CSCE 5300: Intro. to Big Data and Data Science

Project

Fraud Detection in Financial Transactions

Team members

Hari Krishna Sai Rachuri

Vivek Nelluri

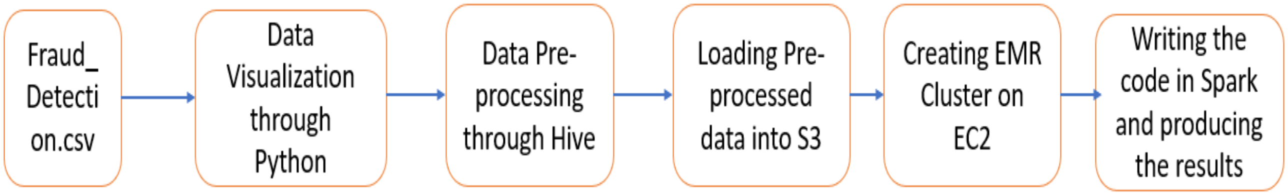
Sathwika Karingu

Krishna Anuhya Regalla

Problem Statement:

Reducing the financial impact of credit card fraud, identity theft, and payment fraud requires constructing a robust fraud detection system. Big data analysis requires the use of sophisticated tools like Hadoop, Spark, and Hive because to the significant increase in digital transactions and the resulting expansion in data. A proficient system must incorporate data processing, machine learning, and real-time analytics to effectively prevent fraud.

Workflow Diagram:


Fraud_ Detection.csv
Data Visualization through Python
Data Pre-processing through Hive
Loading Pre-processed data into S3
Creating EMR Cluster on EC2
Writing the code in Spark and producing the results

Tools used in throughout the project :

- **AWS**
- **HIVE**

Implementation Status Report :

Description :

We are a team of 4 people working on the fraud transactions detection model. We have implemented Random Forest classifier to train our model in pyspark and Python.

Responsibility:

Sathwika – (Team Lead) Helped in building the EMR cluster and working throughout the Notebook.

Anuhya – Worked on the Data Visualization

Vivek – Data Preprocessing through Hive.

Hari - Writing the code for training the model and Hyper tuning parameters.

Contributions:

Sathwika took 25 percent of the project contribution in making the report and the presentation for the project. She is responsible with AWS and Cloudera setup.

Anuhya took 25 percent of the project contribution in making the report and the presentation for the project and also performed Data Visualization using python until the data Preprocessing steps

Vivek- took 25 percent of the project contribution in making the report and the presentation for the project and have written code in SQL.

Hari - took 25 percent of the project contribution in making the report and the presentation for the project and have written code in Python.

Issues/ Concerns:

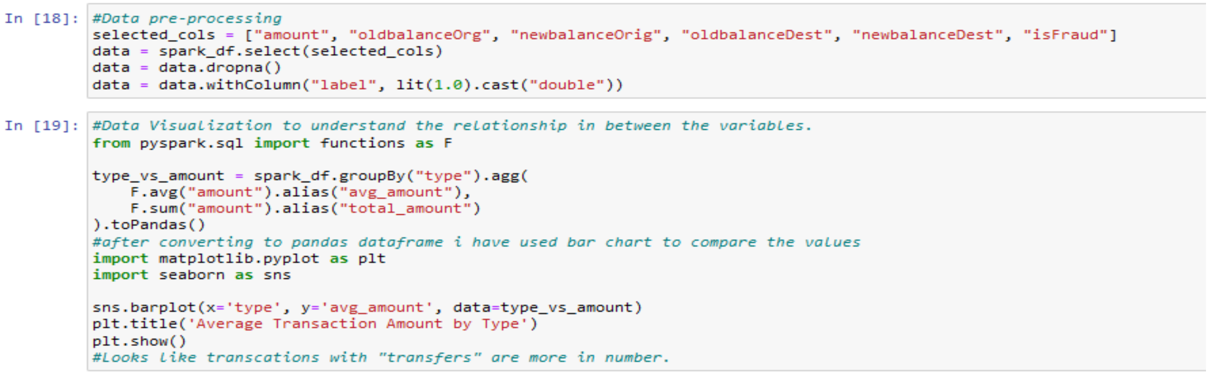
We could have performed more on Hive. We have rather focused on Machine Learning and testing the model instead of working more on preprocessing part. That’s our major concern in this project.

Walking through the Data set :

Step – It’s a integer Data Type which refers to transactions numbering
Type – It’s a String Data Type which contains the payment mode
Amount – Integer or Float value with amount
nameOrig - This column likely represents the originating account or entity's name or identifier. It identifies the source of the transaction.
oldbalanceOrg – Integer value with amount before transaction
newbalanceOrig: Following the transaction, the balance in the originating account is shown in this column. It displays the account balance following a payment or transfer of funds.
nameDest: This field most likely contains the name or identification of the destination account or entity. It indicates who the transaction's recipient or destination is.
oldbalanceDest: The balance in the destination account prior to the transaction is shown in this column. It shows the amount in the recipient's account prior to the funds being sent.
newbalanceDest: Following the transaction, the balance in the destination account is shown in this column. It displays the recipient's account balance following the money transfer.
isFraud- An single integer to state whether the transaction is Fraud or not
isFlaggedFraud – This is also similar to the above variable, It's a potential

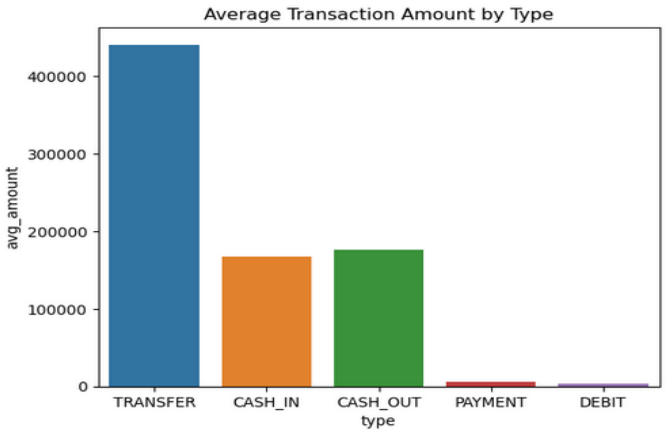
additional fraud detection measure.

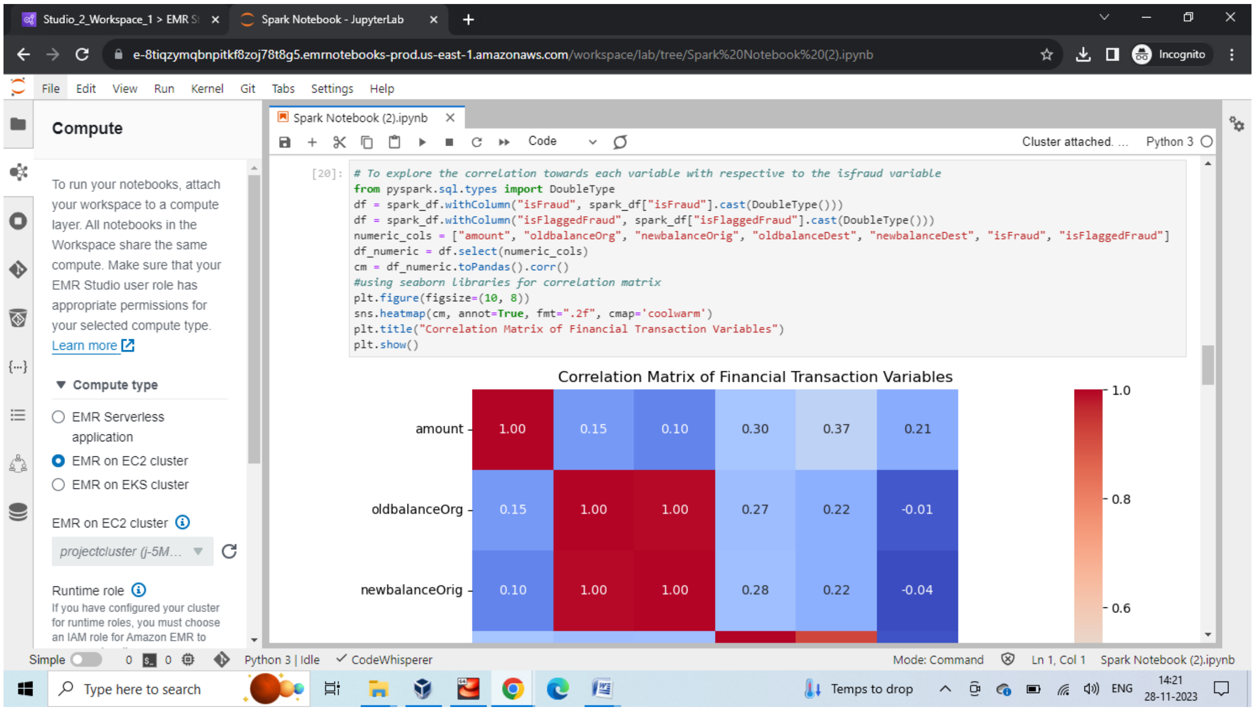
Data Visualization in Python:



```
In [18]: #Data pre-processing
         selected_cols = ["amount", "oldbalanceOrg", "newbalanceOrig", "oldbalanceDest", "newbalanceDest", "isFraud"]
         data = spark_df.select(selected_cols)
         data = data.dropna()
         data = data.withColumn("label", lit(1.0).cast("double"))
```

```
In [19]: #Data Visualization to understand the relationship in between the variables.
         from pyspark.sql import functions as F

         type_vs_amount = spark_df.groupBy("type").agg(
             F.avg("amount").alias("avg_amount"),
             F.sum("amount").alias("total_amount")
         ).toPandas()
         #after converting to pandas dataframe i have used bar chart to compare the values
         import matplotlib.pyplot as plt
         import seaborn as sns

         sns.barplot(x='type', y='avg_amount', data=type_vs_amount)
         plt.title('Average Transaction Amount by Type')
         plt.show()
         #Looks like transcations with "transfers" are more in number.
```




