Real-Time Credit Card Fraud Detection Pipeline on Azure

Problem Statement

In recent years, credit card fraud has become increasingly sophisticated, leveraging advanced techniques such as identity theft, phishing, and AI-driven fraudulent transactions. Traditional fraud detection methods often rely on static rule-based systems, which struggle to keep pace with real-time fraud patterns and large-scale data processing demands. As a result, financial institutions face significant monetary losses, reputational damage, and compromised customer trust.

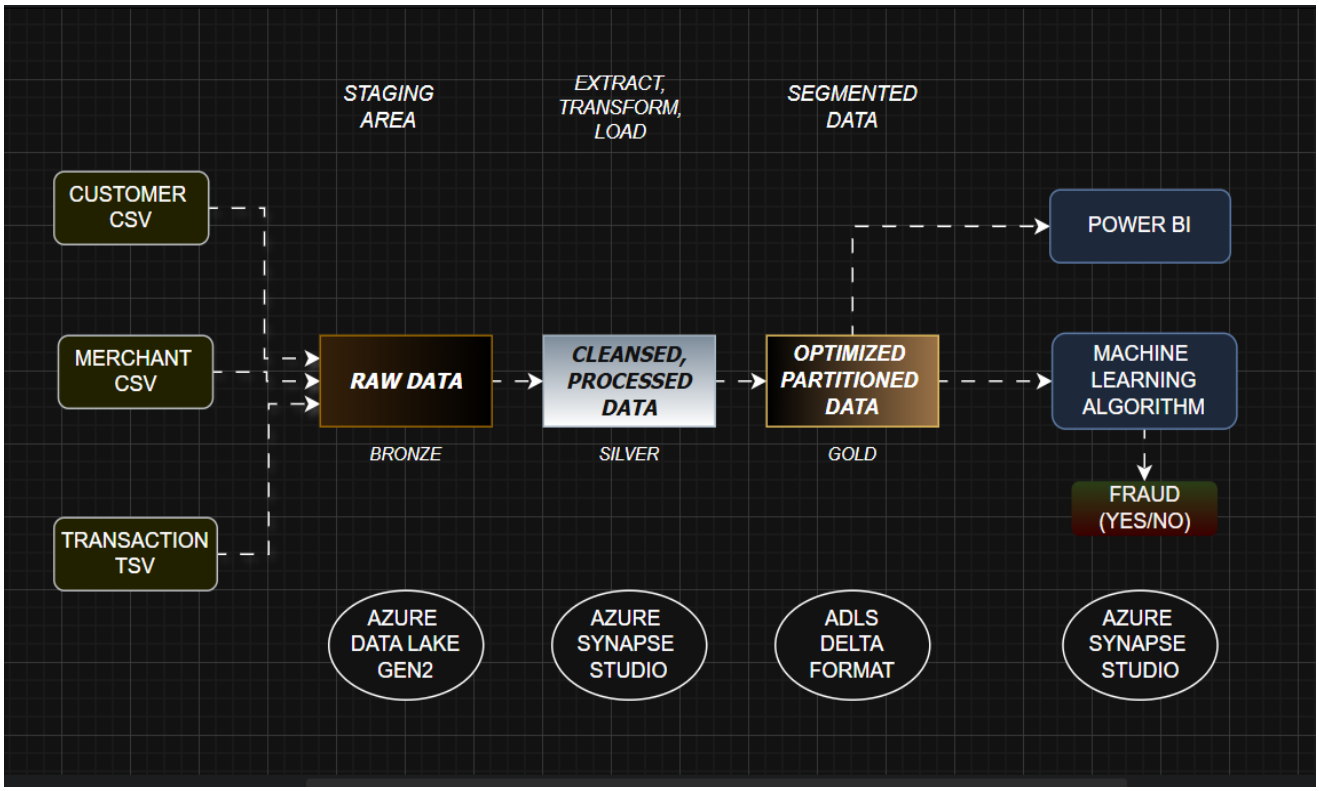
To effectively combat this growing threat, a leading bank seeks to develop a cloud-native, scalable, real-time fraud detection pipeline using Microsoft Azure. The system must be capable of processing massive volumes of transactional data, detecting anomalies instantly, and triggering fraud alerts for investigation, ensuring customer security and financial integrity.

Objectives

The goal of this project is to create a secure, scalable, and optimized data pipeline on Azure to support fraud detection and analysis. The key steps include:

1. **Live Data Streaming & Static Data Ingestion –** Utilize Azure Event Hub to ingest real-time credit card transaction data. Stream incoming transactions into Azure Data Lake Storage Gen2 for processing and analysis**.**
2. **Data Storage & Processing** – Load raw credit card transaction data (CSV & TSV) into Azure Data Lake Storage Gen2, clean it using PySpark in Azure Synapse Studio, and store optimized data in Parquet format for efficiency**.**
3. **Data Optimization –** Improve data retrieval speed by applying bucketing and partitioning, then organize datasets into a gold-layer-data container for advanced analysis.
4. **Fraud Classification using Machine Learning –** Implement a Logistic Regression model in Python within Synapse Studio to classify transactions as fraudulent or legitimate based on historical data.
5. **Visualization & Reporting –** Connect Power BI to the optimized data layer to create interactive dashboards for fraud trend analysis and monitoring.

WORKFLOW:



Technical Stack

1. Data Storage & Processing

- Azure Data Lake Storage Gen2 (ADLS)
- Azure Synapse Studio(Pyspark , ML)

2.Security & Access Management

- Azure Key Vault

3. Visualization & Reporting

- Power BI

4.Live Data Streaming

- Azure event Hub

Dataset Description

This dataset was downloaded from **Kaggle** and contained structured data for **fraud detection** in credit card transactions. It was divided into **three datasets**, each serving a distinct purpose in the fraud detection pipeline.

1. Customer Dataset (15 Columns)

This dataset stored customer demographic details, transaction identifiers, and location-related information. It was used to analyse customer behaviour and assess fraud risk.

Column Name	Description
trans_num	Unique transaction ID associated with the customer.
cc_num	Credit card number used for transactions.
gender	Gender of the customer.
dob	Date of birth of the customer.
job	Occupation of the customer.
street	Street address of the customer.
city	City where the customer resided.
state	State where the customer lived.
zip	Zip code of the customer's location.
lat	Latitude of the customer's location.
long	Longitude of the customer's location.
city_pop	Population of the city where the customer lived.
full_name	Full name of the customer.
random_notes	Additional metadata or random notes.

2. Merchant Dataset (6 Columns)

This dataset recorded merchant-related information linked to transactions. It was primarily used to track merchant categories and identify potential fraud hotspots.

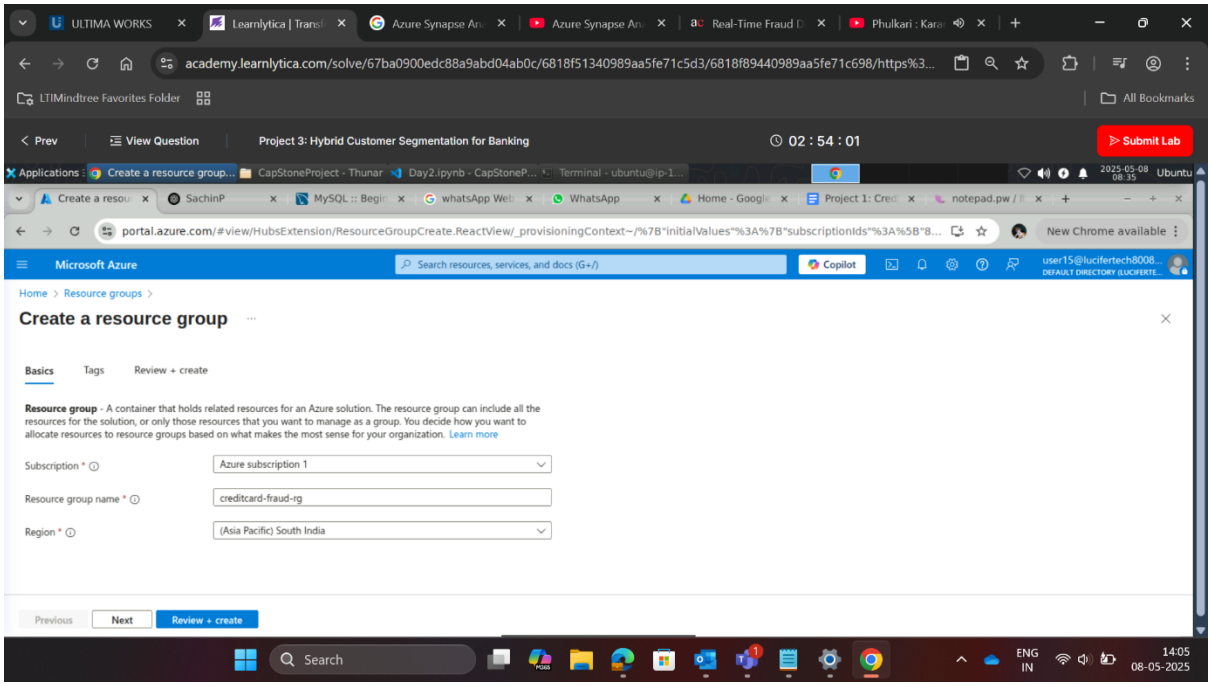
Column Name	Description
trans_num	Unique transaction ID linked to the merchant.
merchant	Name of the merchant involved in the transaction.
category	Business category of the merchant.
merch_lat	Latitude coordinate of the merchant's location.
merch_long	Longitude coordinate of the merchant's location.
merch_zipcode	Zip code of the merchant's business location.

3. Transactions Dataset (7 Columns)

This dataset stored individual **credit card transactions**, including timestamps, transaction amounts, and fraud labels. It was the core dataset for fraud classification using machine learning models.

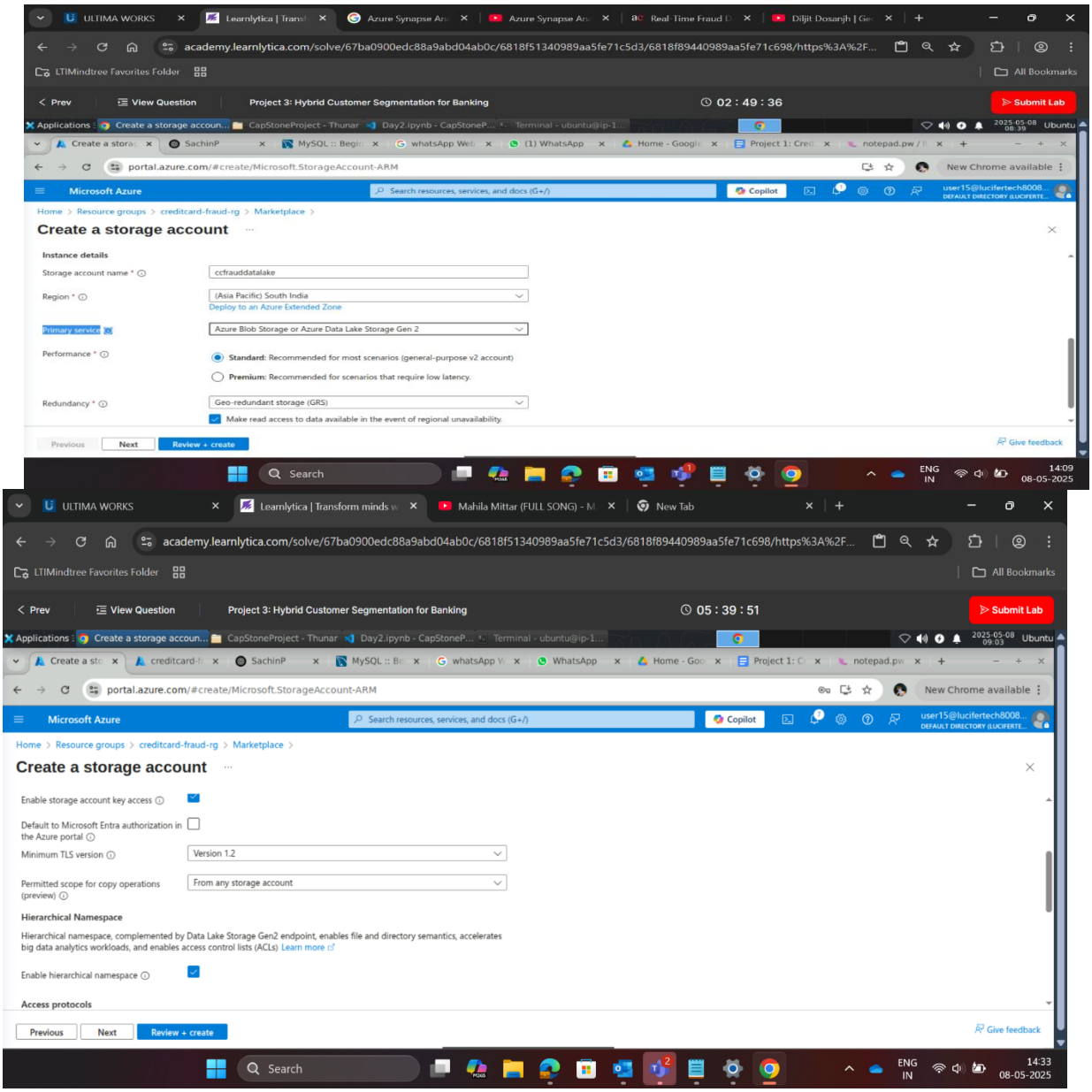
Column Name	Description
trans_num	Unique transaction ID.
trans_date_trans_time	Timestamp of the transaction.
cc_num	Credit card number used for the transaction.
merchant	Merchant involved in the transaction.
amt	Amount spent in the transaction.
unix_time	Unix timestamp for tracking the transaction.
is_fraud	Fraud label (1 = Fraudulent, 0 = Legitimate).

Step 1:

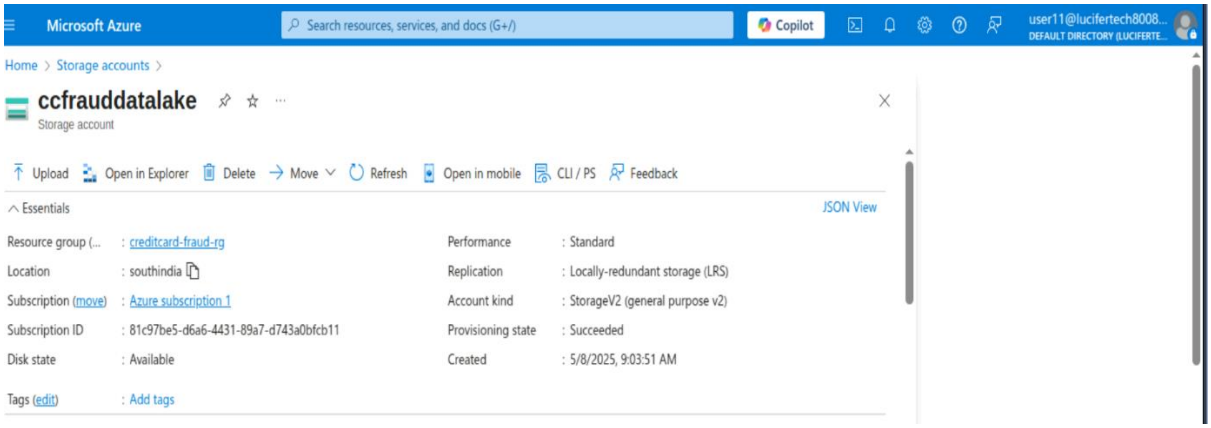


- As part of the Azure infrastructure setup for the credit card fraud detection project, an Azure Resource Group named 'Creditcard-fraud-rg' was created. This resource group acts as a logical container for all related Azure resources, enabling better organization, access control, and cost management. All core services including Azure Synapse , Azure Data Lake Storage Gen2 and Power BI integrations are deployed within this resource group. This structure ensures a centralized and efficient management of all resources associated with the project.

Step 2:

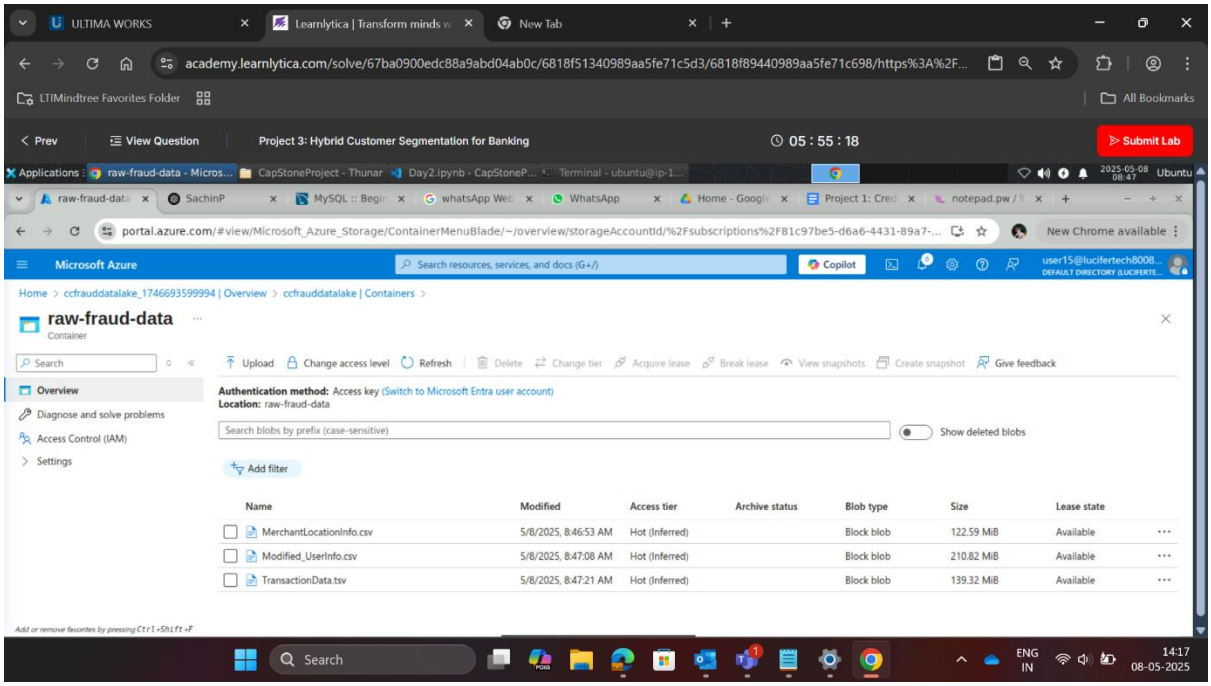


- enable hierarchical namespace, which is currently unchecked.



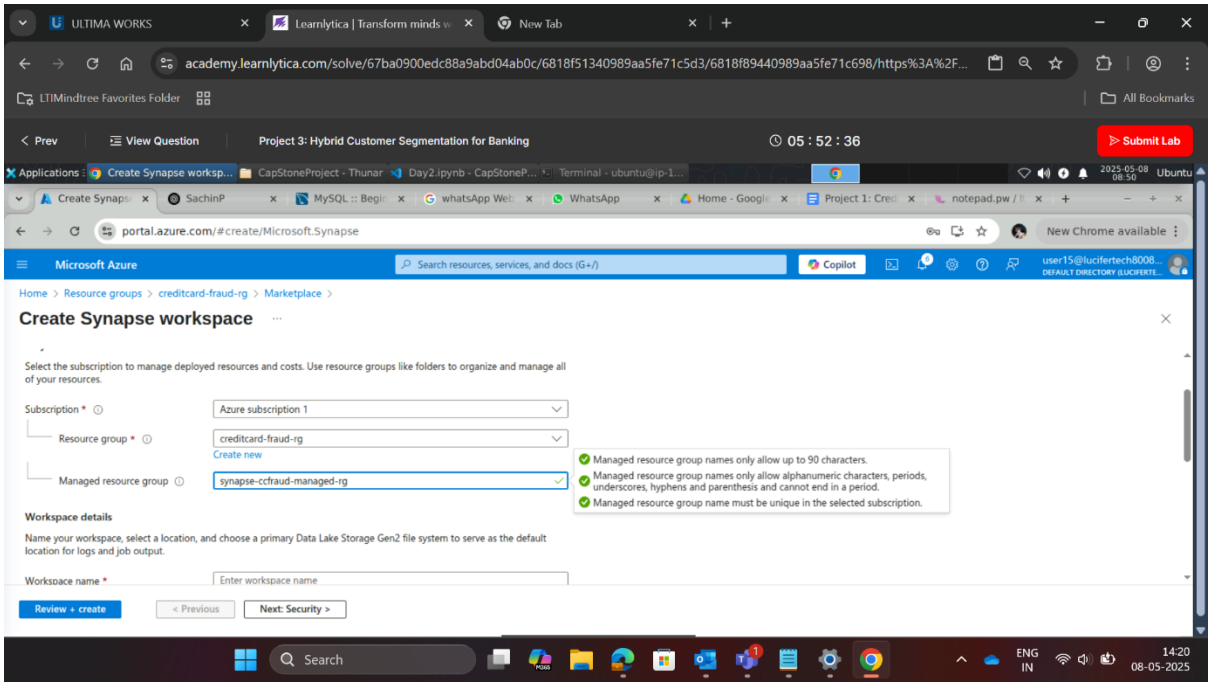
- After setting up the Azure Resource Group 'Creditcard-fraud-rg', an Azure Data Lake Storage account was created to serve as the centralized storage layer for the project's datasets. This storage account is responsible for hosting all raw input data and intermediate processed files.

