DATA ANALYSIS FOR ONLINE SALE

CLOUD COMPUTING AND BIG DATA

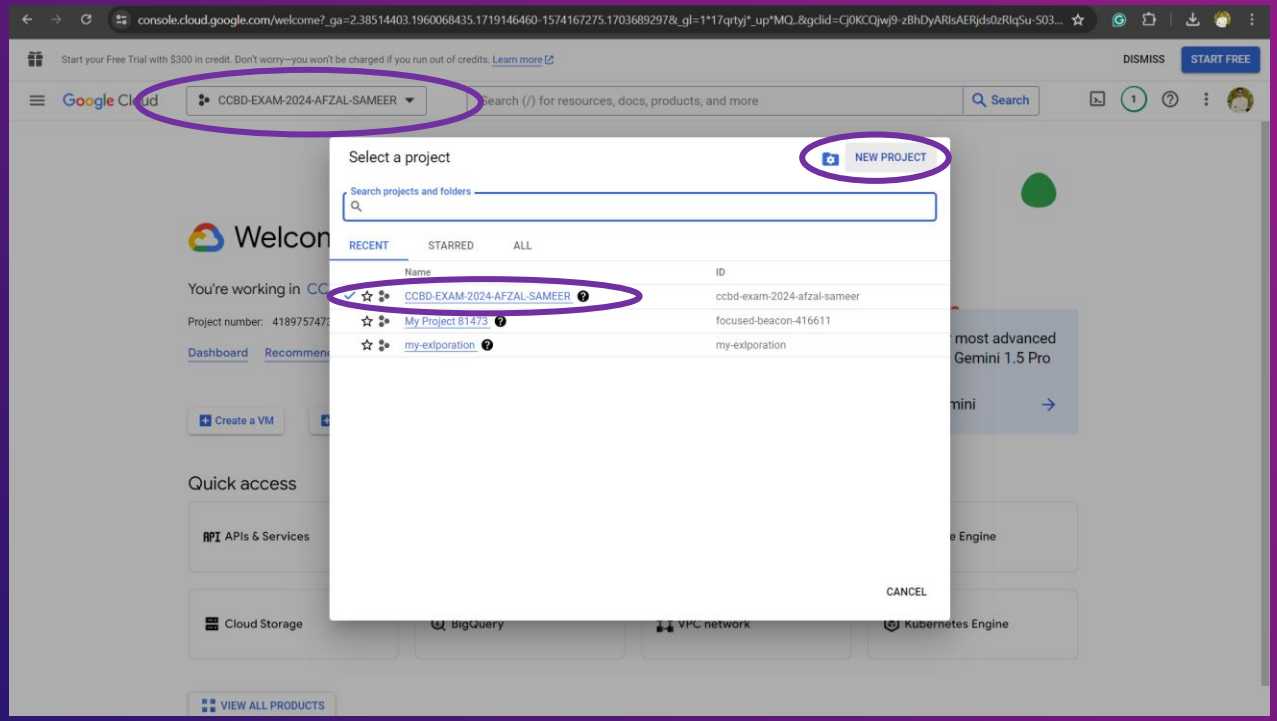
SAMEER AFZAL
1000053143

Professor: Giuseppe PAPPALARDO
Professor: NICOTRA SALVATORE

TABLE OF CONTENT

CLOUD SETUP

CREATE A GOOGLE CLOUD PROJECT



CREATE A CLOUD STORAGE BUCKET

console.cloud.google.com/storage/browser?project=ccbd-exam-2024-afzal-sameer&prefix=&forceOnBucketsSortingFiltering=true

Start your Free Trial with $300 in credit. Don't worry—you won't be charged if you run out of credits. Learn more

DISMISS START FREE

Google Cloud CCBD-EXAM-2024-AFZAL-SAMEER Search (/) for resources, docs, products, and more Search

Buckets CREATE REFRESH GO TO PATH LEARN

Beginning on April 29th, 2024 at-scale policy analysis and advanced IAM recommendation capabilities will require Security Command Center Premium. Learn more DISMISS

Soft Delete is now available

When enabled on a bucket, Soft Delete protects objects from accidental or malicious deletions. A 7 day Soft Delete policy has been added to all existing buckets. This is the default policy for new buckets.

LEARN MORE

Transfer New

Power near real-time analytics and replication with event-driven transfers

You can now capture changes faster at your Google Cloud Storage and Amazon S3 sources via event-driven transfers, enabling you to act on your data in near real time. To get started, create transfer job with a Pub/Sub- or AWS SQS-based event stream configured to send event notific when objects are created or updated.

