from hdfs:///
"Communication Services"        6
"Consumer Discretionary"        24
"Consumer Staples"      16
"Energy"        5
"Financials"    45
"Health Care"   7
"IT"    2
"Industrials"   20
"Materials"     42
"Real Estate"   28
"Utilities"     2
```



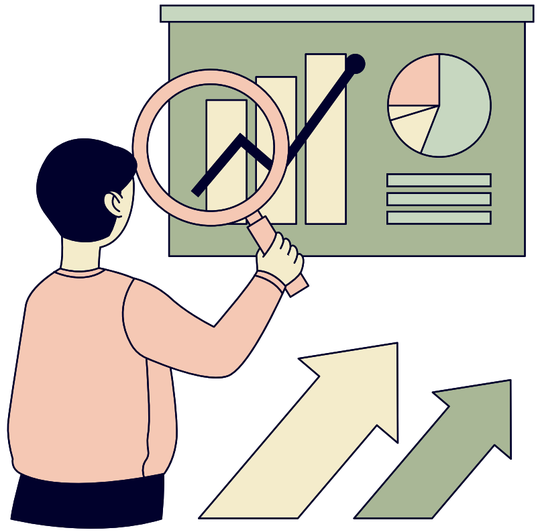


Business domain:

4-count of years to find out the year with the most shares:

In our data, each row expresses the information of one share. Among this information is the year in which this share was offered for trading so we can find the number of shares in each year, this could help Tadawul to understand the change in the stock market.

Third Approach:

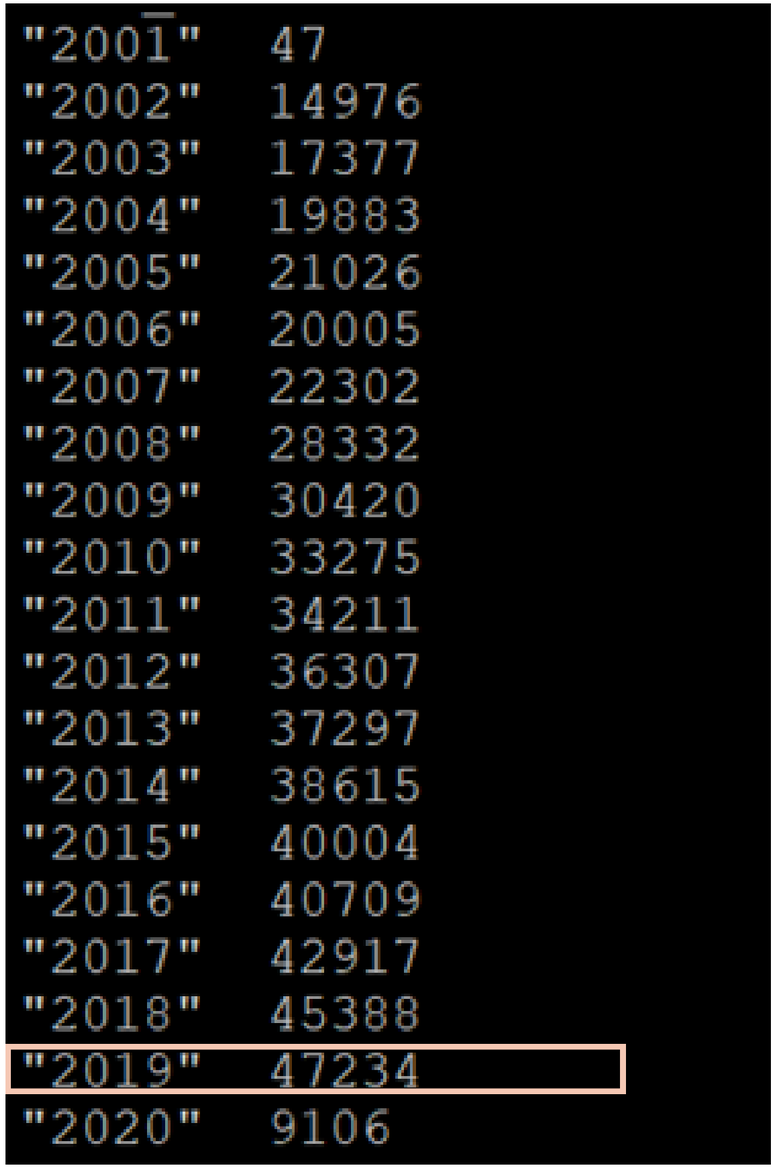


Running Hadoop Cluster on a years data:

```
[maria_dev@sandbox-hdp ~]$ python project_MapReduce.py -r hadoop --hadoop-stream
ing-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar years.csv
```

Output:

-2019 is the most year in which shares were offered for trading

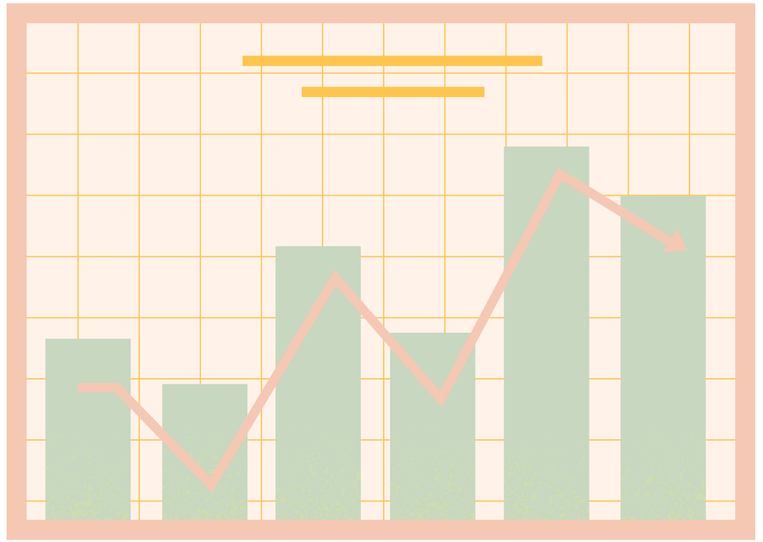


Fourth Business Problem:

4-Count the change categories for each year, month and day:

Our dataset is quite comprehensive and includes all the data necessary to derive information that is helpful to both businesses and investors. What is the ideal time to invest their money is the most