Step 3:



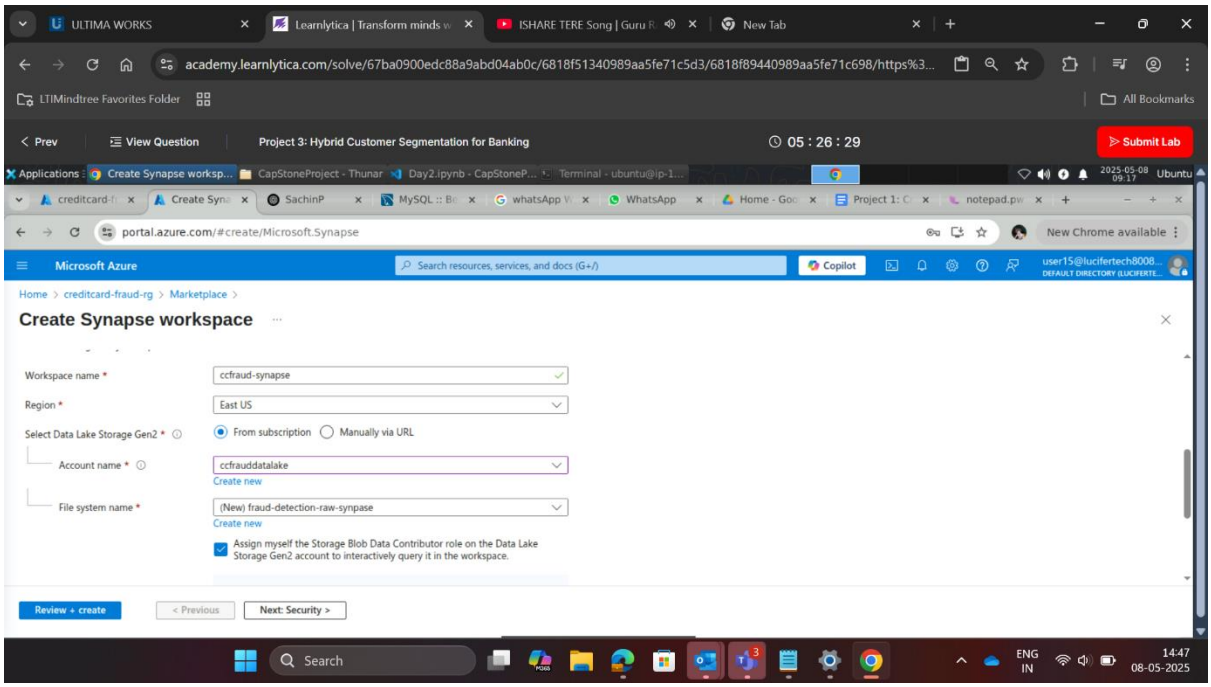
➢ displaying the contents of a container named "raw-fraud-data." The container includes three files: "MerchantLocationInfo.csv," "Modified_UserInfo.csv," and "TransactionData.csv." Each file has details such as the date modified, access tier, archive status, blob type, size, and lease state.

Step 4:

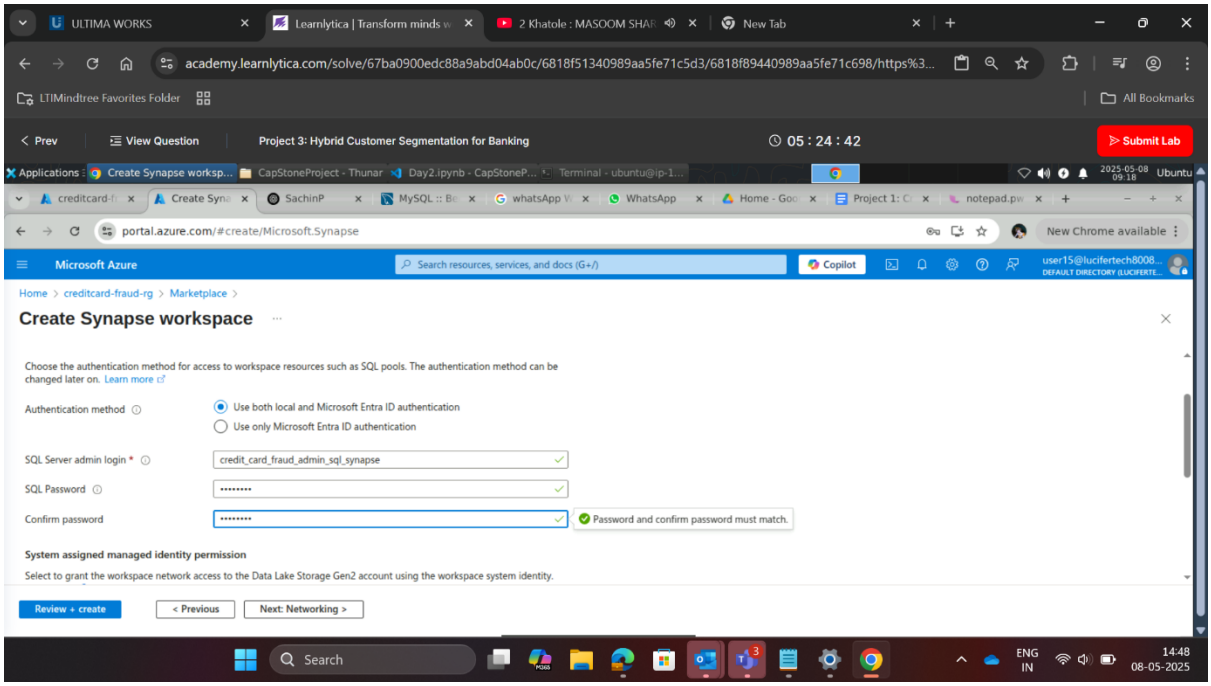


- The page includes options to select the subscription, resource group, and managed resource group for creating a Synapse workspace. The selected subscription is "Azure subscription 1," the resource group option is set to "Create new," and the managed resource group name entered is "synapse-cctfraud-managed-rg." There are also fields for entering a workspace name and selecting a primary Data Lake Storage Gen2 file system.

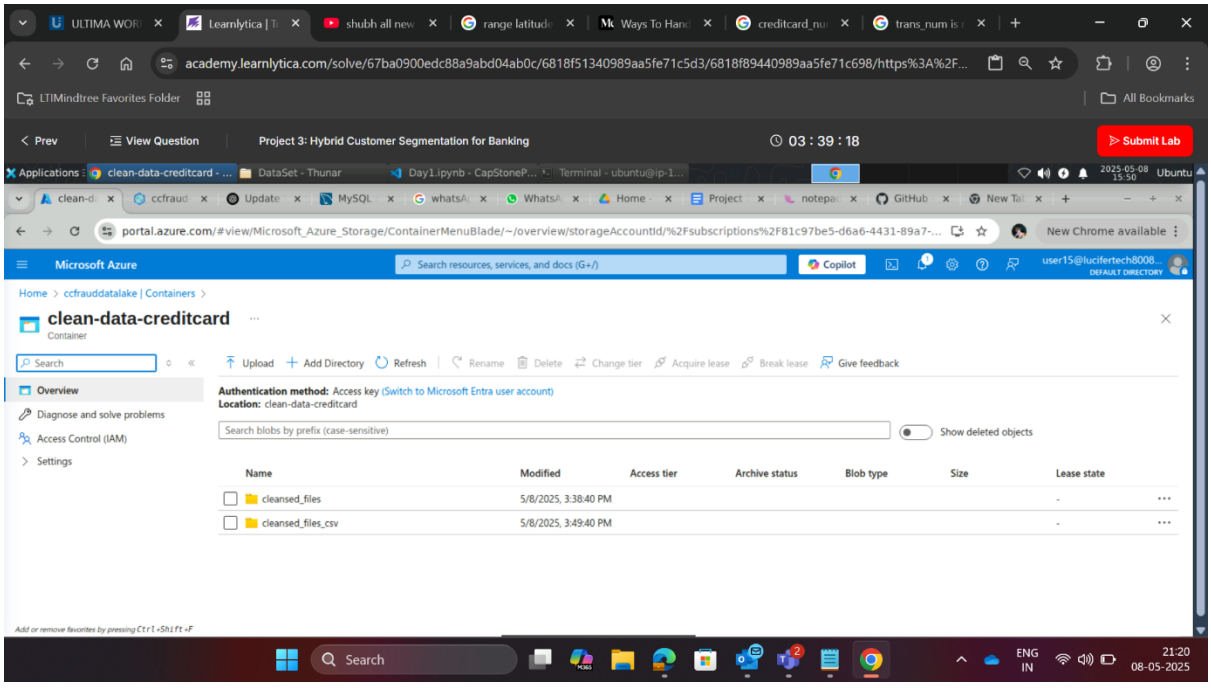
Step 5:



- **Workspace name:** "crtfraud-synapse" Region: "East US"
- **Select Data Lake Storage Gen2**: Options are "From subscription" and "Manually via URL"
- **Account name:** "crtfrauddatalake" with an option to create new File **system name:** New file system named "new-fraud-detection-raw-synapse

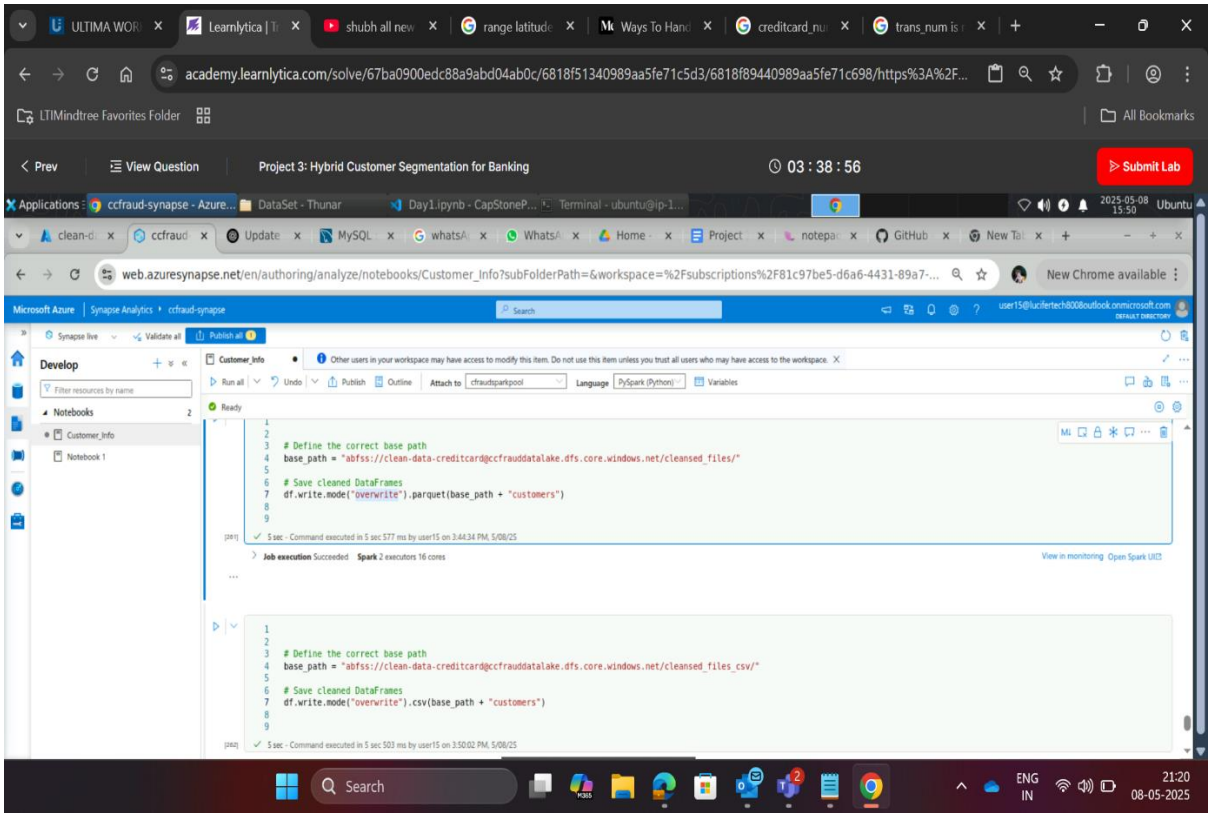


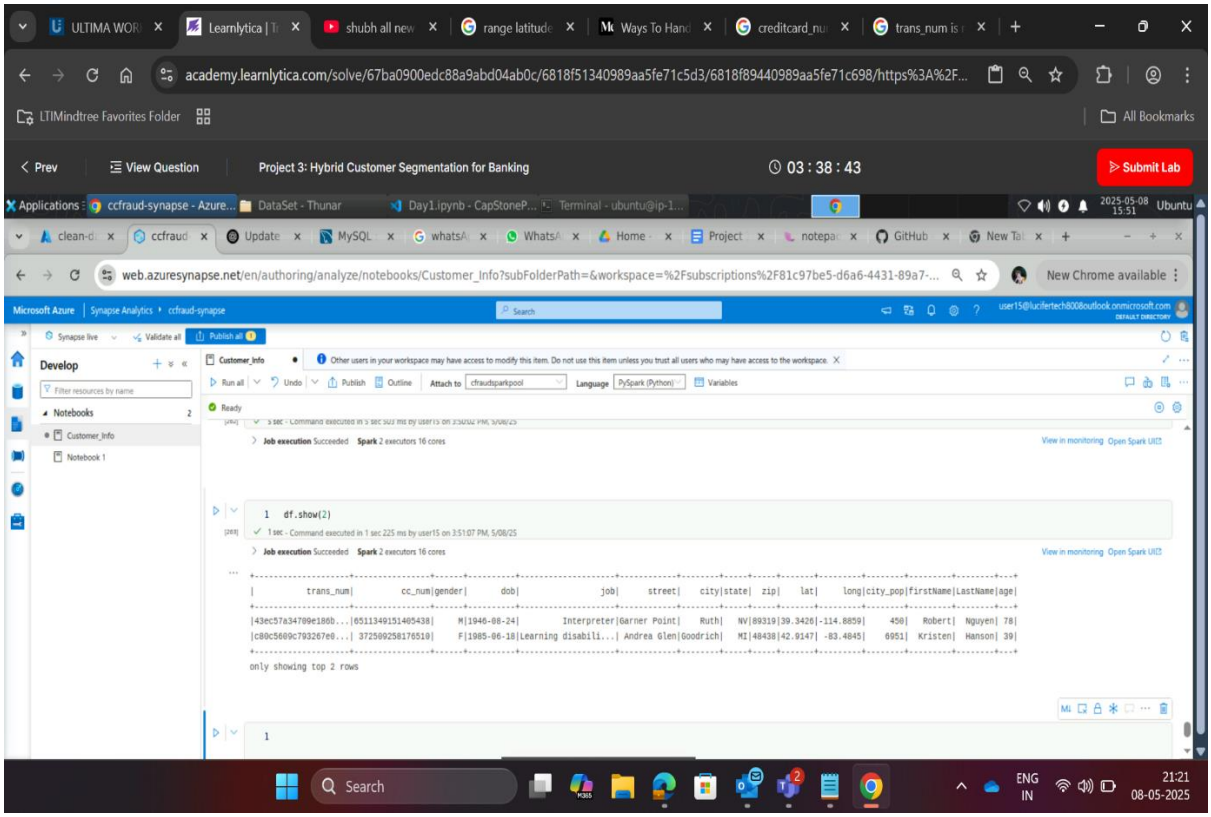
- It includes fields for setting up authentication and SQL server admin login details. The option for authentication method "Use both local and Microsoft Entra ID authentication" is enabled.The SQL Server admin login is set to "credit_card_fraud_admin_sql_synapse," with fields for entering and confirming the password.

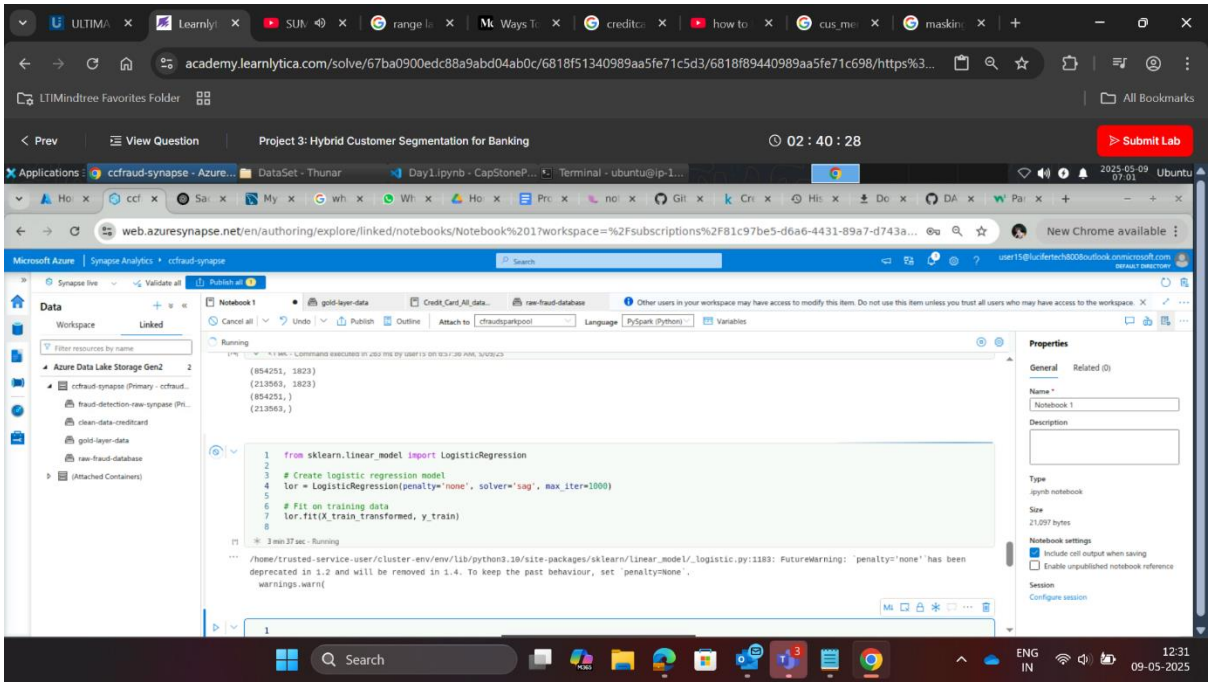


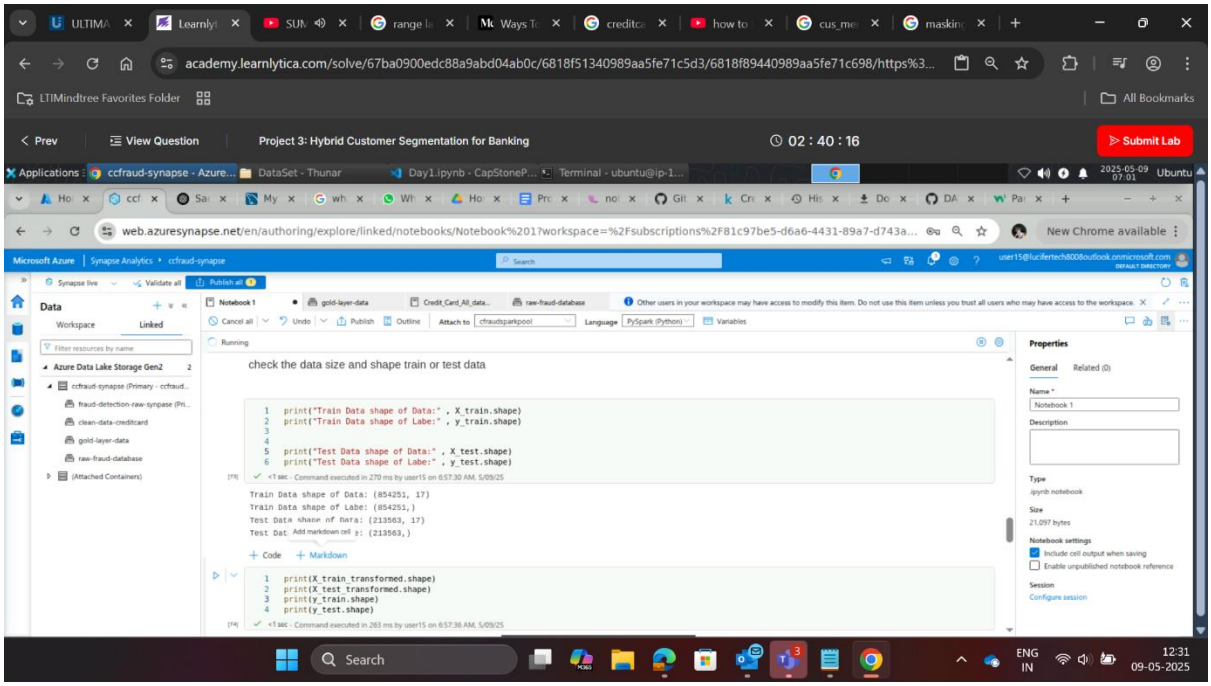
- A container named "clean-data-creditcard" is created inside the ADLS for EDA process.
- The container has two files listed: "cleansed_files" and "cleansed_files.csv".

