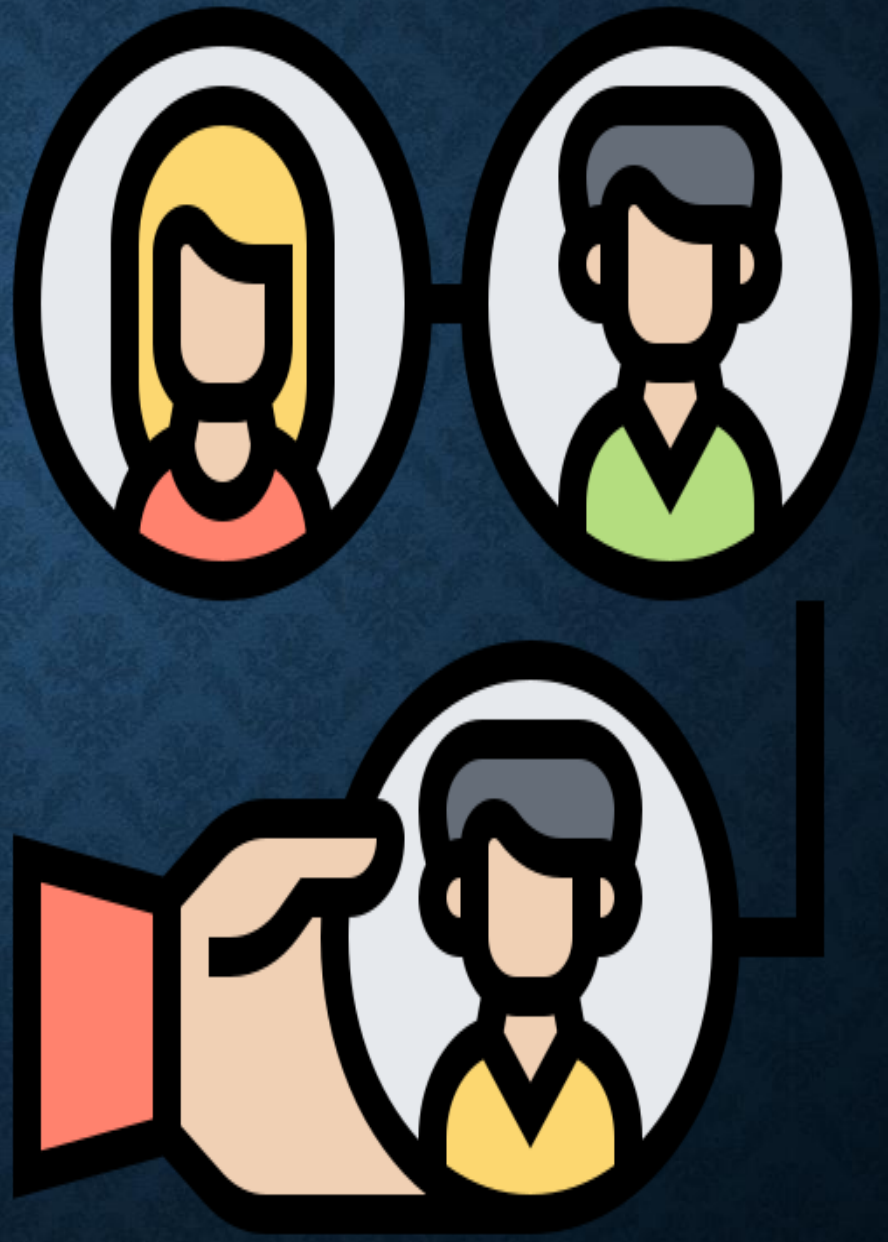
Customer Segmentation



Presented By: Tarek Abdeen

Agenda

1- Overview
2- About Data
3- Tools I used
4- Data Flow
5- Data Ingestion
6- Data Processing With Spark
7- Data Analysis
8- Data Warehousing & Data Lake
9- Data Visualization
10- Codes

Overview

My project aims to divide customers into different groups based on their characteristics and behaviors to understand and serve them better, to improve service, and to satisfy and understand their needs.

About Data

Feature	Description	Values
ID	Id of customer	1,2,3,4,...
Gender	Gender of the customer	(Male & Female)
Ever_Married	Marital status of the customer	(Yes & No)
Age	Age of the customer	10,20, 25, 40,....
Graduated	Is the customer a graduate?	Yes & No)
Profession	Profession of the customer	(Artist, Healthcare,Doctor, Engineer, Lawyer, etc)
Work_Experience	Work Experience	(1:10)
Spending_Score	Spending score of the customer	(Low, Average, High)
Family_Size	Number of family members for the customer (including the customer)	1,5,2,..
Var_1	Variable	(Cat_1, Cat_2, Cat_3, Cat_4)

Tools

Data Flow

Data Ingestion

My data source is CSV files consisting of 11 columns and 8069 rows, and I have stored them in Apache Spark because it is in the memory store and is faster in execution. This makes it easier for me in the operations of cleaning the data, integrate it, and convert it to suitable data types to make data analysis. And machine learning algorithms.

Data cleaning

- Handle data types
- Check duplicates
- Remove Missing data
- Remove Outlier

Data Processing With Spark

These are some of the screenshots of the Spark code, and there is a link to the end with the full code due to the large size of the code.

databricks

Customer Segmentation (Python)

Import Notebook

```
import sys
print(sys.path)
```

```
['/databricks/python_shell/scripts', '/local_disk0/spark-3bfcede1-813c-4416-bacb-65a01940ae9f/userFiles-9196ea64-764e-4f38-be8a-9532f33771ae', '/databricks/spark/python', '/databricks/spark/python/lib/py4j-0.10.9.5-src.zip', '/databricks/jars/spark--driver--driver-spark_3.3_2.12_deploy.jar', '/WSFS_NOTEBOOK_DIR', '/databricks/python_shell', '/usr/lib/python39.zip', '/usr/lib/python3.9', '/usr/lib/python3.9/lib-dynload', '', '/local_disk0/.ephemeral_nfs/envs/pythonEnv-ed265f59-8ace-4383-992a-33a6674d383d/lib/python3.9/site-packages', '/local_disk0/.ephemeral_nfs/cluster_libraries/python/lib/python3.9/site-packages', '/databricks/python/lib/python3.9/site-packages', '/usr/local/lib/python3.9/dist-packages', '/usr/lib/python3/dist-packages']
```

```
import os
print(os.environ['SPARK_HOME'])
print(os.environ['JAVA_HOME'])

print(os.environ['PATH'])
os.environ["pyspark_python"] = "/anconda3/envs/ucBEXtension/bin/python"
```

```
/databricks/spark
/usr/lib/jvm/zulu8-ca-amd64/jre/
/local_disk0/.ephemeral_nfs/envs/pythonEnv-ed265f59-8ace-4383-992a-33a6674d383d/bin:/local_disk0/.ephemeral_nfs/cluster_libraries/python/bin:/databricks/.pyenv/bin:/usr/local/nvidia/bin:/databricks/python3/bin:/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/snap/bin
```

```
pip install findspark
```

```
Python interpreter will be restarted.
Requirement already satisfied: findspark in /local_disk0/.ephemeral_nfs/envs/pythonEnv-ed265f59-8ace-4383-992a-33a6674d383d/lib/python3.9/site-packages (2.0.1)
Python interpreter will be restarted.
```