CREATE TRANSFER JOB LEARN MORE

Filter Filter buckets

Name	Created	Location type	Location	Default storage class	Last modified	Public access	Access control	Protection
ccbd-exam-2023-afzal-sameer	Jun 21, 2024, 2:43:48 PM	Multi-region	eu	Standard	Jun 21, 2024, 2:43:48 PM	Not public	Uniform	Soft Delete

DATA INGESTION

https://www.kaggle.com/datasets/shreyanshverma27/online-sales-dataset-popular-marketplace-data

UPLOAD DATA TO G.C.S PREVIOUSLY DEFINED

console.cloud.google.com/storage/browser/ccbd-exam-2023-afzal-sameer;tab=objects?forceOnBucketsSortingFiltering=true&project=ccbd-exam-2024-afzal-sameer&prefix=&forc...

Start your Free Trial with $300 in credit. Don't worry—you won't be charged if you run out of credits. Learn more

DISMISS START FREE

Google Cloud | CCBD-EXAM-2024-AFZAL-SAMEER | Search (/) for resources, docs, products, and more | Search

Cloud Storage

Buckets

Monitoring

Settings

Bucket details

GO TO PATH REFRESH LEARN

ccbd-exam-2023-afzal-sameer

Location	Storage class	Public access	Protection
eu (multiple regions in European Union)	Standard	Not public	Soft Delete

OBJECTS CONFIGURATION PERMISSIONS PROTECTION LIFECYCLE OBSERVABILITY INVENTORY REPORTS OPERATIONS

Folder browser

ccbd-exam-2023-afzal-sameer

Buckets > ccbd-exam-2023-afzal-sameer

UPLOAD FILES UPLOAD FOLDER CREATE FOLDER TRANSFER DATA MANAGE HOLDS EDIT RETENTION DOWNLOAD

Filter by name prefix only | Filter | Filter objects and folders | Show Live objects only

Name	Size	Type	Created	Storage class	Last modified
OnlineSalesData.csv	2[illegible] KB	text/csv	Jun 21, 2024, 3:14:48 PM	Standard	Jun 21, 2024, 3:14:48