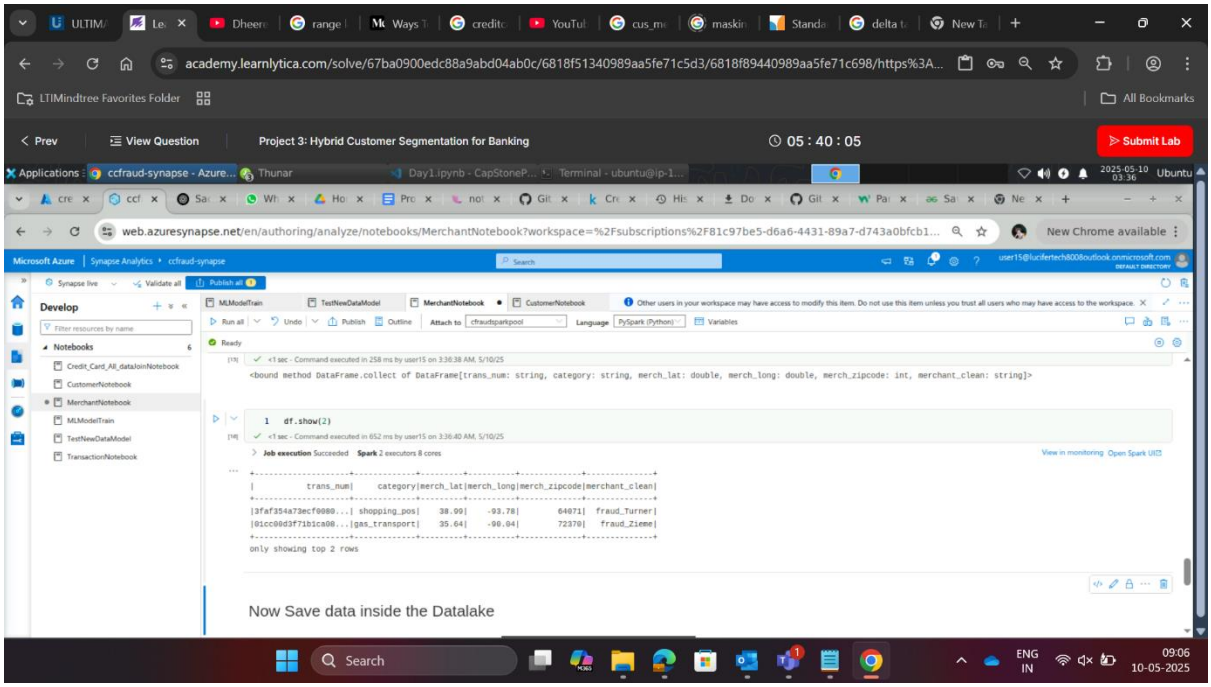
Step 6:

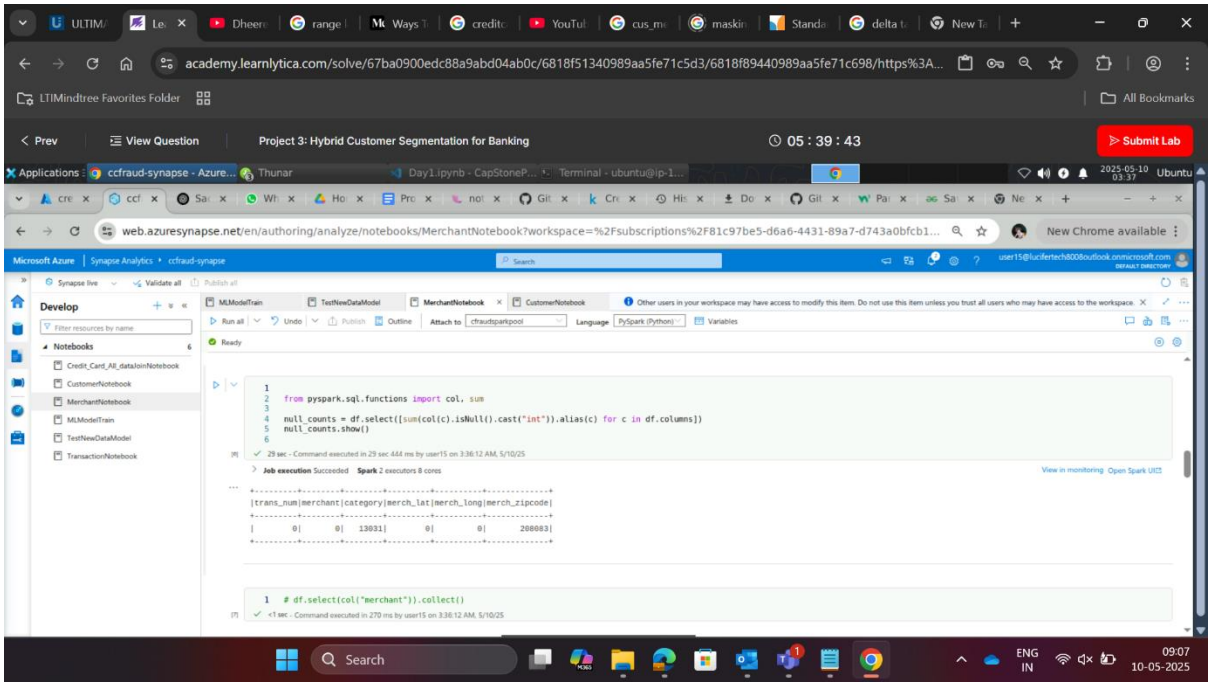


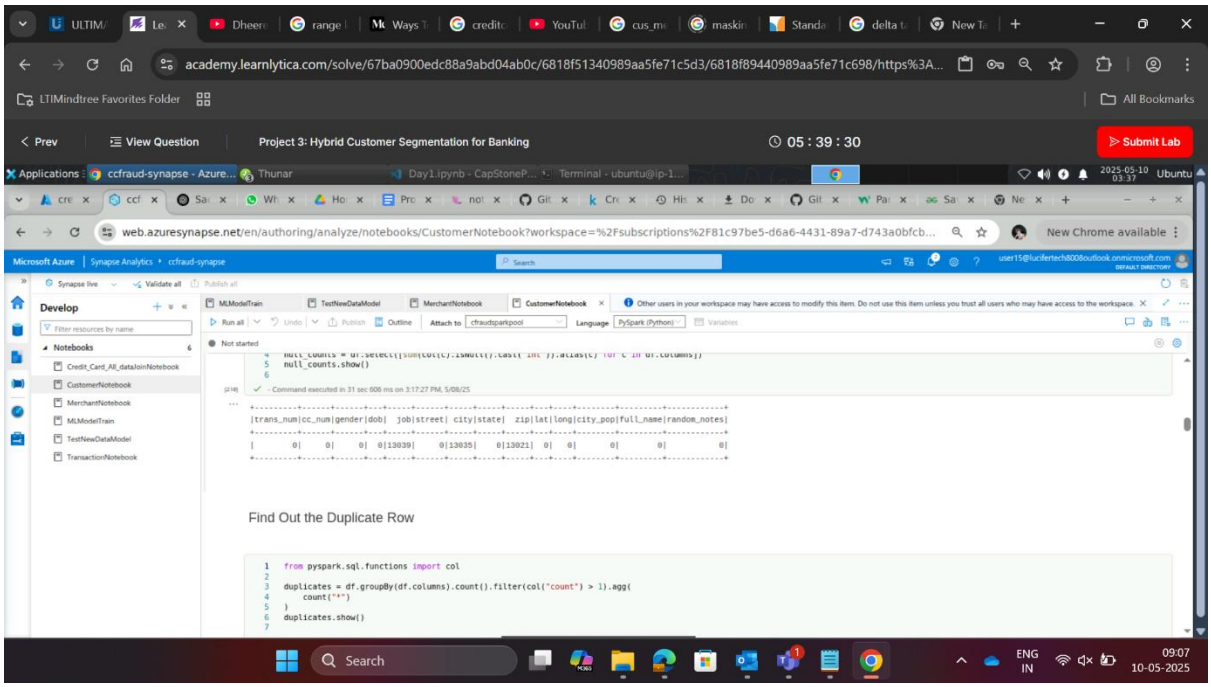

academy.learnlytica.com/solve/67ba0900edc88a9abd04ab0c/6818f51340989aa5fe71c5d3/6818f89440989aa5fe71c698/https%3A%2F...
Prev
View Question
Project 3: Hybrid Customer Segmentation for Banking
03 : 38 : 43
Submit Lab
web.azuresynapse.net/en/authoring/analyze/notebooks/Customer_Info?subFolderPath=&workspace=%2Fsubscriptions%2F81c97be5-d6a6-4431-89a7-...
Develop
Notebooks
Customer_Info
Notebook 1
df.show(2)
only showing top 2 rows
Search
ENG
IN
21:21
08-05-2025

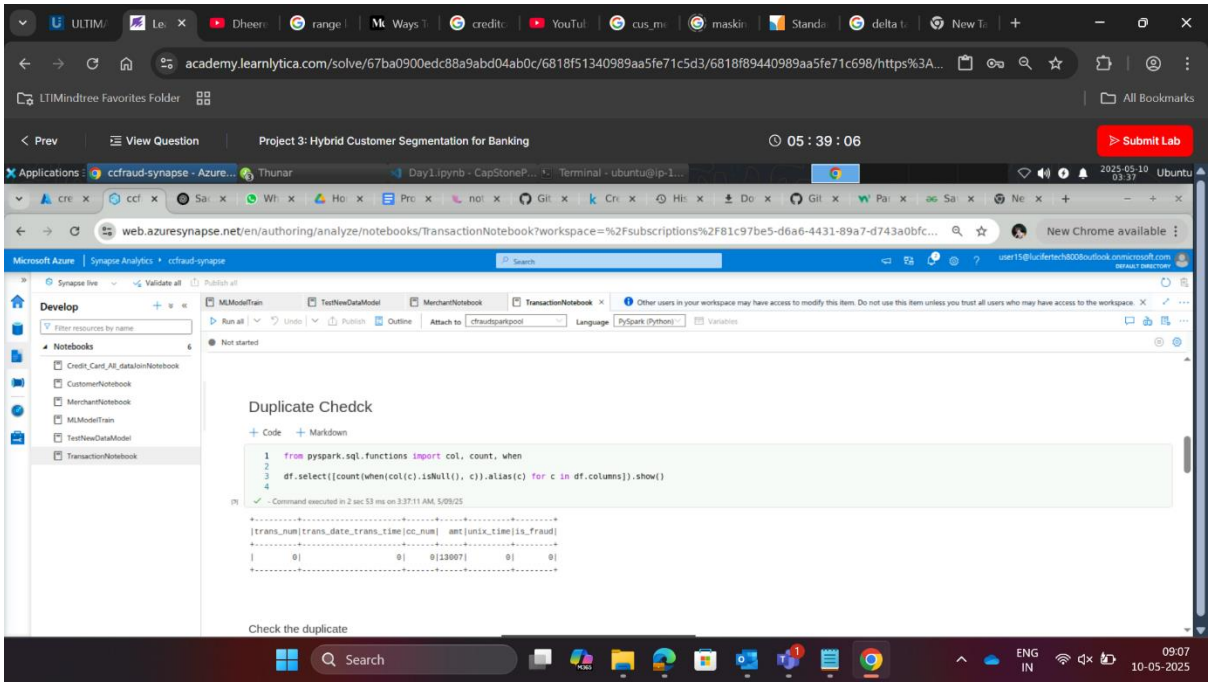

academy.learnlytica.com/solve/67ba0900edc88a9abd04ab0c/6818f51340989aa5fe71c5d3/6818f89440989aa5fe71c698/https%3...
Prev
View Question
Project 3: Hybrid Customer Segmentation for Banking
02 : 40 : 28
Submit Lab
web.azuresynapse.net/en/authoring/explore/linked/notebooks/Notebook%201?workspace=%2Fsubscriptions%2F81c97be5-d6a6-4431-89a7-d743a...
Data
Workspace
Linked
Azure Data Lake Storage Gen2
Notebook 1
Properties
from sklearn.linear_model import LogisticRegression
# Create logistic regression model
lor = LogisticRegression(penalty='none', solver='sag', max_iter=1000)
# Fit on training data
lor.fit(X_train_transformed, y_train)
/home/trusted-service-user/cluster-env/env/lib/python3.10/site-packages/sklearn/linear_model/_logistic.py:1183: FutureWarning: `penalty='none'` has been deprecated in 1.2 and will be removed in 1.4. To keep the past behaviour, set `penalty=None`.
warnings.warn(
Search
ENG
IN
12:31
09-05-2025

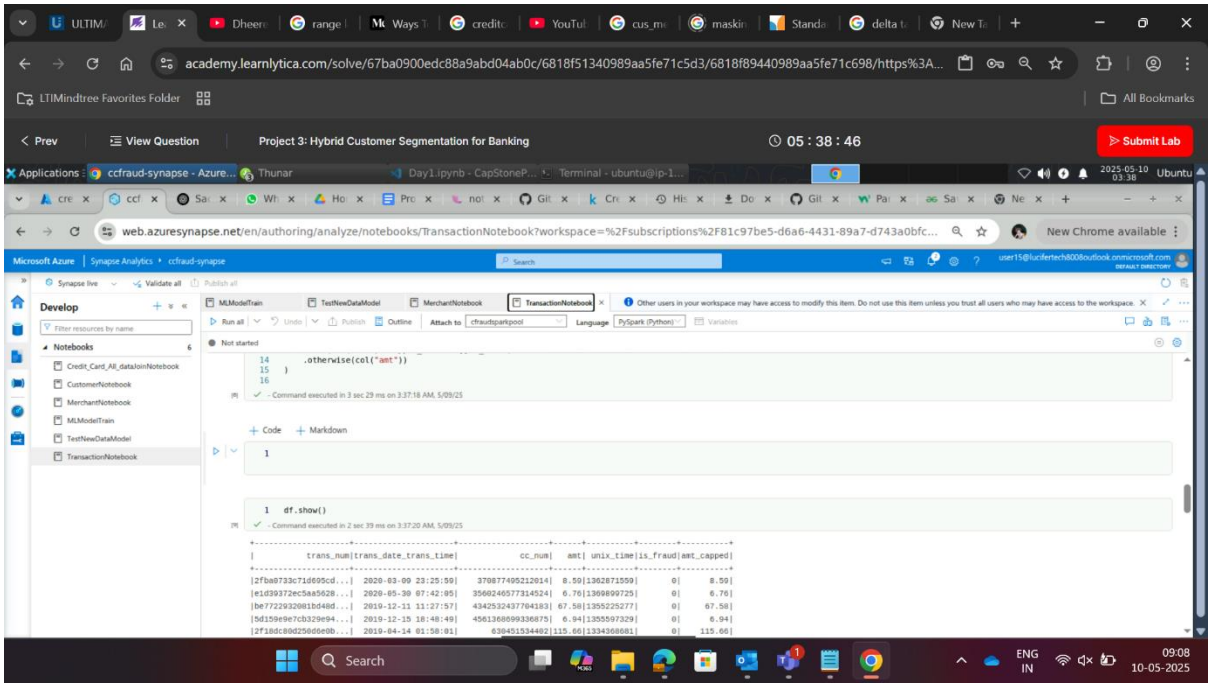

academy.learnlytica.com/solve/67ba0900edc88a9abd04ab0c/6818f51340989aa5fe71c5d3/6818f89440989aa5fe71c698/https%3...
Prev
View Question
Project 3: Hybrid Customer Segmentation for Banking
02 : 40 : 16
Submit Lab
web.azuresynapse.net/en/authoring/explore/linked/notebooks/Notebook%201?workspace=%2Fsubscriptions%2F81c97be5-d6a6-4431-89a7-d743a...
check the data size and shape train or test data
print("Train Data shape of Data:" , X_train.shape)
print("Train Data shape of Labe:" , y_train.shape)
print("Test Data shape of Data:" , X_test.shape)
print("Test Data shape of Labe:" , y_test.shape)
Train Data shape of Data: (854251, 17)
Train Data shape of Labe: (854251,)
Test Data shape of Data: (213563, 17)
Test Data shape of Labe: (213563,)
print(X_train_transformed.shape)
print(X_test_transformed.shape)
print(y_train.shape)
print(y_test.shape)

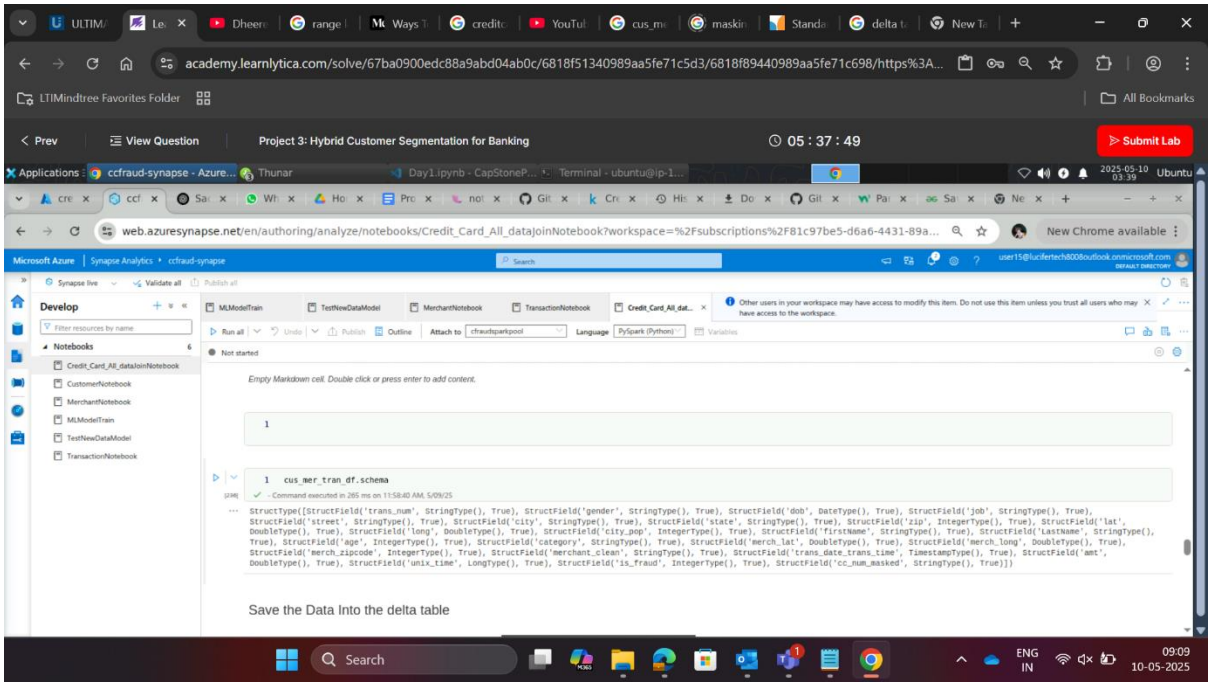

academy.learnlytica.com/solve/67ba0900edc88a9abd04ab0c/6818f51340989aa5fe71c5d3/6818f89440989aa5fe71c698/https%3A...
Prev
View Question
Project 3: Hybrid Customer Segmentation for Banking
05 : 40 : 05
Submit Lab
web.azuresynapse.net/en/authoring/analyze/notebooks/MerchantNotebook?workspace=%2Fsubscriptions%2F81c97be5-d6a6-4431-89a7-d743a0bfcb1...
<bound method DataFrame.collect of DataFrame[trans_num: string, category: string, merch_lat: double, merch_long: double, merch_zipcode: int, merchant_clean: string]>
df.show(2)
trans_num category merch_lat merch_long merch_zipcode merchant_clean
3faf354a73ecf0080... shopping_pos 38.99 -93.78 64071 fraud_Turner
01cc00d3f71b1ca08... gas_transport 35.64 -90.04 72370 fraud_Zieme
only showing top 2 rows
Now Save data inside the Datalake


academy.learnlytica.com/solve/67ba0900edc88a9abd04ab0c/6818f51340989aa5fe71c5d3/6818f89440989aa5fe71c698/https%3A...
LTIMindtree Favorites Folder
All Bookmarks
Prev
View Question
Project 3: Hybrid Customer Segmentation for Banking
05 : 39 : 43
Submit Lab
web.azuresynapse.net/en/authoring/analyze/notebooks/MerchantNotebook?workspace=%2Fsubscriptions%2F81c97be5-d6a6-4431-89a7-d743a0bfcb1...
New Chrome available
Develop
Notebooks
Credit_Card_All_dataJoinNotebook
CustomerNotebook
MerchantNotebook
MLModelTrain
TestNewDataModel
TransactionNotebook
from pyspark.sql.functions import col, sum
null_counts = df.select([sum(col(c).isNull().cast("int")).alias(c) for c in df.columns])
null_counts.show()
29 sec - Command executed in 29 sec 444 ms by user15 on 3:36:12 AM, 5/10/25
Job execution Succeeded Spark 2 executors 8 cores
|trans_num|merchant|category|merch_lat|merch_long|merch_zipcode|
| 0| 0| 13031| 0| 0| 208083|
df.select(col("merchant")).collect()
<1 sec - Command executed in 270 ms by user15 on 3:36:12 AM, 5/10/25
Search
ENG
IN
09:07
10-05-2025