Studio_2_Workspace_1 > EMR St
Spark Notebook - JupyterLab
e-8tiqzymqbnpitkf8zoj78t8g5.emrnotebooks-prod.us-east-1.amazonaws.com/workspace/lab/tree/Spark%20Notebook%20(2).ipynb
Incognito
File
Edit
View
Run
Kernel
Git
Tabs
Settings
Help
Spark Notebook (2).ipynb
Code
Cluster attached. ...
Python 3
Compute

To run your notebooks, attach your workspace to a compute layer. All notebooks in the Workspace share the same compute. Make sure that your EMR Studio user role has appropriate permissions for your selected compute type. Learn more

Compute type

○ EMR Serverless application

● EMR on EC2 cluster

○ EMR on EKS cluster

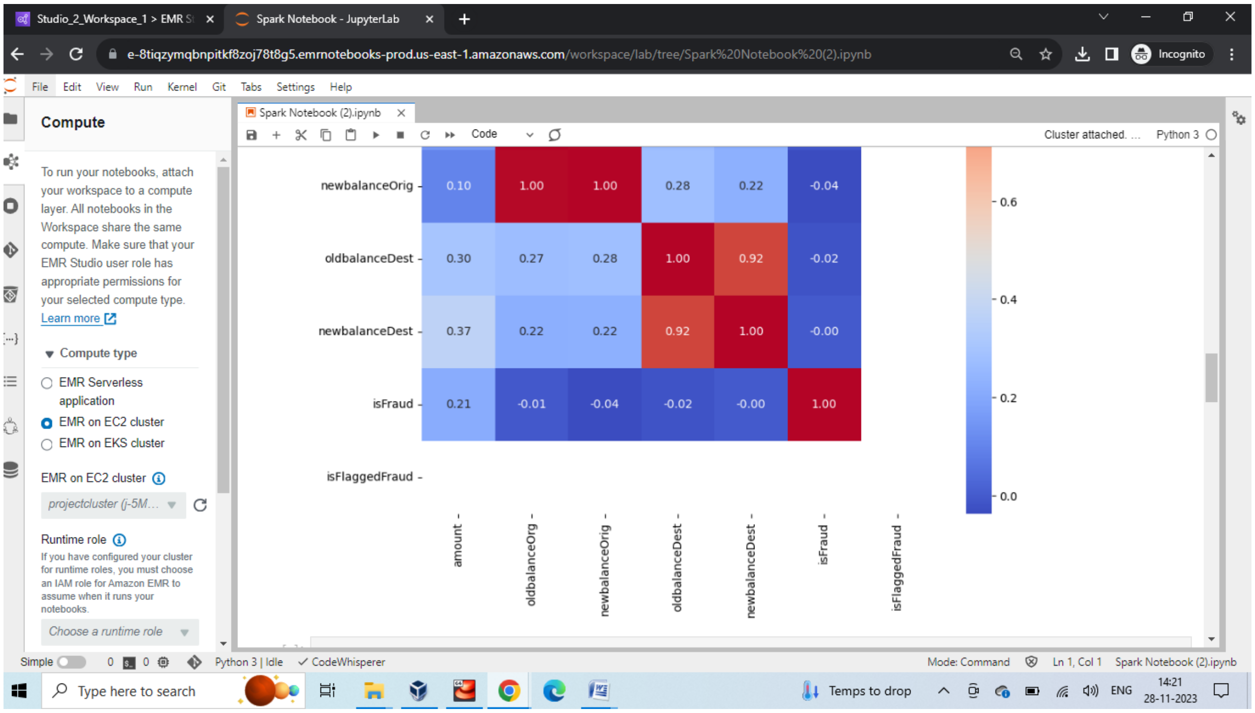
EMR on EC2 cluster: projectcluster (j-5M...

Runtime role

If you have configured your cluster for runtime roles, you must choose an IAM role for Amazon EMR to

```
[20]: # To explore the correlation towards each variable with respective to the isfraud variable
from pyspark.sql.types import DoubleType
df = spark_df.withColumn("isFraud", spark_df["isFraud"].cast(DoubleType()))
df = spark_df.withColumn("isFlaggedFraud", spark_df["isFlaggedFraud"].cast(DoubleType()))
numeric_cols = ["amount", "oldbalanceOrg", "newbalanceOrig", "oldbalanceDest", "newbalanceDest", "isFraud", "isFlaggedFraud"]
df_numeric = df.select(numeric_cols)
cm = df_numeric.toPandas().corr()
#using seaborn libraries for correlation matrix
plt.figure(figsize=(10, 8))
sns.heatmap(cm, annot=True, fmt=".2f", cmap='coolwarm')
plt.title("Correlation Matrix of Financial Transaction Variables")
plt.show()
```

Correlation Matrix of Financial Transaction Variables


Studio_2_Workspace_1 > EMR St
Spark Notebook - JupyterLab
e-8tiqzymqbnpitkf8zoj78t8g5.emrnotebooks-prod.us-east-1.amazonaws.com/workspace/lab/tree/Spark%20Notebook%20(2).ipynb
Incognito
File
Edit
View
Run
Kernel
Git
Tabs
Settings
Help
Spark Notebook (2).ipynb
Code
Cluster attached. ...
Python 3
Compute

To run your notebooks, attach your workspace to a compute layer. All notebooks in the Workspace share the same compute. Make sure that your EMR Studio user role has appropriate permissions for your selected compute type. Learn more

Compute type

○ EMR Serverless application

● EMR on EC2 cluster

○ EMR on EKS cluster

EMR on EC2 cluster: projectcluster (j-5M...

Runtime role

If you have configured your cluster for runtime roles, you must choose an IAM role for Amazon EMR to assume when it runs your notebooks.

Choose a runtime role

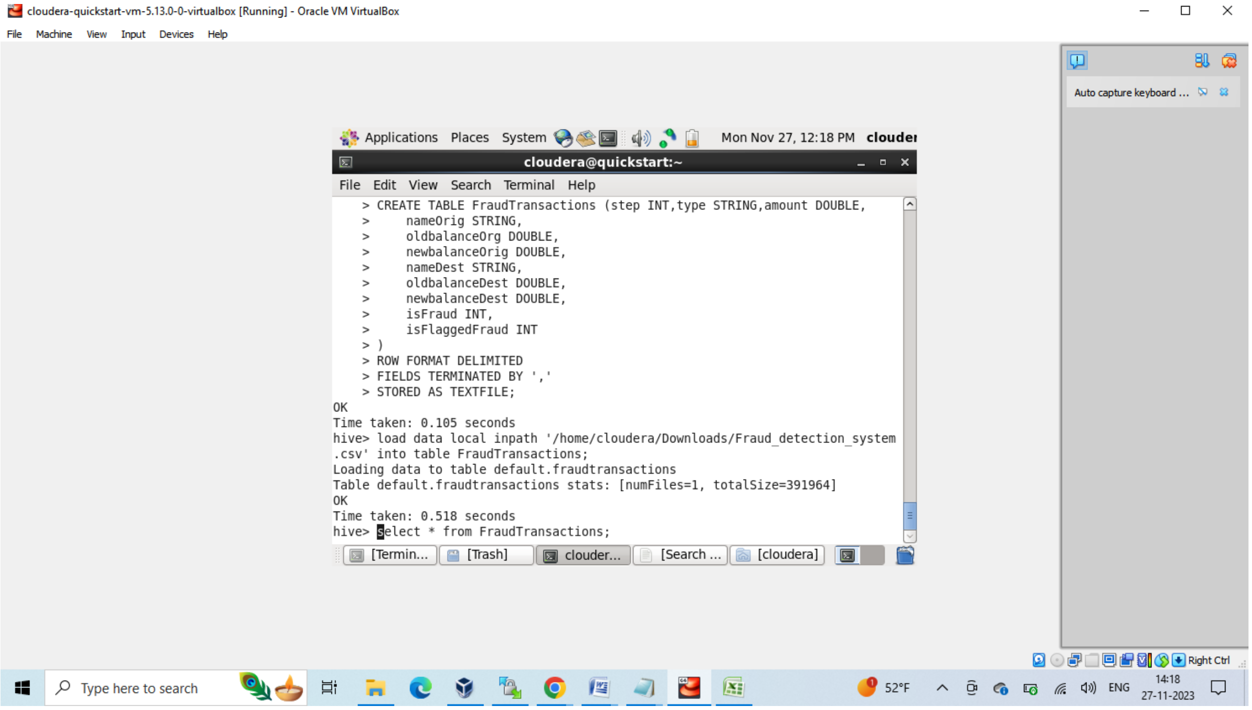
Data Pre-processing in Hive :

Step -01 :
Setup Hive in the Cloudera Platform and create a empty table with the schema as it is present in the csv file.

