Case Study Report

SITI AISYAH BINTI ABD RAZAK

17133184/2

To the best of your understanding, identify ALL the companies provided in the list previously that are belonged to the sector of your choice. Then, explain generally how these companies build their big data solution, and the challenges that they are facing. Support your findings with valid references. (10 marks)

Sector: Financials

Company: JP Morgan, Visa

- **Explain generally how these companies build their big data solution**

Companies like JP Morgan and Visa starts building their big data solution by planning the global strategy needed to start the solution. They will start by having their initial efforts poured into big data solution implementation which can have a beneficial and measurable impact. Develop a big data adoption strategy which will include a roadmap which is based on their business needs and challenges. This step involves defining the problem in the company that needed the solution and whether the problem could be solved using available software or need to use new technologies. Business priorities are involved in which the most important one that formed the most leverage to the economic factor is prioritized (Mousannif et al., 2014).

Then, companies will determine the data that they currently have, and they need for implementing the solution. Consider all the kinds of data they have and finding out which data is most valuable and contain information that is related to the problem and insights that they are trying to solve. It's about examining the company's internal and external data to ensure that it can be used to support decision-making and day-to-day operations (Mousannif et al., 2014). Existing data needed for the solution that will be adopted is identified.

Then, identifying data gaps and silos is the next step in implementing big data solution (Davis, 2021). This step involves finding out where the needed data is stored, whether the data that the company have is missing components and if the missing components could help achieve better result and whether the data that must be used to implement the solution needs to be integrated.

Then, existing data is refined before implementation using big data tools and technologies. Having sufficient good quality data is crucial towards the implementation of big data solution (Chalimov, 2020). Past data stored by the company that resides in database may be in a format that is incompatible or does not contain all the required information. Thus, necessary changes need to be done before building the solution.

Next, choosing the technological stack tools to implement the solution. This step involves choosing which technology will the company used to implement the solution. The technology that is chosen would be tailored to the company needs, scalable to adapt to other parts of organizations and the technology can provide insights in a clear and understandable format.

Then, data preprocessing step is done to raw data (Mousannif et al., 2014). Data obtained from various sources contains imperfections. Potential of missing values and inconsistencies are high given that the data is in various formats and from different sources (Singhal & Jena, 2013). Integrating data, cleaning, and transforming them is important to ensure the data fit the requirement of algorithms for the analysis.Then, implementation is done with the chosen technological tools

Lastly, after the implementation of the solution, assessment and evaluation of the results is important. This step involves investigation on the outcome of the solution and whether the result achieve the set goal, what does the company need to improve to fully utilised the benefit from the efficiency in the data driven process.

- **Challenges that they are facing**

Adopting and managing big data in a company to deliver business solution and value to the company have its challenges. Among the challenges that companies in the financial industry like JP Morgan and Visa faced are:

1. **Technological incompatibilities**

   The adoption of big data solution in financial sector has its limitation and challenges due to its lack of compatibility between the needs in the financial sector and the capabilities of big data tools and technologies (Fang & Zhang, 2016). The current technology possessed by the companies are not enough to handle the high volume, velocity, and variety of the characteristic of big data. To adopt big data solution, changes must be made to the technology used in the company to cater the needs of managing big data. Investments must be made towards the implementation of big data technologies to support their business solution. This includes selecting proper big data tools required and needed by the company and a good understanding of the technologies. The transition from a company former state to a new advancement in big data is surely hard and comes with its complexities.

2. **Siloed storage**

   Companies in financial sector like JP Morgan and Visa stores an enormous amount of customer data and may be one of the most data intensive sectors in the economy. However, due to their siloed storage where their data management usually follows a traditional way where they are focusing on data across system for a specific function, they are not capable to see an overall view of the data they possess on customers or the market (Editorial Team, 2019). The way of storing data have made these company not very good in using their rich data sets to their full potential. This situation may also mean that they have a huge overlap of data across dozens of their legacy data warehouse.

Having data sits in many silos have also made it difficult for the company to centralize their data, and consequently making it difficult adopting big data solution into the company as it requires a lot of data being delivered fast and near real time. Integrating and combining data is a challenge for the companies are also due to not having decentralized departments thus there is insufficient understanding of knowledge on data integration across departments (Sun et al., 2020). Enormous effort in their data storage transition and integration must be done as to overcome this challenge would be extremely challenging and hard after years of traditional data management in siloed storage. Therefore, to adopt big data solution and deliver business value to the company data integration and transition must be made in order for the company to make their big data goals and solution successfully delivered.

3. Security and Privacy Concern

Storing a huge amount of sensitive data, security is a main topic that play an important role in the system. Not having a proper and strong security would pose serious threats to the company. The consequences of the event of breach or security attack would be devastating and harsh to a lot of people and aspect of the company. The huge and increasingly growing amount of big data in the company which contains, private sensitive information has made security and privacy more important than ever. Adopting big data solution which requires centralized data and transition from traditional database management system have made data no longer in their well protected siloed storage. This situation has made data security risks double and amplified.

In 2014, JP Morgan had a broad security attack in which a billion passwords and username are breached. It took the company security team two months to detect the attack before it was stopped (Perlroth, 2014). It is clear that forming a strong and reliable security policy for data is a necessity for a company but also a real challenge.

4. Inadequate knowledge

Handling big data tools and technologies require skilled data professionals. The fact that data processing tools and technologies have advanced rapidly, but most experts have not resulted in lack of skilled data professionals that is knowledgeable in handling tools in big data. Handling big data requires new skill and techniques using Hadoop, NoSQL, Map Reduce which is new and foreign to the financial companies (Jaiswal & Bagale, 2017).

Not only they have to be knowledge in technical skills in data analytics and IT, but it is very important that the data professional in the process of big data solution adoption have the ability to effectively communicate with the decision makers in the company. Current lack of professional that possess both quality have made companies having difficulty acquiring the right employee for the job (Massachusetts Institute of Technology, 2012) thus Companies have to hire a team which have the combination skills required to implement the big data solution (Thomson Reuters, 2014).

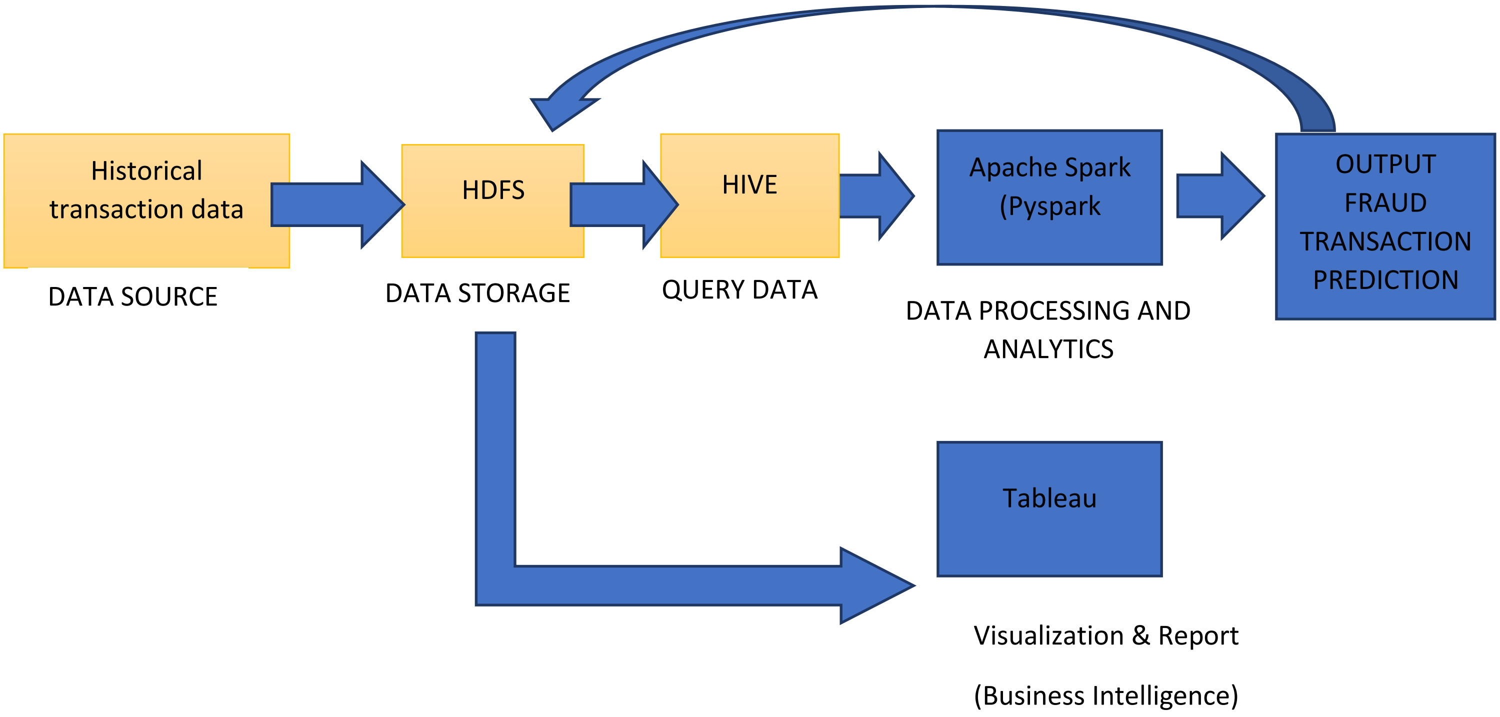
5. **Data Quality**

Lack of data quality are also contributing challenges to company to adopt big data solution. Data that is being generated for years in the company are not analysed because of the risk connected to sensitive financial information, lack of competent big data analysts and lack of awareness have inhibited the company from having good quality of data (Jaiswal & Bagale, 2017). Not having good quality of data would prevent the company from successfully implementing the big data solution.