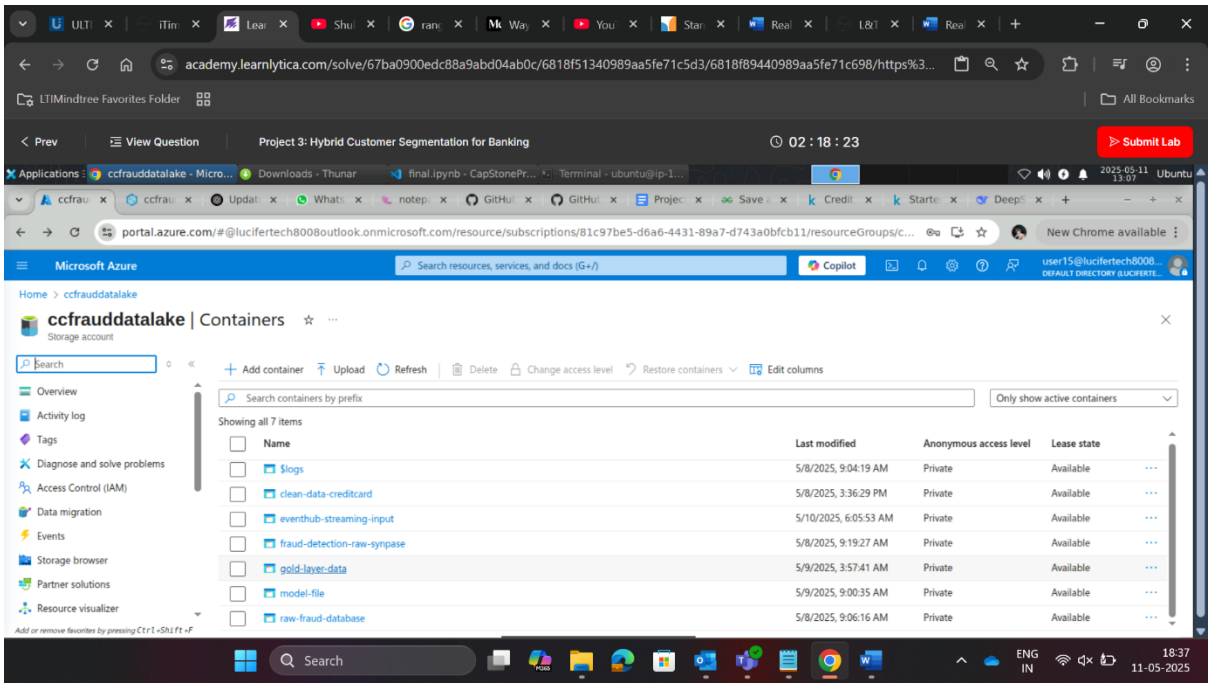

academy.learnlytica.com/solve/67ba0900edc88a9abd04ab0c/6818f51340989aa5fe71c5d3/6818f89440989aa5fe71c698/https%3A...
LTIMindtree Favorites Folder
All Bookmarks
Prev
View Question
Project 3: Hybrid Customer Segmentation for Banking
05 : 39 : 30
Submit Lab
web.azuresynapse.net/en/authoring/analyze/notebooks/CustomerNotebook?workspace=%2Fsubscriptions%2F81c97be5-d6a6-4431-89a7-d743a0bfcb...
New Chrome available
Develop
Notebooks
Credit_Card_All_dataJoinNotebook
CustomerNotebook
MerchantNotebook
MLModelTrain
TestNewDataModel
TransactionNotebook
Not started
- Command executed in 31 sec 006 ms on 3:17:27 PM, 5/08/25
|trans_num|cc_num|gender|dob| job|street| city|state| zip|lat|long|city_pop|full_name|random_notes|
Find Out the Duplicate Row
from pyspark.sql.functions import col
duplicates = df.groupBy(df.columns).count().filter(col("count") > 1).agg(
count("*")
)
duplicates.show()
Search
ENG
IN
09:07
10-05-2025

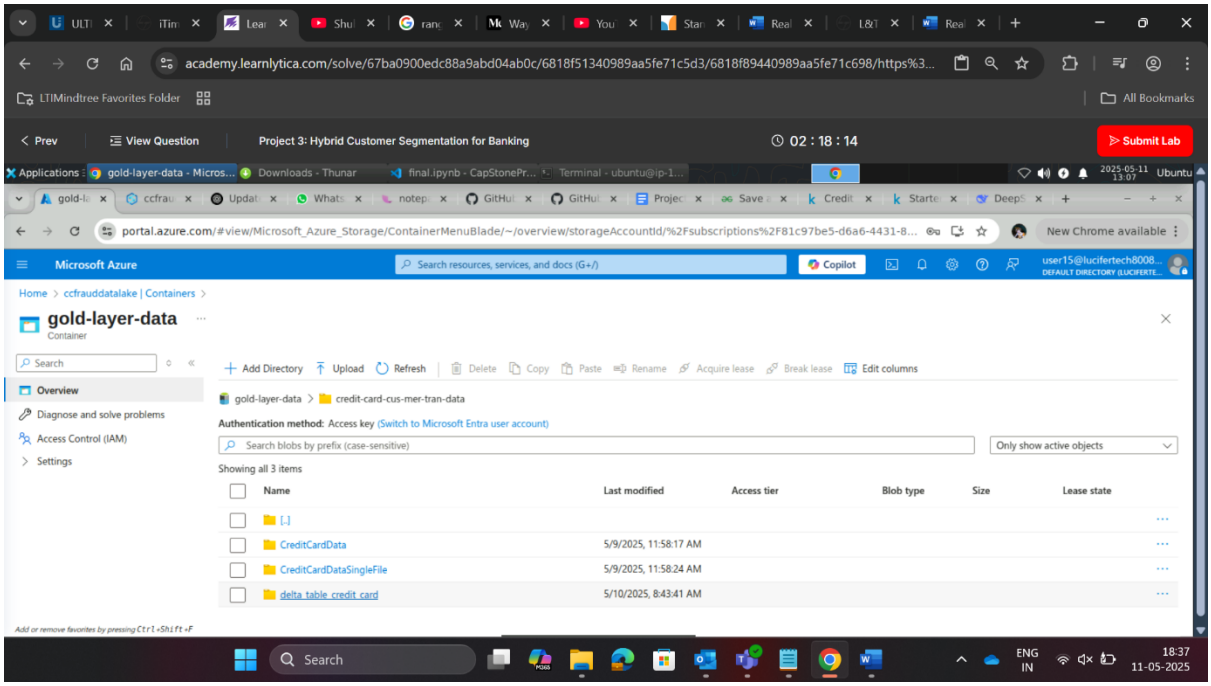

academy.learnlytica.com/solve/67ba0900edc88a9abd04ab0c/6818f51340989aa5fe71c5d3/6818f89440989aa5fe71c698/https%3A...
Prev
View Question
Project 3: Hybrid Customer Segmentation for Banking
05 : 39 : 06
Submit Lab
web.azuresynapse.net/en/authoring/analyze/notebooks/TransactionNotebook?workspace=%2Fsubscriptions%2F81c97be5-d6a6-4431-89a7-d743a0bfc...
Duplicate Chedck
Check the duplicate
09:07
10-05-2025


academy.learnlytica.com/solve/67ba0900edc88a9abd04ab0c/6818f51340989aa5fe71c5d3/6818f89440989aa5fe71c698/https%3A...
Prev
View Question
Project 3: Hybrid Customer Segmentation for Banking
05 : 38 : 46
Submit Lab
web.azuresynapse.net/en/authoring/analyze/notebooks/TransactionNotebook?workspace=%2Fsubscriptions%2F81c97be5-d6a6-4431-89a7-d743a0bfc...
09:08
10-05-2025



Creating another container

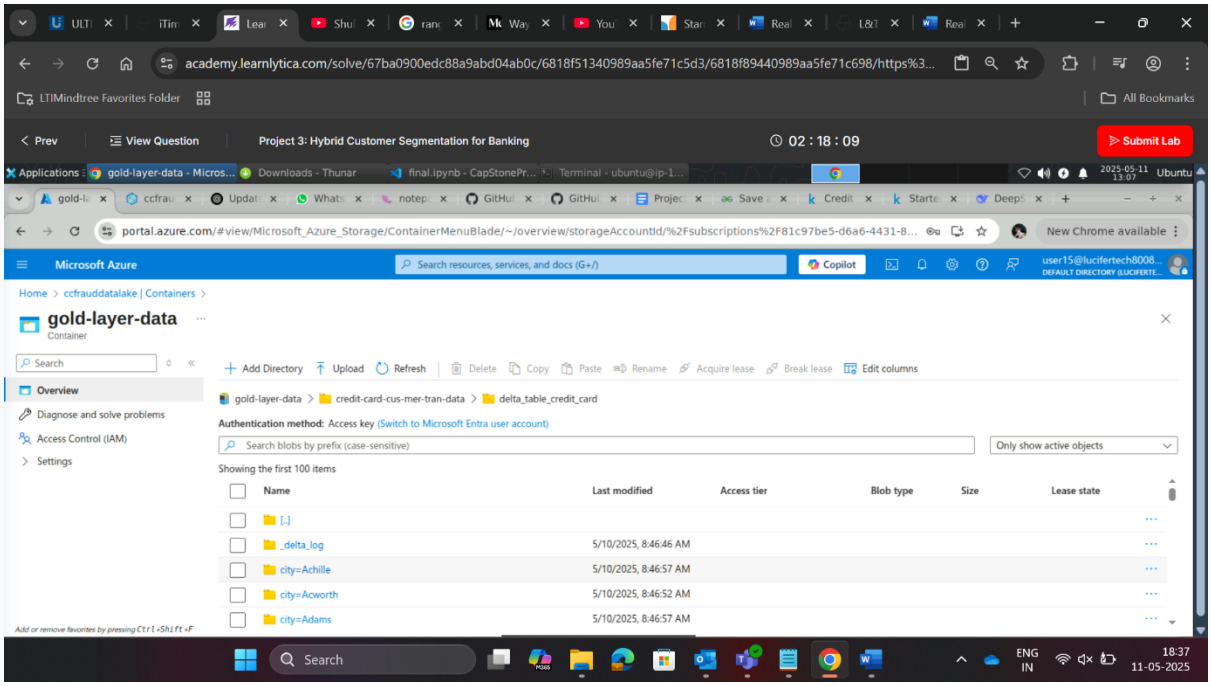




- The "gold-layer-data" container within the adls contains crucial datasets for the credit card fraud detection project.

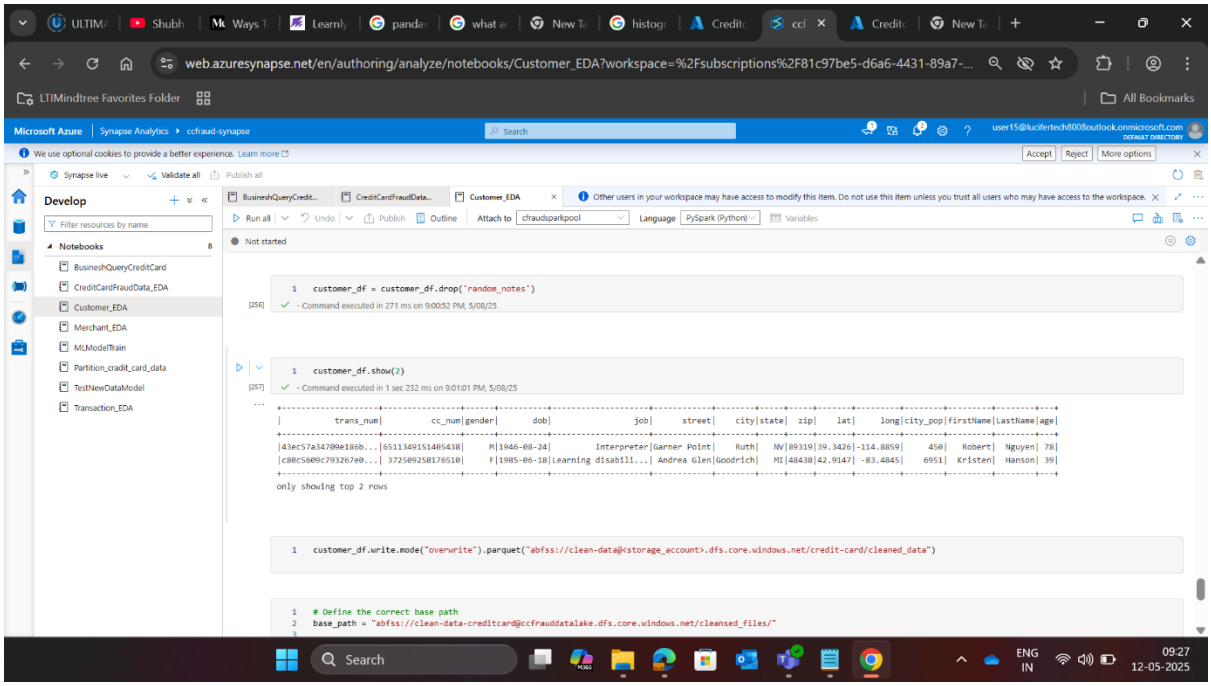
- It has three folders

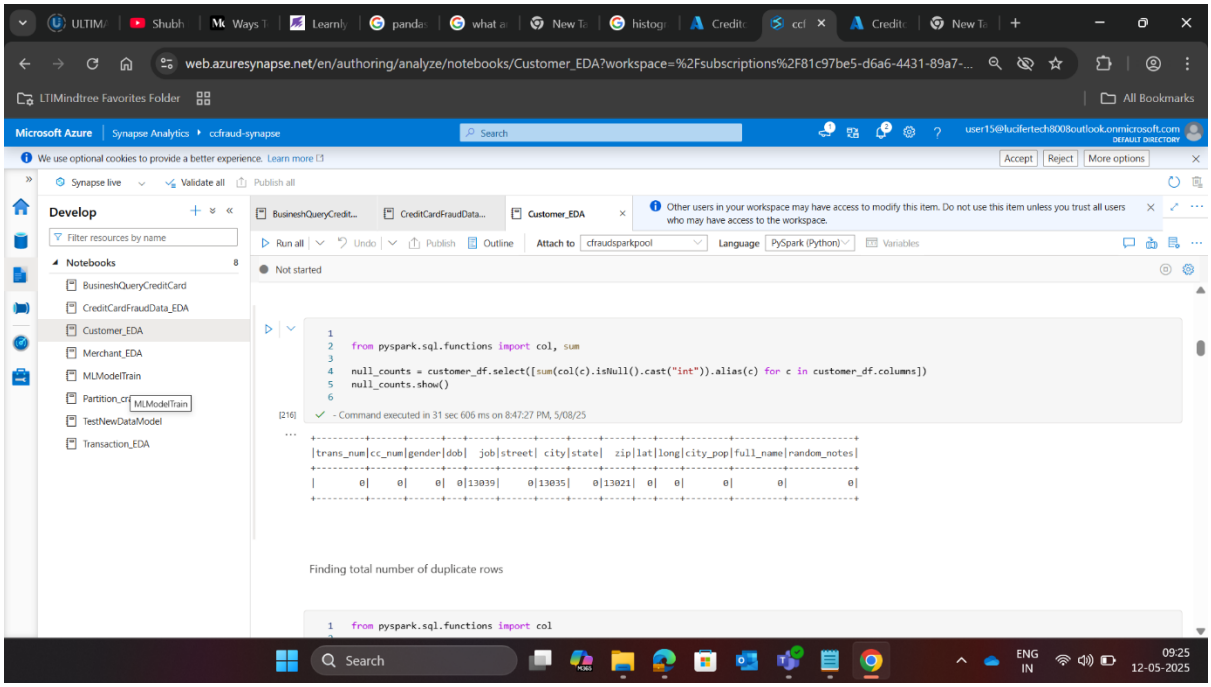
- CreditCardData
- CreditCardDataSingleFile
- delta_table_credit_card

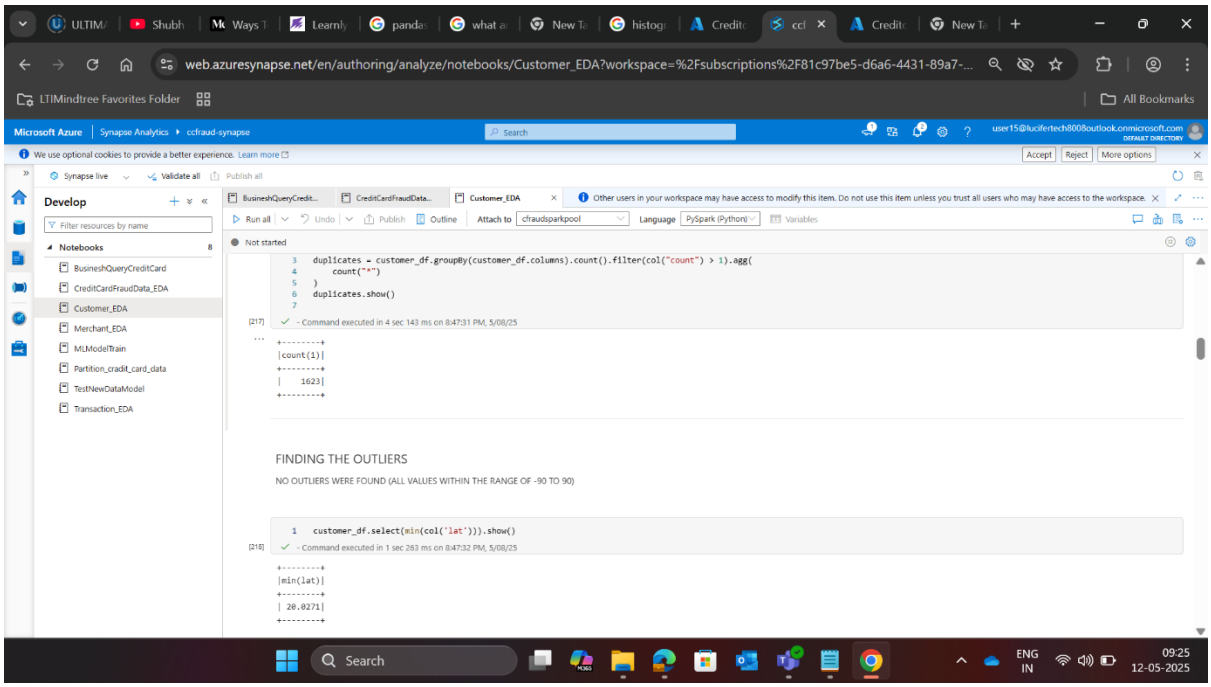


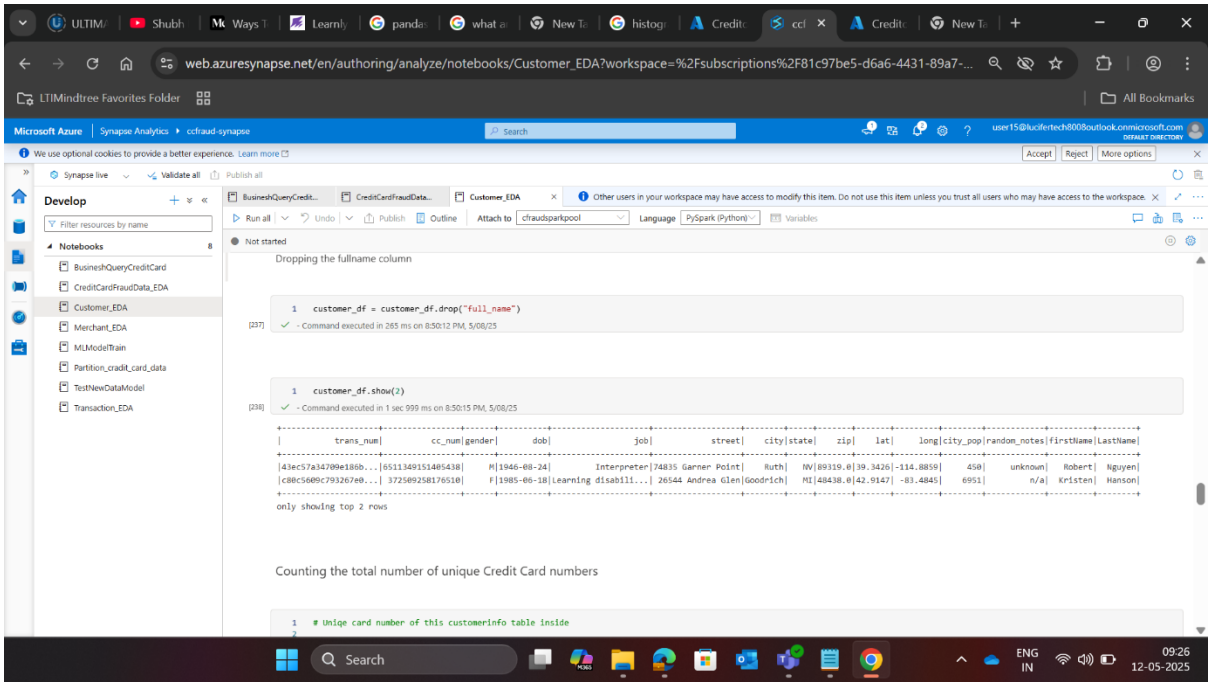
Code:

1.Customer EDA

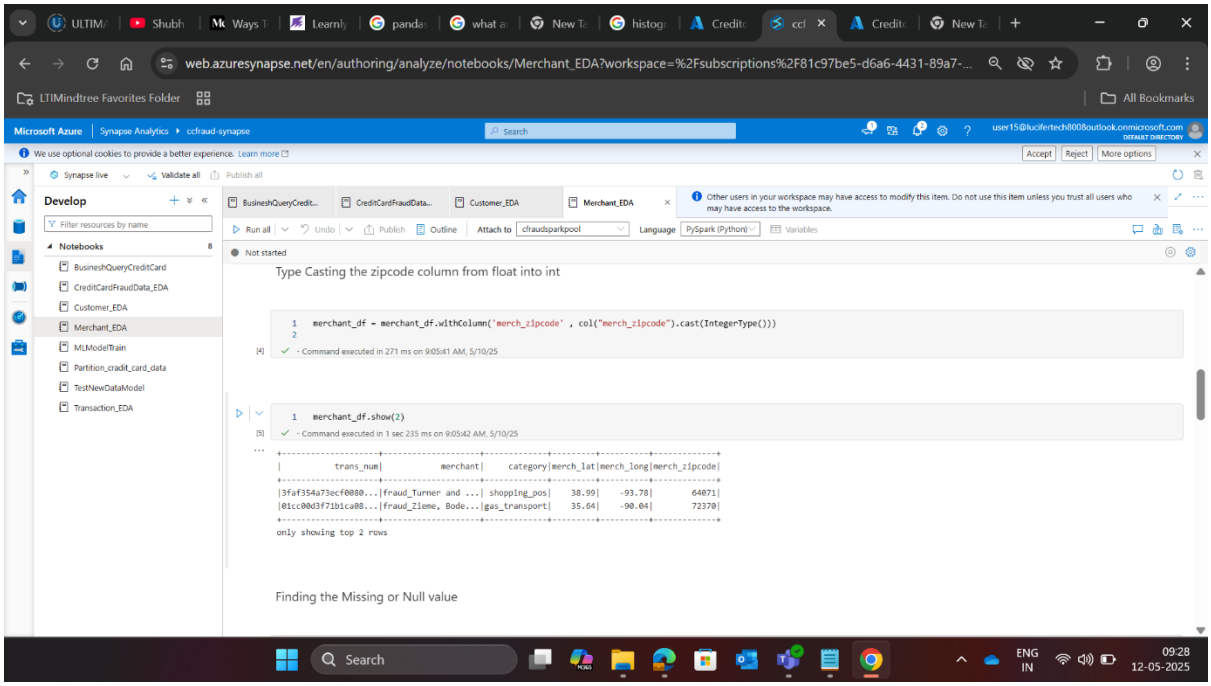


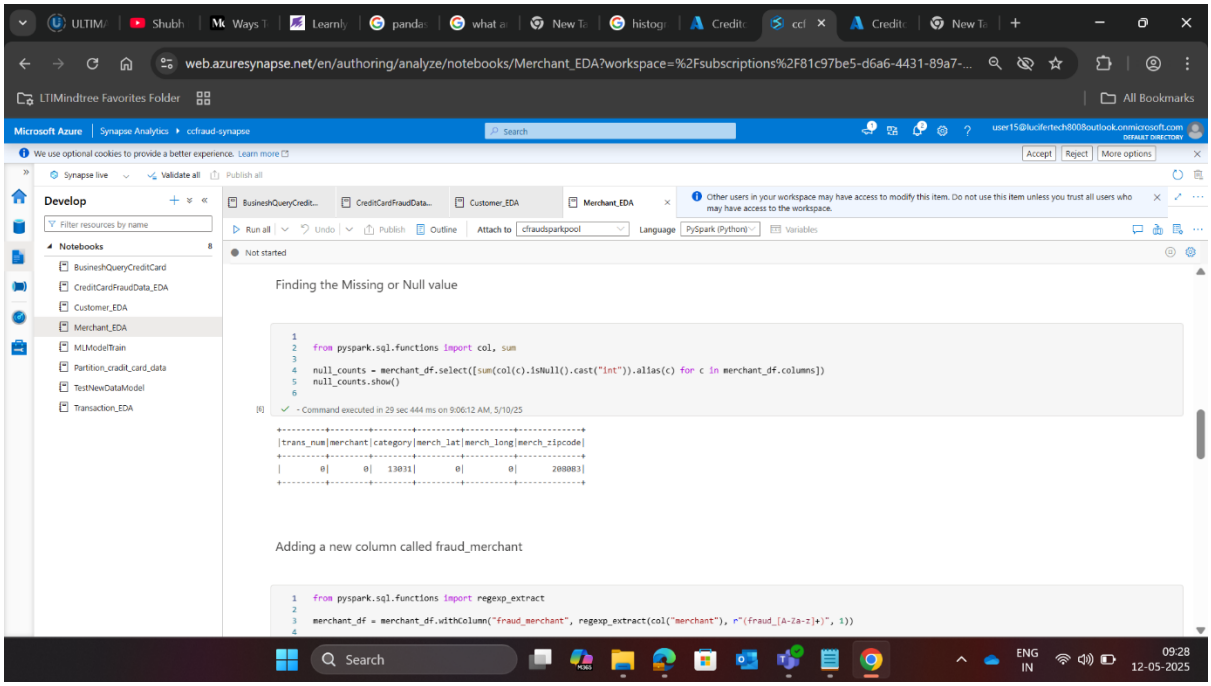

web.azuresynapse.net/en/authoring/analyze/notebooks/Customer_EDA?workspace=%2Fsubscriptions%2F81c97be5-d6a6-4431-89a7-...
Microsoft Azure | Synapse Analytics | ccfraud-synapse
Develop
Notebooks
BusineshQueryCreditCard
CreditCardFraudData_EDA
Customer_EDA
Merchant_EDA
MLModelTrain
Partition_cr
TestNewDataModel
Transaction_EDA
Run all | Undo | Publish | Outline | Attach to cfraudsparkpool | Language PySpark (Python) | Variables
Not started
from pyspark.sql.functions import col, sum
null_counts = customer_df.select([sum(col(c).isNull().cast("int")).alias(c) for c in customer_df.columns])
null_counts.show()
[216] - Command executed in 31 sec 606 ms on 8:47:27 PM, 5/08/25
|trans_num|cc_num|gender|dob| job|street| city|state| zip|lat|long|city_pop|full_name|random_notes|
| 0| 0| 0| 0|13039| 0|13035| 0|13021| 0| 0| 0| 0| 0|
Finding total number of duplicate rows
from pyspark.sql.functions import col
09:25
12-05-2025

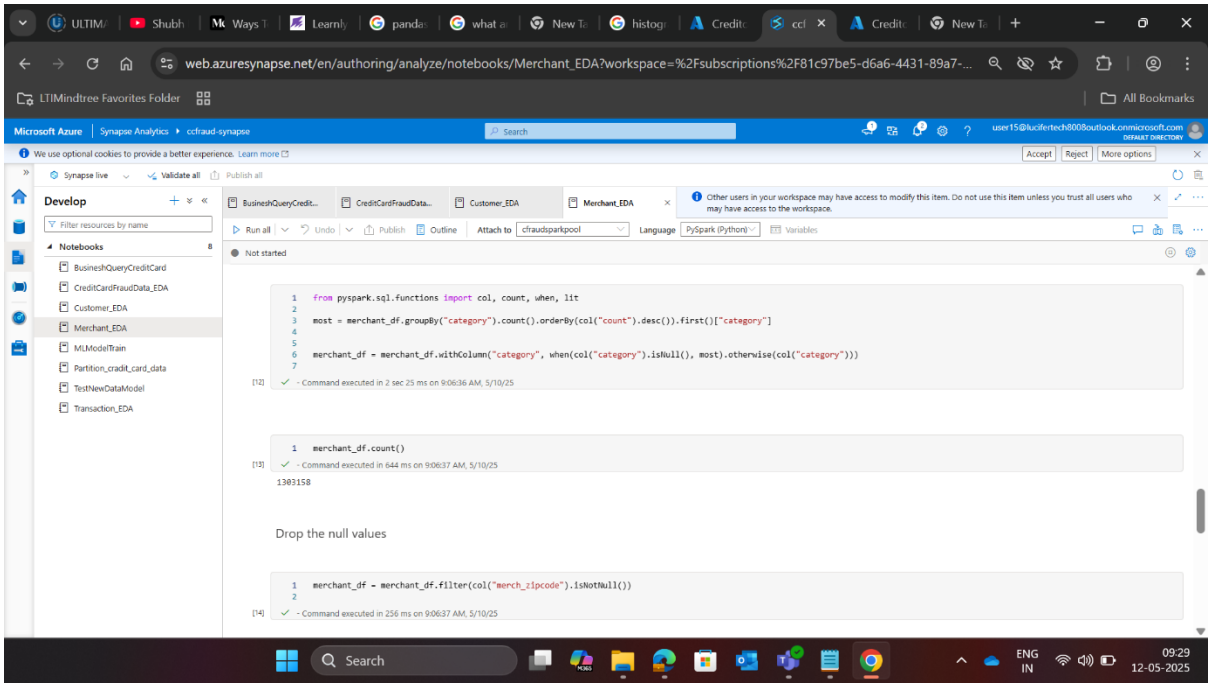

web.azuresynapse.net/en/authoring/analyze/notebooks/Customer_EDA?workspace=%2Fsubscriptions%2F81c97be5-d6a6-4431-89a7-...
Microsoft Azure | Synapse Analytics | ccfraud-synapse
Develop
Notebooks
BusineshQueryCreditCard
CreditCardFraudData_EDA
Customer_EDA
Merchant_EDA
MLModelTrain
Partition_cradit_card_data
TestNewDataModel
Transaction_EDA
Run all | Undo | Publish | Outline | Attach to cfraudsparkpool | Language PySpark (Python) | Variables
Not started
duplicates = customer_df.groupBy(customer_df.columns).count().filter(col("count") > 1).agg(
count("*")
)
duplicates.show()
[217] - Command executed in 4 sec 143 ms on 8:47:31 PM, 5/08/25
|count(1)|
| 1623|
FINDING THE OUTLIERS
NO OUTLIERS WERE FOUND (ALL VALUES WITHIN THE RANGE OF -90 TO 90)
customer_df.select(min(col('lat'))).show()
[218] - Command executed in 1 sec 263 ms on 8:47:32 PM, 5/08/25
|min(lat)|
| 20.0271|
09:25
12-05-2025

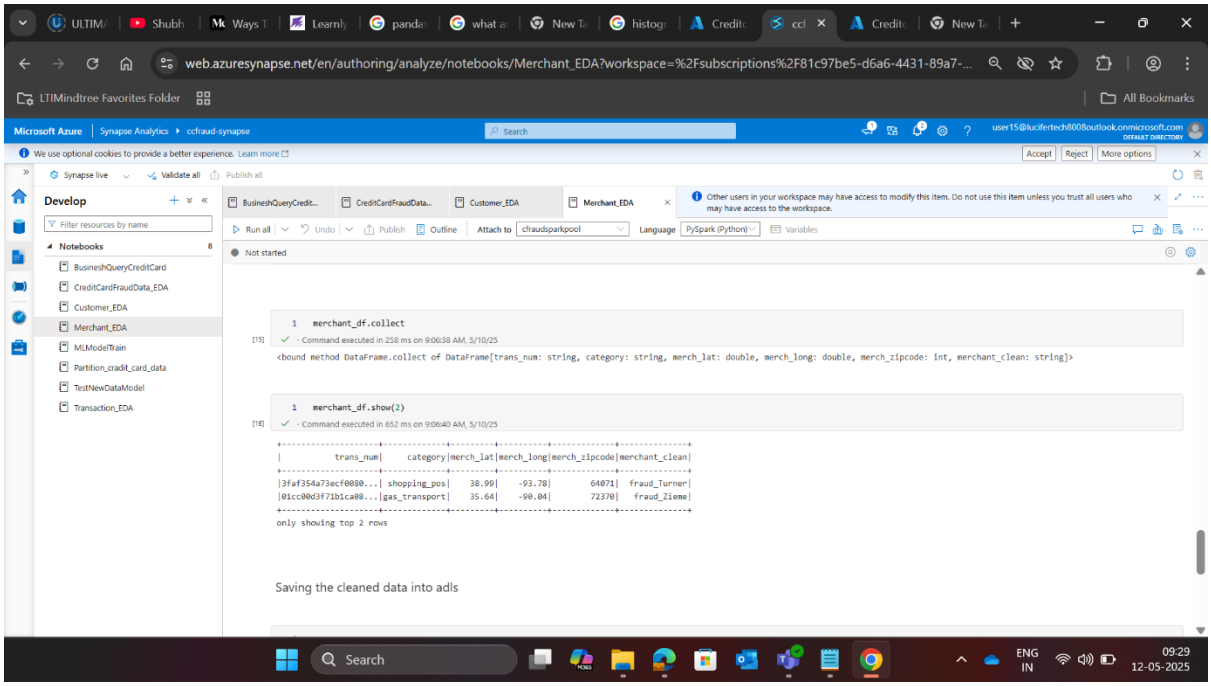


2.Merchant EDA

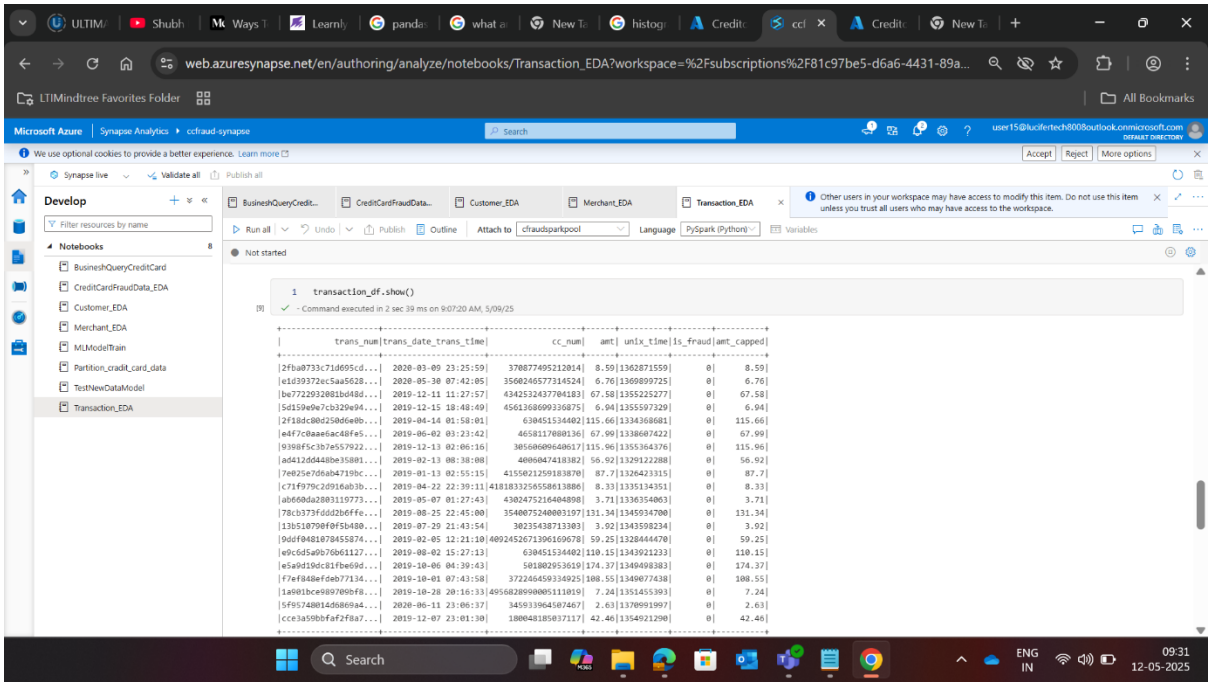


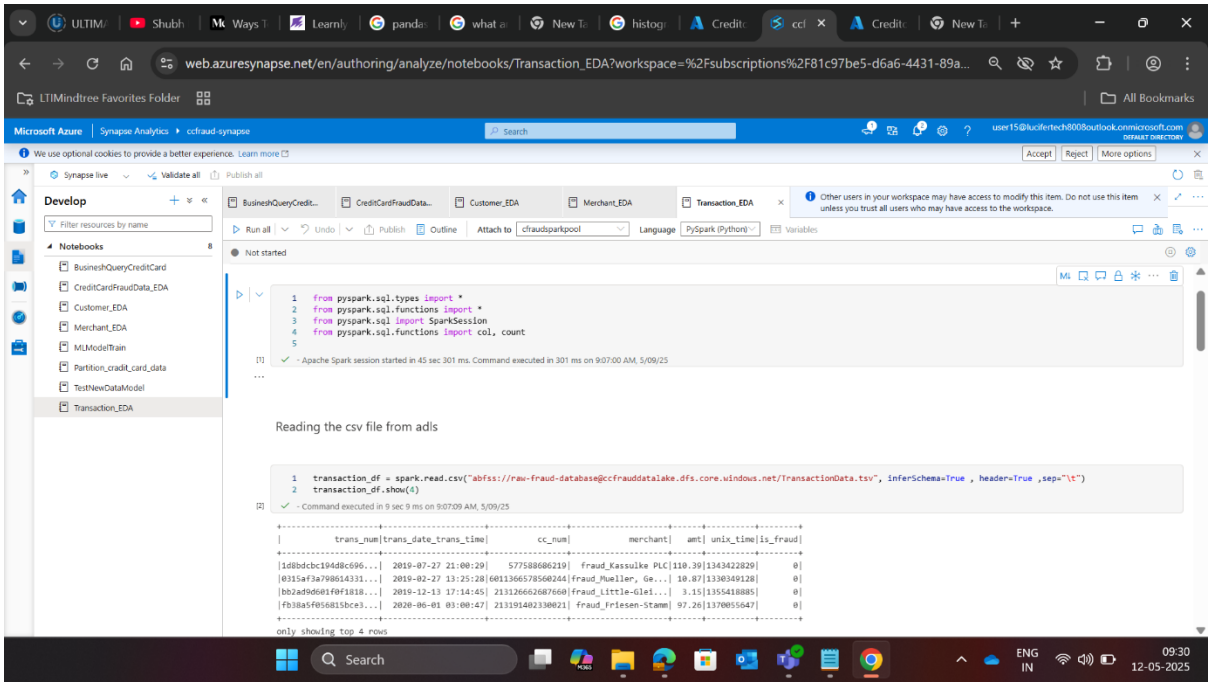

web.azuresynapse.net/en/authoring/analyze/notebooks/Merchant_EDA?workspace=%2Fsubscriptions%2F81c97be5-d6a6-4431-89a7-...
Microsoft Azure | Synapse Analytics | ccfraud-synapse
Develop
Notebooks
BusineshQueryCreditCard
CreditCardFraudData_EDA
Customer_EDA
Merchant_EDA
MLModelTrain
Partition_cradit_card_data
TestNewDataModel
Transaction_EDA
Finding the Missing or Null value
from pyspark.sql.functions import col, sum
null_counts = merchant_df.select([sum(col(c).isNull().cast("int")).alias(c) for c in merchant_df.columns])
null_counts.show()
- Command executed in 29 sec 444 ms on 9:06:12 AM, 5/10/25
|trans_num|merchant|category|merch_lat|merch_long|merch_zipcode|
| 0| 0| 13031| 0| 0| 208083|
Adding a new column called fraud_merchant
from pyspark.sql.functions import regexp_extract
merchant_df = merchant_df.withColumn("fraud_merchant", regexp_extract(col("merchant"), r"(fraud_[A-Za-z]+)", 1))
Search
ENG
IN
09:28
12-05-2025