1 file successfully uploaded

Uploads and CCBD-EXAM-2024-AFZA... operations

OnlineSalesData.csv Complete

Marketplace

Release Notes

USE CLOUD SHELL TO LIST FILE UPLOADED

shell.cloud.google.com/?hl=en_US&fromcloudshell=true&show=terminal

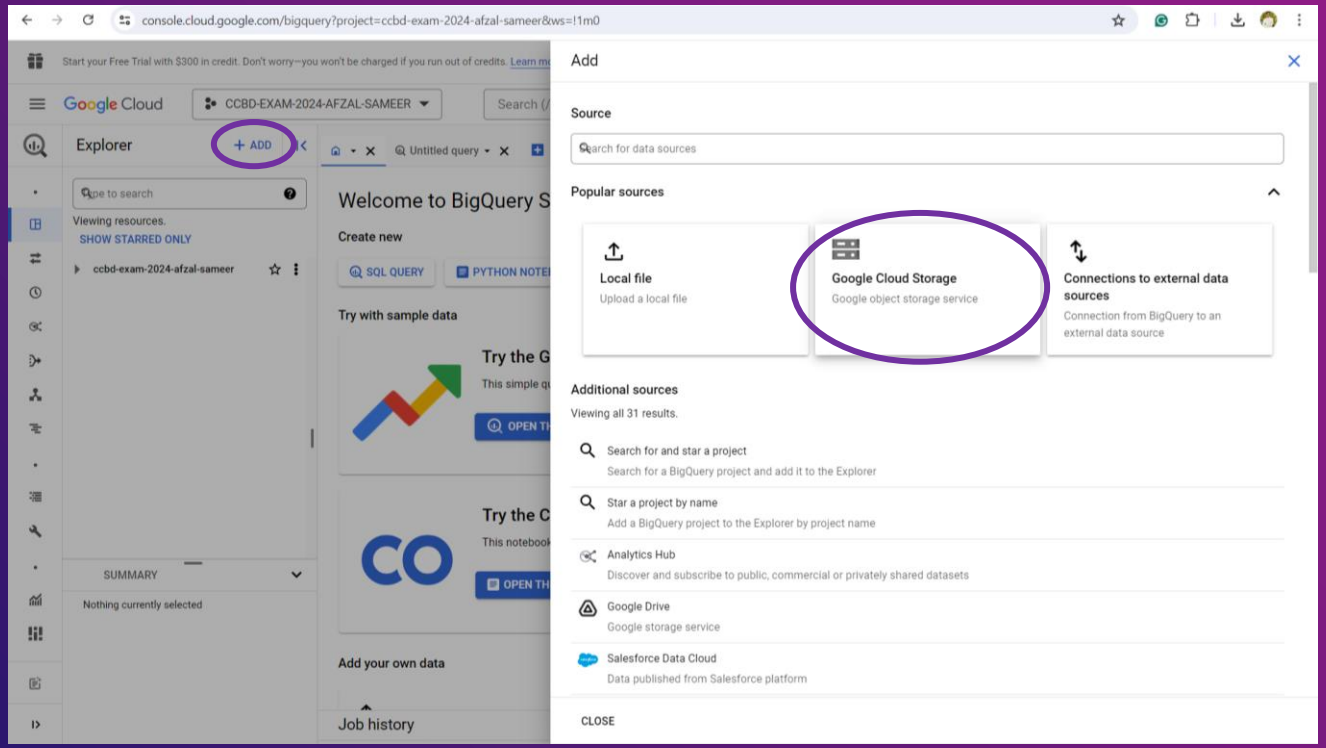
Cloud Shell Editor

(ccbd-exam-2024-afzal-sameer)