crucial piece of knowledge they require. So in the following MapReduce functions, we count the occurrence of good change, stable change, and bad change with each year, each month, and each day.

Fourth Approach:



1-year

- Running Hadoop Cluster on a year_change data:

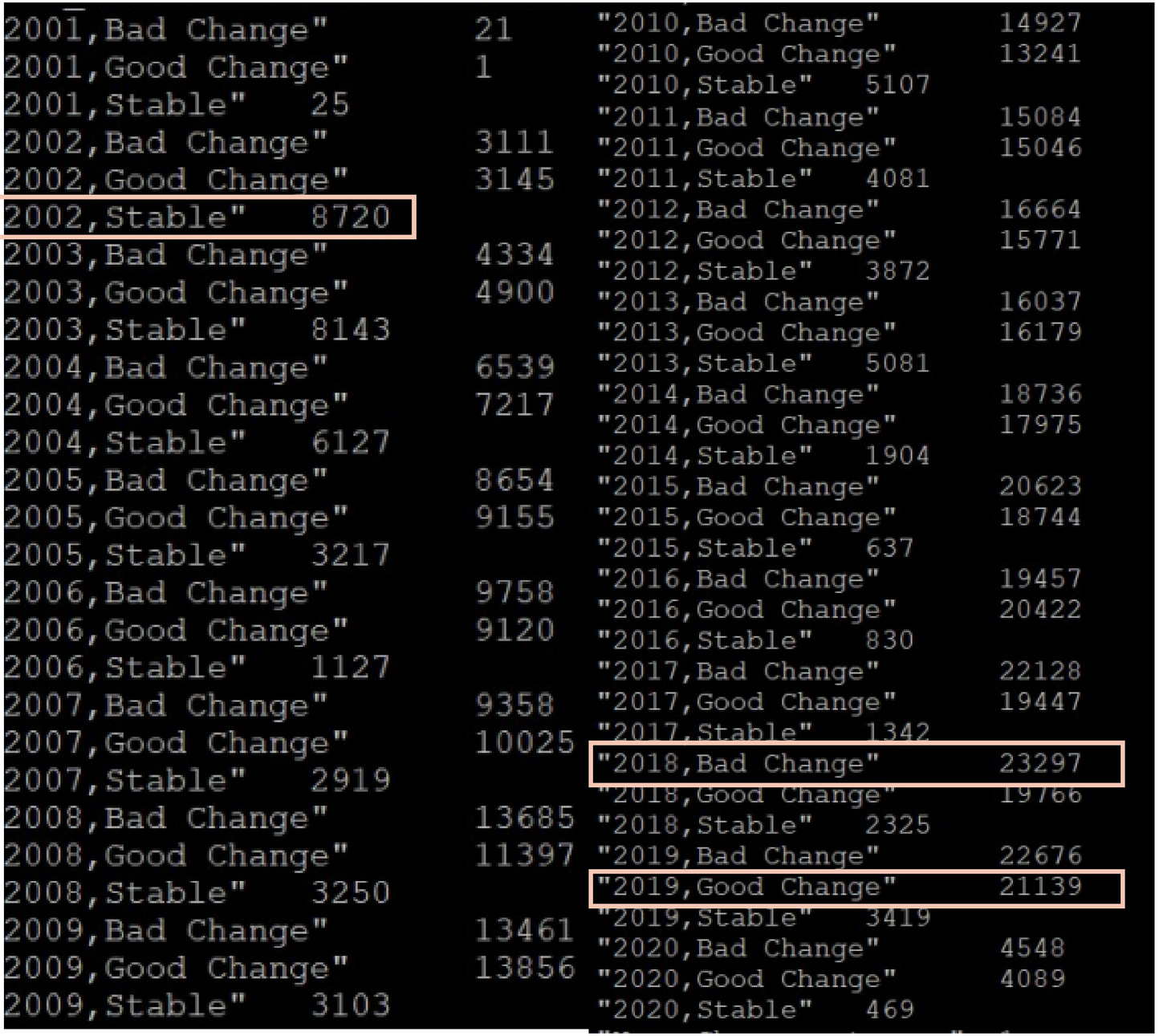
```
[maria_dev@sandbox-hdp ~]$ python project_MapReduce.py -r hadoop --hadoop-stream
ing-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar years_chan
ge.csv
```

Output:

- **2019 is the year with the highest Good change (21139)**

- **2018 is the year with the highest Bad Change (23297)**

- **2002 is the most stable year (8720)**