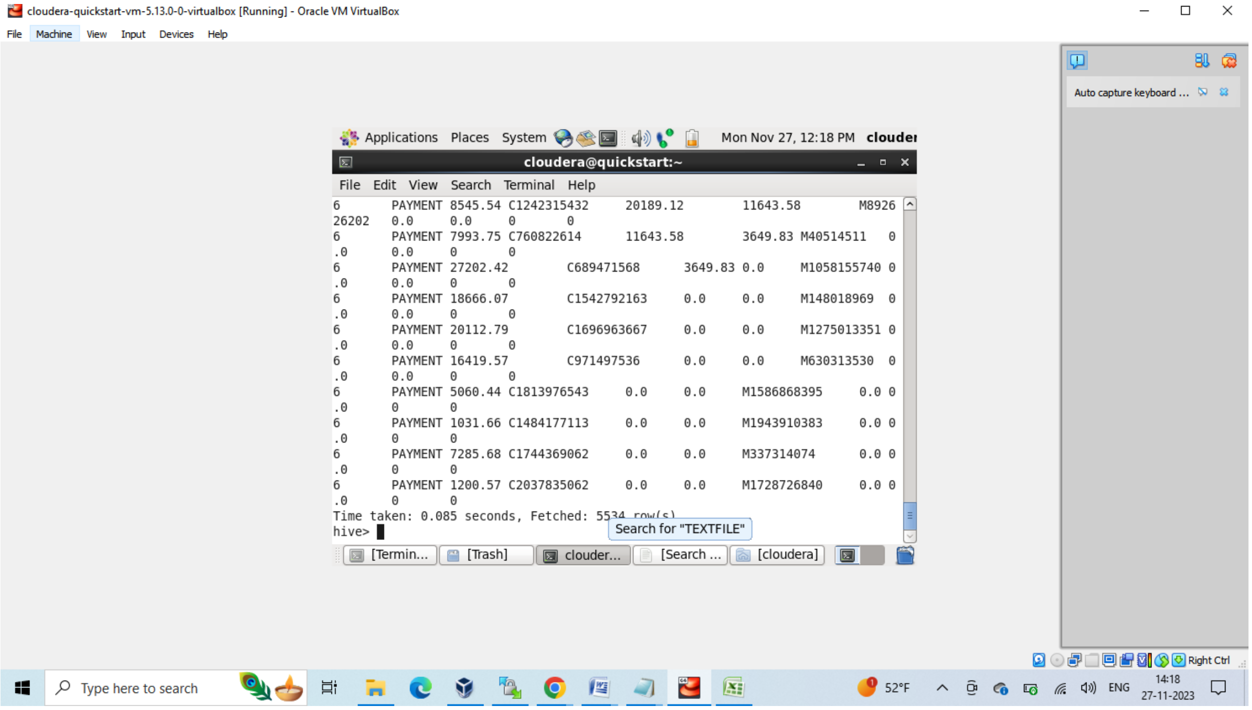
Code to load the data from CSV to Hive Table:

```
CREATE TABLE FraudTransactions (step INT,type STRING,amount DOUBLE,
    nameOrig STRING,
    oldbalanceOrg DOUBLE,
    newbalanceOrig DOUBLE,
    nameDest STRING,
    oldbalanceDest DOUBLE,
    newbalanceDest DOUBLE,
isFraud INT,
    isFlaggedFraud INT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE;
```

Output:


cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Mon Nov 27, 12:18 PM
cloudera@quickstart:~
File Edit View Search Terminal Help
[Termin... [Trash] clouder... [Search ... [cloudera]
Auto capture keyboard ...
Right Ctrl
Type here to search
52°F
ENG
14:18
27-11-2023

```
    > CREATE TABLE FraudTransactions (step INT,type STRING,amount DOUBLE,
    >     nameOrig STRING,
    >     oldbalanceOrg DOUBLE,
    >     newbalanceOrig DOUBLE,
    >     nameDest STRING,
    >     oldbalanceDest DOUBLE,
    >     newbalanceDest DOUBLE,
    >     isFraud INT,
    >     isFlaggedFraud INT
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > STORED AS TEXTFILE;
OK
Time taken: 0.105 seconds
hive> load data local inpath '/home/cloudera/Downloads/Fraud_detection_system
.csv' into table FraudTransactions;
Loading data to table default.fraudtransactions
Table default.fraudtransactions stats: [numFiles=1, totalSize=391964]
OK
Time taken: 0.518 seconds
hive> select * from FraudTransactions;
```


cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Mon Nov 27, 12:18 PM
cloudera@quickstart:~
File Edit View Search Terminal Help
Search for "TEXTFILE"
[Termin... [Trash] clouder... [Search ... [cloudera]
Auto capture keyboard ...
Right Ctrl
Type here to search
52°F
ENG
14:18
27-11-2023

```
6       PAYMENT 8545.54 C1242315432     20189.12        11643.58        M8926
26202   0.0     0.0     0       0
6       PAYMENT 7993.75 C760822614      11643.58        3649.83 M40514511     0
.0      0.0     0       0
6       PAYMENT 27202.42        C689471568      3649.83 0.0     M1058155740 0
.0      0.0     0       0
6       PAYMENT 18666.07        C1542792163     0.0     0.0     M148018969  0
.0      0.0     0       0
6       PAYMENT 20112.79        C1696963667     0.0     0.0     M1275013351 0
.0      0.0     0       0
6       PAYMENT 16419.57        C971497536      0.0     0.0     M630313530  0
.0      0.0     0       0
6       PAYMENT 5060.44 C1813976543     0.0     0.0     M1586868395     0.0 0
.0      0       0
6       PAYMENT 1031.66 C1484177113     0.0     0.0     M1943910383     0.0 0
.0      0       0
6       PAYMENT 7285.68 C1744369062     0.0     0.0     M337314074      0.0 0
.0      0       0
6       PAYMENT 1200.57 C2037835062     0.0     0.0     M1728726840     0.0 0
.0      0       0
Time taken: 0.085 seconds, Fetched: 5534 row(s)
hive> 
```

we can execute several Hive Queries to understand the data and perform necessary analysis in Hive: The following are the queries used to extract insights from the queries :

1) Total Number of transactions by type :

```
query: SELECT type, AVG(amount) AS avg_transaction_amount FROM Fraud
GROUP BY type;
```

2) Money laundering transactions may involve multiple transfers with small amounts. You can identify such patterns by grouping transactions by nameOrig and calculating the total number of outgoing transactions and the total amount transferred for each account

```
SELECT nameOrig, COUNT(*) AS num_outgoing_transactions, SUM(amount)
AS total_outgoing_amount

FROM your_table_name

WHERE type IN ('TRANSFER', 'CASH_OUT')

GROUP BY nameOrig

HAVING num_outgoing_transactions > 1 AND total_outgoing_amount <1000;
```

3) Based on your fraud detection criteria, you can flag transactions that meet certain conditions as potentially fraudulent. For example, transactions where isFraud is not flagged but the amount is significantly higher than average.