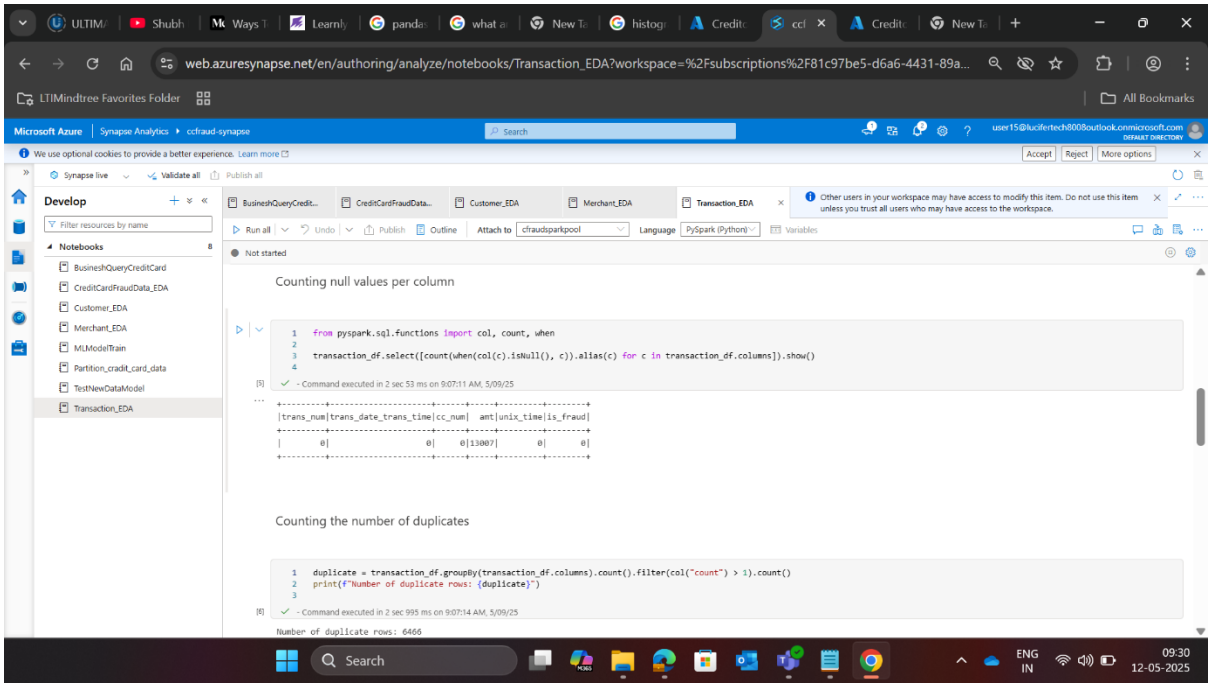

web.azuresynapse.net/en/authoring/analyze/notebooks/Merchant_EDA?workspace=%2Fsubscriptions%2F81c97be5-d6a6-4431-89a7-...
Microsoft Azure | Synapse Analytics | ccfraud-synapse
Develop
Notebooks
BusineshQueryCreditCard
CreditCardFraudData_EDA
Customer_EDA
Merchant_EDA
MLModelTrain
Partition_cradit_card_data
TestNewDataModel
Transaction_EDA
from pyspark.sql.functions import col, count, when, lit
most = merchant_df.groupBy("category").count().orderBy(col("count").desc()).first()["category"]
merchant_df = merchant_df.withColumn("category", when(col("category").isNull(), most).otherwise(col("category")))
- Command executed in 2 sec 25 ms on 9:06:36 AM, 5/10/25
merchant_df.count()
- Command executed in 644 ms on 9:06:37 AM, 5/10/25
1303158
Drop the null values
merchant_df = merchant_df.filter(col("merch_zipcode").isNotNull())
- Command executed in 256 ms on 9:06:37 AM, 5/10/25
Search
ENG
IN
09:29
12-05-2025

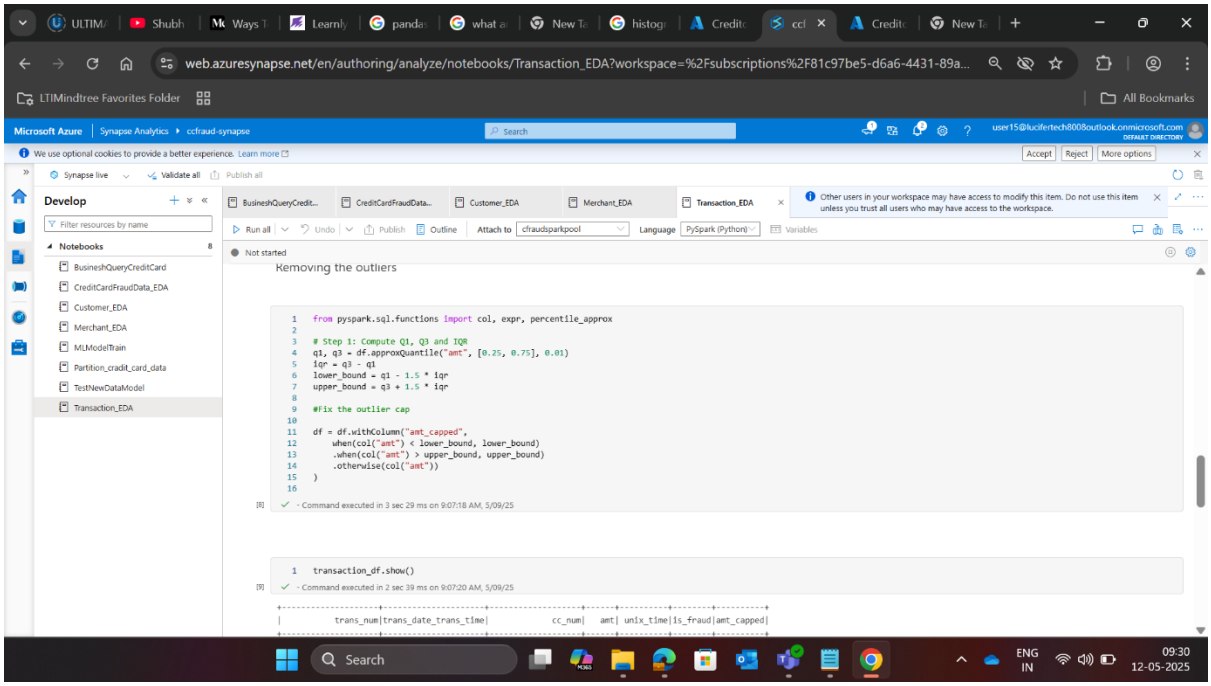


3.Transaction EDA

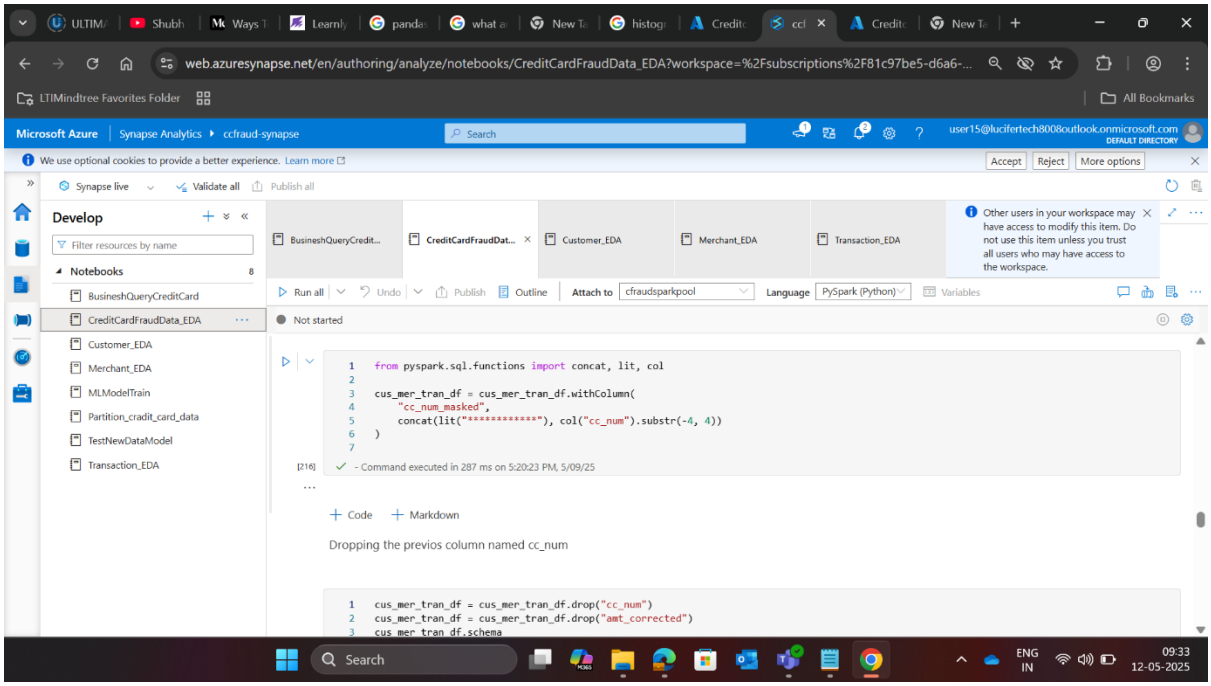


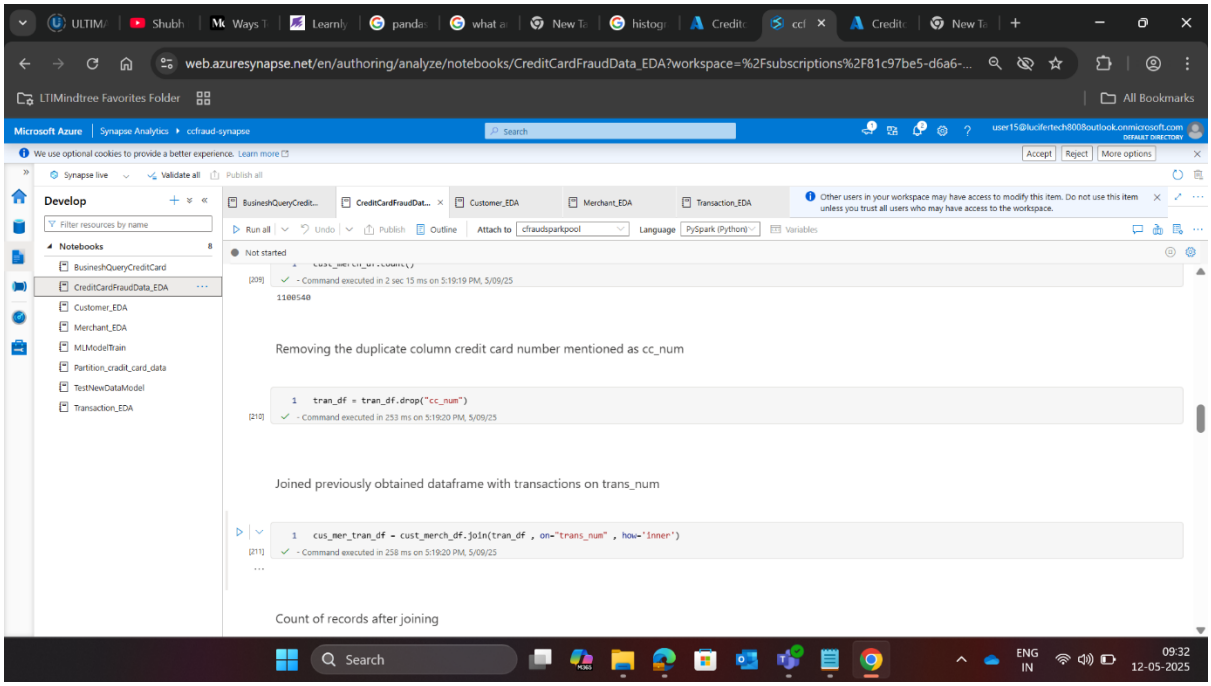

web.azuresynapse.net/en/authoring/analyze/notebooks/Transaction_EDA?workspace=%2Fsubscriptions%2F81c97be5-d6a6-4431-89a...
Microsoft Azure | Synapse Analytics | ccfraud-synapse
Develop
Notebooks 8
BusineshQueryCreditCard
CreditCardFraudData_EDA
Customer_EDA
Merchant_EDA
MLModelTrain
Partition_cradit_card_data
TestNewDataModel
Transaction_EDA
Attach to cfraudsparkpool
Language PySpark (Python)
Not started
from pyspark.sql.types import *
from pyspark.sql.functions import *
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, count
- Apache Spark session started in 45 sec 301 ms. Command executed in 301 ms on 9:07:00 AM, 5/09/25
Reading the csv file from adls
transaction_df = spark.read.csv("abfss://raw-fraud-database@ccfrauddatalake.dfs.core.windows.net/TransactionData.tsv", inferSchema=True , header=True ,sep="\t")
transaction_df.show(4)
- Command executed in 9 sec 9 ms on 9:07:09 AM, 5/09/25
| trans_num|trans_date_trans_time| cc_num| merchant| amt| unix_time|is_fraud|
|1d8bdcbc194d8c696...| 2019-07-27 21:00:29| 577588686219| fraud_Kassulke PLC|110.39|1343422829| 0|
|0315af3a790614331...| 2019-02-27 13:25:28|601136578560244|fraud_Mueller, Ge...| 10.87|1330349128| 0|
|bb2ad9d601f0f1818...| 2019-12-13 17:14:45| 213126662687660|fraud_Little-Gle1...| 3.15|1355418885| 0|
|fb38a5f056815bce3...| 2020-06-01 03:00:47| 213191402330021| fraud_Friesen-Stamm| 97.26|1370055647| 0|
only showing top 4 rows
Search
ENG IN
09:30 12-05-2025

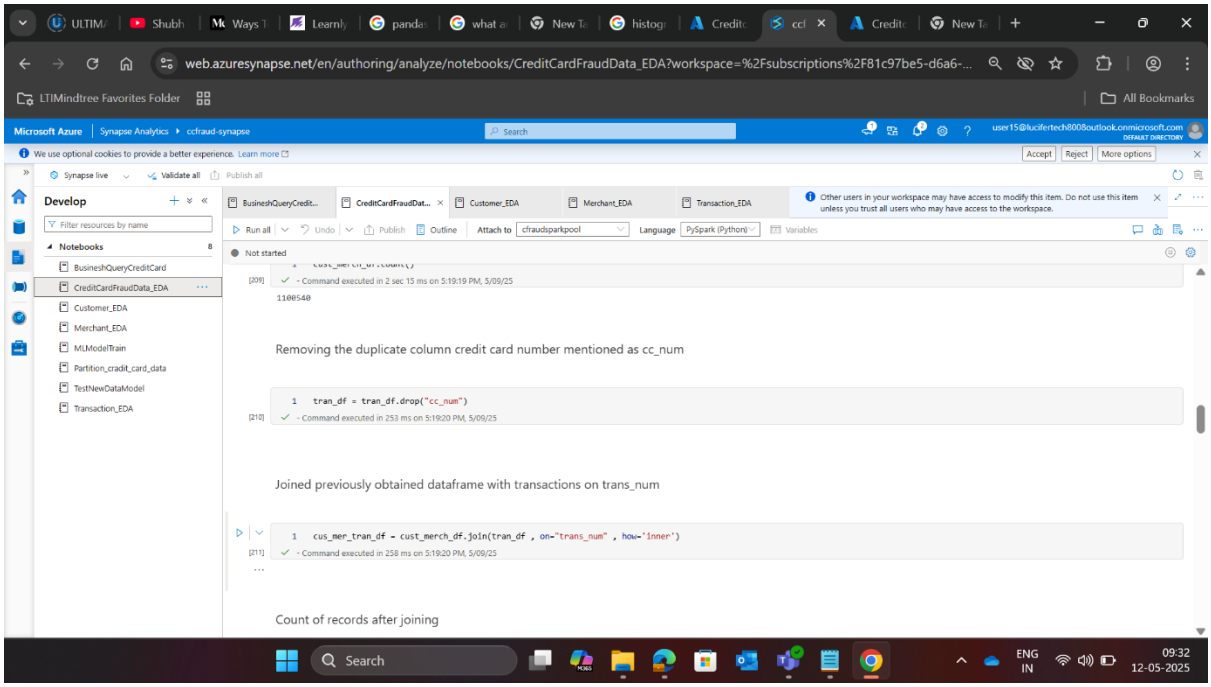

web.azuresynapse.net/en/authoring/analyze/notebooks/Transaction_EDA?workspace=%2Fsubscriptions%2F81c97be5-d6a6-4431-89a...
Microsoft Azure | Synapse Analytics | ccfraud-synapse
Develop
Notebooks 8
BusineshQueryCreditCard
CreditCardFraudData_EDA
Customer_EDA
Merchant_EDA
MLModelTrain
Partition_cradit_card_data
TestNewDataModel
Transaction_EDA
Attach to cfraudsparkpool
Language PySpark (Python)
Not started
Counting null values per column
from pyspark.sql.functions import col, count, when
transaction_df.select([count(when(col(c).isNull(), c)).alias(c) for c in transaction_df.columns]).show()
- Command executed in 2 sec 53 ms on 9:07:11 AM, 5/09/25
|trans_num|trans_date_trans_time|cc_num| amt|unix_time|is_fraud|
| 0| 0| 0|13007| 0| 0|
Counting the number of duplicates
duplicate = transaction_df.groupBy(transaction_df.columns).count().filter(col("count") > 1).count()
print(f"Number of duplicate rows: {duplicate}")
- Command executed in 2 sec 995 ms on 9:07:14 AM, 5/09/25
Number of duplicate rows: 6466
Search
ENG IN
09:30 12-05-2025

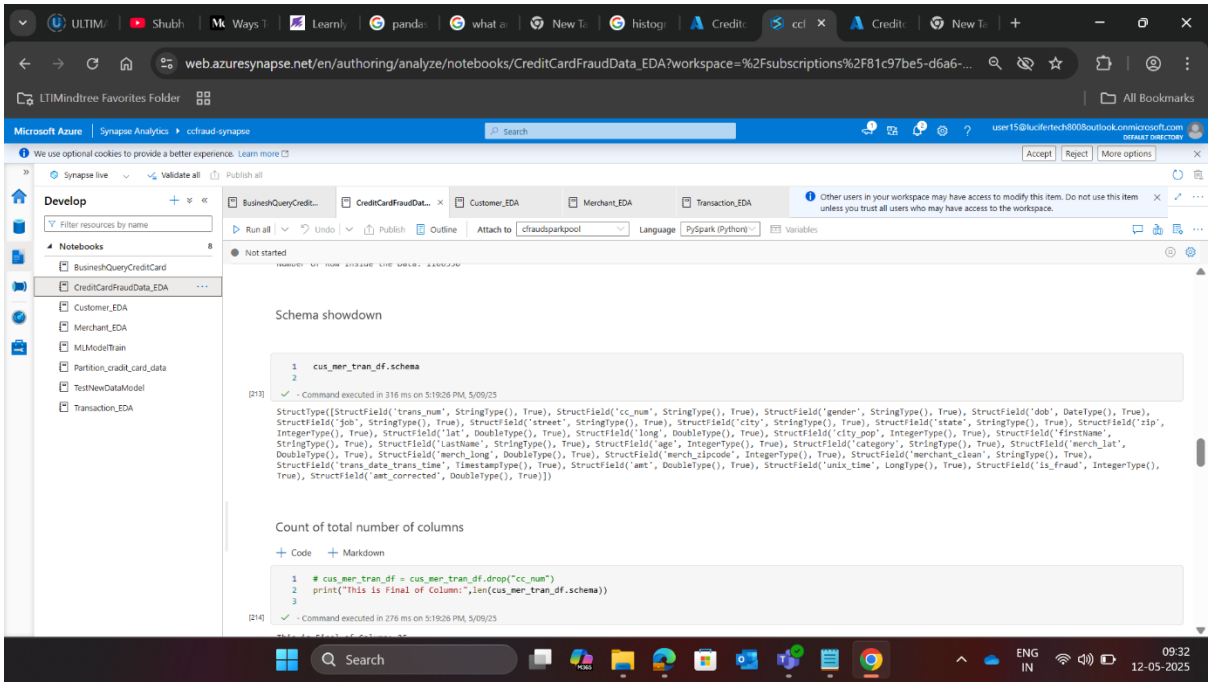


Optimization:

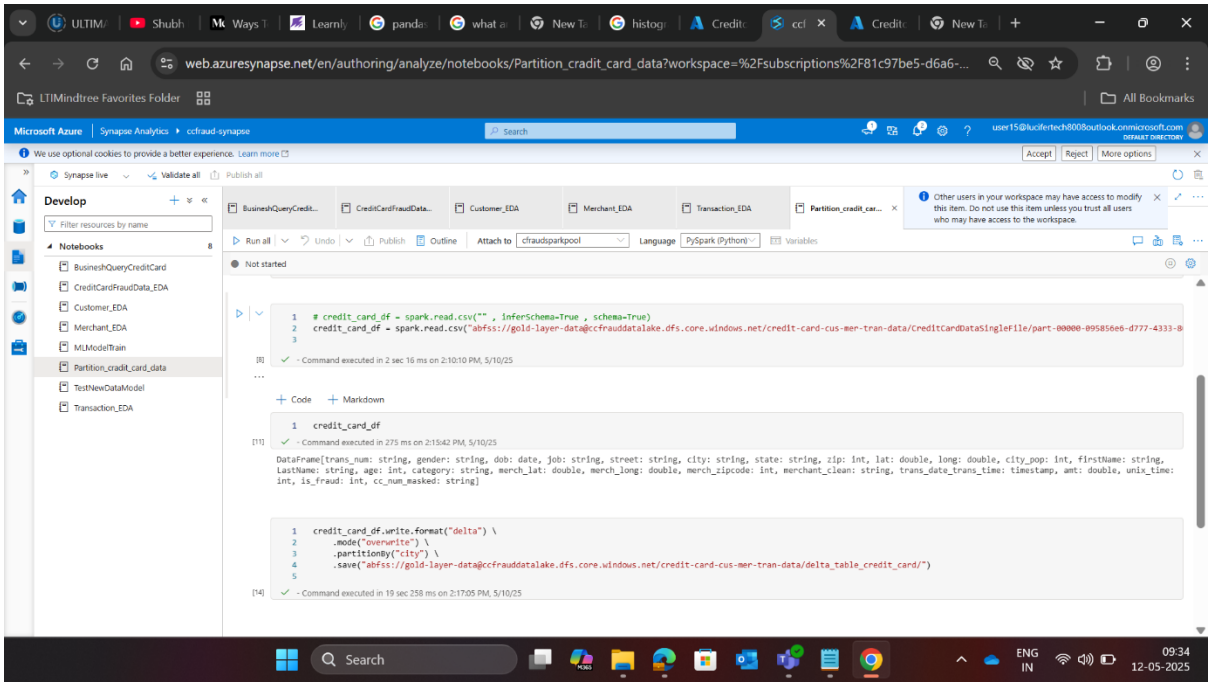



web.azuresynapse.net/en/authoring/analyze/notebooks/CreditCardFraudData_EDA?workspace=%2Fsubscriptions%2F81c97be5-d6a6-...
Microsoft Azure | Synapse Analytics | ccfraud-synapse
Develop
Notebooks 8
BusineshQueryCreditCard
CreditCardFraudData_EDA
Customer_EDA
Merchant_EDA
MLModelTrain
Partition_cradit_card_data
TestNewDataModel
Transaction_EDA
Attach to cfraudsparkpool
Language PySpark (Python)
Not started
- Command executed in 2 sec 15 ms on 5:19:19 PM, 5/09/25
1100540
Removing the duplicate column credit card number mentioned as cc_num
tran_df = tran_df.drop("cc_num")
- Command executed in 253 ms on 5:19:20 PM, 5/09/25
Joined previously obtained dataframe with transactions on trans_num
cus_mer_tran_df = cust_merch_df.join(tran_df , on="trans_num" , how='inner')
- Command executed in 258 ms on 5:19:20 PM, 5/09/25
Count of records after joining
Search
ENG IN
09:32 12-05-2025

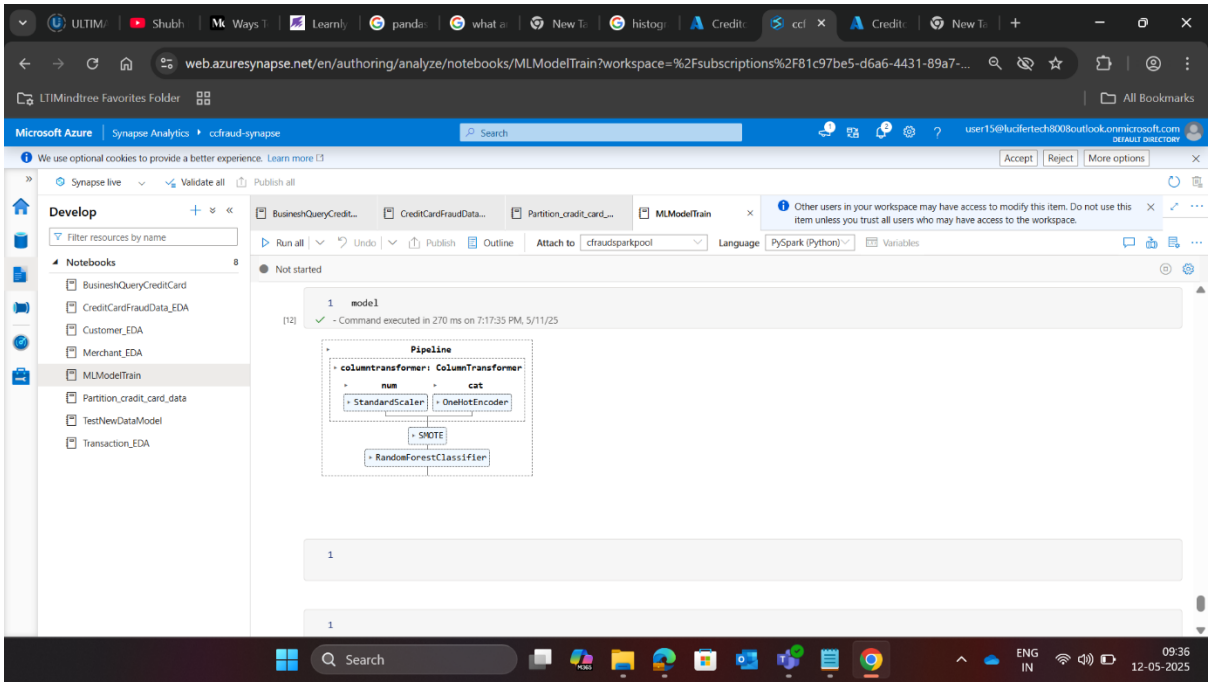

web.azuresynapse.net/en/authoring/analyze/notebooks/CreditCardFraudData_EDA?workspace=%2Fsubscriptions%2F81c97be5-d6a6-...
Microsoft Azure | Synapse Analytics | ccfraud-synapse
Develop
Notebooks 8
BusineshQueryCreditCard
CreditCardFraudData_EDA
Customer_EDA
Merchant_EDA
MLModelTrain
Partition_cradit_card_data
TestNewDataModel
Transaction_EDA
Attach to cfraudsparkpool
Language PySpark (Python)
Not started
- Command executed in 2 sec 15 ms on 5:19:19 PM, 5/09/25
1100540
Removing the duplicate column credit card number mentioned as cc_num
tran_df = tran_df.drop("cc_num")
- Command executed in 253 ms on 5:19:20 PM, 5/09/25
Joined previously obtained dataframe with transactions on trans_num
cus_mer_tran_df = cust_merch_df.join(tran_df , on="trans_num" , how='inner')
- Command executed in 258 ms on 5:19:20 PM, 5/09/25
Count of records after joining
Search
ENG IN
09:32 12-05-2025

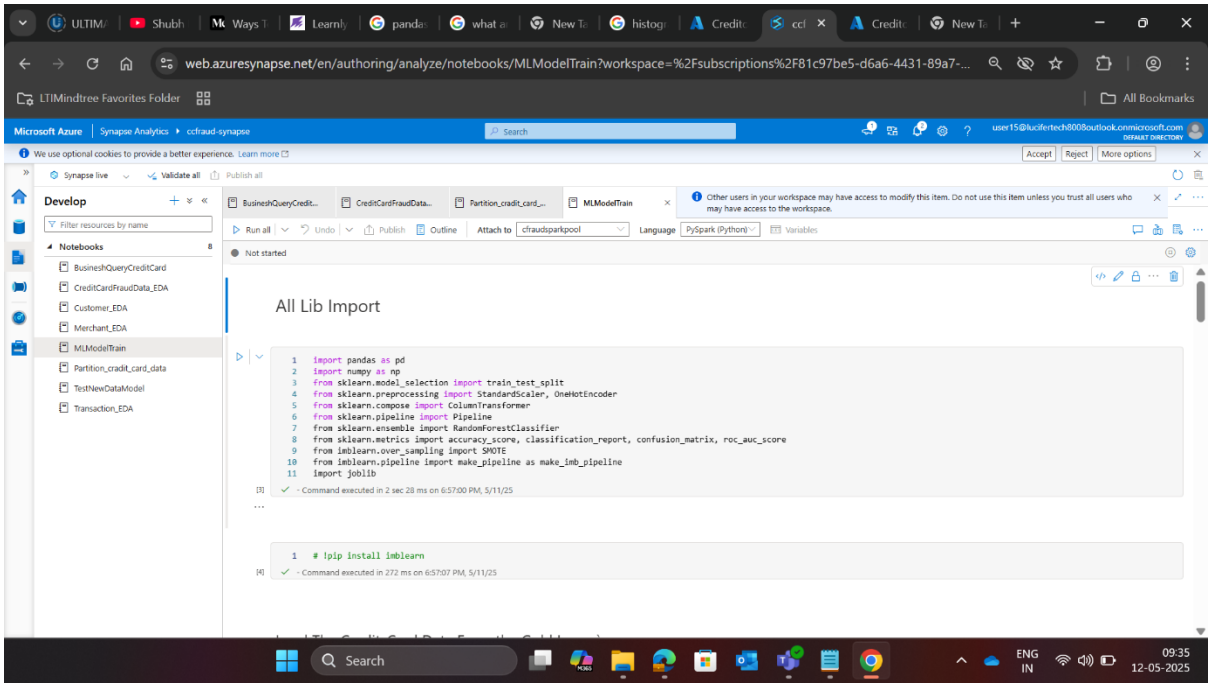


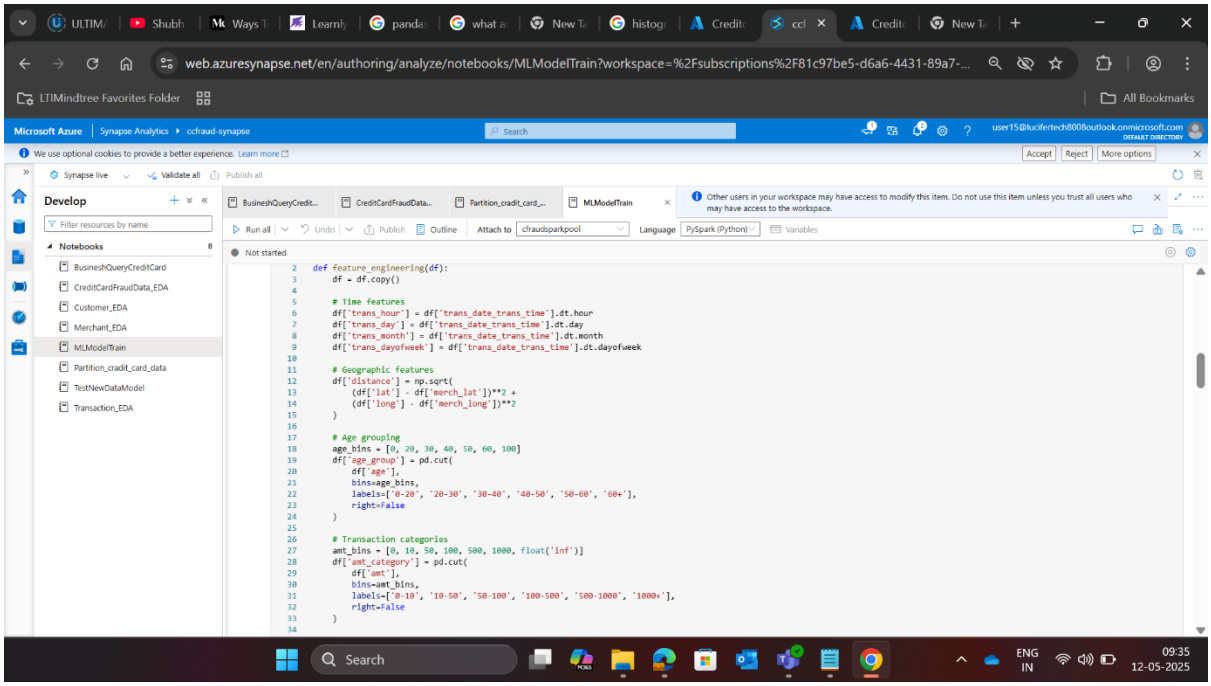
Partition And delta table optimization:

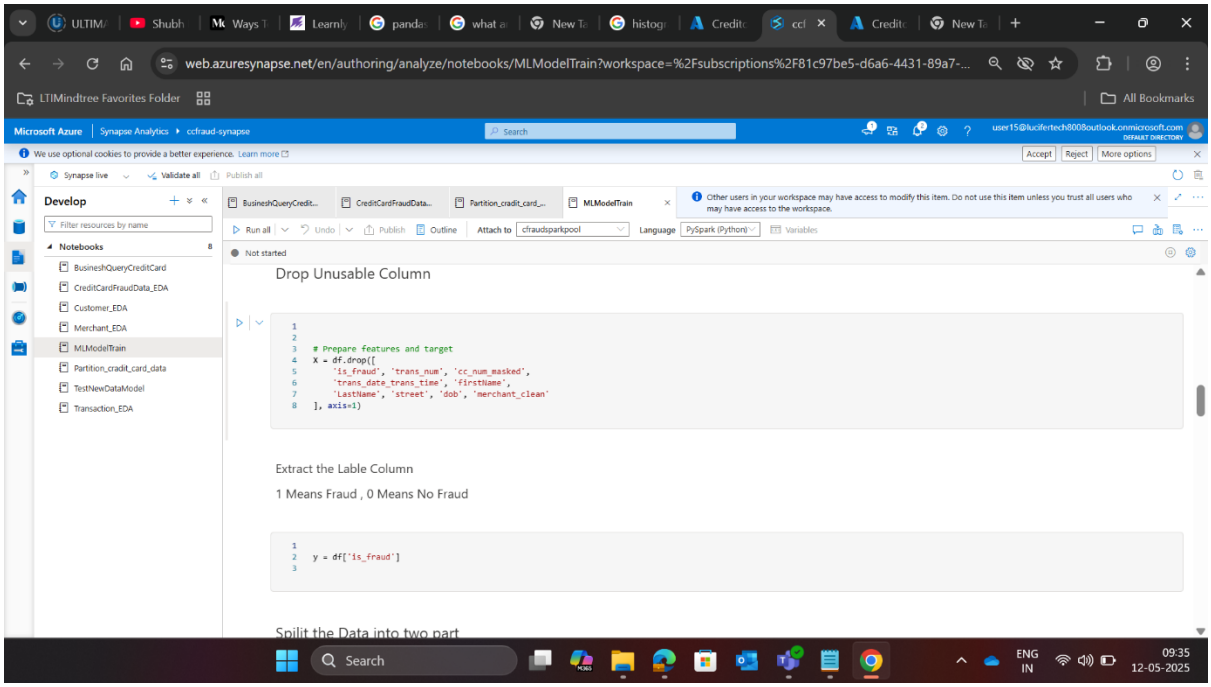


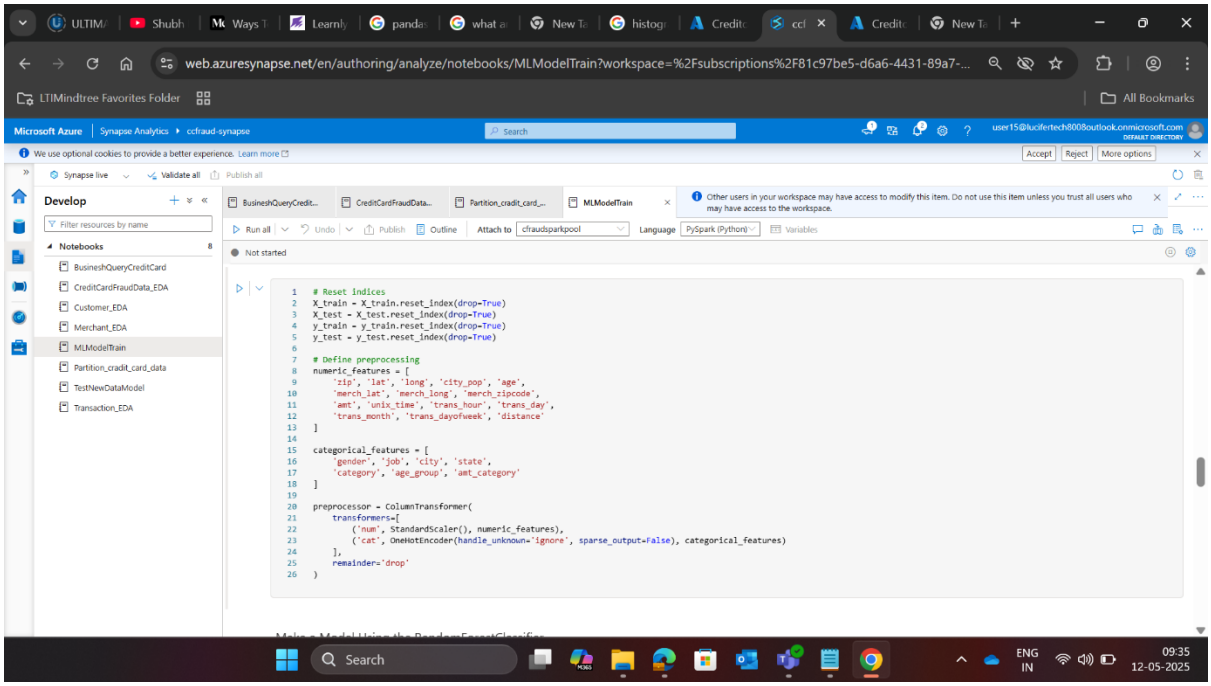
Machine Learning to detect Fraudsters:

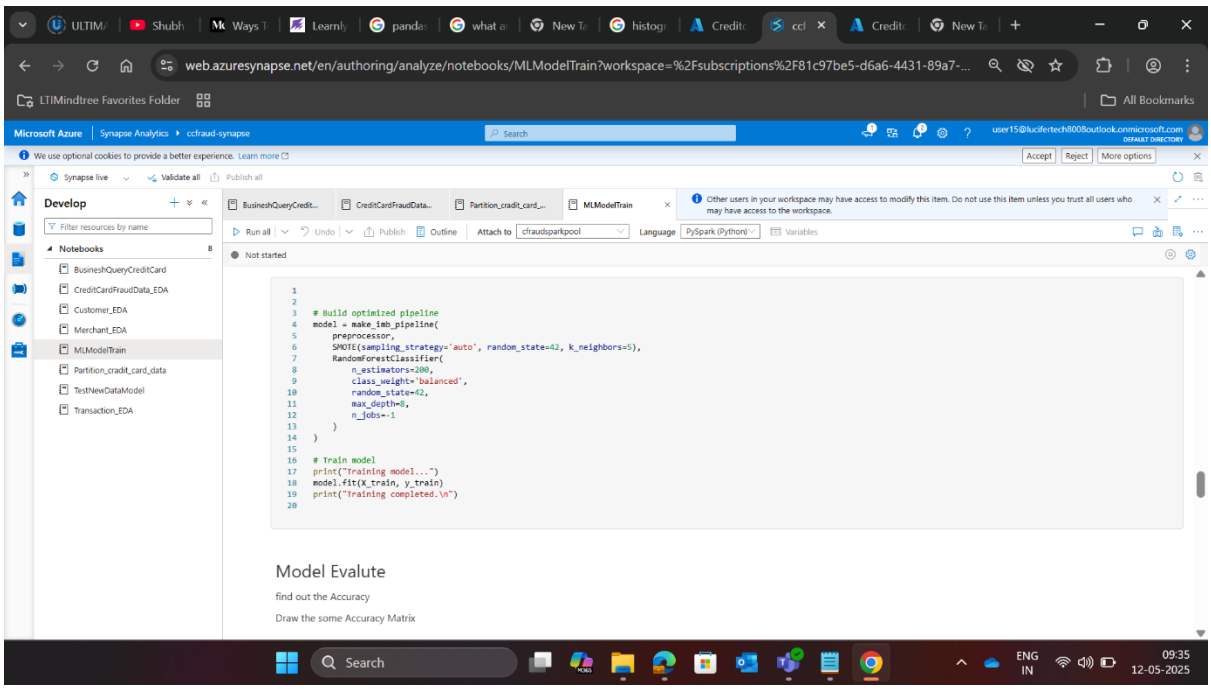


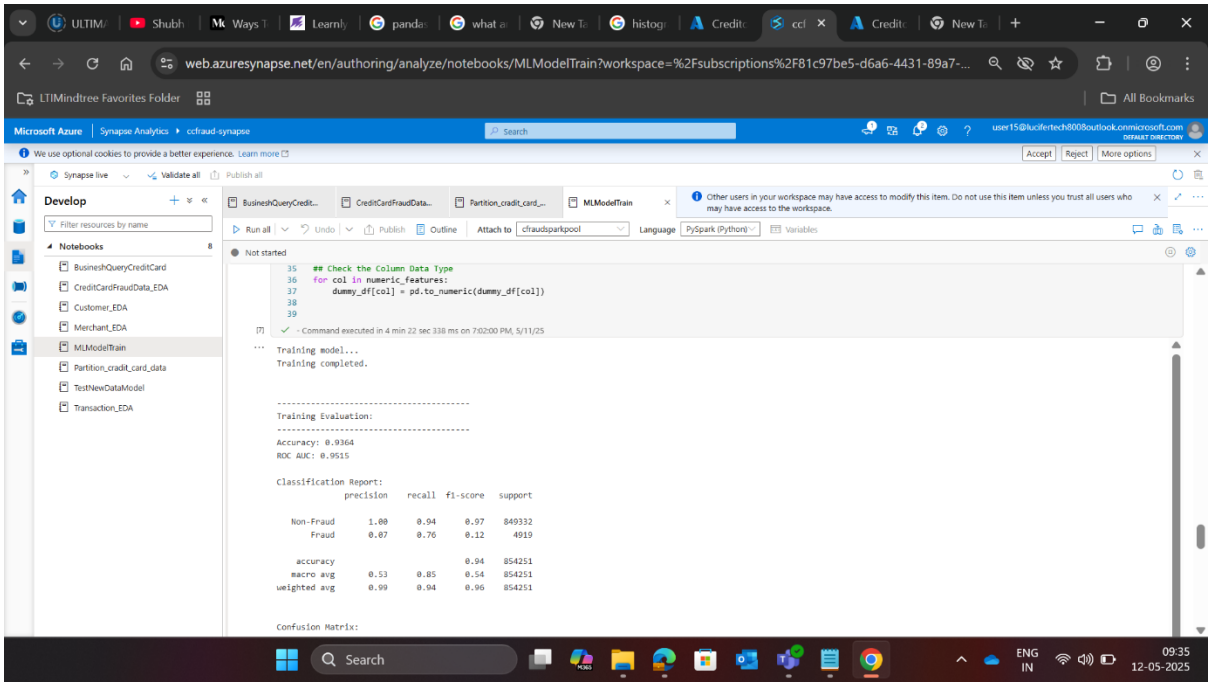


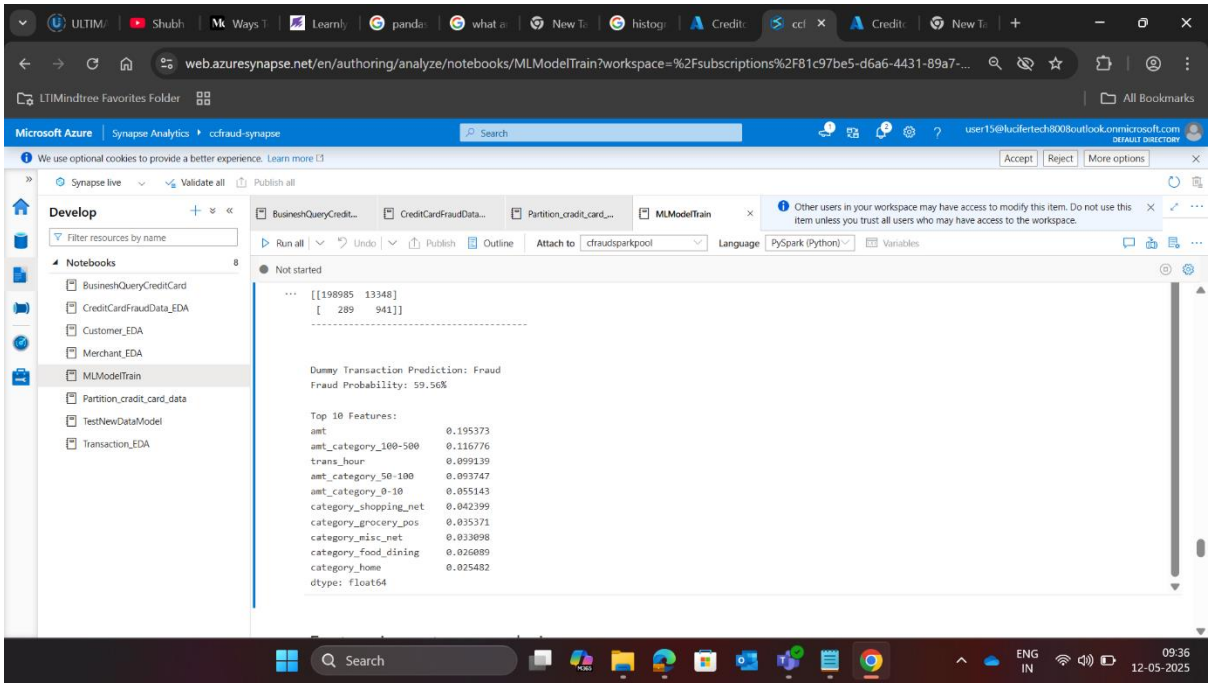

web.azuresynapse.net/en/authoring/analyze/notebooks/MLModelTrain?workspace=%2Fsubscriptions%2F81c97be5-d6a6-4431-89a7-...
Microsoft Azure | Synapse Analytics | ccfraud-synapse
Develop
Notebooks
BusineshQueryCreditCard
CreditCardFraudData_EDA
Customer_EDA
Merchant_EDA
MLModelTrain
Partition_cradit_card_data
TestNewDataModel
Transaction_EDA
Attach to cfraudsparkpool
Language PySpark (Python)
Not started
def feature_engineering(df):
df = df.copy()
# Time features
df['trans_hour'] = df['trans_date_trans_time'].dt.hour
df['trans_day'] = df['trans_date_trans_time'].dt.day
df['trans_month'] = df['trans_date_trans_time'].dt.month
df['trans_dayofweek'] = df['trans_date_trans_time'].dt.dayofweek
Geographic features
df['distance'] = np.sqrt(
(df['lat'] - df['merch_lat'])**2 +
(df['long'] - df['merch_long'])**2
)
# Age grouping
age_bins = [0, 20, 30, 40, 50, 60, 100]
df['age_group'] = pd.cut(
df['age'],
bins=age_bins,
labels=['0-20', '20-30', '30-40', '40-50', '50-60', '60+'],
right=False
)
# Transaction categories
amt_bins = [0, 10, 50, 100, 500, 1000, float('inf')]
df['amt_category'] = pd.cut(
df['amt'],
bins=amt_bins,
labels=['0-10', '10-50', '50-100', '100-500', '500-1000', '1000+'],
right=False
)
Search
ENG
IN
09:35
12-05-2025


web.azuresynapse.net/en/authoring/analyze/notebooks/MLModelTrain?workspace=%2Fsubscriptions%2F81c97be5-d6a6-4431-89a7-...
Microsoft Azure | Synapse Analytics | ccfraud-synapse
Develop
Notebooks
BusineshQueryCreditCard
CreditCardFraudData_EDA
Customer_EDA
Merchant_EDA
MLModelTrain
Partition_cradit_card_data
TestNewDataModel
Transaction_EDA
Attach to cfraudsparkpool
Language PySpark (Python)
Not started
Drop Unusable Column
# Prepare features and target
X = df.drop([
'is_fraud', 'trans_num', 'cc_num_masked',
'trans_date_trans_time', 'firstName',
'lastName', 'street', 'dob', 'merchant_clean'
], axis=1)
Extract the Lable Column
1 Means Fraud , 0 Means No Fraud
y = df['is_fraud']
Spilit the Data into two part
Search
ENG
IN
09:35
12-05-2025

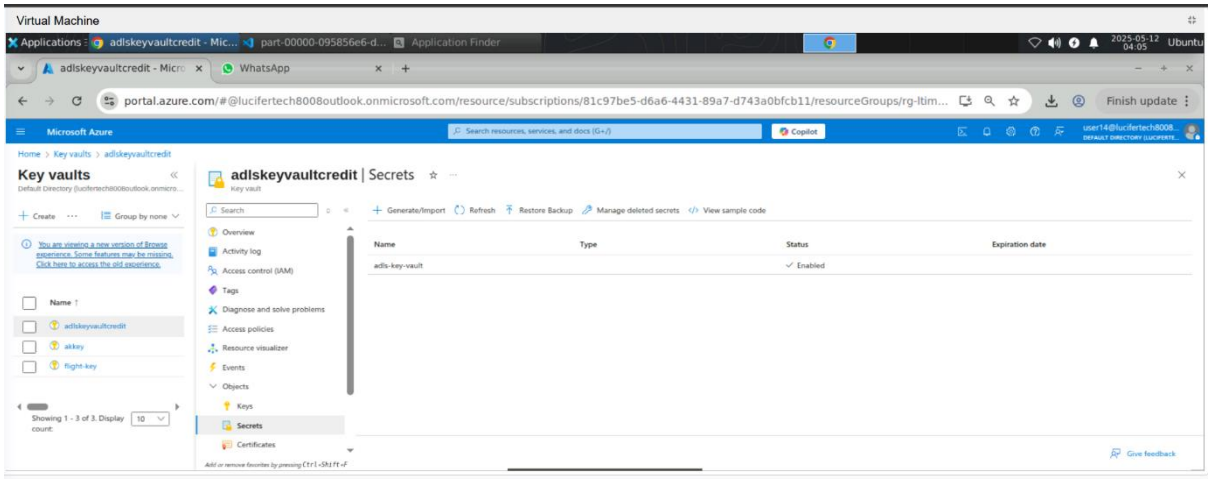

web.azuresynapse.net/en/authoring/analyze/notebooks/MLModelTrain?workspace=%2Fsubscriptions%2F81c97be5-d6a6-4431-89a7-...
Microsoft Azure | Synapse Analytics | ccfraud-synapse
Develop
Notebooks 8
BusineshQueryCreditCard
CreditCardFraudData_EDA
Customer_EDA
Merchant_EDA
MLModelTrain
Partition_cradit_card_data
TestNewDataModel
Transaction_EDA
Attach to cfraudsparkpool
Language PySpark (Python)
Not started
# Reset indices
X_train = X_train.reset_index(drop=True)
X_test = X_test.reset_index(drop=True)
y_train = y_train.reset_index(drop=True)
y_test = y_test.reset_index(drop=True)
# Define preprocessing
numeric_features = [
'zip', 'lat', 'long', 'city_pop', 'age',
'merch_lat', 'merch_long', 'merch_zipcode',
'amt', 'unix_time', 'trans_hour', 'trans_day',
'trans_month', 'trans_dayofweek', 'distance'
]
categorical_features = [
'gender', 'job', 'city', 'state',
'category', 'age_group', 'amt_category'
]
preprocessor = ColumnTransformer(
transformers=[
('num', StandardScaler(), numeric_features),
('cat', OneHotEncoder(handle_unknown='ignore', sparse_output=False), categorical_features)
],
remainder='drop'
)
Search
ENG
IN
09:35
12-05-2025

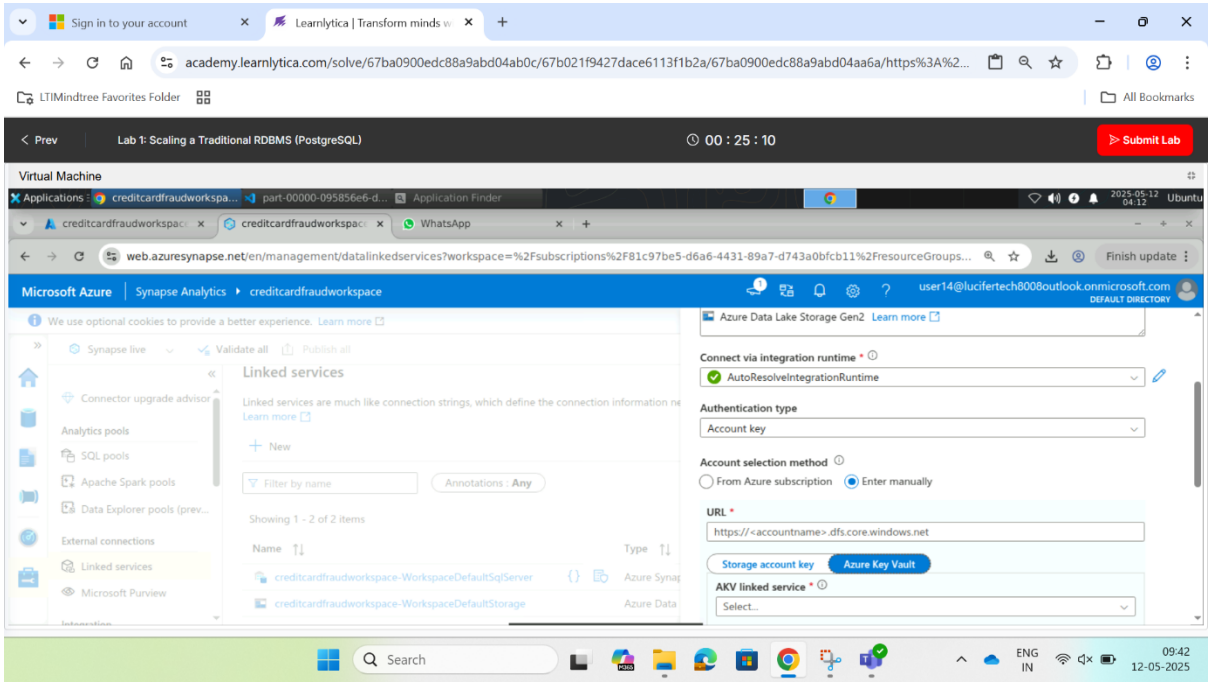

web.azuresynapse.net/en/authoring/analyze/notebooks/MLModelTrain?workspace=%2Fsubscriptions%2F81c97be5-d6a6-4431-89a7-...
Microsoft Azure | Synapse Analytics | ccfraud-synapse
Develop
Notebooks 8
BusineshQueryCreditCard
CreditCardFraudData_EDA
Customer_EDA
Merchant_EDA
MLModelTrain
Partition_cradit_card_data
TestNewDataModel
Transaction_EDA
Attach to cfraudsparkpool
Language PySpark (Python)
Not started
# Build optimized pipeline
model = make_imb_pipeline(
preprocessor,
SMOTE(sampling_strategy='auto', random_state=42, k_neighbors=5),
RandomForestClassifier(
n_estimators=200,
class_weight='balanced',
random_state=42,
max_depth=8,
n_jobs=-1
)
)
# Train model
print("Training model...")
model.fit(X_train, y_train)
print("Training completed.\n")
Model Evalute
find out the Accuracy
Draw the some Accuracy Matrix
Search
ENG
IN
09:35
12-05-2025


web.azuresynapse.net/en/authoring/analyze/notebooks/MLModelTrain?workspace=%2Fsubscriptions%2F81c97be5-d6a6-4431-89a7-...
Microsoft Azure | Synapse Analytics | ccfraud-synapse
Develop
Notebooks 8
BusineshQueryCreditCard
CreditCardFraudData_EDA
Customer_EDA
Merchant_EDA
MLModelTrain
Partition_cradit_card_data
TestNewDataModel
Transaction_EDA
Attach to cfraudsparkpool
Language PySpark (Python)
Not started
Check the Column Data Type
for col in numeric_features:
dummy_df[col] = pd.to_numeric(dummy_df[col])
Command executed in 4 min 22 sec 338 ms on 7:02:00 PM, 5/11/25
Training model...
Training completed.
Training Evaluation:
Accuracy: 0.9364
ROC AUC: 0.9515
precision recall f1-score support
Non-Fraud 1.00 0.94 0.97 849332
Fraud 0.07 0.76 0.12 4919
accuracy 0.94 854251
macro avg 0.53 0.85 0.54 854251
weighted avg 0.99 0.94 0.96 854251
Confusion Matrix:
Search
ENG IN
09:35 12-05-2025


web.azuresynapse.net/en/authoring/analyze/notebooks/MLModelTrain?workspace=%2Fsubscriptions%2F81c97be5-d6a6-4431-89a7-...
Microsoft Azure | Synapse Analytics | ccfraud-synapse
Develop
Notebooks 8
BusineshQueryCreditCard
CreditCardFraudData_EDA
Customer_EDA
Merchant_EDA
MLModelTrain
Partition_cradit_card_data
TestNewDataModel
Transaction_EDA
Attach to cfraudsparkpool
Language PySpark (Python)
Not started
[[198985 13348]
[289 941]]
Dummy Transaction Prediction: Fraud
Fraud Probability: 59.56%
Top 10 Features:
amt 0.195373
amt_category_100-500 0.116776
trans_hour 0.099139
amt_category_50-100 0.093747
amt_category_0-10 0.055143
category_shopping_net 0.042399
category_grocery_pos 0.035371
category_misc_net 0.033098
category_food_dining 0.026089
category_home 0.025482
dtype: float64
Search
ENG IN
09:36 12-05-2025

Azure Key Vault:





Step 7:

1. **Total Records and Fraud Cases**
    - Total Records: 1M
    - Total Fraud Cases: 6K
2. **Fraudsters per State per City**
    - A bar chart showing fraud counts for different cities (Achille, Acworth, Adams, Afton, Akron) across various states (TX, PA, NY, FL).
3. **Fraud Cases Greater Than 500 by Month**
    - May: 722
    - June: 544
    - January: 692
    - February: 686
    - December: 616
    - April: 558
    - Total Fraud: 4540

4. **Amount of Fraud Segregated by Gender**
    - A bar chart displaying the sum of fraud amounts for genders Female (F) and Male (M).
5. **Fraud Count per Credit Card**
    - A bar chart showing fraud counts for different credit card numbers.
6. **Detecting Seasonal Trends in Fraud Activity**
    - A line graph illustrating trends in various categories (entertainment, food_dining, gas_transport) over months from January to December.

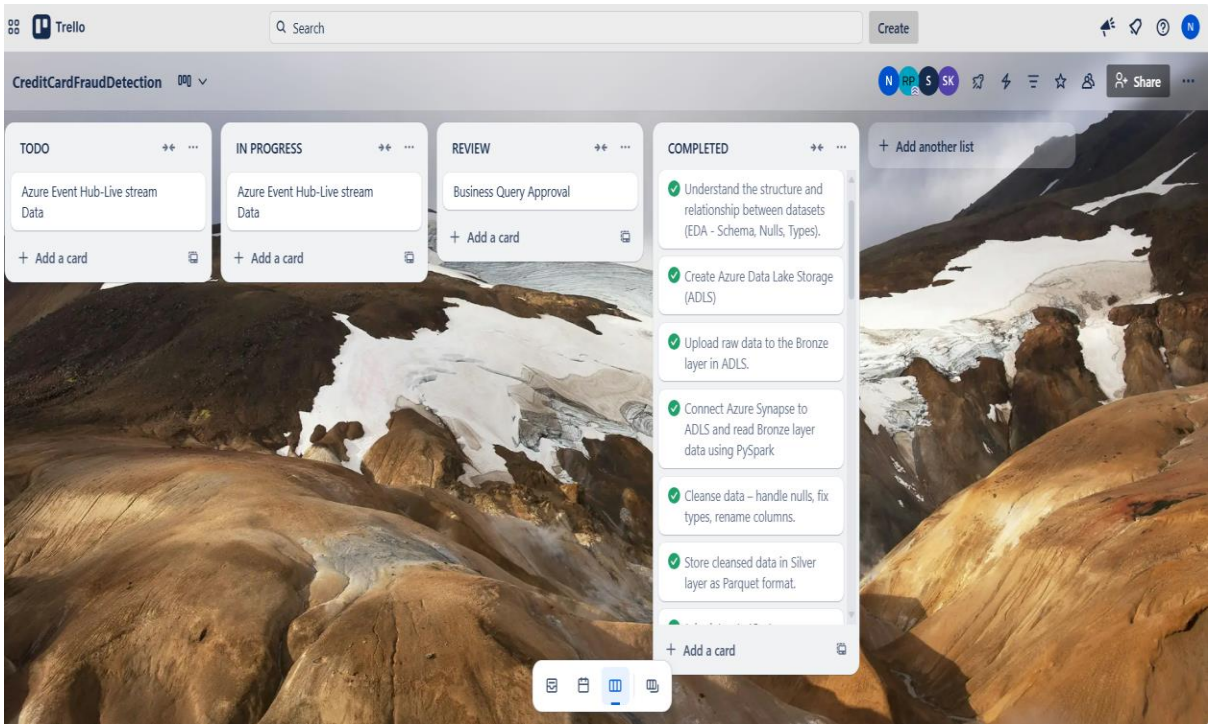
Trello

In our project, we utilized Trello for effective project management and collaboration. Trello helped us organize tasks, track progress, and ensure timely completion of project milestones.

Project Completion Status

- **Review:** 8% of the project is currently under review. This includes tasks that have been completed but are awaiting final approval and quality checks.
- **Completed:** 90% of the project tasks have been successfully completed. This includes all major deliverables and milestones.
- **Future Work:** 2% of the project is planned for future work. These tasks are identified as potential improvements or additional features that can be implemented in subsequent phases.

Trello:



FUTURE ENHANCEMENT:

- Live Stream Data: Implement live stream data processing using Azure Event Hub.
- Monitor user behavior to spot subtle fraud patterns.
- Integrate additional data sources for more context.
- Implement automatic actions for detected fraud.
- Optimize for handling more data efficiently.
- Improve tools for detailed insights and trends.