2) Based on the sector that you choose and the challenges you discussed in question 1, identify ONE (1) problem statement that is valid generally across the sector. Justify your answer.

Credit card fraud has been on the rise, as well as financial losses because of fraud are skyrocketing. In 2019, 28.65 billion dollars loss due to card fraud (Lee, 2021). The consequences of falling prey to schemes in fraud cases is severe. Losing customer confidence, ruined company reputation, and suffer losses are some implications of fraud. Implementation of fraud detection to combat fraud in financial sector is critical as ever.   Before this, company are only analysing small number of transactions to detect fraud (Cavanillas et al., 2016). Fraud management that is effective is essential for financial institutions. To prevent fraud, initiative involving AI, predictive analytics and machine learning could help us combat fraud and the utilization of big data in fraud detection surely would help identifying trend in fraud and effectively detecting fraudulent transactions

3) Using at least THREE (3) different applications (tools) of your choice, form a big data pipeline. Its analytical outcome(s) should be able to solve the problem you identified in question 2



4) Implement the big data pipeline that you formed in question 3. Step-by-steps include (but not limited to) installation, command lines, user interfaces, and/or pseudocodes can be included in the answer to improve the clarity of the implementation.

Installation big data tools

- Hadoop in Ubuntu
-

Prerequisites:

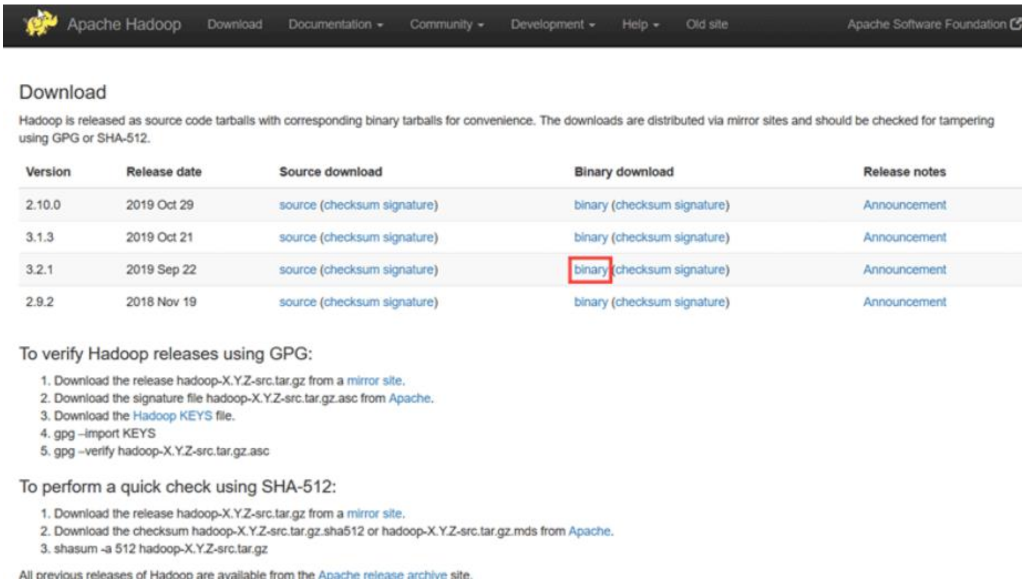
- Ensure that you have java 8 installed in your system
- Your machine must have at least 4GB of RAM with 60 GB storage in hard disk to ensure faster performance

1. Start terminal
2. Ensure your system is updated, update your system by typing the following commands

- sudo apt-get update
- sudo apt-get upgrade
- sudo apt-get install openssh-server

3. Check Java installation and version
   sudo apt-get install openjdk-8-jdk
    java -version

4. Download the Hadoop package, select the version



Download

Hadoop is released as source code tarballs with corresponding binary tarballs for convenience. The downloads are distributed via mirror sites and should be checked for tampering using GPG or SHA-512.

Version	Release date	Source download	Binary download	Release notes
2.10.0	2019 Oct 29	source (checksum signature)	binary (checksum signature)	Announcement
3.1.3	2019 Oct 21	source (checksum signature)	binary (checksum signature)	Announcement
3.2.1	2019 Sep 22	source (checksum signature)	binary (checksum signature)	Announcement
2.9.2	2018 Nov 19	source (checksum signature)	binary (checksum signature)	Announcement

To verify Hadoop releases using GPG:

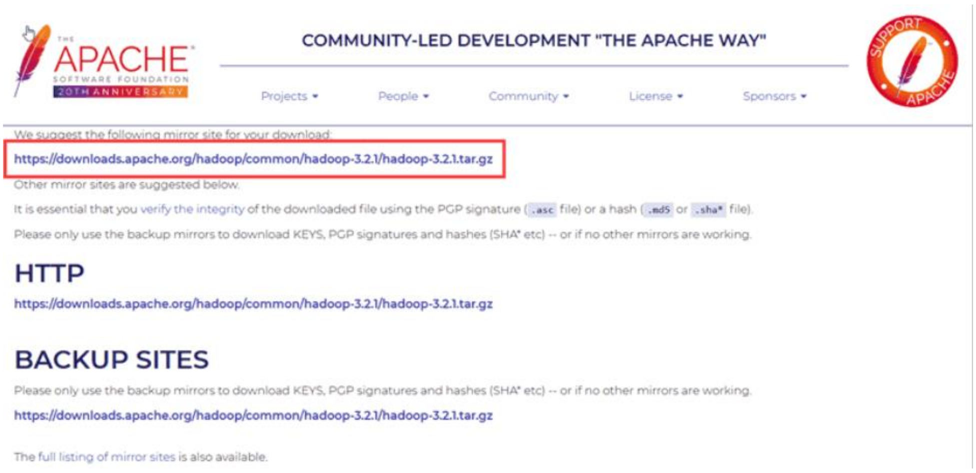
1. Download the release hadoop-X.Y.Z-src.tar.gz from a mirror site.
2. Download the signature file hadoop-X.Y.Z-src.tar.gz.asc from Apache.
3. Download the Hadoop KEYS file.
4. gpg –import KEYS
5. gpg –verify hadoop-X.Y.Z-src.tar.gz.asc

To perform a quick check using SHA-512:

1. Download the release hadoop-X.Y.Z-src.tar.gz from a mirror site.
2. Download the checksum hadoop-X.Y.Z-src.tar.gz.sha512 or hadoop-X.Y.Z-src.tar.gz.mds from Apache.
3. shasum -a 512 hadoop-X.Y.Z-src.tar.gz

All previous releases of Hadoop are available from the Apache release archive site.

Select the link and then you will be redirected to the following mirror link



THE APACHE SOFTWARE FOUNDATION 20TH ANNIVERSARY
COMMUNITY-LED DEVELOPMENT "THE APACHE WAY"
Projects ▾
People ▾
Community ▾
License ▾
Sponsors ▾
SUPPORT APACHE

We suggest the following mirror site for your download:

https://downloads.apache.org/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz

Other mirror sites are suggested below.

It is essential that you verify the integrity of the downloaded file using the PGP signature (`.asc` file) or a hash (`.md5` or `.sha*` file).

Please only use the backup mirrors to download KEYS, PGP signatures and hashes (SHA* etc) -- or if no other mirrors are working.

HTTP

https://downloads.apache.org/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz

BACKUP SITES

Please only use the backup mirrors to download KEYS, PGP signatures and hashes (SHA* etc) -- or if no other mirrors are working.

https://downloads.apache.org/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz

The full listing of mirror sites is also available.

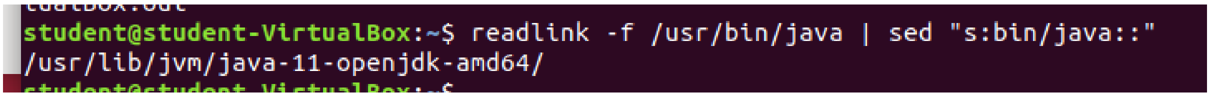
5. Untar the folder using the following command

   tar -xzf hadoop-3.2.1.tar.gz

   sudo mv hadoop-3.2.1 /home/student/hadoop/

6. Configure Hadoop JAVA HOME

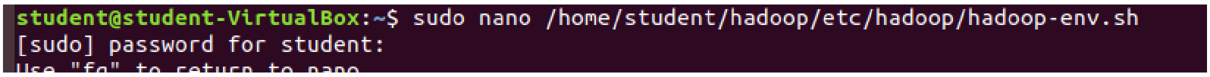
- Run the below command to know the path to JAVA and to setup Hadoop JAVA home value



```
student@student-VirtualBox:~$ readlink -f /usr/bin/java | sed "s:bin/java::"
/usr/lib/jvm/java-11-openjdk-amd64/
student@student-VirtualBox:~$
```

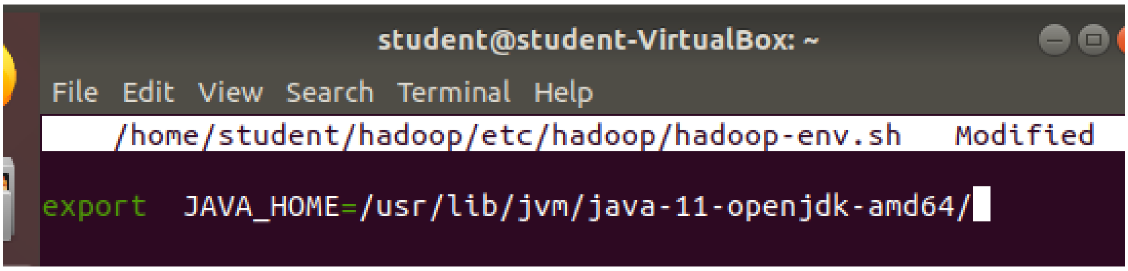
- Write the command to proceed to Hadoop-env.sh
-

  sudo nano /home/student/hadoop/etc/hadoop/hadoop-env.sh



```
student@student-VirtualBox:~$ sudo nano /home/student/hadoop/etc/hadoop/hadoop-env.sh
[sudo] password for student:
Use "fg" to return to nano
```