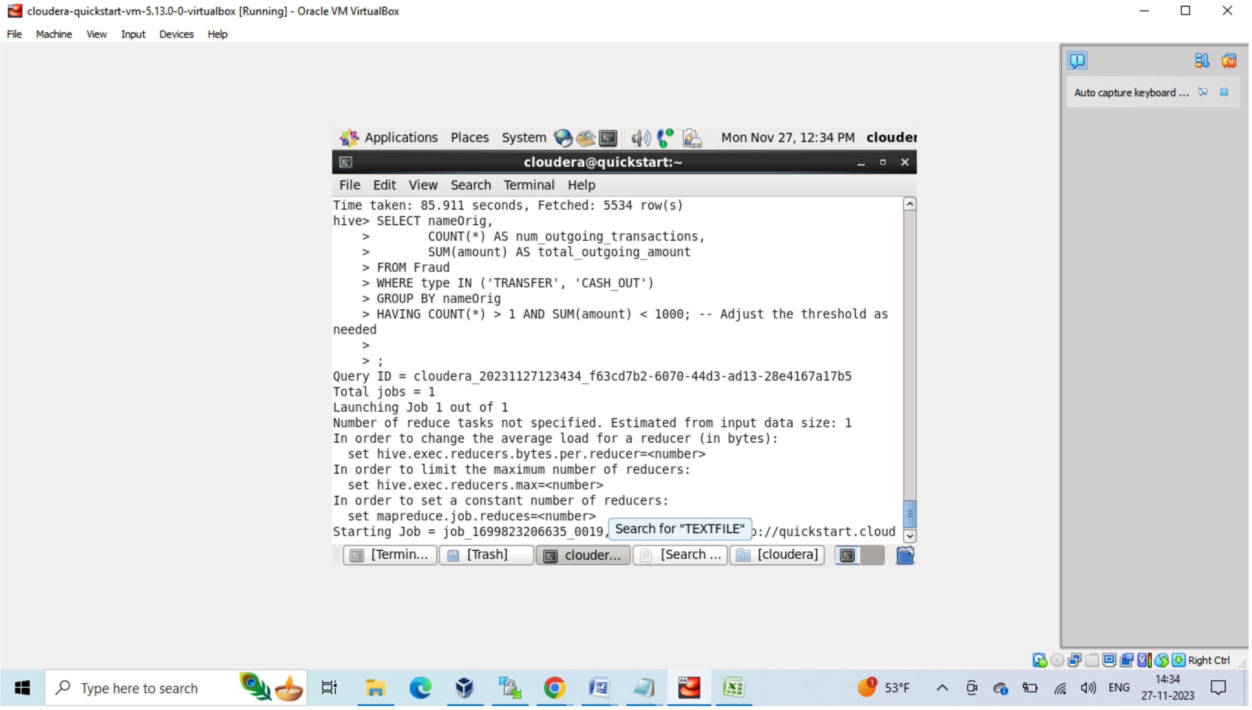
```
query: SELECT F.*,

    CASE
```

```
          WHEN F.isFraud = 0 AND F.amount > A.avg_amount THEN 1
          ELSE 0
      END AS flagged_as_potential_fraud
FROM FraudTransactions F
CROSS JOIN (SELECT AVG(amount) as avg_amount FROM Fraud) A;
```

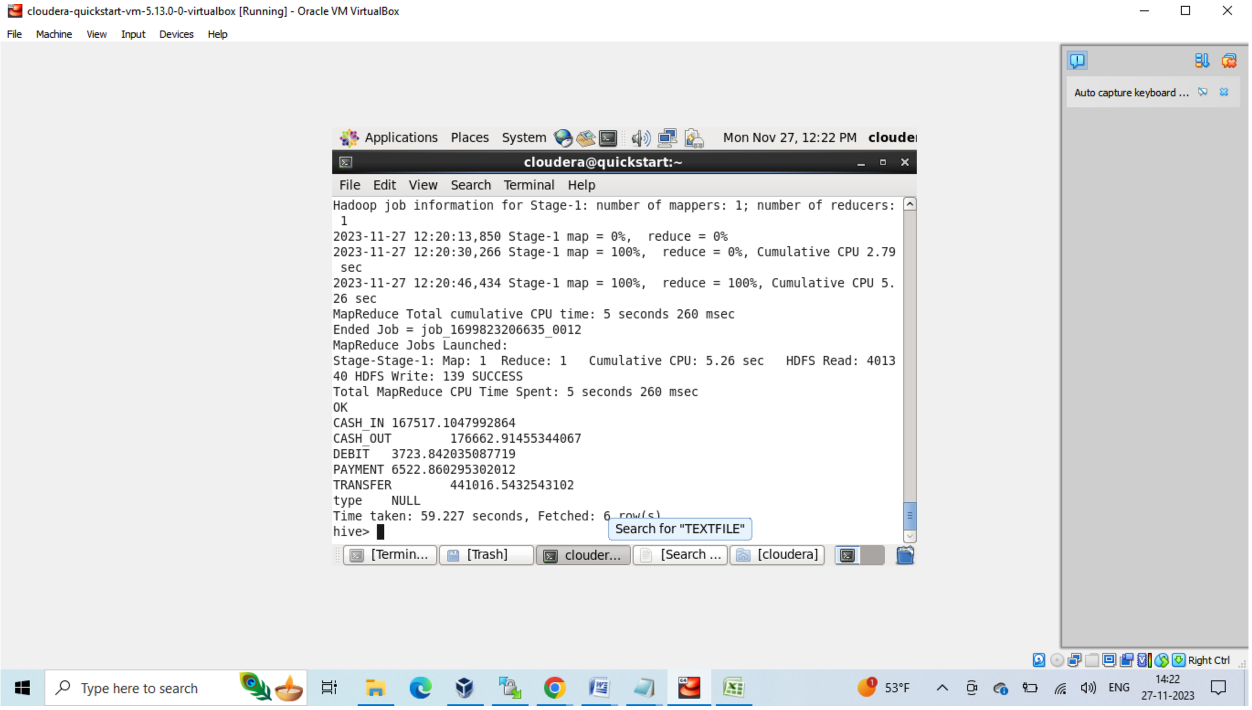
Insights:

- The last query will yield a new column which says potential fraud. This will flag transactions that meet certain criteria as potentially fraudulent.
- It will return all columns from the FraudTransactions table along with a new column flagged_as_potential_fraud. Transactions where isFraud is 0 (not flagged as fraud) and the amount is significantly higher than the average amount in the entire Fraud dataset will be flagged as 1; otherwise, they will be flagged as 0.
- Based on the isFraud and Potential_fraud transactions, there isn’t much difference. So it’s better to go with the same dataset we have taken.

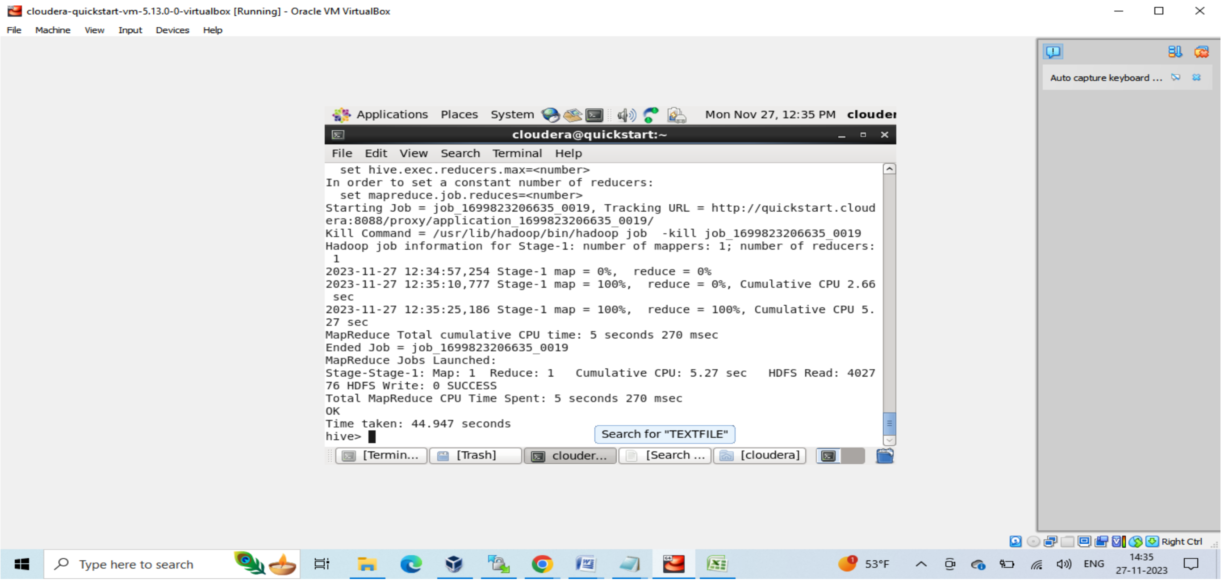
Output :


cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Mon Nov 27, 12:34 PM clouder
cloudera@quickstart:~
File Edit View Search Terminal Help
Search for "TEXTFILE"
[Termin... [Trash] clouder... [Search ... [cloudera]
Auto capture keyboard ...
Right Ctrl
Type here to search
53°F
ENG
14:34
27-11-2023