databricks

Customer Segmentation (Python)

Import Notebook

```
import findspark
findspark.init()
```

```
import pyspark
from pyspark import SparkContext, SparkConf
from pyspark.sql import SparkSession, types
from pyspark.sql import Row
from pyspark.sql import SQLContext
```

```
from pyspark.sql import SparkSession

spark = SparkSession.builder \
    .appName("KafkaSparkIntegration") \
    .config("spark.driver.cores", "1") \
    .config("spark.driver.memory", "1g") \
    .config("spark.executor.cores", "2") \
    .config("spark.executor.memory", "1gb") \
    .config("spark.num.executors", "2") \
    .enableHiveSupport() \
    .getOrCreate()

sc = spark.sparkContext
```



```
spark.sparkContext.getConf().getAll()
```

```
spark.sparkContext.getConf().getAll()
```

```
Out[4]: [('spark.databricks.preemption.enabled', 'true'),
 ('spark.sql.hive.metastore.jars', '/databricks/databricks-hive/*'),
 ('spark.driver.tempDirectory', '/local_disk0/tmp'),
 ('spark.sql.warehouse.dir', 'dbfs:/user/hive/warehouse'),
 ('spark.databricks.managedCatalog.clientClassName',
  'com.databricks.managedcatalog.ManagedCatalogClientImpl'),
 ('spark.databricks.credential.scope.fs.gs.auth.access.tokenProviderClassName',
  'com.databricks.backend.daemon.driver.credentials.CredentialScopeGCPTokenProvider'),
 ('spark.hadoop.fs.fcfs-s3.impl.disable.cache', 'true'),
 ('spark.sql.streaming.checkpointFileManagerClass',
  'com.databricks.spark.sql.streaming.DatabricksCheckpointFileManager'),
 ('spark.databricks.service.dbutils.repl.backend',
  'com.databricks.dbconnect.ReplDBUtils'),
 ('spark.hadoop.databricks.s3.verifyBucketExists.enabled', 'false'),
 ('spark.streaming.driver.writeAheadLog.allowBatching', 'true'),
 ('spark.databricks.clusterSource', 'UI'),
 ('spark.hadoop.hive.server2.transport.mode', 'http'),
 ('spark.executor.memory', '8278m'),
 ('spark.hadoop.fs.cpfs-adl.impl.disable.cache', 'true'),
 ('spark.databricks.clusterUsageTags.hailEnabled', 'false'),
 ('spark.databricks.clusterUsageTags.clusterLogDeliveryEnabled', 'false'),
```

```
sqlcontext = SQLContext(sc)
```

```
/databricks/spark/python/pyspark/sql/context.py:117: FutureWarning: Deprecated in 3.0.0. Use SparkSession.builder.getOrCreate() instead.
  warnings.warn(
```

```
type(data)
```

```
Out[8]: pyspark.sql.dataframe.DataFrame
```

```
data.collect()
```

```
Out[9]: [Row(ID=462809, Gender='Male', Ever_Married='No', Age=22, Graduated='No', Profession='Healthcare', Work_Experience=1.0, Spending_Score='Low', Family_Size=4.0, Var_1='Cat_
4', Segmentation='D'),
 Row(ID=462643, Gender='Female', Ever_Married='Yes', Age=38, Graduated='Yes', Profession='Engineer', Work_Experience=None, Spending_Score='Average', Family_Size=3.0, Var_1='Cat_
4', Segmentation='A'),
 Row(ID=466315, Gender='Female', Ever_Married='Yes', Age=67, Graduated='Yes', Profession='Engineer', Work_Experience=1.0, Spending_Score='Low', Family_Size=1.0, Var_1='Cat_6', Se
gmentation='B'),
 Row(ID=461735, Gender='Male', Ever_Married='Yes', Age=67, Graduated='Yes', Profession='Lawyer', Work_Experience=0.0, Spending_Score='High', Family_Size=2.0, Var_1='Cat_6', Segme
ntation='B'),
 Row(ID=462669, Gender='Female', Ever_Married='Yes', Age=40, Graduated='Yes', Profession='Entertainment', Work_Experience=None, Spending_Score='High', Family_Size=6.0, Var_1='Cat
_6', Segmentation='A'),
 Row(ID=461319, Gender='Male', Ever_Married='Yes', Age=56, Graduated='No', Profession='Artist', Work_Experience=0.0, Spending_Score='Average', Family_Size=2.0, Var_1='Cat_6', Seg
mentation='C'),
 Row(ID=460156, Gender='Male', Ever_Married='No', Age=32, Graduated='Yes', Profession='Healthcare', Work_Experience=1.0, Spending_Score='Low', Family_Size=3.0, Var_1='Cat_6', Seg
mentation='C'),
 Row(ID=464347, Gender='Female', Ever_Married='No', Age=33, Graduated='Yes', Profession='Healthcare', Work_Experience=1.0, Spending_Score='Low', Family_Size=3.0, Var_1='Cat_6', S
egmentation='D'),
 Row(ID=465015, Gender='Female', Ever_Married='Yes', Age=61, Graduated='Yes', Profession='Engineer', Work_Experience=0.0, Spending_Score='Low', Family_Size=3.0, Var_1='Cat_7', Se
gmentation='D'),
 Row(ID=465176, Gender='Female', Ever_Married='Yes', Age=55, Graduated='Yes', Profession='Artist', Work_Experience=1.0, Spending_Score='Average', Family_Size=4.0, Var_1='Cat_6',
Segmentation='C'),
 Row(ID=464041, Gender='Female', Ever_Married='No', Age=26, Graduated='Yes', Profession='Engineer', Work_Experience=1.0, Spending_Score='Low', Family_Size=3.0, Var_1='Cat_6', Seg
```