2-month

Running Hadoop Cluster on a month_change data:

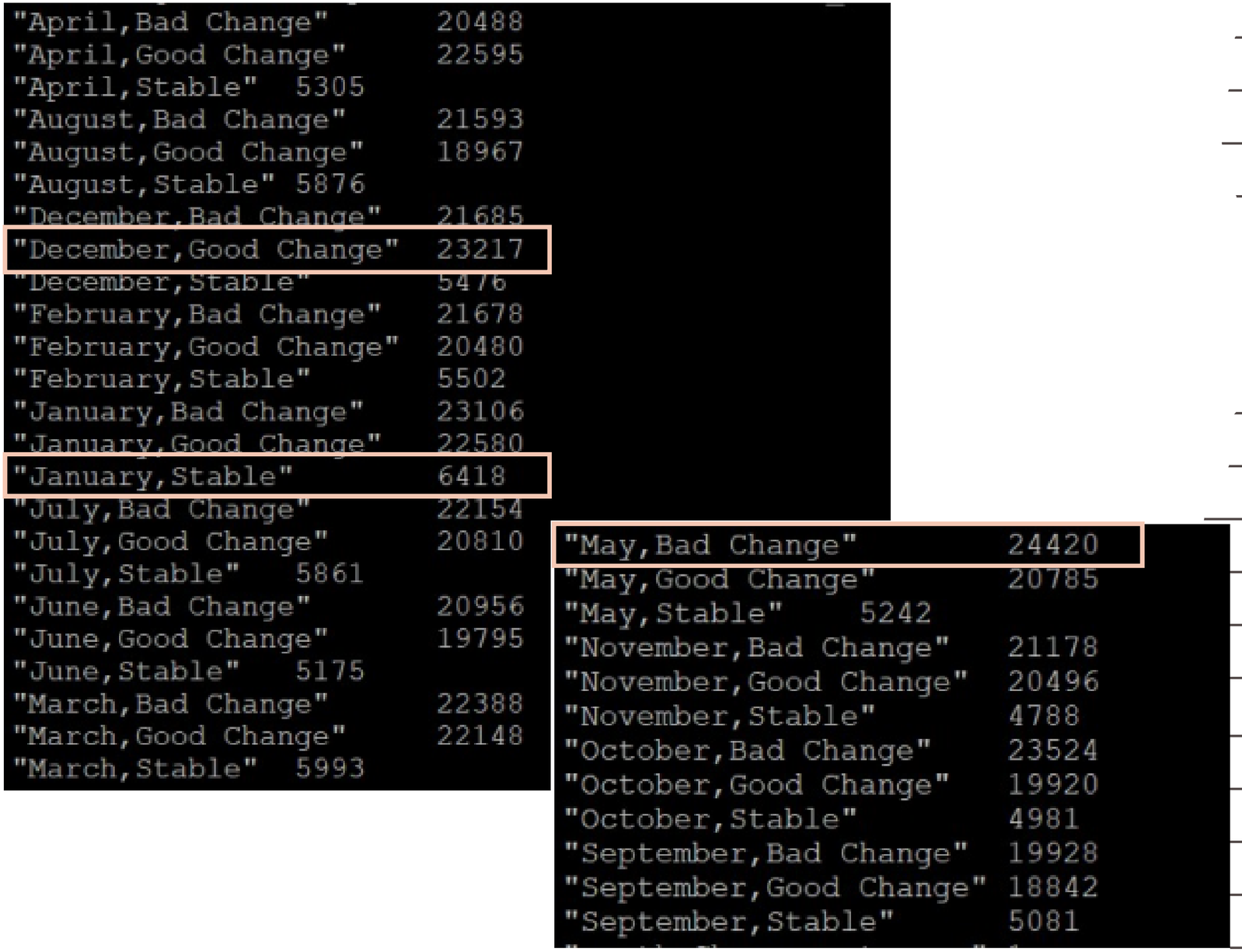
```
[maria_dev@sandbox-hdp ~]$ python project_MapReduce.py -r hadoop --hadoop-stream
ing-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar months_cha
nge.csv
```

Output:

-December is the month with the highest Good change (23217)

- May is the month with the highest Bad Change (24420)

- January is the most stable month (6418)



3-day

- Running Hadoop Cluster on a days_change data:

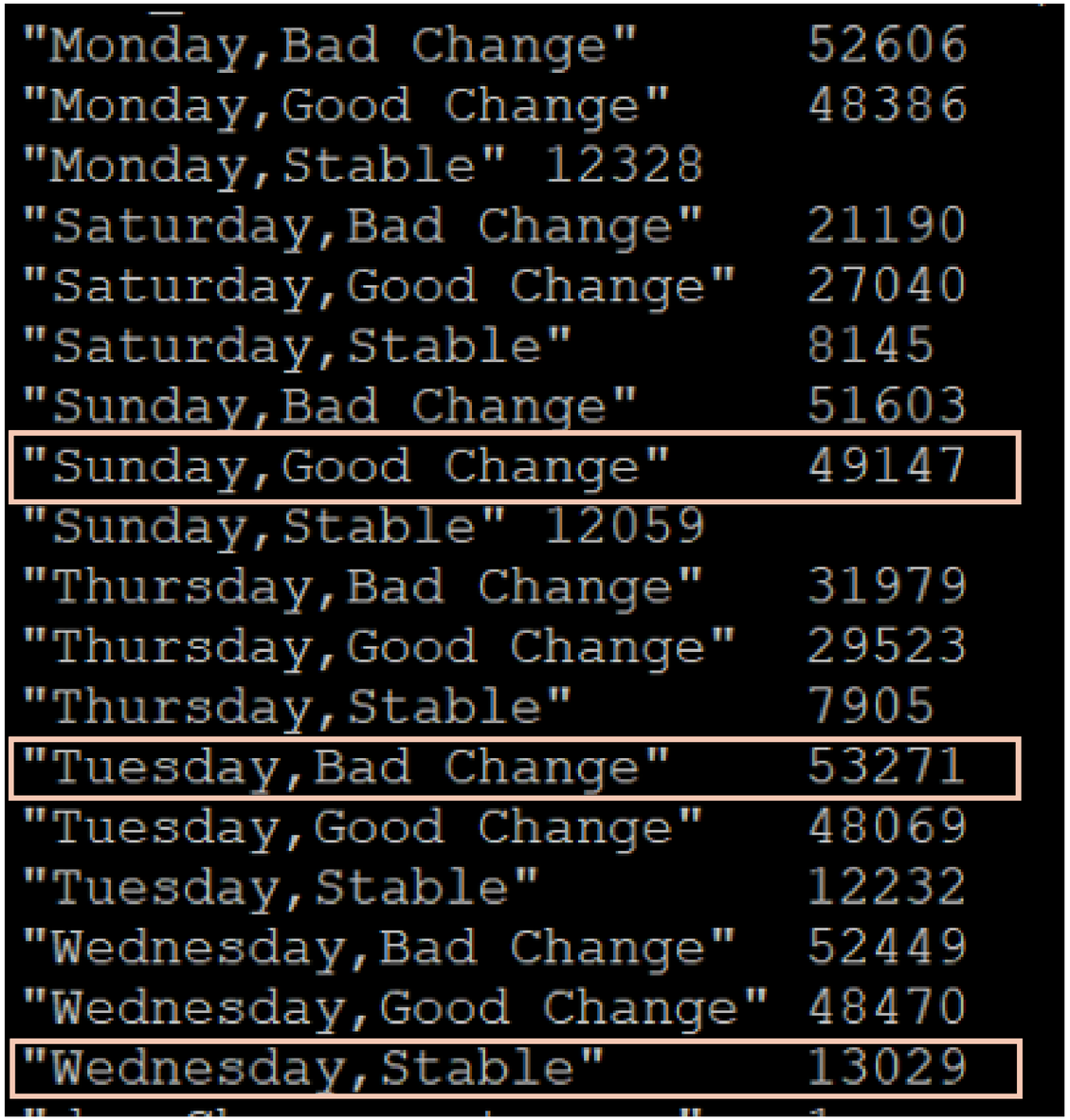
```
[maria_dev@sandbox-hdp ~]$ python project_MapReduce.py -r hadoop --hadoop-stream