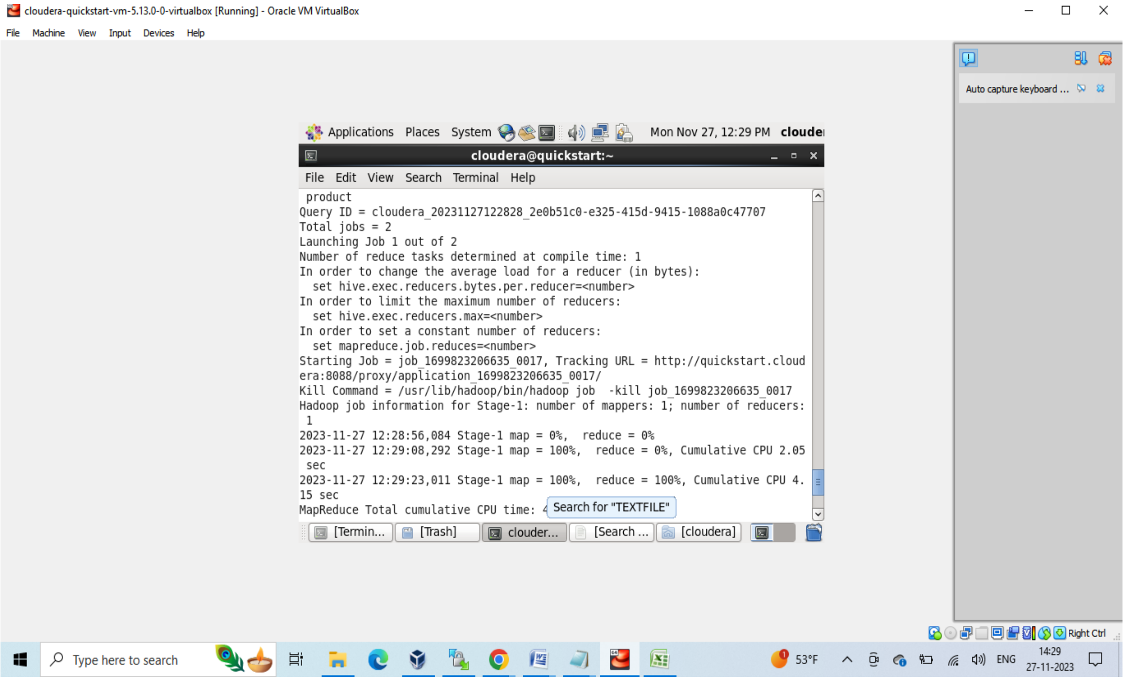
```
Time taken: 85.911 seconds, Fetched: 5534 row(s)
hive> SELECT nameOrig,
    >        COUNT(*) AS num_outgoing_transactions,
    >        SUM(amount) AS total_outgoing_amount
    > FROM Fraud
    > WHERE type IN ('TRANSFER', 'CASH_OUT')
    > GROUP BY nameOrig
    > HAVING COUNT(*) > 1 AND SUM(amount) < 1000; -- Adjust the threshold as
needed
    >
    > ;
Query ID = cloudera_20231127123434_f63cd7b2-6070-44d3-ad13-28e4167a17b5
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1699823206635_0019,                    p://quickstart.cloud
```


cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Mon Nov 27, 12:22 PM clouder
cloudera@quickstart:~
File Edit View Search Terminal Help
Search for "TEXTFILE"
[Termin... [Trash] clouder... [Search ... [cloudera]
Auto capture keyboard ...
Right Ctrl
Type here to search
53°F
ENG
14:22
27-11-2023

```
Hadoop job information for Stage-1: number of mappers: 1; number of reducers:
 1
2023-11-27 12:20:13,850 Stage-1 map = 0%,  reduce = 0%
2023-11-27 12:20:30,266 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.79
 sec
2023-11-27 12:20:46,434 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.
26 sec
MapReduce Total cumulative CPU time: 5 seconds 260 msec
Ended Job = job_1699823206635_0012
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.26 sec   HDFS Read: 4013
40 HDFS Write: 139 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 260 msec
OK
CASH_IN 167517.1047992864
CASH_OUT        176662.91455344067
DEBIT   3723.842035087719
PAYMENT 6522.860295302012
TRANSFER        441016.5432543102
type    NULL
Time taken: 59.227 seconds, Fetched: 6 row(s)
hive> 
```


cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Mon Nov 27, 12:35 PM clouder
cloudera@quickstart:~
File Edit View Search Terminal Help
Search for "TEXTFILE"
[Termin... [Trash] clouder... [Search ... [cloudera]
Auto capture keyboard ...
Right Ctrl
Type here to search
53°F ENG 14:35 27-11-2023

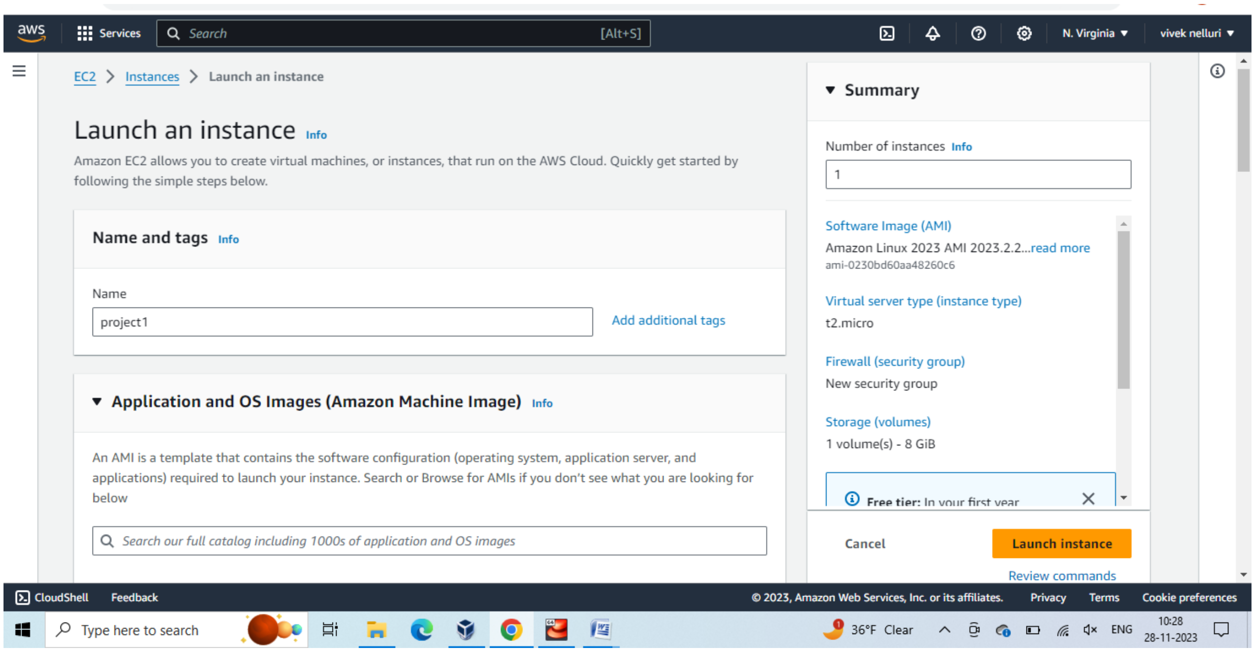
```
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1699823206635_0019, Tracking URL = http://quickstart.cloud
era:8088/proxy/application_1699823206635_0019/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1699823206635_0019
Hadoop job information for Stage-1: number of mappers: 1; number of reducers:
 1
2023-11-27 12:34:57,254 Stage-1 map = 0%,  reduce = 0%
2023-11-27 12:35:10,777 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.66
 sec
2023-11-27 12:35:25,186 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.
27 sec
MapReduce Total cumulative CPU time: 5 seconds 270 msec
Ended Job = job_1699823206635_0019
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.27 sec   HDFS Read: 4027
76 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 270 msec
OK
Time taken: 44.947 seconds
hive> 
```


cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Mon Nov 27, 12:29 PM clouder
cloudera@quickstart:~
File Edit View Search Terminal Help
Search for "TEXTFILE"
[Termin... [Trash] clouder... [Search ... [cloudera]
Auto capture keyboard ...
Right Ctrl
Type here to search
53°F ENG 14:29 27-11-2023

```
 product
Query ID = cloudera_20231127122828_2e0b51c0-e325-415d-9415-1088a0c47707
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1699823206635_0017, Tracking URL = http://quickstart.cloud
era:8088/proxy/application_1699823206635_0017/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1699823206635_0017
Hadoop job information for Stage-1: number of mappers: 1; number of reducers:
 1
2023-11-27 12:28:56,084 Stage-1 map = 0%,  reduce = 0%
2023-11-27 12:29:08,292 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.05
 sec
2023-11-27 12:29:23,011 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.
15 sec
MapReduce Total cumulative CPU time: 4
```

EMR Cluster Creation on EC2 instance (AWS) :

EC2 instance Launch :



Launch an instance

Amazon EC2 allows you to create virtual machines, or instances, that run on the AWS Cloud. Quickly get started by following the simple steps below.