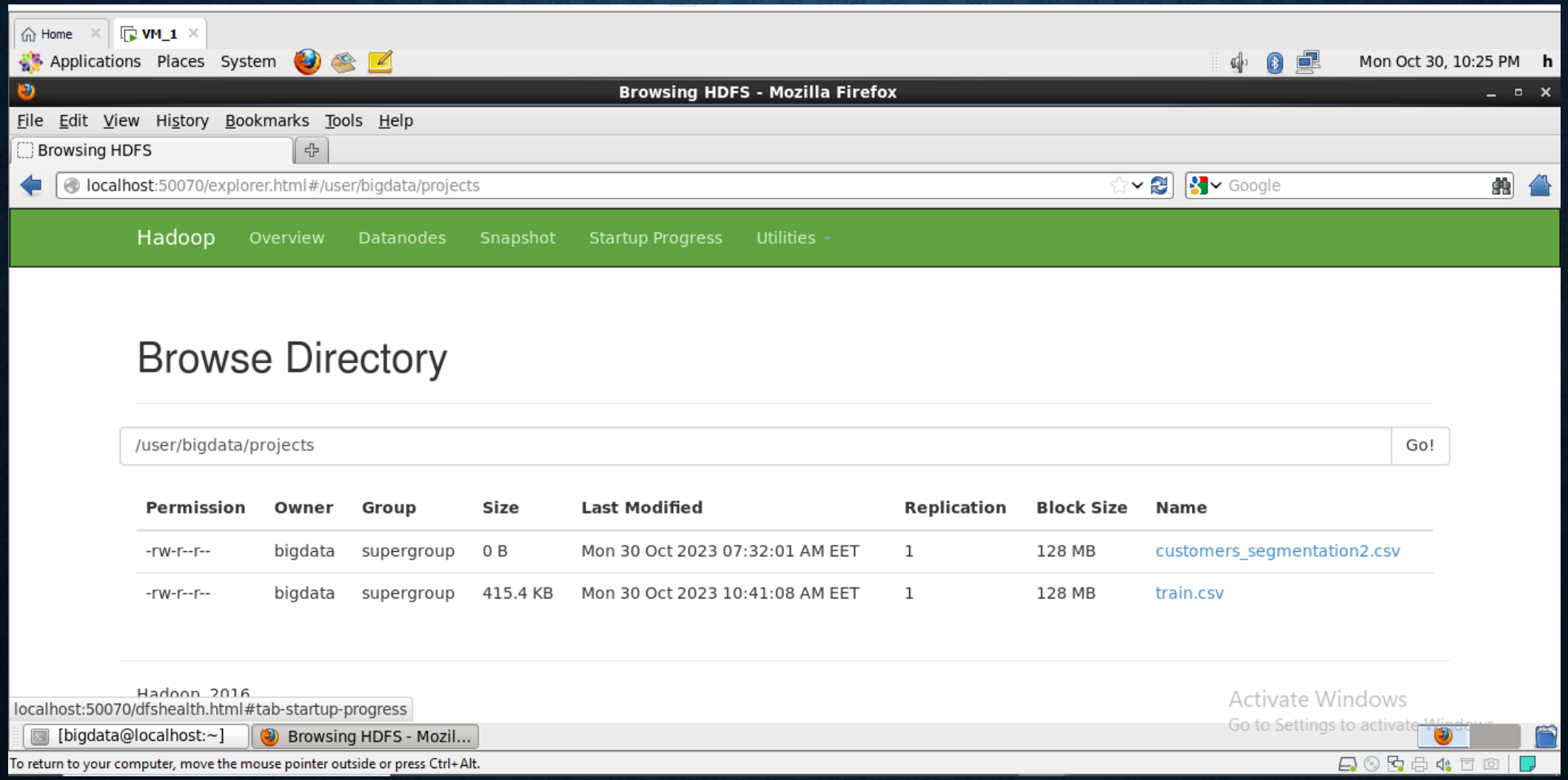
HDFS

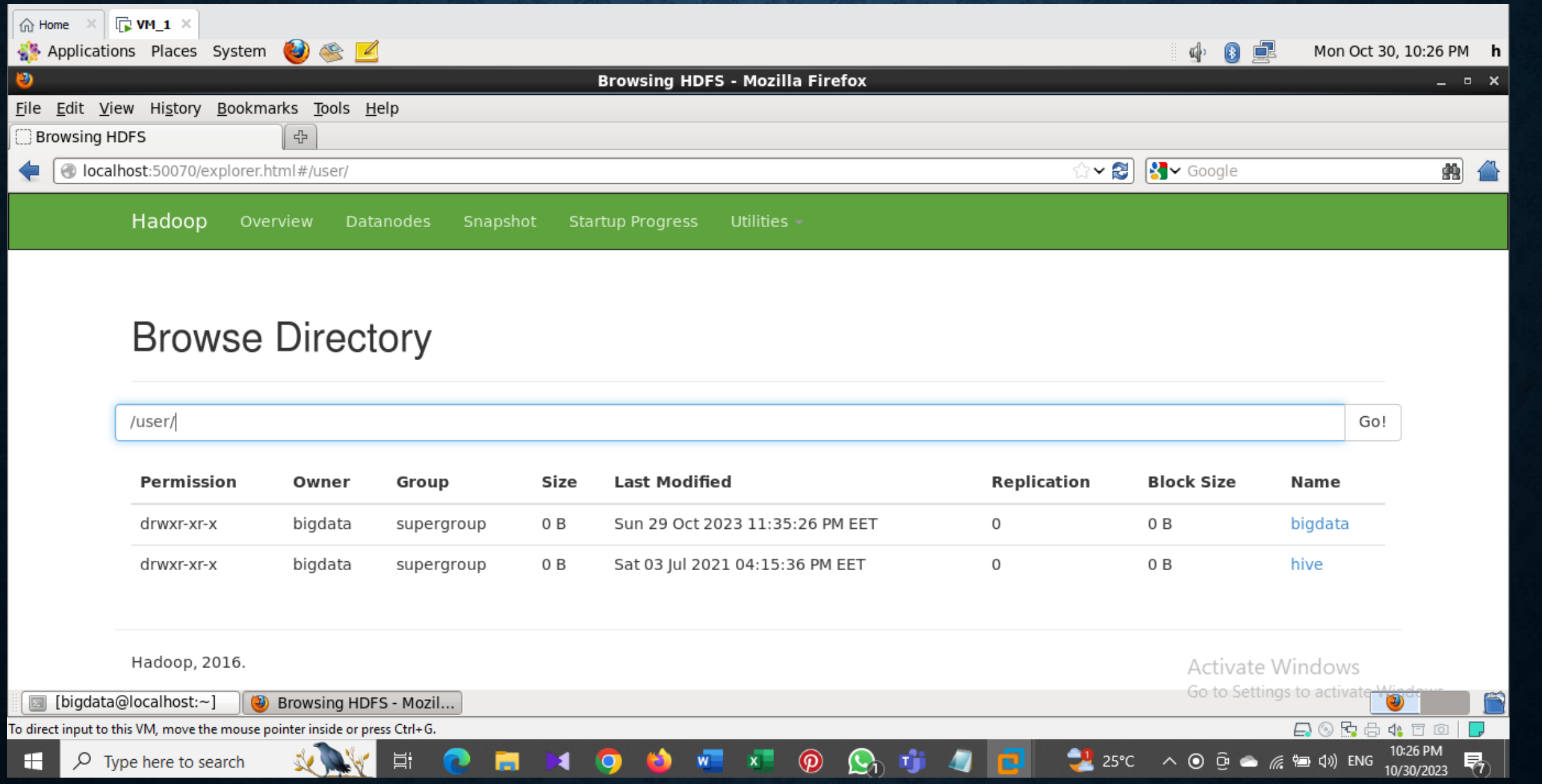
```
cd
start-all.sh
 hadoop fs -cat /user/bigdata/projects/customers_segmentation2.csv
hadoop fs -ls /user/bigdata/projects/
hadoop fs -ls /user/bigdata/
hadoop fs -ls /user
hadoop fs -ls /projects/
 hdfs dfs -ls
hdfs dfs -mkdir /user/bigdata/projects/
hadoop fs -touch /user/bigdata/projects/customers_segmentation2.csv
hdfs dfs -cat /user/bigdata/projects/customers_segmentation2.csv | head -n 5
```

HIVE

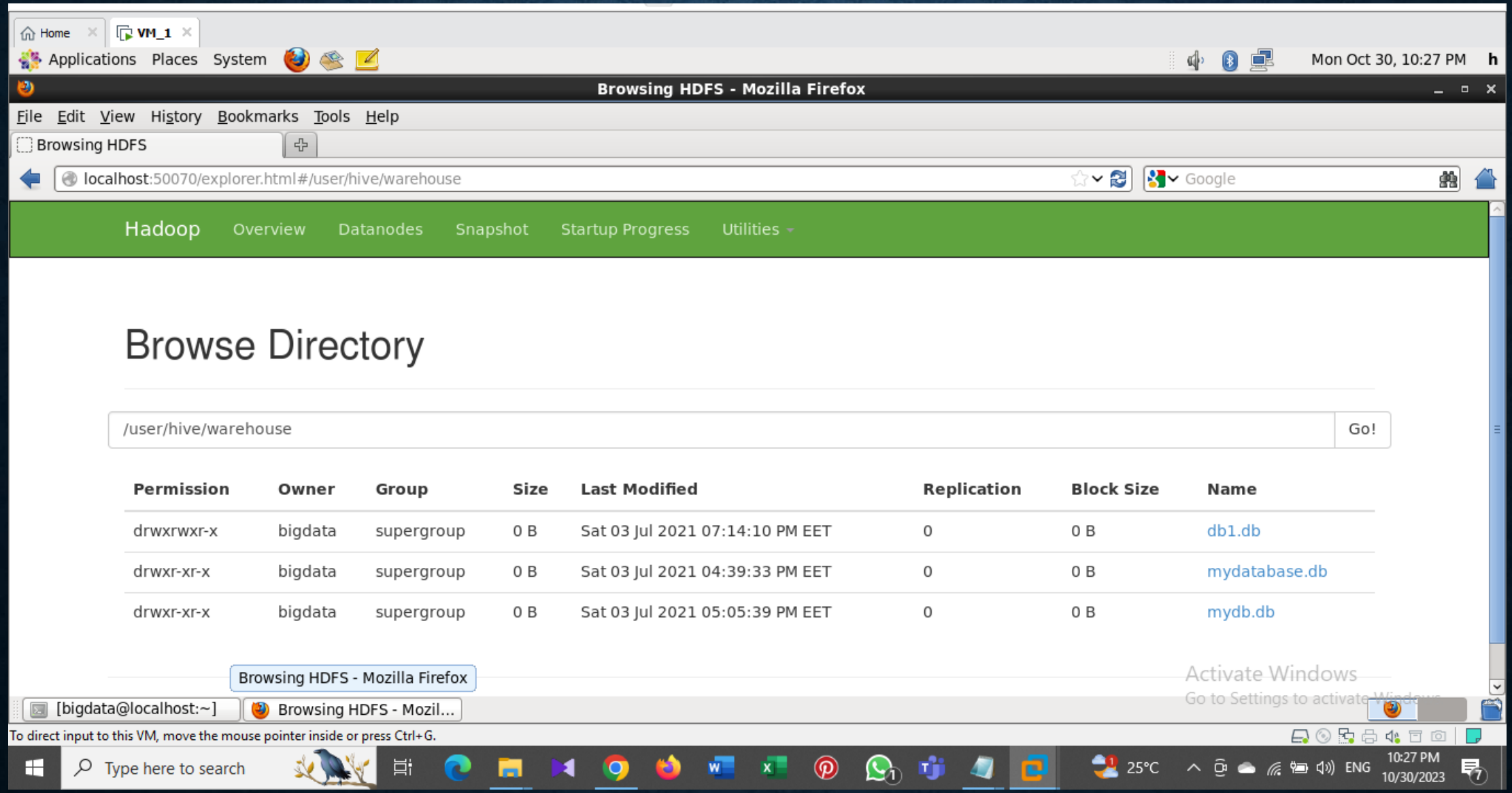
```
hive

hadoop fs -ls /user/hive/

create external table customers(id int, gender varchar(5),ever_married
varchar(8),age string,

graduated boolean,work_experience int,spending_score
varchar(10),family_size int,var_1 text,segmentation varchar(4))

    row format delimited fields terminated by ',' stored as textfile
location'/home/bigdata/customers';


hadoop fs -copyFromLocal /user/bigdata/projects/train.csv /user/hive

SELECT * FROM customers;

show databases;

show tables;

use default;
```