- Setup JAVA Home value in Hadoop-env.sh



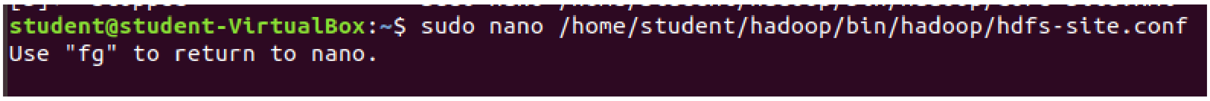
```
    /home/student/hadoop/etc/hadoop/hadoop-env.sh    Modified

export   JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64/
```

7. Run Hadoop

   /home/student/hadoop/bin/Hadoop

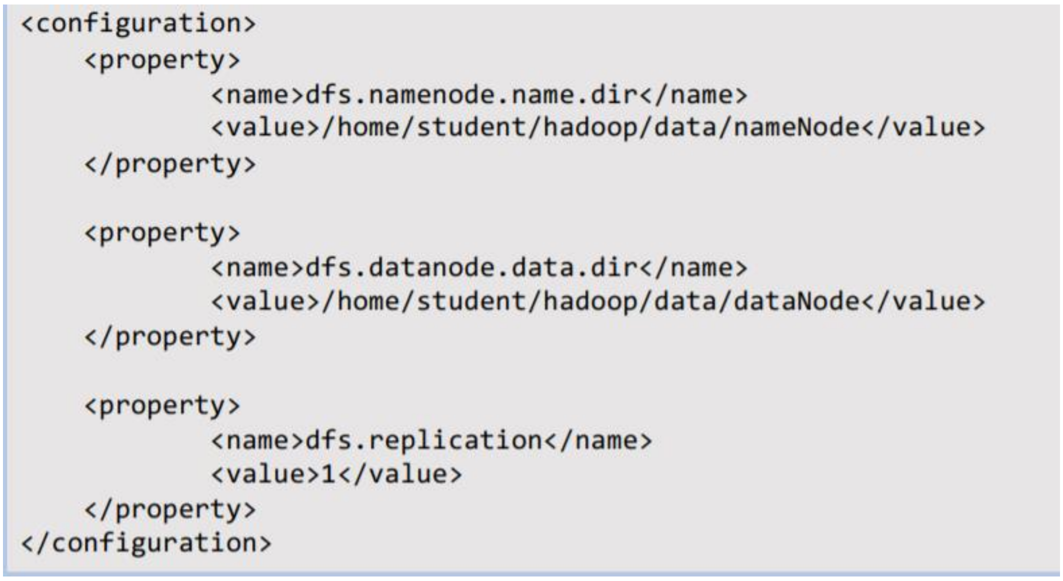
8. Configure HDFS AND YARN for Hadoop

- Set namenode location



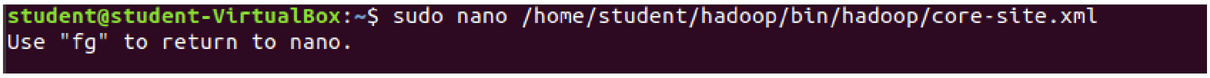
```
student@student-VirtualBox:~$ sudo nano /home/student/hadoop/bin/hadoop/hdfs-site.conf
Use "fg" to return to nano.
```

Update hdfs-site.conf with the following



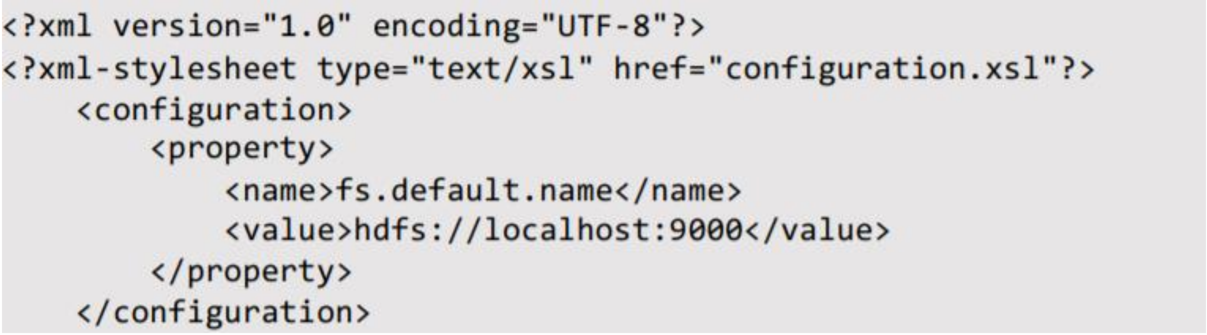
```
<configuration>
    <property>
            <name>dfs.namenode.name.dir</name>
            <value>/home/student/hadoop/data/nameNode</value>
    </property>

    <property>
            <name>dfs.datanode.data.dir</name>
            <value>/home/student/hadoop/data/dataNode</value>
    </property>

    <property>
            <name>dfs.replication</name>
            <value>1</value>
    </property>
</configuration>
```

- Configure HDFS



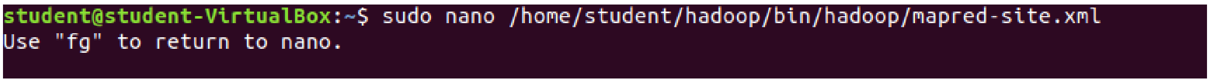
```
student@student-VirtualBox:~$ sudo nano /home/student/hadoop/bin/hadoop/core-site.xml
Use "fg" to return to nano.
```

- , update core-site.xml with the following



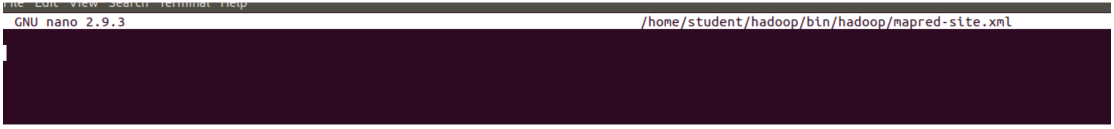
```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
    <configuration>
        <property>
            <name>fs.default.name</name>
            <value>hdfs://localhost:9000</value>
        </property>
    </configuration>
```

- **Configure YARN**
- **Open mapred-site.xml**



```
student@student-VirtualBox:~$ sudo nano /home/student/hadoop/bin/hadoop/mapred-site.xml
Use "fg" to return to nano.
```

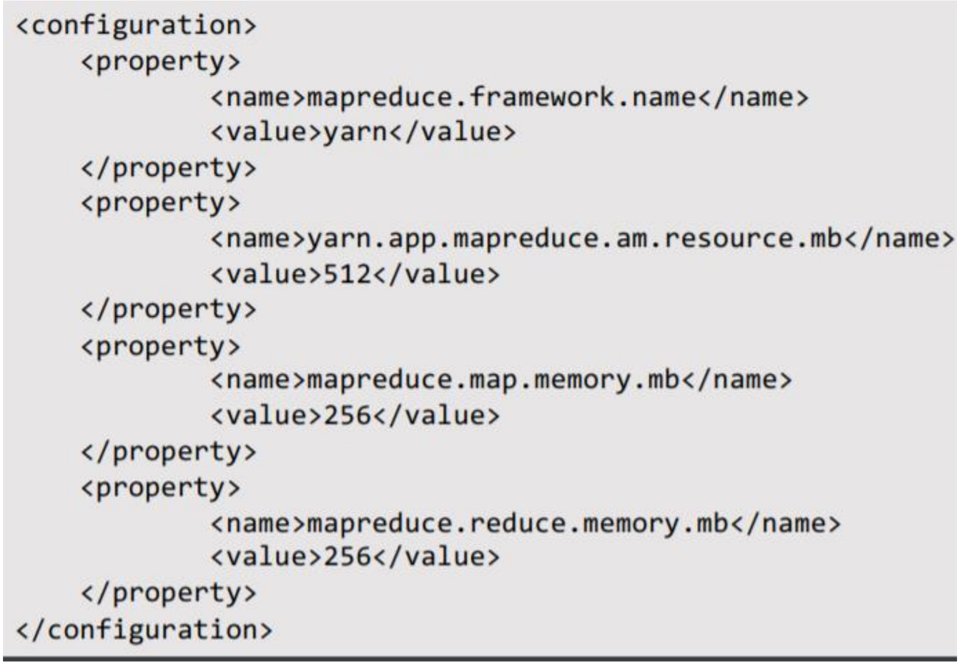
-



```
  GNU nano 2.9.3                                   /home/student/hadoop/bin/hadoop/mapred-site.xml
```