ing-jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar days_chang
e.csv
```

Output:

- **Sunday is the day with the highest Good change (49147)**

- **Tuesday is the day with the highest Bad Change (53271)**

- **Wednesday is the most stable day (13029)**

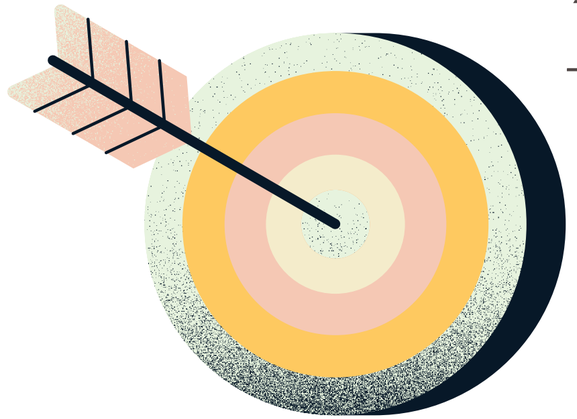


Results

What Can BeImproved?

- combining problem 1 and 2 to which will count the number of companies in each change category in each sector as follow:
  ((Change category, count of change category), number of companies (reduced)).

- Compare the time for running this problem and the one that we did in the project and choose the faster one.



Thank you

Any questions?