Name and tags

Name: project1

Application and OS Images (Amazon Machine Image)

An AMI is a template that contains the software configuration (operating system, application server, and applications) required to launch your instance. Search or Browse for AMIs if you don't see what you are looking for below

Search our full catalog including 1000s of application and OS images

Summary

Number of instances: 1

Software Image (AMI)
Amazon Linux 2023 AMI 2023.2.2...read more
ami-0230bd60aa48260c6

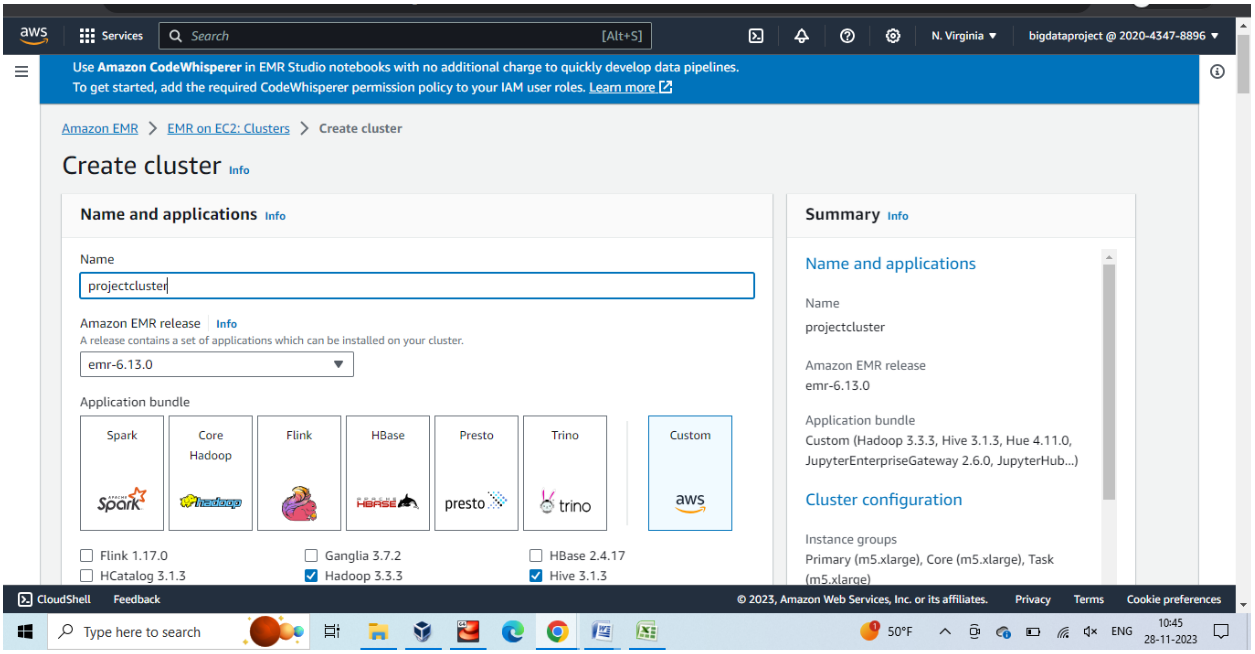
Virtual server type (instance type)
t2.micro

Firewall (security group)
New security group

Storage (volumes)
1 volume(s) - 8 GiB

Free tier: In your first year

Cluster Creation :



Use **Amazon CodeWhisperer** in EMR Studio notebooks with no additional charge to quickly develop data pipelines.
To get started, add the required CodeWhisperer permission policy to your IAM user roles. Learn more

Amazon EMR > EMR on EC2: Clusters > Create cluster

Create cluster Info

Name and applications Info

Name: projectcluster

Amazon EMR release Info
A release contains a set of applications which can be installed on your cluster.
emr-6.13.0

Application bundle: Spark | Core Hadoop | Flink | HBase | Presto | Trino | Custom

☐ Flink 1.17.0 ☐ Ganglia 3.7.2 ☐ HBase 2.4.17
☐ HCatalog 3.1.3 ☑ Hadoop 3.3.3 ☑ Hive 3.1.3

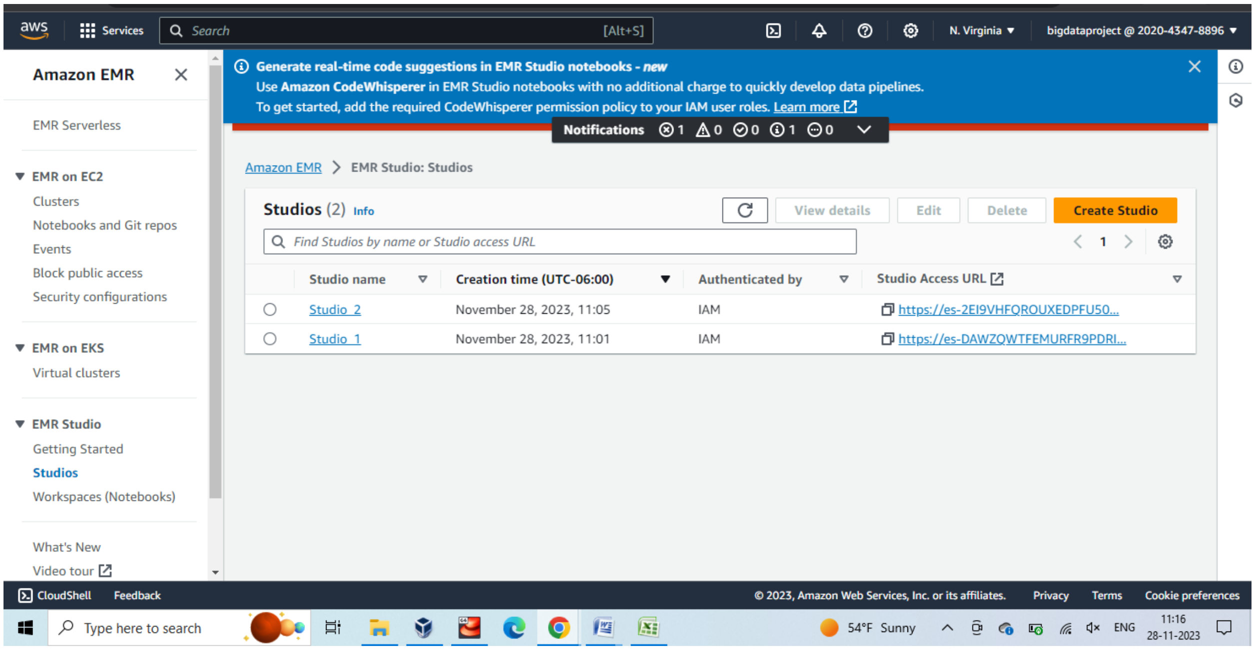
Summary Info

Name and applications
Name: projectcluster
Amazon EMR release: emr-6.13.0
Application bundle: Custom (Hadoop 3.3.3, Hive 3.1.3, Hue 4.11.0, JupyterEnterpriseGateway 2.6.0, JupyterHub...)

Cluster configuration
Instance groups: Primary (m5.xlarge), Core (m5.xlarge), Task (m5.xlarge)

© 2023, Amazon Web Services, Inc. or its affiliates.

Workspace Creation :



Amazon EMR

- EMR Serverless
- EMR on EC2
  - Clusters
  - Notebooks and Git repos
  - Events
  - Block public access
  - Security configurations
- EMR on EKS
  - Virtual clusters
- EMR Studio
  - Getting Started
  - Studios
  - Workspaces (Notebooks)
- What's New
- Video tour

Generate real-time code suggestions in EMR Studio notebooks - *new*
Use **Amazon CodeWhisperer** in EMR Studio notebooks with no additional charge to quickly develop data pipelines.
To get started, add the required CodeWhisperer permission policy to your IAM user roles. Learn more