HDFS Localhost

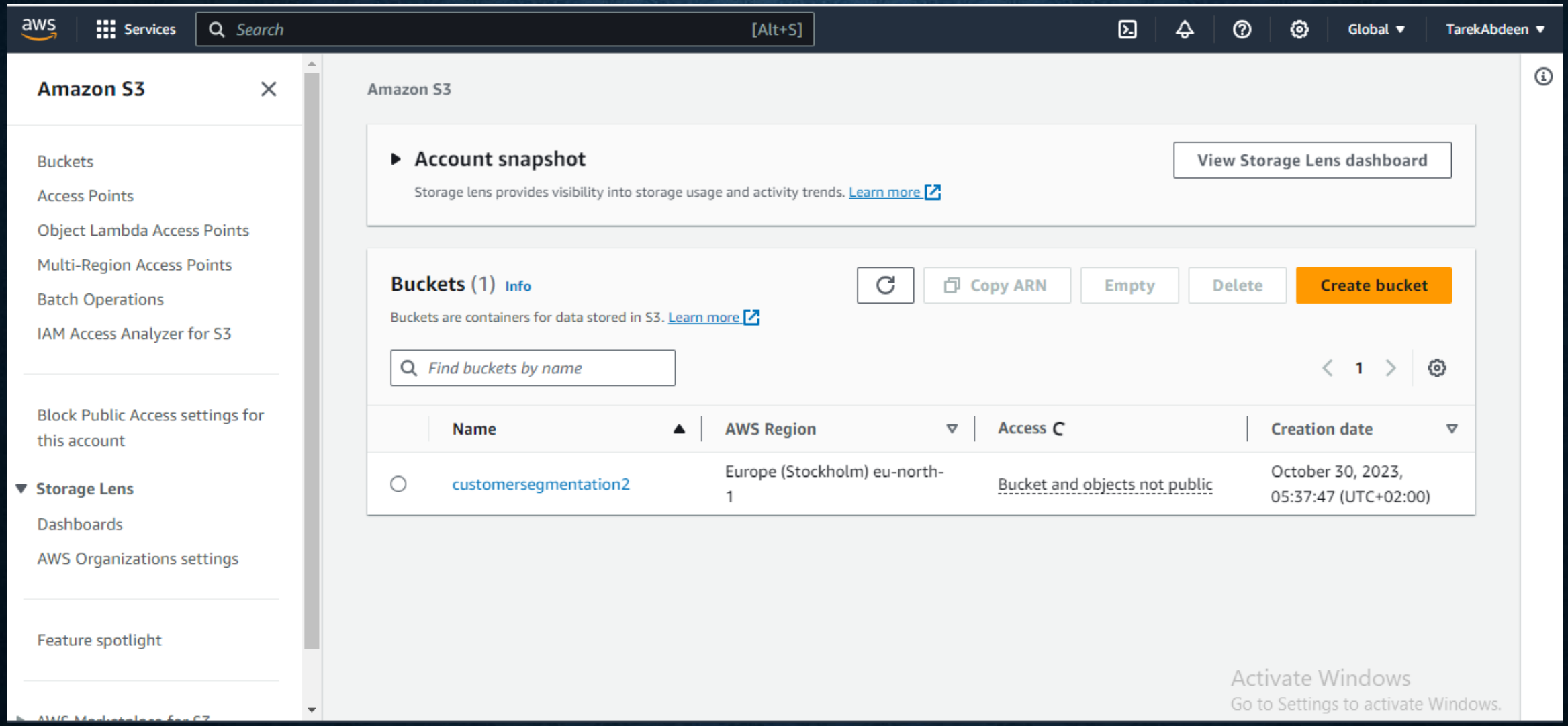


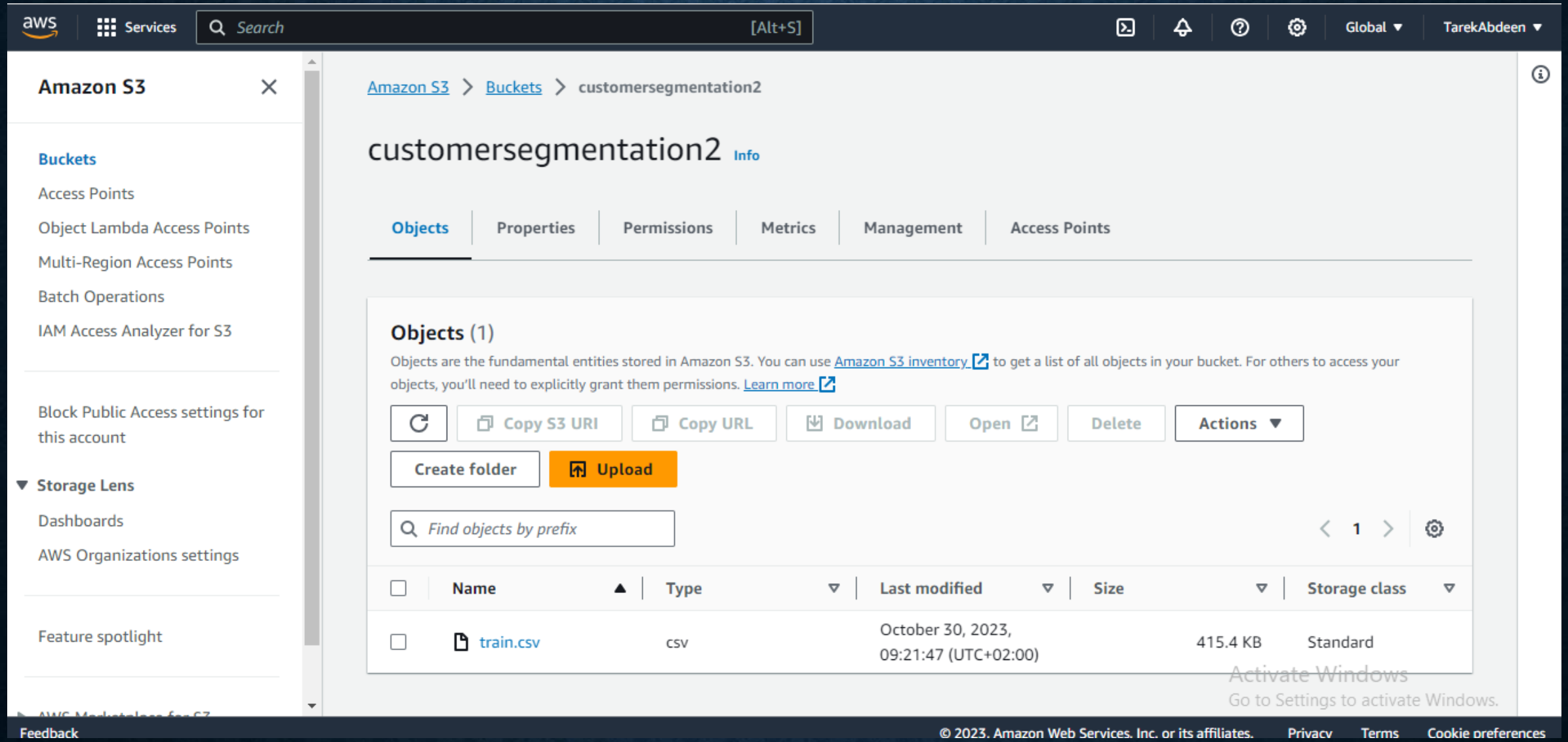

Home
VM_1
Applications Places System
Mon Oct 30, 10:26 PM
Browsing HDFS - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Browsing HDFS
localhost:50070/explorer.html#/user/
Google
Hadoop Overview Datanodes Snapshot Startup Progress Utilities
Browse Directory
/user/
Go!
Permission Owner Group Size Last Modified Replication Block Size Name
drwxr-xr-x bigdata supergroup 0 B Sun 29 Oct 2023 11:35:26 PM EET 0 0 B bigdata
drwxr-xr-x bigdata supergroup 0 B Sat 03 Jul 2021 04:15:36 PM EET 0 0 B hive
Hadoop, 2016.
Activate Windows
Go to Settings to activate Windows.
[bigdata@localhost:~]
Browsing HDFS - Mozil...