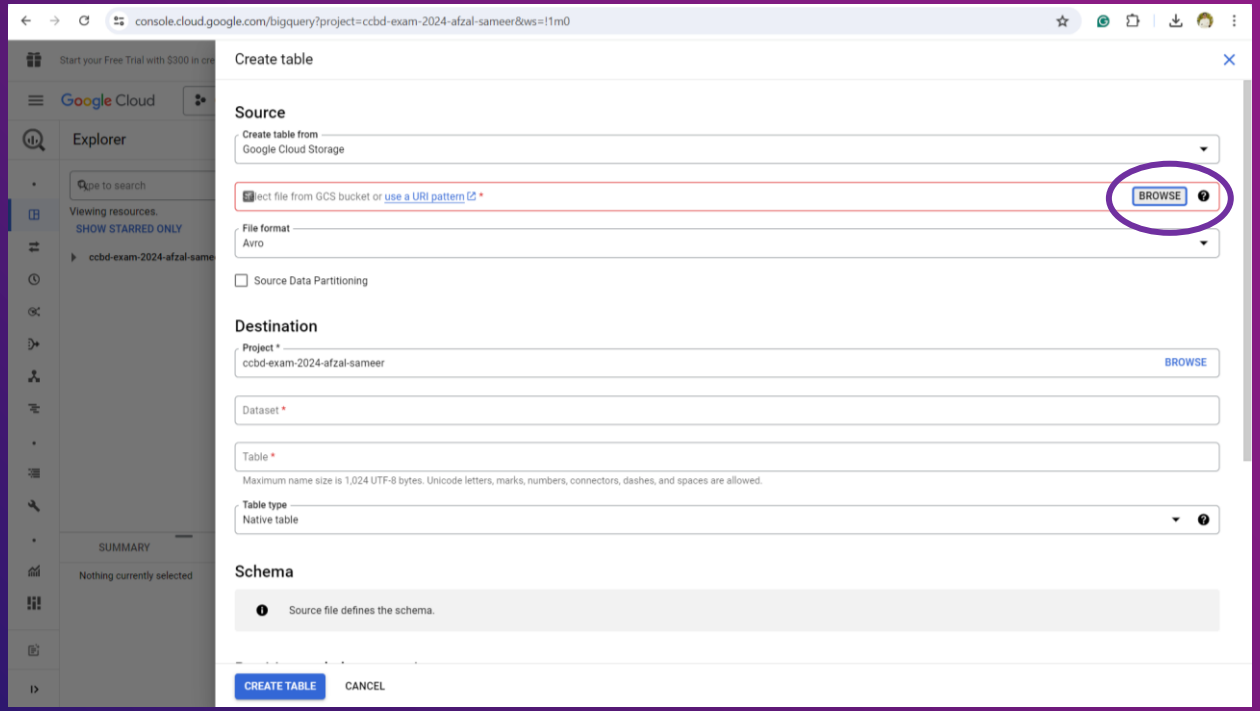
```
Welcome to Cloud Shell! Type "help" to get started.
Your Cloud Platform project in this session is set to ccbd-exam-2024-afzal-sameer.
Use "gcloud config set project [PROJECT_ID]" to change to a different project.
afzal1998sameer@cloudshell:~ (ccbd-exam-2024-afzal-sameer)$ gsutil ls gs://ccbd-exam-2023-afzal-sameer/
gs://ccbd-exam-2023-afzal-sameer/OnlineSalesData.csv
afzal1998sameer@cloudshell:~ (ccbd-exam-2024-afzal-sameer)$ ~head OnlineSalesData.csv
-bash: ~head: command not found
afzal1998sameer@cloudshell:~ (ccbd-exam-2024-afzal-sameer)$ head OnlineSalesData.csv
head: cannot open 'OnlineSalesData.csv' for reading: No such file or directory
afzal1998sameer@cloudshell:~ (ccbd-exam-2024-afzal-sameer)$ gsutil cp gs://ccbd-exam-2023-afzal-samee/OnlineSalesData.csv
CommandException: Wrong number of arguments for "cp" command.
afzal1998sameer@cloudshell:~ (ccbd-exam-2024-afzal-sameer)$ gsutil cat gs://ccbd-exam-2023-afzal-samee/OnlineSalesData.csv | head
BucketNotFoundException: 404 gs://ccbd-exam-2023-afzal-samee bucket does not exist.
afzal1998sameer@cloudshell:~ (ccbd-exam-2024-afzal-sameer)$ gsutil cat gs://ccbd-exam-2023-afzal-sameer/OnlineSalesData.csv | head
Transaction ID,Date,Product Category,Product Name,Units Sold,Unit Price,Total Revenue,Region,Payment Method
10001,2024-01-01,Electronics,iPhone 14 Pro,2,999.99,1999.98,North America,Credit Card
10002,2024-01-02,Home Appliances,Dyson V11 Vacuum,1,499.99,499.99,Europe,PayPal
10003,2024-01-03,Clothing,Levi's 501 Jeans,3,69.99,209.97,Asia,Debit Card
10004,2024-01-04,Books,The Da Vinci Code,4,15.99,63.96,North America,Credit Card
10005,2024-01-05,Beauty Products,Neutrogena Skincare Set,1,89.99,89.99,Europe,PayPal
10006,2024-01-06,Sports,Wilson Evolution Basketball,5,29.99,149.95,Asia,Credit Card
10007,2024-01-07,Electronics,MacBook Pro 16-inch,1,2499.99,2499.99,North America,Credit Card
10008,2024-01-08,Home Appliances,Blueair Classic 480i,2,599.99,1199.98,Europe,PayPal
10009,2024-01-09,Clothing,Nike Air Force 1,6,89.99,539.94,Asia,Debit Card
afzal1998sameer@cloudshell:~ (ccbd-exam-2024-afzal-sameer)$
```