Notifications ⊗1 ⚠0 ⊘0 ⓘ1 ⊖0

Amazon EMR > EMR Studio: Studios

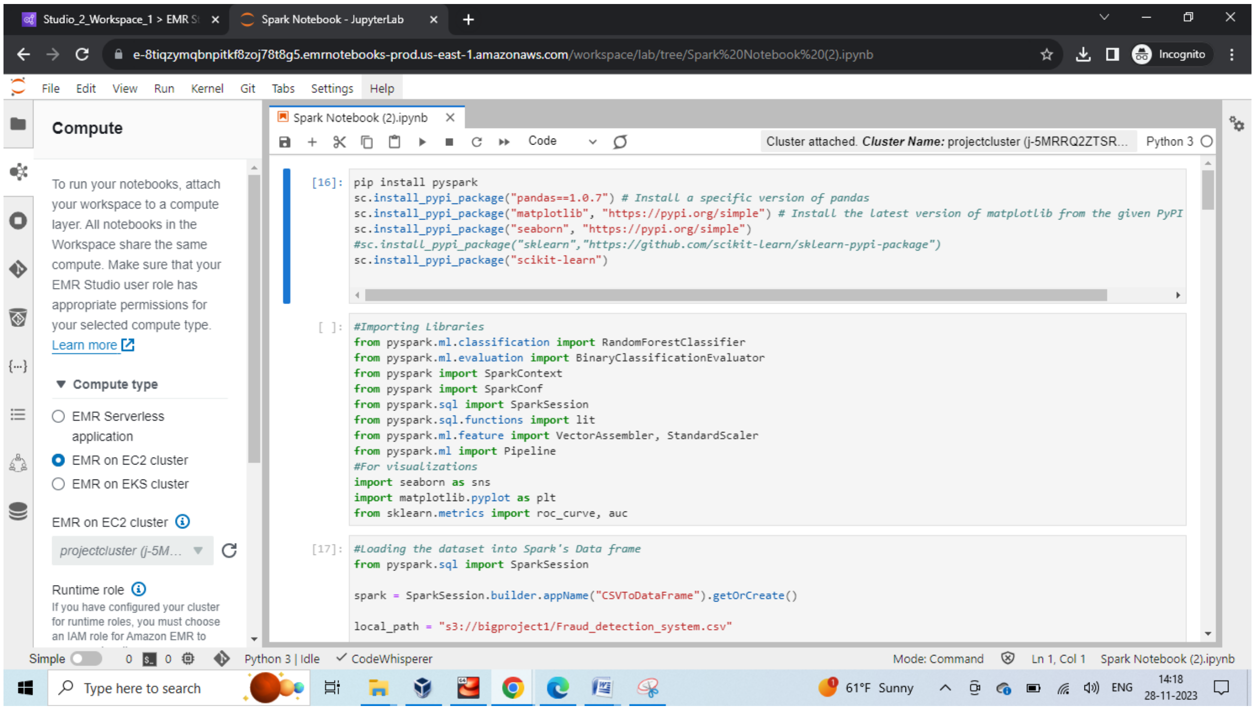
Studios (2) Info

Find Studios by name or Studio access URL

	Studio name	Creation time (UTC-06:00)	Authenticated by	Studio Access URL
○	Studio_2	November 28, 2023, 11:05	IAM	https://es-2EI9VHFQROUXEDPFU50...
○	Studio_1	November 28, 2023, 11:01	IAM	https://es-DAWZQWTFEMURFR9PDRI...

© 2023, Amazon Web Services, Inc. or its affiliates.

Loading the Notebook for writing the code in pyspark :



Cluster attached. ***Cluster Name:*** projectcluster (j-5MRRQ2ZTSR...

Compute

To run your notebooks, attach your workspace to a compute layer. All notebooks in the Workspace share the same compute. Make sure that your EMR Studio user role has appropriate permissions for your selected compute type. Learn more

Compute type

○ EMR Serverless application
● EMR on EC2 cluster
○ EMR on EKS cluster

EMR on EC2 cluster: *projectcluster (j-5M…*

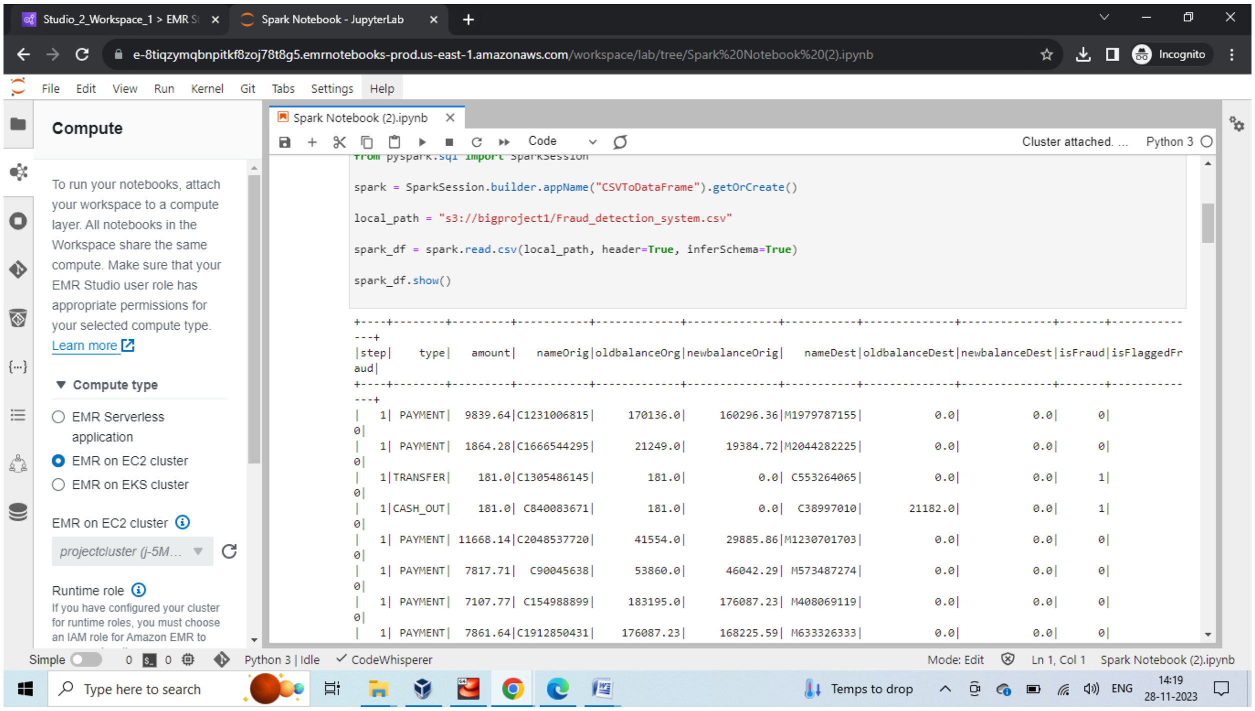
Runtime role
If you have configured your cluster for runtime roles, you must choose an IAM role for Amazon EMR to

[16]:
```
pip install pyspark
sc.install_pypi_package("pandas==1.0.7") # Install a specific version of pandas
sc.install_pypi_package("matplotlib", "https://pypi.org/simple") # Install the latest version of matplotlib from the given PyPI
sc.install_pypi_package("seaborn", "https://pypi.org/simple")
#sc.install_pypi_package("sklearn","https://github.com/scikit-learn/sklearn-pypi-package")
sc.install_pypi_package("scikit-learn")
```

[ ]:
```
#Importing Libraries
from pyspark.ml.classification import RandomForestClassifier
from pyspark.ml.evaluation import BinaryClassificationEvaluator
from pyspark import SparkContext
from pyspark import SparkConf
from pyspark.sql import SparkSession
from pyspark.sql.functions import lit
from pyspark.ml.feature import VectorAssembler, StandardScaler
from pyspark.ml import Pipeline
#For visualizations
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.metrics import roc_curve, auc
```

[17]:
```
#Loading the dataset into Spark's Data frame
from pyspark.sql import SparkSession

spark = SparkSession.builder.appName("CSVToDataFrame").getOrCreate()

local_path = "s3://bigproject1/Fraud_detection_system.csv"
```

Loading the data into Spark Data Frame :



Compute

To run your notebooks, attach your workspace to a compute layer. All notebooks in the Workspace share the same compute. Make sure that your EMR Studio user role has appropriate permissions for your selected compute type. Learn more

Compute type

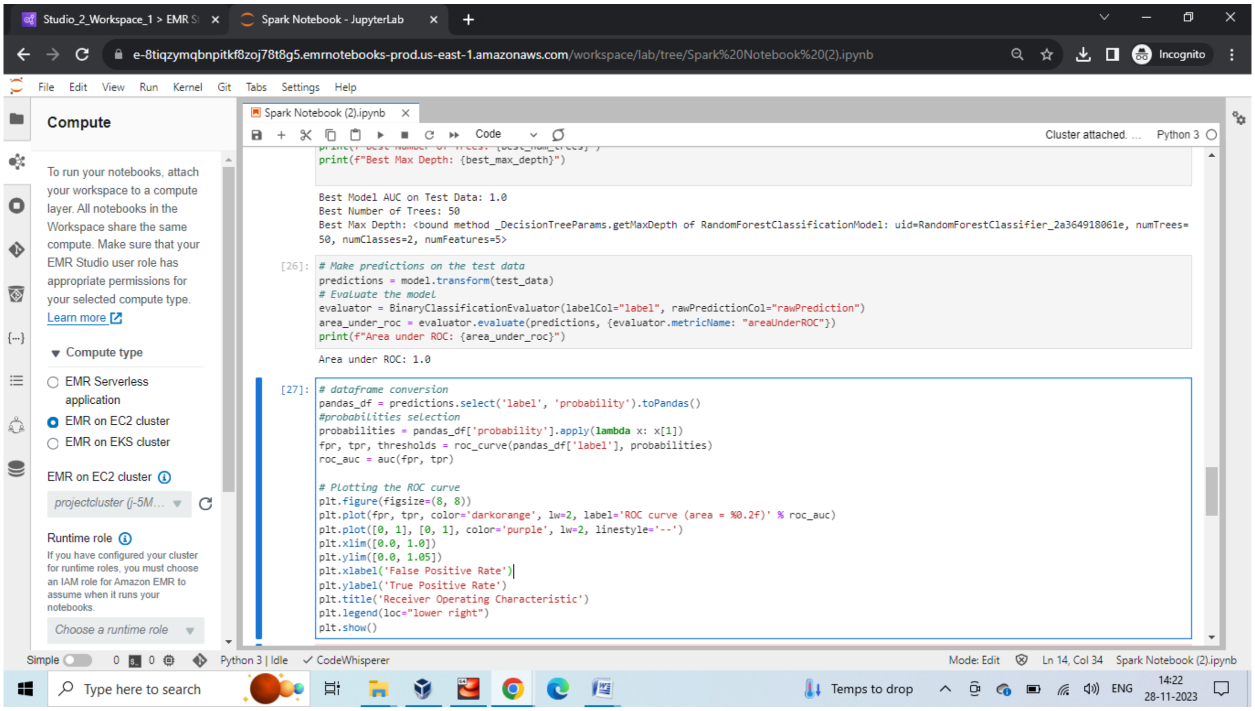
○ EMR Serverless application
● EMR on EC2 cluster
○ EMR on EKS cluster