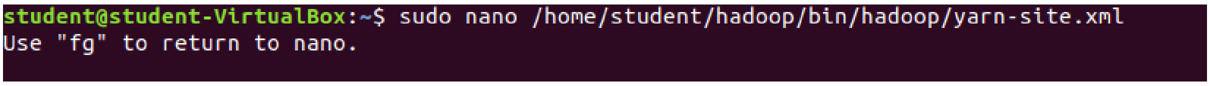
-

- Update mapred-site with the following



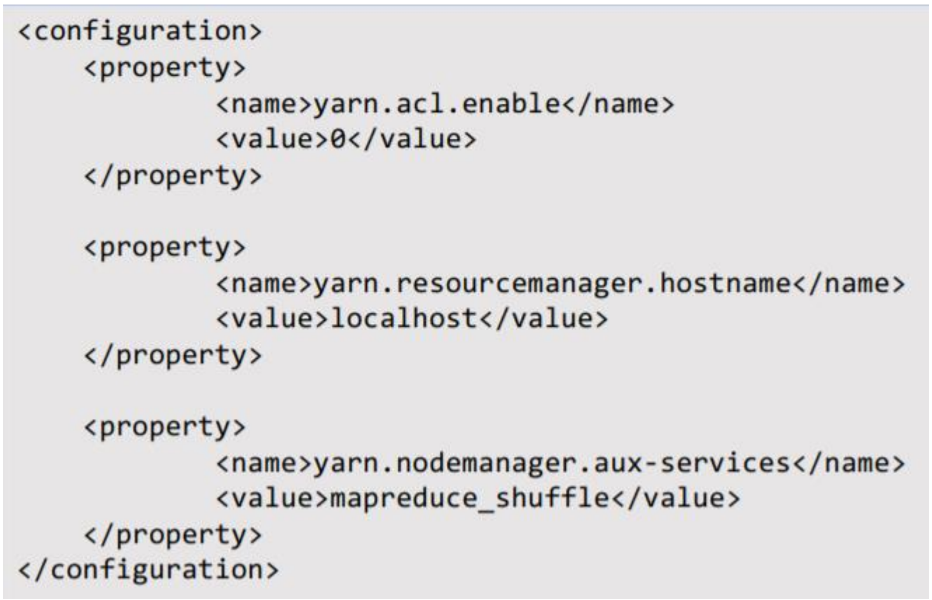
```
<configuration>
    <property>
            <name>mapreduce.framework.name</name>
            <value>yarn</value>
    </property>
    <property>
            <name>yarn.app.mapreduce.am.resource.mb</name>
            <value>512</value>
    </property>
    <property>
            <name>mapreduce.map.memory.mb</name>
            <value>256</value>
    </property>
    <property>
            <name>mapreduce.reduce.memory.mb</name>
            <value>256</value>
    </property>
</configuration>
```

- Open yarn-site.xml

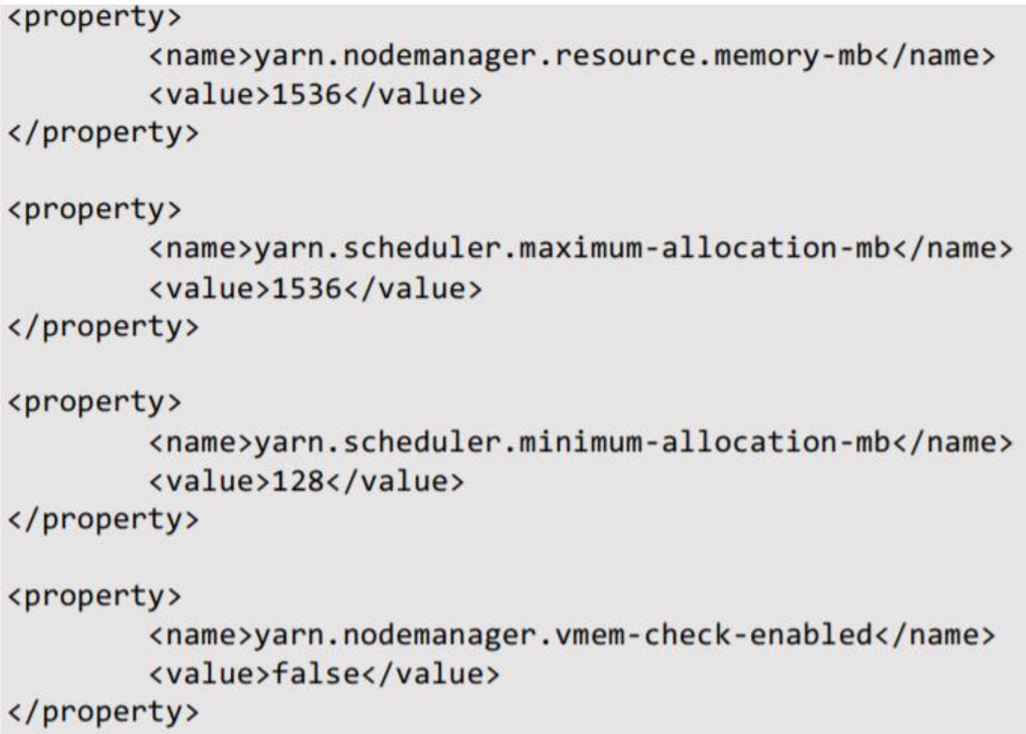


```
student@student-VirtualBox:~$ sudo nano /home/student/hadoop/bin/hadoop/yarn-site.xml
Use "fg" to return to nano.
```

- update yarn-site.xml with the following



```
<configuration>
    <property>
            <name>yarn.acl.enable</name>
            <value>0</value>
    </property>

    <property>
            <name>yarn.resourcemanager.hostname</name>
            <value>localhost</value>
    </property>

    <property>
            <name>yarn.nodemanager.aux-services</name>
            <value>mapreduce_shuffle</value>
    </property>
</configuration>
```



```
<property>
        <name>yarn.nodemanager.resource.memory-mb</name>
        <value>1536</value>
</property>

<property>
        <name>yarn.scheduler.maximum-allocation-mb</name>
        <value>1536</value>
</property>

<property>
        <name>yarn.scheduler.minimum-allocation-mb</name>
        <value>128</value>
</property>

<property>
        <name>yarn.nodemanager.vmem-check-enabled</name>
        <value>false</value>
</property>
```

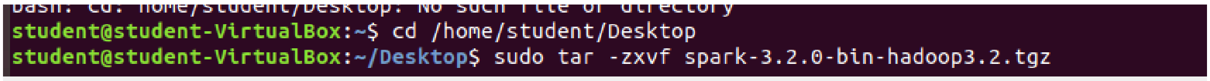
- **Format hdfs**

  hdfs namenode -format

2. Apache Spark (Pyspark)

- Download Spark from link : https://spark.apache.org/downloads.html

- Go to directory where spark is installed and untar the file



```
bash: cd: home/student/Desktop: No such file or directory
student@student-VirtualBox:~$ cd /home/student/Desktop
student@student-VirtualBox:~/Desktop$ sudo tar -zxvf spark-3.2.0-bin-hadoop3.2.tgz
```

- Configure environment variables for spark

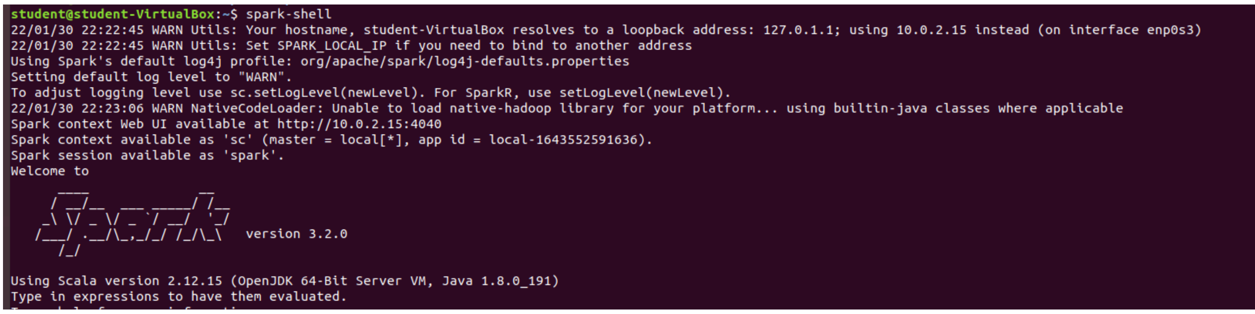
  vim ~/.bashrc

- Add the following line

```
export SPARK_HOME=~/Desktop/spark-3.2.0-bin-hadoop3.2.tgz
export PATH=$PATH:$SPARK_HOME/bin
export PATH=$PATH:~/anaconda3/bin
export PYTHONPATH=$SPARK_HOME/python:$PYTHONPATH
export PYSPARK_DRIVER_PYTHON="jupyter"
export PYSPARK_DRIVER_PYTHON_OPTS="notebook"
export PYSPARK_PYTHON=python3
export PATH=$PATH:$JAVA_HOME/jre/bin
```

- Save and exit
- Load .bashrc file again by running the following

  source ~/.bashrc
- Run Spark shell with following line:
  spark-shell
  You should see:



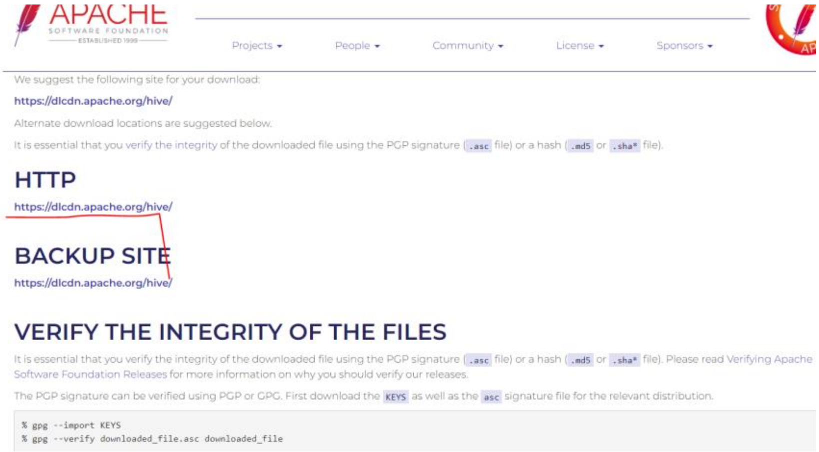
```
student@student-VirtualBox:~$ spark-shell
22/01/30 22:22:45 WARN Utils: Your hostname, student-VirtualBox resolves to a loopback address: 127.0.1.1; using 10.0.2.15 instead (on interface enp0s3)
22/01/30 22:22:45 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/01/30 22:23:06 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Spark context Web UI available at http://10.0.2.15:4040
Spark context available as 'sc' (master = local[*], app id = local-1643552591636).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.2.0
      /_/

Using Scala version 2.12.15 (OpenJDK 64-Bit Server VM, Java 1.8.0_191)
Type in expressions to have them evaluated.
```

- Download findspark and pyspark to use in jupyter notebook: Run this in the terminal
  Pip install findspark
  Pip install pyspark

3. Hive Installation

1. Go to link and download hive from apache mirror download link

   https://www.apache.org/dyn/closer.cgi/hive/



We suggest the following site for your download:

https://dlcdn.apache.org/hive/

Alternate download locations are suggested below.