To direct input to this VM, move the mouse pointer inside or press Ctrl+G.
Type here to search
25°C
ENG
10:26 PM
10/30/2023

HIVE Localhost



S3

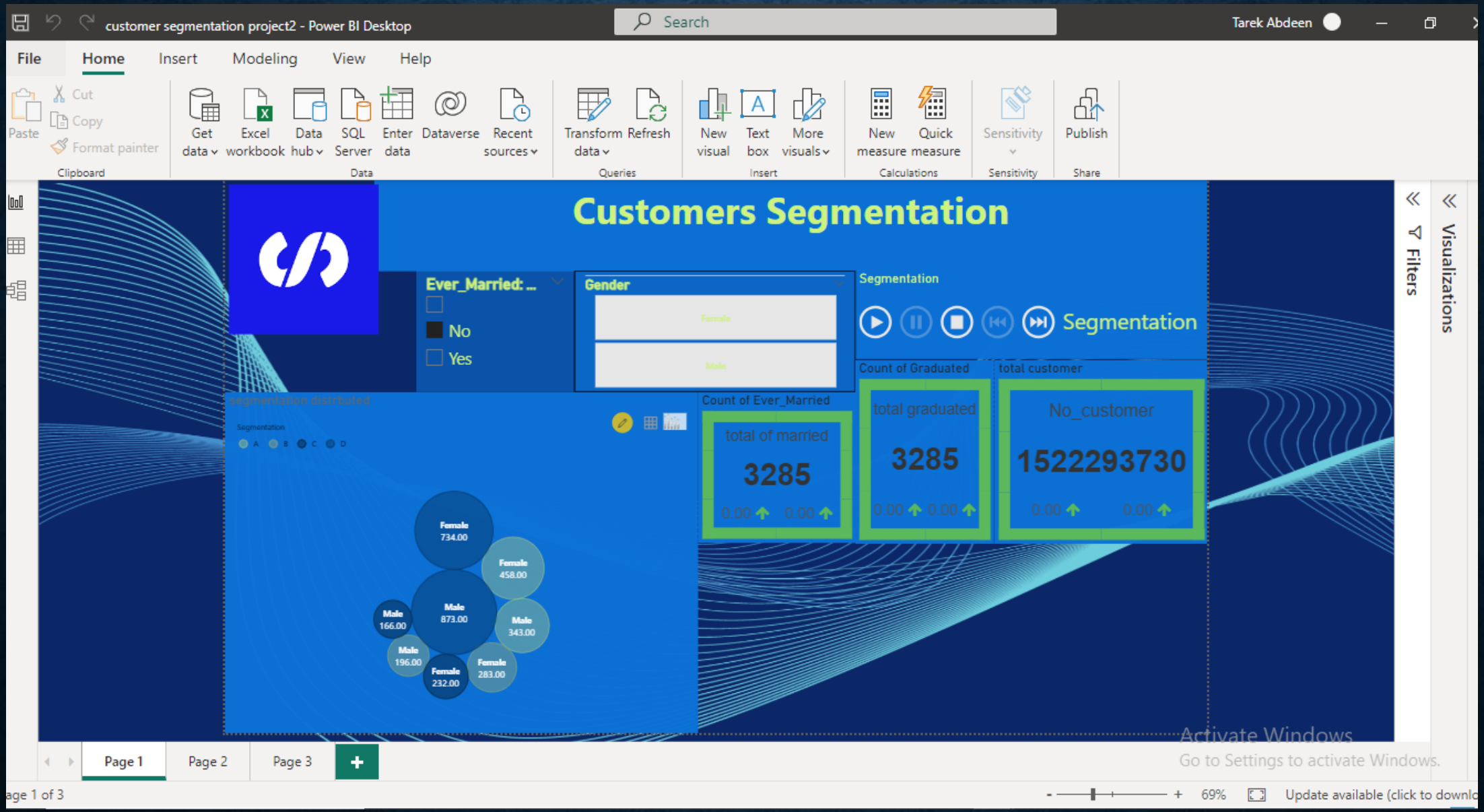

aws
Services
Search
[Alt+S]
Global
TarekAbdeen
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
Amazon S3
Account snapshot
Storage lens provides visibility into storage usage and activity trends. Learn more
View Storage Lens dashboard
Buckets (1) Info
Buckets are containers for data stored in S3. Learn more
Copy ARN
Empty
Delete
Create bucket
Find buckets by name
1
Name
AWS Region
Access
Creation date
customersegmentation2
Europe (Stockholm) eu-north-1
Bucket and objects not public
October 30, 2023, 05:37:47 (UTC+02:00)
Activate Windows
Go to Settings to activate Windows.