EMR on EC2 cluster: projectcluster (j-5M…

Runtime role: If you have configured your cluster for runtime roles, you must choose an IAM role for Amazon EMR to

```
from pyspark.sql import SparkSession

spark = SparkSession.builder.appName("CSVToDataFrame").getOrCreate()

local_path = "s3://bigproject1/Fraud_detection_system.csv"

spark_df = spark.read.csv(local_path, header=True, inferSchema=True)

spark_df.show()
```

```
+----+--------+--------+-----------+-------------+--------------+-----------+--------------+--------------+-------+-----------
---+
|step|    type|  amount|   nameOrig|oldbalanceOrg|newbalanceOrig|   nameDest|oldbalanceDest|newbalanceDest|isFraud|isFlaggedFr
aud|
+----+--------+--------+-----------+-------------+--------------+-----------+--------------+--------------+-------+-----------
---+
|   1| PAYMENT| 9839.64|C1231006815|     170136.0|     160296.36|M1979787155|           0.0|           0.0|      0|
0|
|   1| PAYMENT| 1864.28|C1666544295|      21249.0|      19384.72|M2044282225|           0.0|           0.0|      0|
0|
|   1|TRANSFER|   181.0|C1305486145|        181.0|           0.0| C553264065|           0.0|           0.0|      1|
0|
|   1|CASH_OUT|   181.0| C840083671|        181.0|           0.0|  C38997010|       21182.0|           0.0|      1|
0|
|   1| PAYMENT|11668.14|C2048537720|      41554.0|      29885.86|M1230701703|           0.0|           0.0|      0|
0|
|   1| PAYMENT| 7817.71|  C90045638|      53860.0|      46042.29| M573487274|           0.0|           0.0|      0|
0|
|   1| PAYMENT| 7107.77| C154988899|     183195.0|     176087.23| M408069119|           0.0|           0.0|      0|
0|
|   1| PAYMENT| 7861.64|C1912850431|    176087.23|     168225.59| M633326333|           0.0|           0.0|      0|
```



Compute

To run your notebooks, attach your workspace to a compute layer. All notebooks in the Workspace share the same compute. Make sure that your EMR Studio user role has appropriate permissions for your selected compute type. Learn more

Compute type

○ EMR Serverless application
● EMR on EC2 cluster
○ EMR on EKS cluster

EMR on EC2 cluster: projectcluster (j-5M…

Runtime role: If you have configured your cluster for runtime roles, you must choose an IAM role for Amazon EMR to assume when it runs your notebooks.

Choose a runtime role

```
print(f"Best Number of Trees: {best_num_trees}")
print(f"Best Max Depth: {best_max_depth}")
```

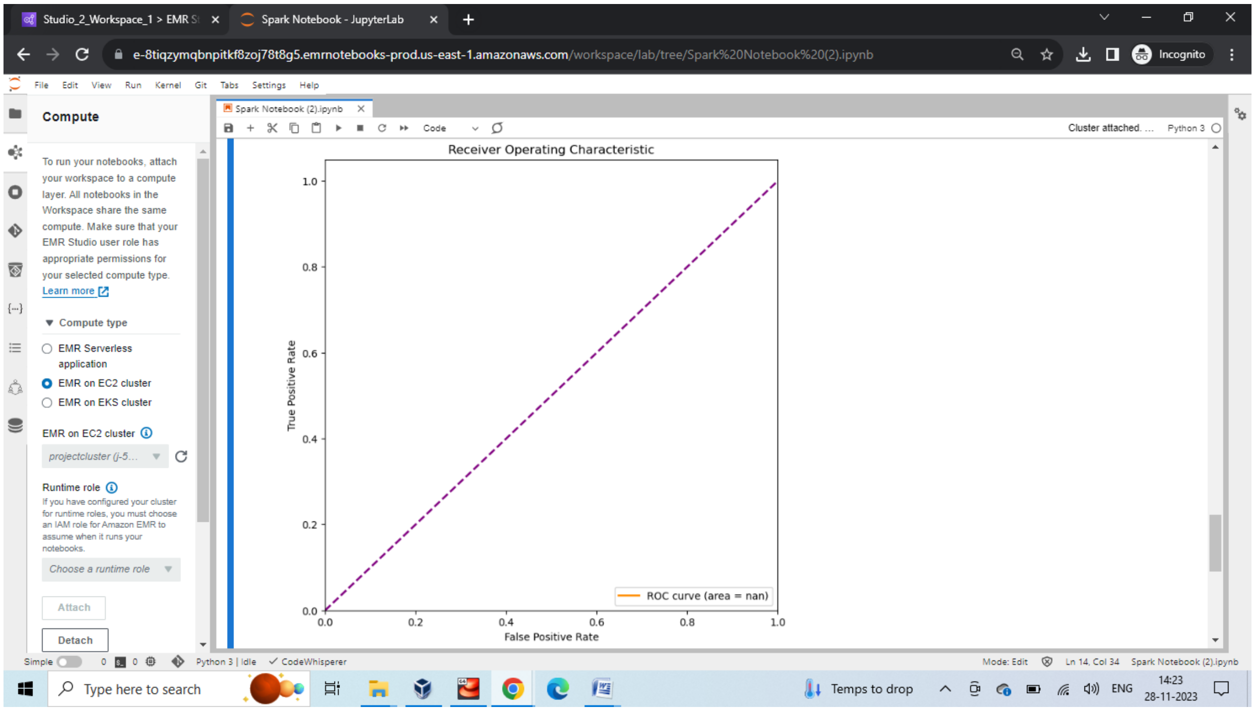
```
Best Model AUC on Test Data: 1.0
Best Number of Trees: 50
Best Max Depth: <bound method _DecisionTreeParams.getMaxDepth of RandomForestClassificationModel: uid=RandomForestClassifier_2a364918061e, numTrees=
50, numClasses=2, numFeatures=5>
```

```
[26]: # Make predictions on the test data
predictions = model.transform(test_data)
# Evaluate the model
evaluator = BinaryClassificationEvaluator(labelCol="label", rawPredictionCol="rawPrediction")
area_under_roc = evaluator.evaluate(predictions, {evaluator.metricName: "areaUnderROC"})
print(f"Area under ROC: {area_under_roc}")
```

```
Area under ROC: 1.0
```

```
[27]: # dataframe conversion
pandas_df = predictions.select('label', 'probability').toPandas()
#probabilities selection
probabilities = pandas_df['probability'].apply(lambda x: x[1])
fpr, tpr, thresholds = roc_curve(pandas_df['label'], probabilities)
roc_auc = auc(fpr, tpr)

# Plotting the ROC curve
plt.figure(figsize=(8, 8))
plt.plot(fpr, tpr, color='darkorange', lw=2, label='ROC curve (area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], color='purple', lw=2, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic')
plt.legend(loc="lower right")
plt.show()
```

Studio_2_Workspace_1 > EMR S
Spark Notebook - JupyterLab
e-8tiqzymqbnpitkf8zoj78t8g5.emrnotebooks-prod.us-east-1.amazonaws.com/workspace/lab/tree/Spark%20Notebook%20(2).ipynb
Incognito
Spark Notebook (2).ipynb
Code
Cluster attached. ...
Python 3
Python 3 | Idle
CodeWhisperer
Mode: Edit
Ln 14, Col 34
Type here to search
Temps to drop
ENG
14:22
28-11-2023



Compute

To run your notebooks, attach your workspace to a compute layer. All notebooks in the Workspace share the same compute. Make sure that your EMR Studio user role has appropriate permissions for your selected compute type. Learn more

Compute type

○ EMR Serverless application
● EMR on EC2 cluster
○ EMR on EKS cluster

EMR on EC2 cluster: projectcluster (j-5...

Runtime role

If you have configured your cluster for runtime roles, you must choose an IAM role for Amazon EMR to assume when it runs your notebooks.

Choose a runtime role

Attach

Detach

- **Final Conclusion:**

- The Receiver Operating Characteristic (ROC) curve offers a graphical depiction of how well a model can differentiate between two categories, such as positive and negative instances. This curve enables the evaluation of the model's proficiency in distinguishing true positives (accurately identified positive cases) from false positives (wrongly identified positive cases), over a range of probability thresholds.
- The area under ROC curve indicates that 1 which indicates a perfect classifier
- The “evaluate method” of the evaluator computes the metric specified for the model predictions provided.
- In this case, {evaluator.metricName: "areaUnderROC"} sets the metric to the Area Under the ROC Curve (AUC - ROC).

- The AUC - ROC is a single scalar value that summarizes the performance of the binary classification model across all classification thresholds.
- It essentially measures the ability of the model to discriminate between the two classes.