It is essential that you verify the integrity of the downloaded file using the PGP signature (.asc file) or a hash (.md5 or .sha* file).

HTTP

https://dlcdn.apache.org/hive/

BACKUP SITE

https://dlcdn.apache.org/hive/

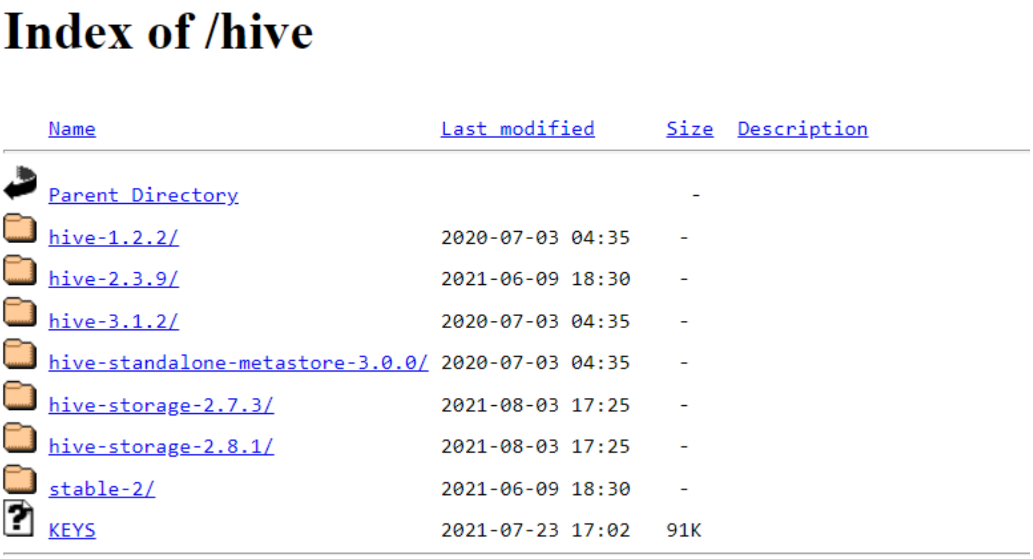
VERIFY THE INTEGRITY OF THE FILES

It is essential that you verify the integrity of the downloaded file using the PGP signature (.asc file) or a hash (.md5 or .sha* file). Please read Verifying Apache Software Foundation Releases for more information on why you should verify our releases.

The PGP signature can be verified using PGP or GPG. First download the KEYS as well as the asc signature file for the relevant distribution.

```
% gpg --import KEYS
% gpg --verify downloaded_file.asc downloaded_file
```

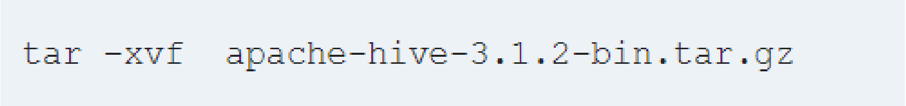
- Download the file



Index of /hive

```
    Name                               Last modified      Size  Description
    Parent Directory                                        -
    hive-1.2.2/                        2020-07-03 04:35    -
    hive-2.3.9/                        2021-06-09 18:30    -
    hive-3.1.2/                        2020-07-03 04:35    -
    hive-standalone-metastore-3.0.0/   2020-07-03 04:35    -
    hive-storage-2.7.3/                2021-08-03 17:25    -
    hive-storage-2.8.1/                2021-08-03 17:25    -
    stable-2/                          2021-06-09 18:30    -
    KEYS                               2021-07-23 17:02   91K
```

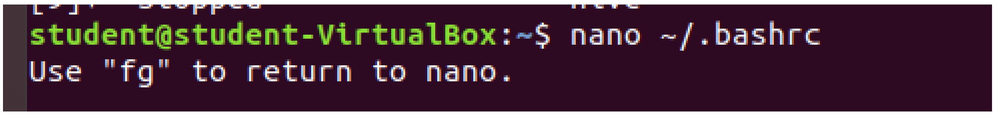
2. Go to directory and untar the hive file



```
tar -xvf  apache-hive-3.1.2-bin.tar.gz
```

3. Mention hive path in bashrc

- Open bash rc



```
[3]+  Stopped                 hive
student@student-VirtualBox:~$ nano ~/.bashrc
Use "fg" to return to nano.
```

4. Mention Hive home path i.e., HIVE_HOME path in bashrc file and export it, run the following code

```
export HIVE_HOME="/home/student/apache-hive-3.1.2-bin"
export PATH=$PATH:$HIVE_HOME/bin
```

5. Exporting **Hadoop path in Hive-config.sh**

- **In hive-config.sh, insert the following code of Hadoop**
- Export HADOOP_HOME =/home/student/hadoop

6. **Create hive warehouse**

-hadoop fs -mkdir /user/hive/warehouse
-hadoop fs -chmod 765 /user/hive/warehouse

7. Insert the following in bin folder

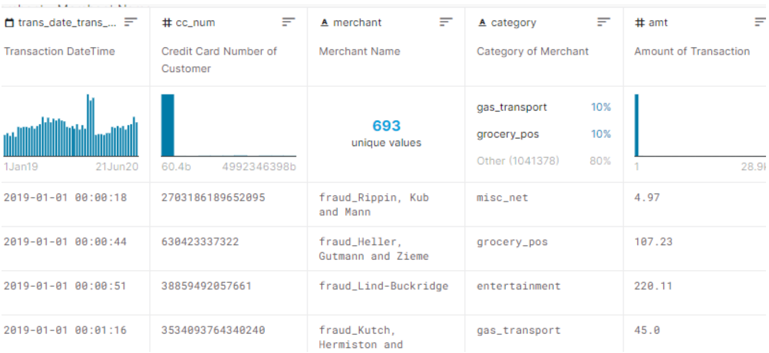
- ./schematool -initSchema -dbType derby

8.Type “hive” in terminal to run hive

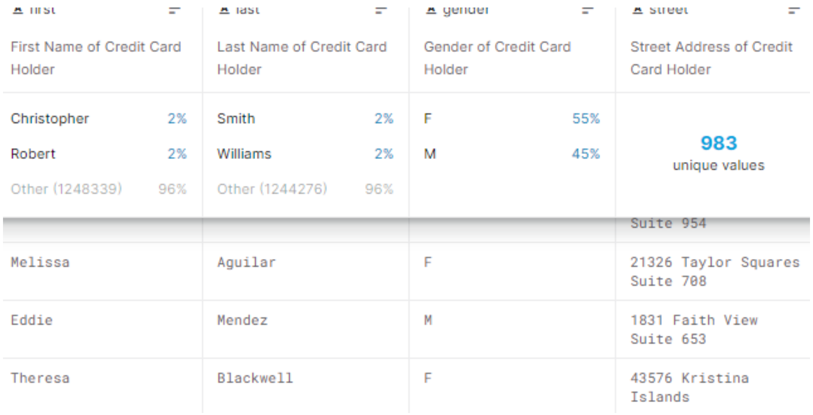
Implementation

1. Obtain data from data sources (historical transaction, demographic data)

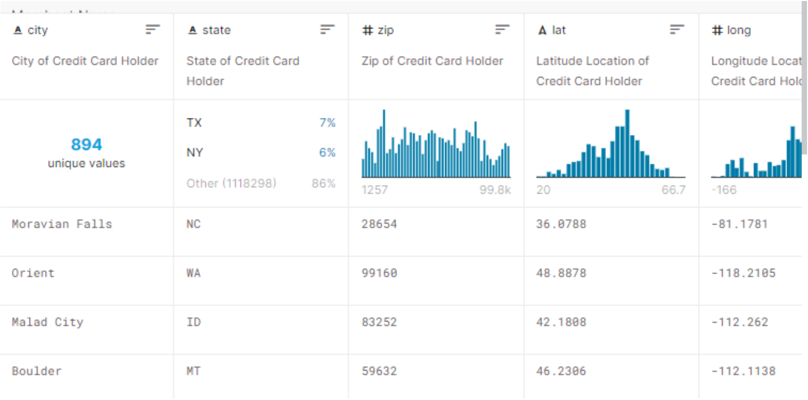
Link: https://www.kaggle.com/kartik2112/fraud-detection?select=fraudTrain.csv



trans_date_trans_...	cc_num	merchant	category	amt
Transaction DateTime	Credit Card Number of Customer	Merchant Name	Category of Merchant	Amount of Transaction
1Jan19 21Jun20	60.4b 4992346398b	693 unique values	gas_transport 10% grocery_pos 10% Other (1041378) 80%	1 28.9k
2019-01-01 00:00:18	2703186189652095	fraud_Rippin, Kub and Mann	misc_net	4.97
2019-01-01 00:00:44	630423337322	fraud_Heller, Gutmann and Zieme	grocery_pos	107.23
2019-01-01 00:00:51	38859492057661	fraud_Lind-Buckridge	entertainment	220.11
2019-01-01 00:01:16	3534093764340240	fraud_Kutch, Hermiston and	gas_transport	45.0



first	last	gender	street
First Name of Credit Card Holder	Last Name of Credit Card Holder	Gender of Credit Card Holder	Street Address of Credit Card Holder
Christopher 2% Robert 2% Other (1248339) 96%	Smith 2% Williams 2% Other (1244276) 96%	F 55% M 45%	983 unique values
			Suite 954
Melissa	Aguilar	F	21326 Taylor Squares Suite 708
Eddie	Mendez	M	1831 Faith View Suite 653
Theresa	Blackwell	F	43576 Kristina Islands