DATA MANIPULATION USING DATABASES

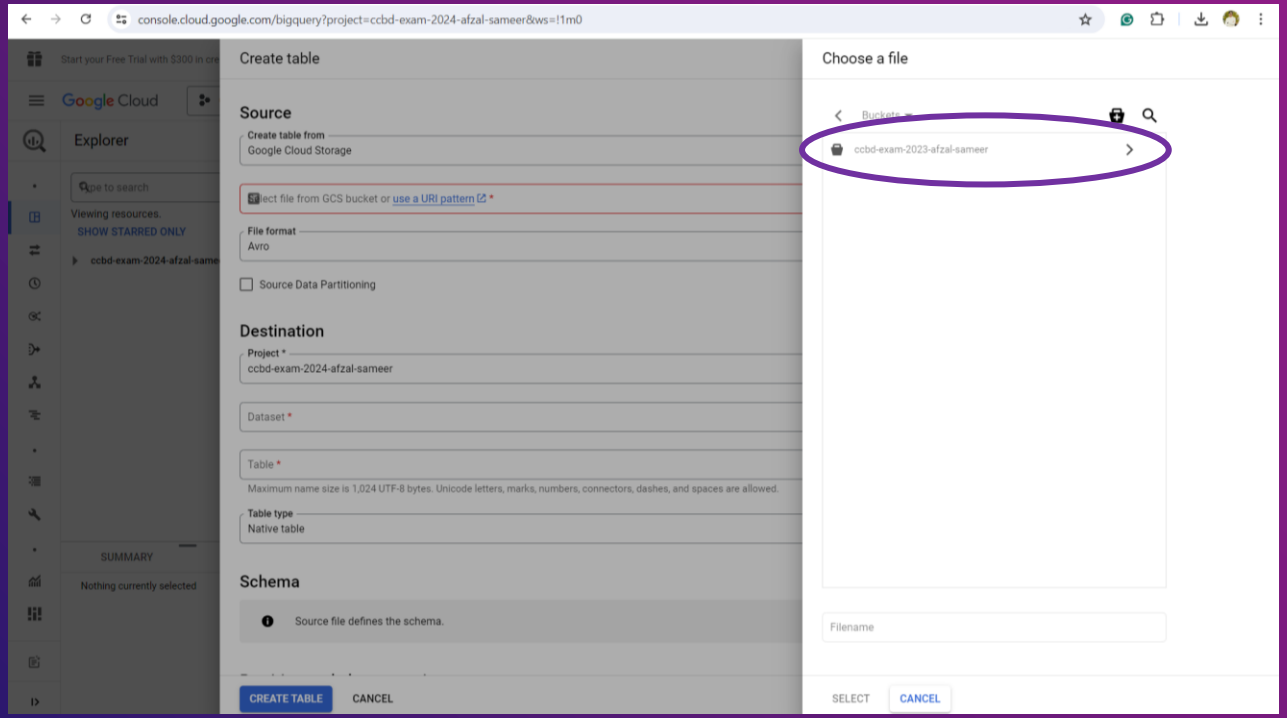
BIG QUERY WEB CONSOLE



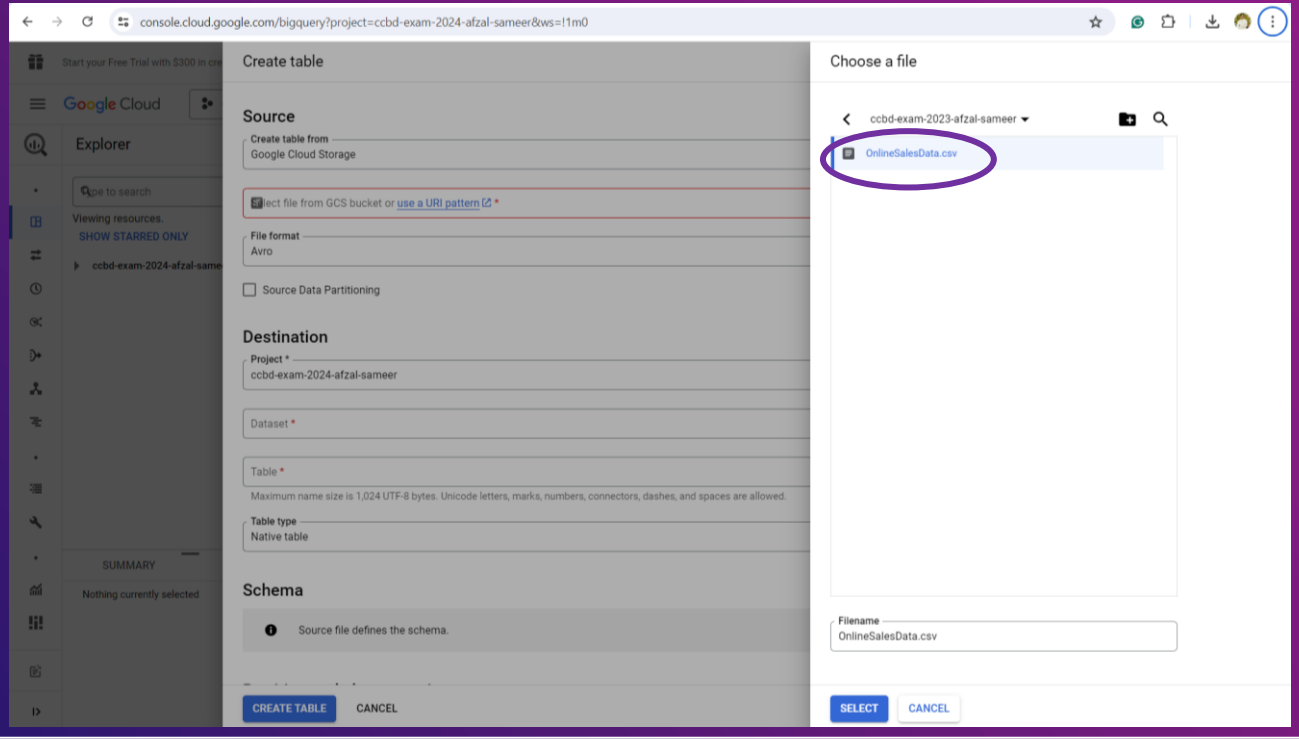
LOAD YOUR DATA FROM G.C.S INTO BIG QUERY