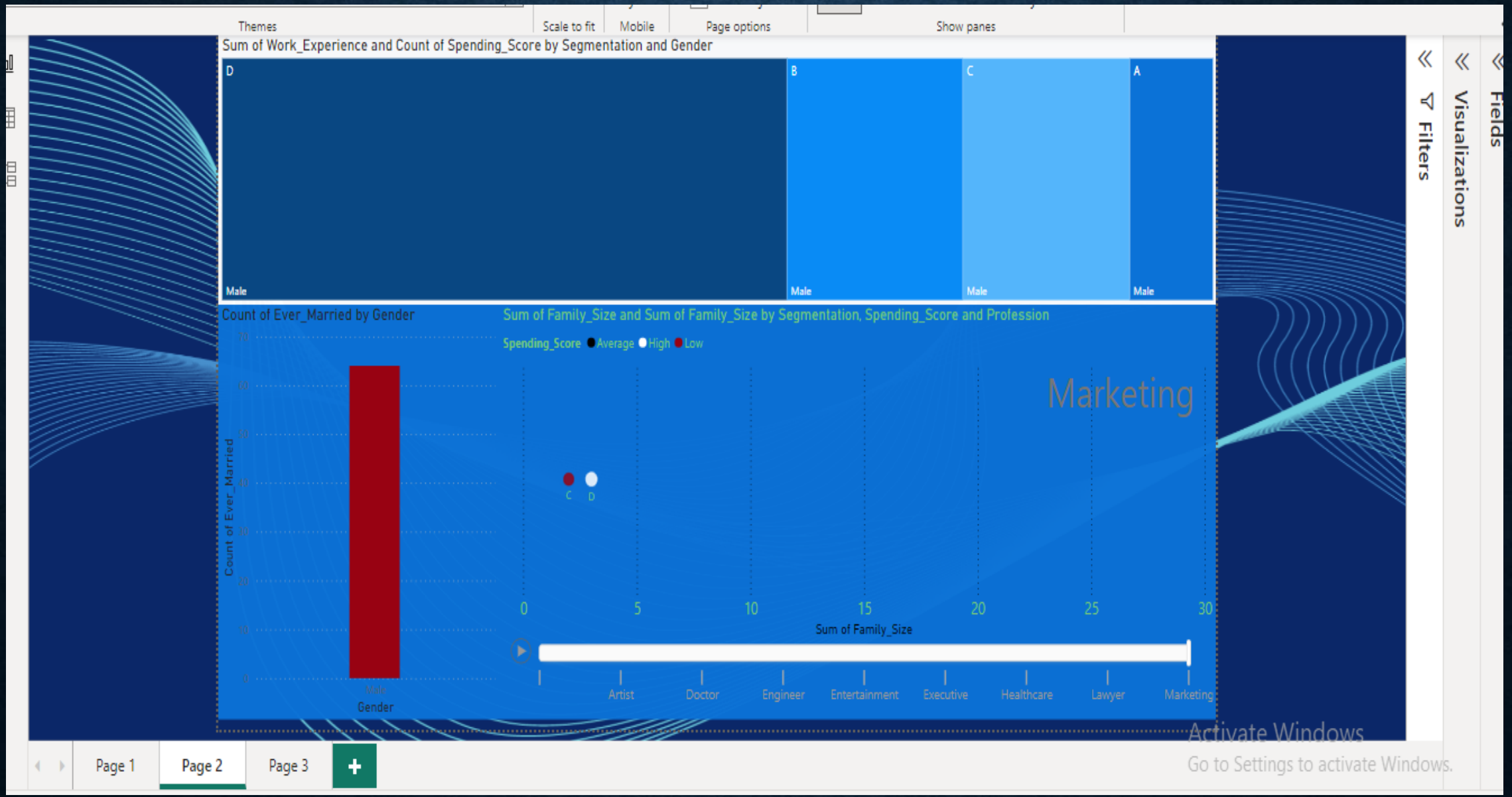

aws
Services
Search
[Alt+S]
Global
TarekAbdeen
Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
Amazon S3 > Buckets > customersegmentation2
customersegmentation2 Info
Objects
Properties
Permissions
Metrics
Management
Access Points
Objects (1)
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
1
Name
Type
Last modified
Size
Storage class
train.csv
csv
October 30, 2023, 09:21:47 (UTC+02:00)
415.4 KB
Standard
Activate Windows
Go to Settings to activate Windows.
Feedback
© 2023, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