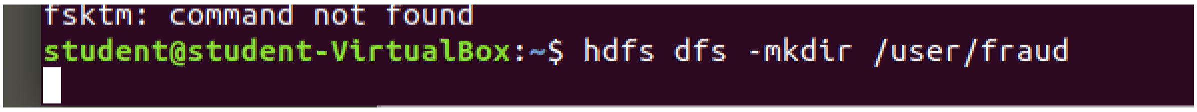


city	state	zip	lat	long
City of Credit Card Holder	State of Credit Card Holder	Zip of Credit Card Holder	Latitude Location of Credit Card Holder	Longitude Locat Credit Card Hold
894 unique values	TX 7% NY 6% Other (1118298) 86%	1257 99.8k	20 66.7	-166
Moravian Falls	NC	28654	36.0788	-81.1781
Orient	WA	99160	48.8878	-118.2105
Malad City	ID	83252	42.1808	-112.262
Boulder	MT	59632	46.2306	-112.1138

2. Hadoop for Data Storage and Hive for data query

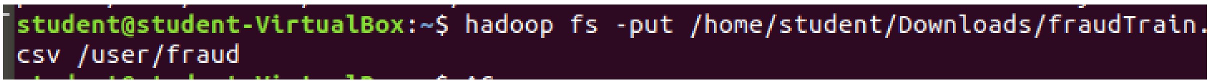
i. Store csv file in Hadoop hdfs

✓ Create directory in hdfs for data files

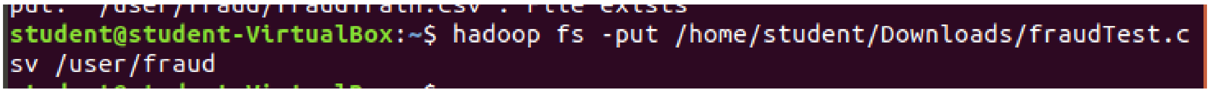


```
fsktm: command not found
student@student-VirtualBox:~$ hdfs dfs -mkdir /user/fraud
```

✓ Load data into hdfs directory

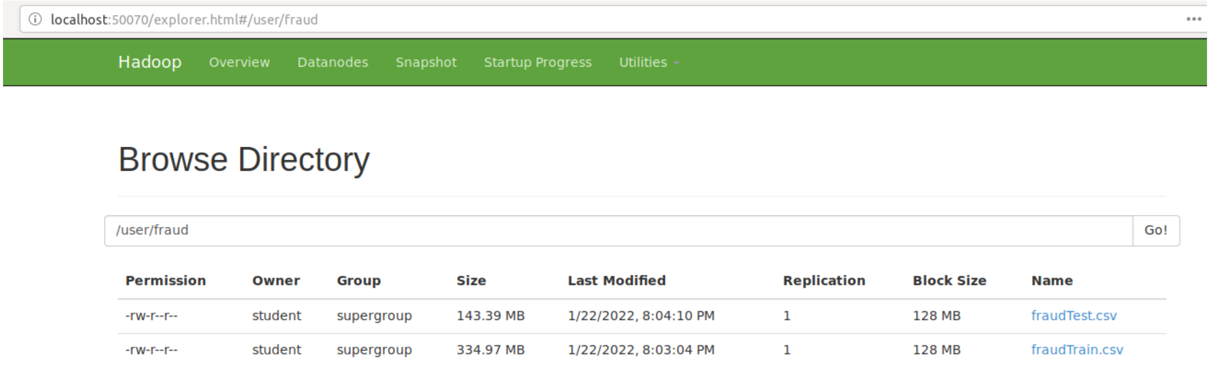


```
student@student-VirtualBox:~$ hadoop fs -put /home/student/Downloads/fraudTrain.
csv /user/fraud
```



```
put: `/user/fraud/fraudTrain.csv': File exists
student@student-VirtualBox:~$ hadoop fs -put /home/student/Downloads/fraudTest.c
sv /user/fraud
```

✓ Check the data is loaded into hdfs

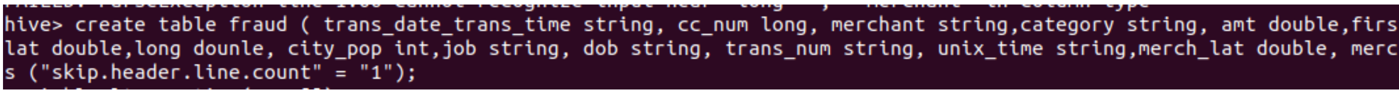


Browse Directory

/user/fraud

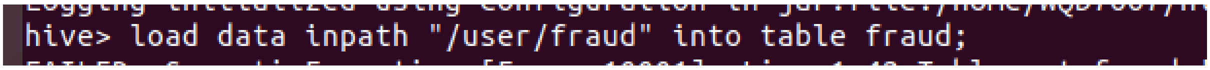
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	student	supergroup	143.39 MB	1/22/2022, 8:04:10 PM	1	128 MB	fraudTest.csv
-rw-r--r--	student	supergroup	334.97 MB	1/22/2022, 8:03:04 PM	1	128 MB	fraudTrain.csv

- create table in hive



```
hive> create table fraud ( trans_date_trans_time string, cc_num long, merchant string,category string, amt double,firs
lat double,long dounle, city_pop int,job string, dob string, trans_num string, unix_time string,merch_lat double, merc
s ("skip.header.line.count" = "1");
```

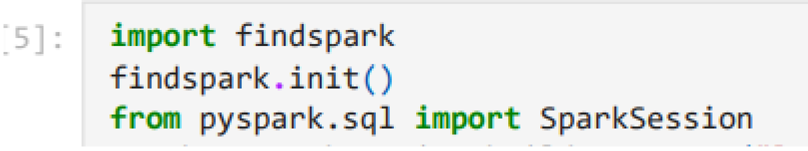
- **Load data in hdfs into hive table**



```
hive> load data inpath "/user/fraud" into table fraud;
```

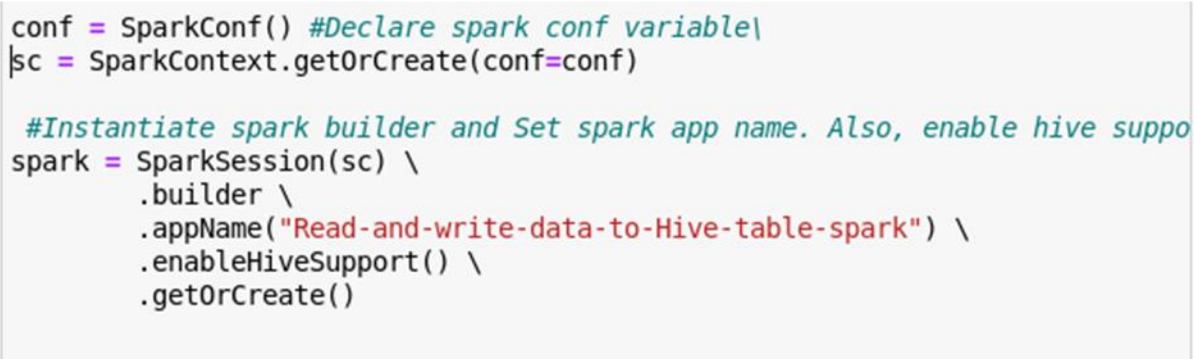
3. Pyspark for Data Preprocessing and fraud prediction

i. Import necessary package for running pyspark in notebook



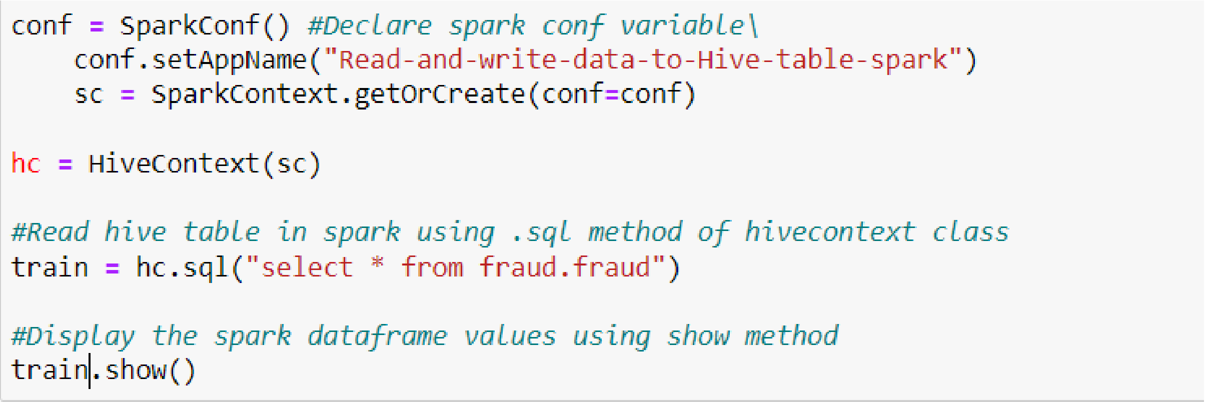
```
[5]: import findspark
     findspark.init()
     from pyspark.sql import SparkSession
```

ii. Create a SparkSession App object

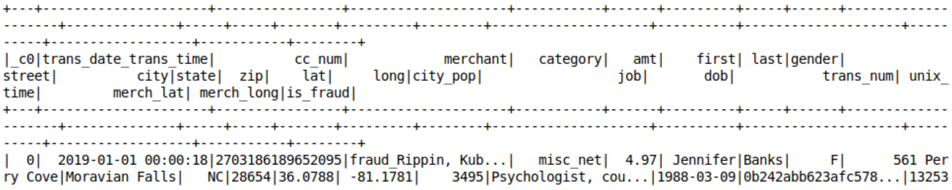


```
conf = SparkConf() #Declare spark conf variable\
sc = SparkContext.getOrCreate(conf=conf)

 #Instantiate spark builder and Set spark app name. Also, enable hive suppo
spark = SparkSession(sc) \
        .builder \
        .appName("Read-and-write-data-to-Hive-table-spark") \
        .enableHiveSupport() \
        .getOrCreate()
```