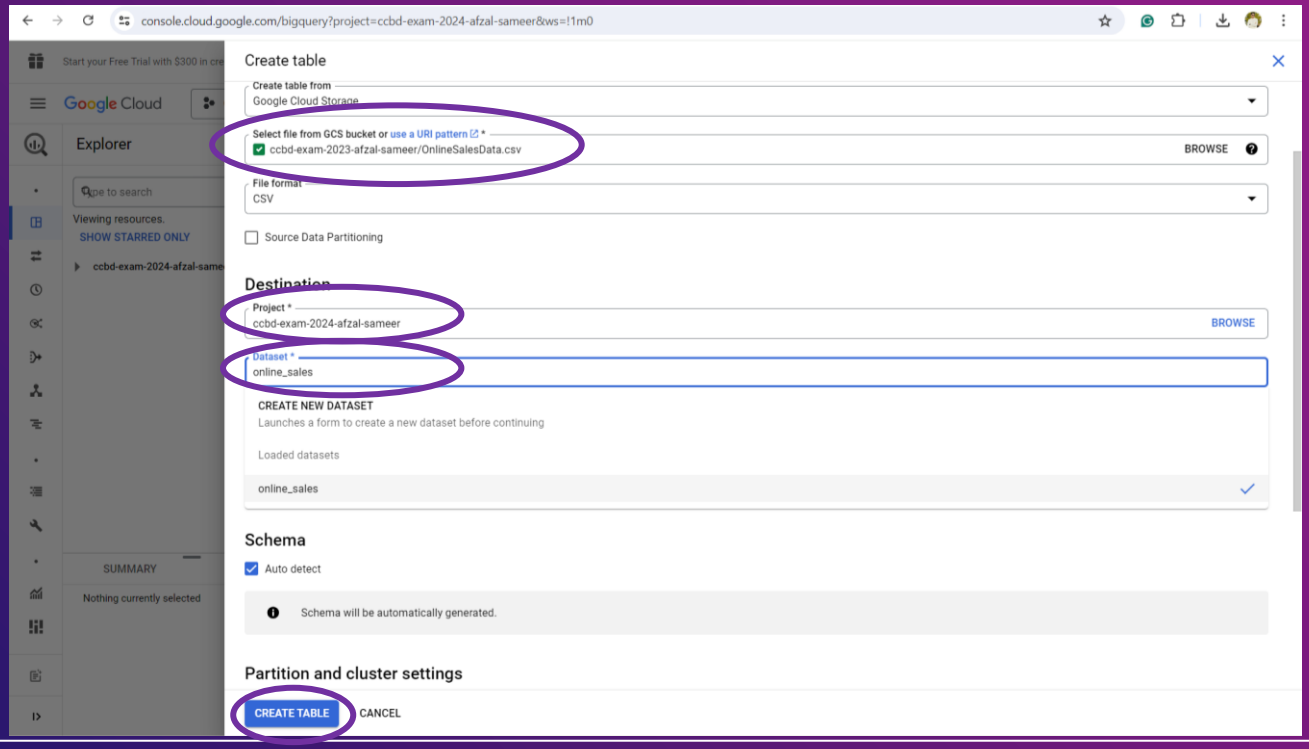
LOAD YOUR DATA FROM G.C.S INTO BIG QUERY