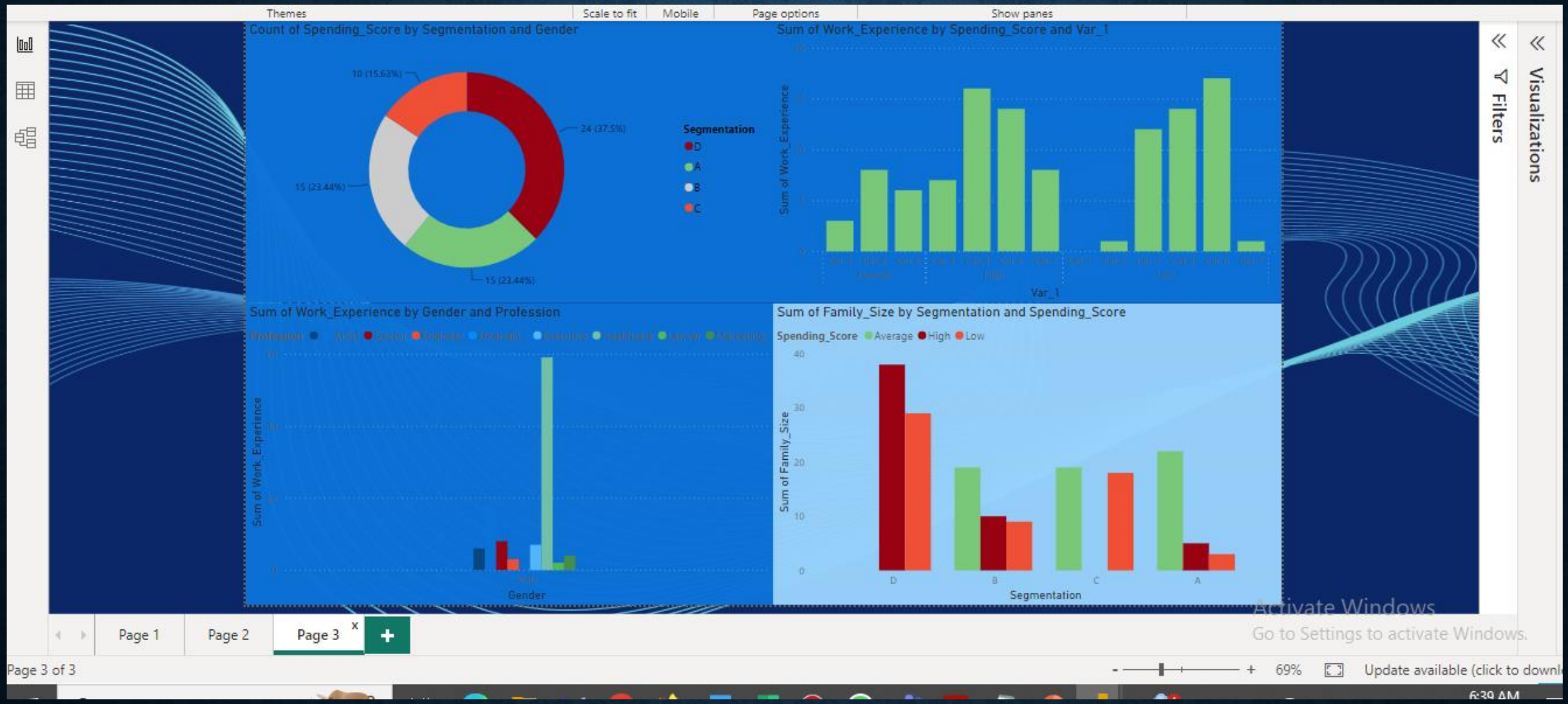
Power BI-DAX

```
female graduated = CALCULATE(SUM(train[Graduated]) , train[Gender] = "female")
male graduated = CALCULATE(SUM(train[Graduated]) , train[Gender] = "male")
single = CALCULATE(SUM(train[Ever_Married]), train[Ever_Married] = "no")
total customer = SUM((train[ID]))
total customer from female = CALCULATE([total customer] , train[Gender] =
"female")
total customer from male = CALCULATE([total customer], train[Gender]="male")
total graduated = SUM(train[Graduated])
total married = CALCULATE(SUM(train[Ever_Married]), train[Ever_Married] =
"yes")
total_gender = SUM((train[Gender]))
```

Power BI


customer segmentation project2 - Power BI Desktop
Search
Tarek Abdeen
File
Home
Insert
Modeling
View
Help
Paste
Cut
Copy
Format painter
Clipboard
Get data
Excel workbook
Data hub
SQL Server
Enter data
Dataverse
Recent sources
Data
Transform data
Refresh
Queries
New visual
Text box
More visuals
Insert
New measure
Quick measure
Calculations
Sensitivity
Sensitivity
Publish
Share
Customers Segmentation
Ever_Married: ...
No
Yes
Gender
Female
Male
Segmentation
Segmentation
Count of Graduated
total customer
segmentation distributed
Segmentation
A
B
C
D
Count of Ever_Married
total of married
3285
total graduated
3285
No_customer
1522293730
Female 734.00
Female 458.00
Male 873.00
Male 166.00
Male 343.00
Male 196.00
Female 232.00
Female 283.00
Filters
Visualizations
Page 1
Page 2
Page 3
Activate Windows
Go to Settings to activate Windows.
age 1 of 3
69%
Update available (click to downlo


Themes
Scale to fit
Mobile
Page options
Show panes
Sum of Work_Experience and Count of Spending_Score by Segmentation and Gender
D
B
C
A
Male
Male
Male
Male
Count of Ever_Married by Gender
Sum of Family_Size and Sum of Family_Size by Segmentation, Spending_Score and Profession
Spending_Score
Average
High
Low
Marketing
Count of Ever_Married
Male
Gender
C
D
0
5
10
15
20
25
30
Sum of Family_Size
Artist
Doctor
Engineer
Entertainment
Executive
Healthcare
Lawyer
Marketing
Filters
Visualizations
Fields
Page 1
Page 2
Page 3
Activate Windows
Go to Settings to activate Windows.


Themes
Scale to fit
Mobile
Page options
Show panes
Count of Spending_Score by Segmentation and Gender
10 (15.63%)
24 (37.5%)
15 (23.44%)
15 (23.44%)
Segmentation
D
A
B
C
Sum of Work_Experience by Spending_Score and Var_1
Sum of Work_Experience
Var_1
Sum of Work_Experience by Gender and Profession
Profession
Sum of Work_Experience
Gender
Sum of Family_Size by Segmentation and Spending_Score
Spending_Score
Average
High
Low
Sum of Family_Size
Segmentation
Filters
Visualizations
Page 1
Page 2
Page 3
Page 3 of 3
69%
Update available (click to downl
Activate Windows
Go to Settings to activate Windows.
6:39 AM

Links

databricks link :

https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/6296595295968836/3140546493539922/8239006622875743/latest.html

##s3 link:

https://s3.console.aws.amazon.com/s3/home?region=eu-north-1#

##power bi project link :

https://app.powerbi.com/links/GXPf_cZ3Ft?ctid=7a5fce51-6864-4873-8d87-4d24bbfc93f6&pbi_source=linkShare&bookmarkGuid=6e118312-4261-412c-b84b-67603c23378d

problems

1-Why, when I finished writing the code on pyspark and wanted to send the file to hdfs, the code did not give anything at all, but it was only running and did not finish it.

2- I worked on S3 and it was fine, but

I stopped when I wanted to transfer data from it or why in the data sync part

Because I don't know how to do the activation key

THANK YOU