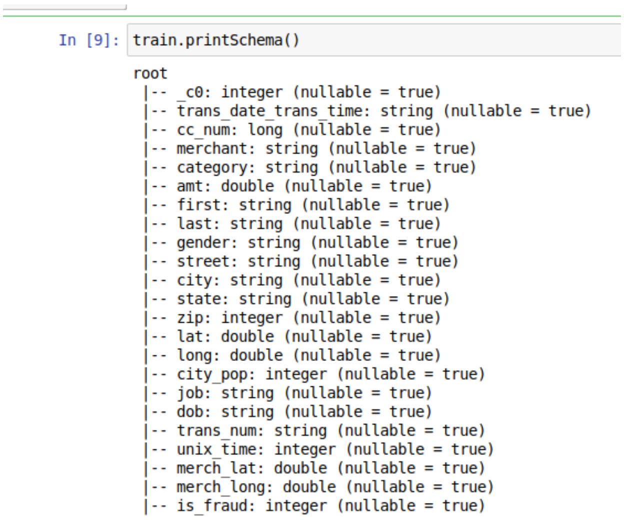
iii. Read hive table into a dataframe and examine



```
conf = SparkConf() #Declare spark conf variable\
    conf.setAppName("Read-and-write-data-to-Hive-table-spark")
    sc = SparkContext.getOrCreate(conf=conf)

hc = HiveContext(sc)

#Read hive table in spark using .sql method of hivecontext class
train = hc.sql("select * from fraud.fraud")

#Display the spark dataframe values using show method
train.show()
```



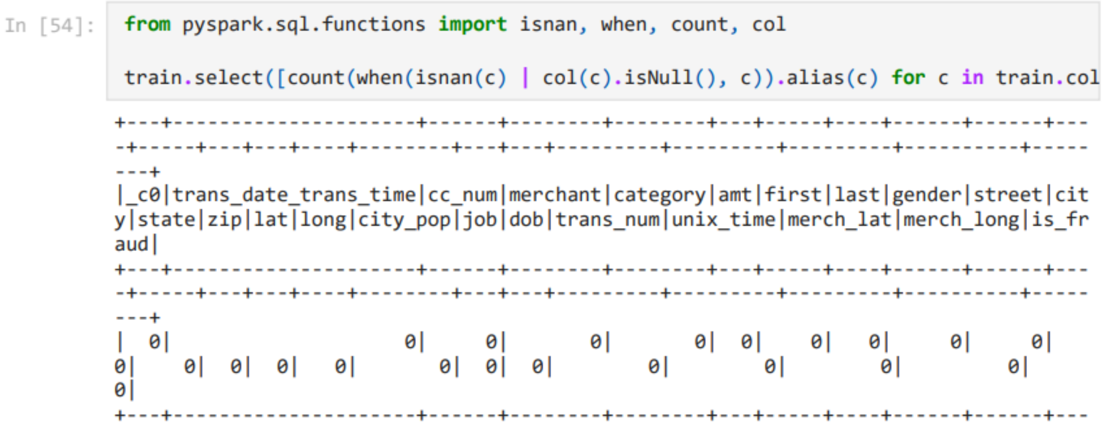
```
+---+-------------------+----------------+--------------------+----------+------+---------+-----+------+------------
-------+-------------+-----+-----+-------+---------+-------+--------------------+----------+--------------------+-----
-----+-----------------+-----------+--------+
|_c0|trans_date_trans_time|          cc_num|            merchant|   category|   amt|     first| last|gender|
street|           city|state|  zip|    lat|      long|city_pop|                   job|       dob|           trans_num| unix_
time|         merch_lat| merch_long|is_fraud|
+---+-------------------+----------------+--------------------+----------+------+---------+-----+------+------------
-------+-------------+-----+-----+-------+---------+-------+--------------------+----------+--------------------+-----
-----+-----------------+-----------+--------+
|  0|  2019-01-01 00:00:18|2703186189652095|fraud_Rippin, Kub...|   misc_net|  4.97| Jennifer|Banks|     F|      561 Per
ry Cove|Moravian Falls|   NC|28654|36.0788| -81.1781|    3495|Psychologist, cou...|1988-03-09|0b242abb623afc578...|13253
```



```
In [9]: train.printSchema()

        root
         |-- _c0: integer (nullable = true)
         |-- trans_date_trans_time: string (nullable = true)
         |-- cc_num: long (nullable = true)
         |-- merchant: string (nullable = true)
         |-- category: string (nullable = true)
         |-- amt: double (nullable = true)
         |-- first: string (nullable = true)
         |-- last: string (nullable = true)
         |-- gender: string (nullable = true)
         |-- street: string (nullable = true)
         |-- city: string (nullable = true)
         |-- state: string (nullable = true)
         |-- zip: integer (nullable = true)
         |-- lat: double (nullable = true)
         |-- long: double (nullable = true)
         |-- city_pop: integer (nullable = true)
         |-- job: string (nullable = true)
         |-- dob: string (nullable = true)
         |-- trans_num: string (nullable = true)
         |-- unix_time: integer (nullable = true)
         |-- merch_lat: double (nullable = true)
         |-- merch_long: double (nullable = true)
         |-- is_fraud: integer (nullable = true)
```

iv. Explore data

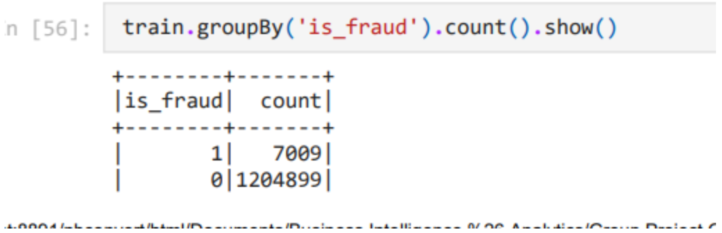
- Check for missing values in data



```
In [54]: from pyspark.sql.functions import isnan, when, count, col

         train.select([count(when(isnan(c) | col(c).isNull(), c)).alias(c) for c in train.col
```

```
+---+---------------------+------+--------+--------+---+-----+----+------+------+---
-+-----+---+---+----+--------+---+---+---------+---------+---------+----------+-----
---+
|_c0|trans_date_trans_time|cc_num|merchant|category|amt|first|last|gender|street|cit
y|state|zip|lat|long|city_pop|job|dob|trans_num|unix_time|merch_lat|merch_long|is_fr
aud|
+---+---------------------+------+--------+--------+---+-----+----+------+------+---
-+-----+---+---+----+--------+---+---+---------+---------+---------+----------+-----
---+
|  0|                    0|     0|       0|       0|  0|    0|   0|     0|     0|   
0|    0|  0|  0|   0|       0|  0|  0|        0|        0|        0|         0|       
0|
+---+---------------------+------+--------+--------+---+-----+----+------+------+---
```

- Check output label distribution



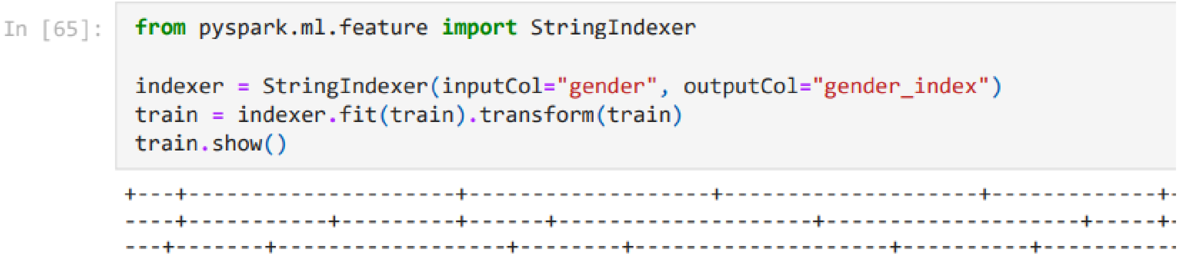
```
n [56]: train.groupBy('is_fraud').count().show()
```

```
+--------+-------+
|is_fraud|  count|
+--------+-------+
|       1|   7009|
|       0|1204899|
```

t:8891/nbconvert/html/Documents/Business Intelligence %26 Analytics/Group Project (

v. Pre-process data

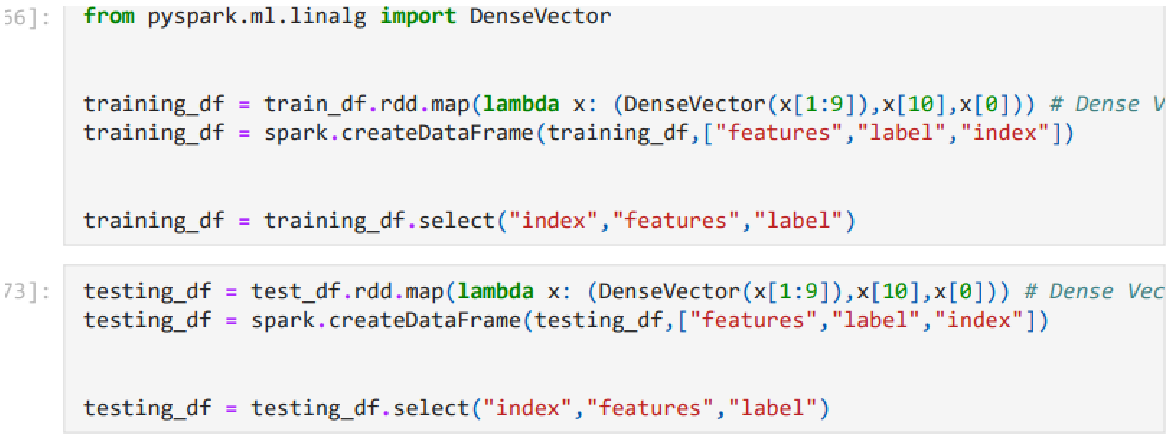
- Label encode gender column



```
In [65]: from pyspark.ml.feature import StringIndexer

         indexer = StringIndexer(inputCol="gender", outputCol="gender_index")
         train = indexer.fit(train).transform(train)
         train.show()
```

```
+---+-------------------+------------------+-------------------+------------+-
----+----------+--------+------+-------------------+-------------------+-----+-
---+-------+-----------------+--------+-------------------+---------+----------
```