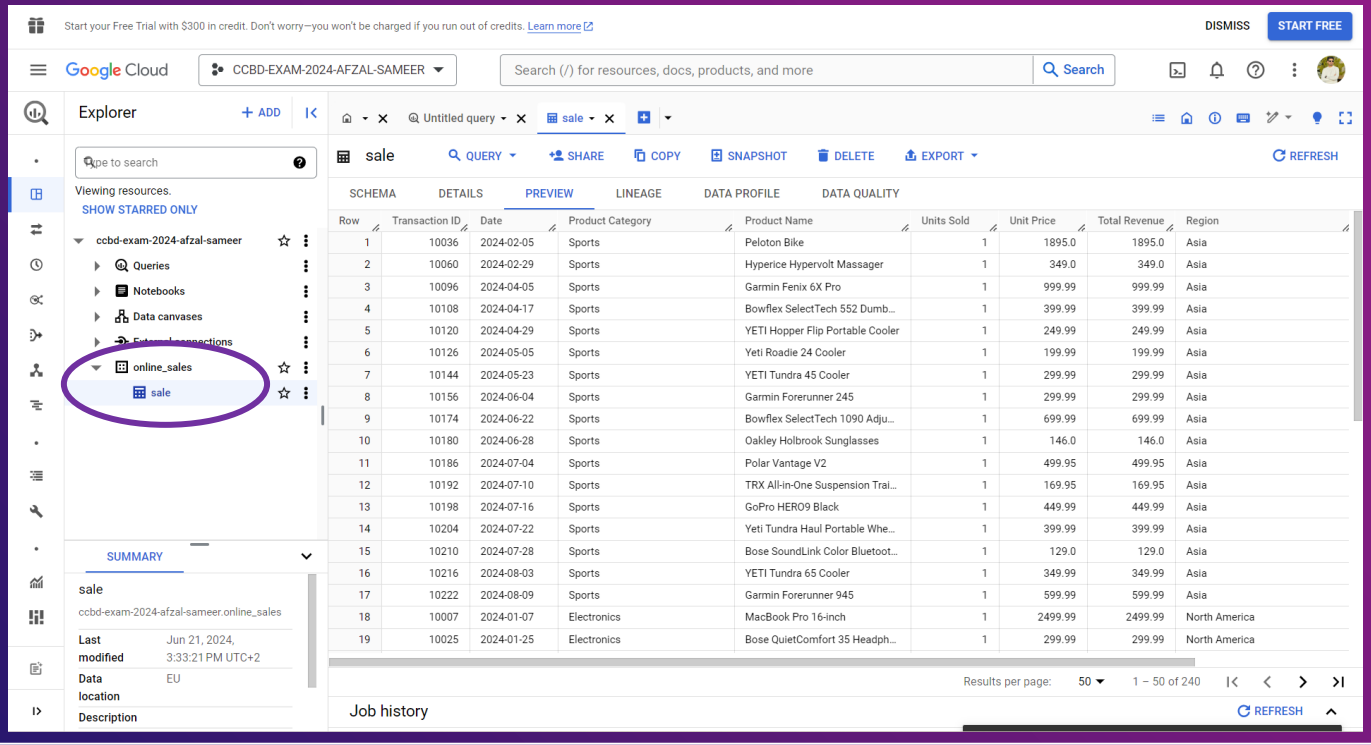
LOAD YOUR DATA FROM G.C.S INTO BIG QUERY



LOAD YOUR DATA FROM G.C.S INTO BIG QUERY



LOAD YOUR DATA FROM G.C.S INTO BIG QUERY



Row	Transaction ID	Date	Product Category	Product Name	Units Sold	Unit Price	Total Revenue	Region
1	10036	2024-02-05	Sports	Peloton Bike	1	1895.0	1895.0	Asia
2	10060	2024-02-29	Sports	Hyperice Hypervolt Massager	1	349.0	349.0	Asia
3	10096	2024-04-05	Sports	Garmin Fenix 6X Pro	1	999.99	999.99	Asia
4	10108	2024-04-17	Sports	Bowflex SelectTech 552 Dumb...	1	399.99	399.99	Asia
5	10120	2024-04-29	Sports	YETI Hopper Flip Portable Cooler	1	249.99	249.99	Asia
6	10126	2024-05-05	Sports	Yeti Roadie 24 Cooler	1	199.99	199.99	Asia
7	10144	2024-05-23	Sports	YETI Tundra 45 Cooler	1	299.99	299.99	Asia
8	10156	2024-06-04	Sports	Garmin Forerunner 245	1	299.99	299.99	Asia
9	10174	2024-06-22	Sports	Bowflex SelectTech 1090 Adju...	1	699.99	699.99	Asia
10	10180	2024-06-28	Sports	Oakley Holbrook Sunglasses	1	146.0	146.0	Asia
11	10186	2024-07-04	Sports	Polar Vantage V2	1	499.95	499.95	Asia
12	10192	2024-07-10	Sports	TRX All-in-One Suspension Trai...	1	169.95	169.95	Asia
13	10198	2024-07-16	Sports	GoPro HERO9 Black	1	449.99	449.99	Asia
14	10204	2024-07-22	Sports	Yeti Tundra Haul Portable Whe...	1	399.99	399.99	Asia
15	10210	2024-07-28	Sports	Bose SoundLink Color Bluetoot...	1	129.0	129.0	Asia
16	10216	2024-08-03	Sports	YETI Tundra 65 Cooler	1	349.99	349.99	Asia
17	10222	2024-08-09	Sports	Garmin Forerunner 945	1	599.99	599.99	Asia
18	10007	2024-01-07	Electronics	MacBook Pro 16-inch	1	2499.99	2499.99	North America
19	10025	2024-01-25	Electronics	Bose QuietComfort 35 Headph...	1	299.99	299.99	North America