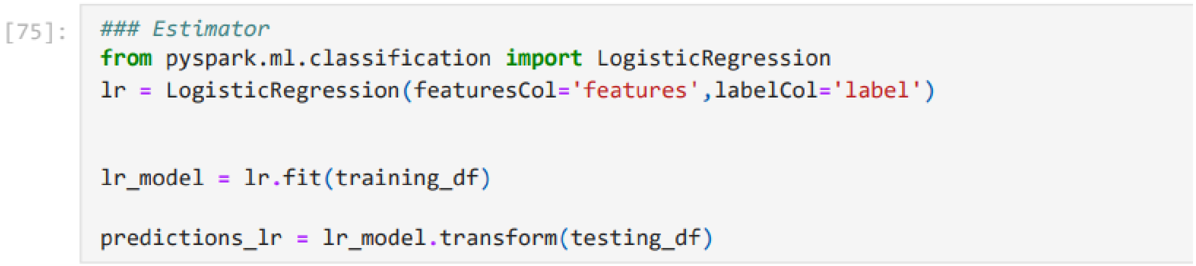
vi. Convert dataframe into RDD Map



```python
from pyspark.ml.linalg import DenseVector

training_df = train_df.rdd.map(lambda x: (DenseVector(x[1:9]),x[10],x[0])) # Dense V
training_df = spark.createDataFrame(training_df,["features","label","index"])

training_df = training_df.select("index","features","label")
```

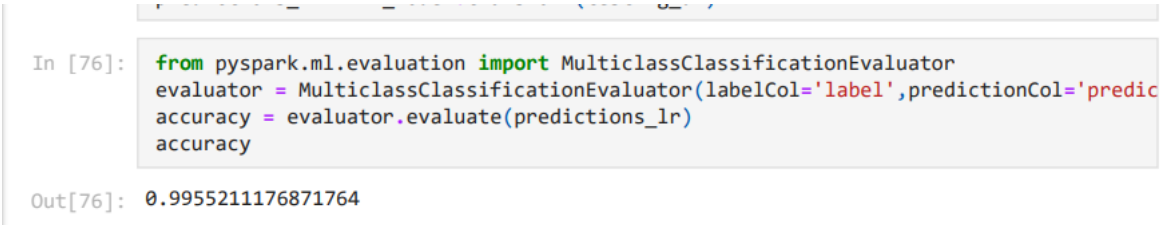
```python
testing_df = test_df.rdd.map(lambda x: (DenseVector(x[1:9]),x[10],x[0])) # Dense Vec
testing_df = spark.createDataFrame(testing_df,["features","label","index"])

testing_df = testing_df.select("index","features","label")
```

vii. Train model with training set and make predictions on test set



```python
### Estimator
from pyspark.ml.classification import LogisticRegression
lr = LogisticRegression(featuresCol='features',labelCol='label')

lr_model = lr.fit(training_df)

predictions_lr = lr_model.transform(testing_df)
```

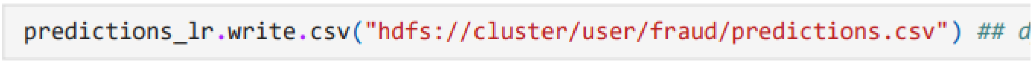
viii. Evaluate the model to check if the performance is sufficiently good



```python
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
evaluator = MulticlassClassificationEvaluator(labelCol='label',predictionCol='predic
accuracy = evaluator.evaluate(predictions_lr)
accuracy
```

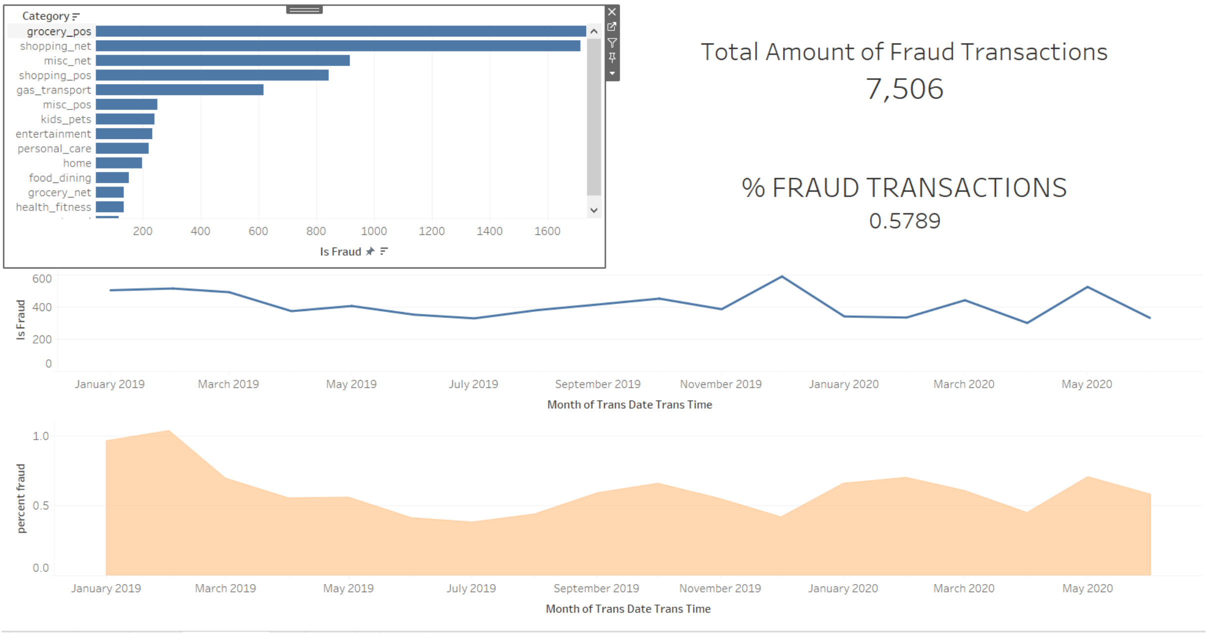
```
Out[76]: 0.9955211176871764
```

ix. Write result prediction data into hdfs



```python
predictions_lr.write.csv("hdfs://cluster/user/fraud/predictions.csv") ## d
```

TABLEAU VISUALIZATION FOR INSIGHTS



REFERENCE

Cavanillas, J. M., Curry, E., & Wahlster, W. (Eds.). (2016). New Horizons for a Data-Driven Economy. *A Roadmap for Usage and Exploitation of Big Data in Europe*. https://doi.org/10.1007/978-3-319-21569-3

Chalimov, A. (2020, December 1). *7 Big Data Solutions Examples and a Roadmap for Their Implementation*. Eastern Peak - Technology Consulting & Development Company. https://easternpeak.com/blog/7-big-data-solutions-examples-and-a-roadmap-for-their-implementation/

Davis, T. (2021, November 29). *Roadmap for Implementing Big Data Analytics at Your Organization*. 3Pillar Global. https://www.3pillarglobal.com/insights/roadmap-for-implementing-big-data-analytics-at-your-organization/

Editorial Team. (2019, September 9). *Big Data in the Financial Services Industry - From data to insights*. Finextra Research. https://www.finextra.com/blogposting/17847/big-data-in-the-financial-services-industry---from-data-to-insights

Fang, B., & Zhang, P. (2016). Big Data in Finance. *Big Data Concepts, Theories, and Applications*, 391–412. https://doi.org/10.1007/978-3-319-27763-9_11

Jaiswal, A., & Bagale, P. (2017). A Survey on Big Data in Financial Sector. *2017 International Conference on Networking and Network Applications (NaNA)*. https://doi.org/10.1109/nana.2017.46

Lee, N. (2021, February 1). Credit card fraud will increase due to the Covid pandemic, experts warn. *CNBC*. https://www.cnbc.com/2021/01/27/credit-card-fraud-is-on-the-rise-due-to-covid-pandemic.html

Massachusetts Institute of Technology. (2012, July 30). *How 'Big Data' Is Different*. MIT Sloan Management Review. https://sloanreview.mit.edu/article/how-big-data-is-different/

Mousannif, H., Sabah, H., Douiji, Y., & Sayad, Y. O. (2014). From Big Data to Big Projects: A Step-by-Step Roadmap. *2014 International Conference on Future Internet of Things and Cloud*. https://doi.org/10.1109/ficloud.2014.66

Perlroth, M. G. A. N. (2014, November 3). Luck Played Role in Discovery of Data Breach at JPMorgan Affecting Millions. *DealBook*. https://dealbook.nytimes.com/2014/10/31/discovery-of-jpmorgan-cyberattack-aided-by-company-that-runs-race-website-for-bank/

Sun, H., Rabbani, M. R., Sial, M. S., Yu, S., Filipe, J. A., & Cherian, J. (2020). Identifying Big Data's Opportunities, Challenges, and Implications in Finance. *Mathematics*, *8*(10), 1738. https://doi.org/10.3390/math8101738

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, *70*, 263–286. https://doi.org/10.1016/j.jbusres.2016.08.001

Thomson Reuters. (2014, August 13). *"Big Data in Capital Markets" Survey Results | Scoop News* [Press release]. https://www.scoop.co.nz/stories/WO1408/S00155/big-data-in-capital-markets-survey-results.htm