BUSINESS QUESTIONS

01

TOTAL REVENUE BY PRODUCT CATEGORY

What is the total revenue generated by each product category, and which product category has the highest total revenue?

02

MOST POPULAR PAYMENT METHOD BY REGION

What is the most popular payment method in each region?

03

TOTAL REVENUE BY YEAR AND MONTH

What is the total revenue generated in each month of each year, and how does revenue trend over time?

04

UNITS SOLD BY PRODUCT AND REGION

How many units of each product were sold in each region, and which product has the highest number of units sold in each region?

05

AVERAGE UNIT PRICE BY PRODUCT CATEGORY

What is the average unit price of products in each product category, and which category has the highest average unit price?

EXECUTE QUERIES ON BIG QUERY UI (WEB CONSOLE) (1/5)

TOTAL REVENUE BY PRODUCT CATEGORY

```sql
SELECT
  `Product Category`,
  SUM(`Total Revenue`) AS `Total Revenue`
FROM
  `online_sales.sale`
GROUP BY
  `Product Category`
ORDER BY
  `Total Revenue` DESC;
```

Row	Product Category	Total Revenue
1	Electronics	34982.41
2	Home Appliances	18646.15999999...
3	Sports	14326.51999999...
4	Clothing	8128.929999999...
5	Beauty Products	2621.899999999...
6	Books	1861.930000000...

EXECUTE QUERIES ON BIG QUERY UI (WEB CONSOLE) (2/5)

MOST POPULAR PAYMENT METHOD BY REGION

```
SELECT
  `Region`,
  `Payment Method`,
  COUNT(*) AS `Number of Transactions`
FROM
  `online_sales.sale`
GROUP BY
  `Region`,
  `Payment Method`
ORDER BY
  `Region`,
  `Number of Transactions` DESC;
```

Row	Region	Payment Method	Number of Transacti
1	Asia	Credit Card	40
2	Asia	Debit Card	40
3	Europe	PayPal	80
4	North America	Credit Card	80

EXECUTE QUERIES ON BIG QUERY UI (WEB CONSOLE) (3/5)

UNITS SOLD BY PRODUCT AND REGION

```sql
SELECT
  `Region`,
  `Product Name`,
  SUM(`Units Sold`) AS `Total Units Sold`
FROM
  `online_sales.sale`
GROUP BY
  `Region`,
  `Product Name`
ORDER BY
  `Region`,
  `Total Units Sold` DESC;
```

Row	Region	Product Name	Total Units Sold
1	Asia	Hanes ComfortSoft T-Shirt	10
2	Asia	Yeti Rambler Tumbler	6
3	Asia	Spalding NBA Street Basketball	6
4	Asia	Nike Air Force 1	6
5	Asia	Gap Essential Crewneck T-Shirt	6
6	Asia	Wilson Evolution Basketball	5
7	Asia	Titleist Pro V1 Golf Balls	5

EXECUTE QUERIES ON BIG QUERY UI (WEB CONSOLE) (4/5)

TOTAL REVENUE BY YEAR AND MONTH

```
SELECT
  EXTRACT(YEAR FROM `Date`) AS `Year`,
  EXTRACT(MONTH FROM `Date`) AS `Month`,
  SUM(`Total Revenue`) AS `Total Revenue`
FROM
  `online_sales.sale`
GROUP BY
  `Year`,
  `Month`
ORDER BY
  `Year`,
  `Month`;
```

Row	Year	Month	Total Revenue
1	2024	1	14548.31999999...
2	2024	2	10803.36999999...
3	2024	3	12849.23999999...
4	2024	4	12451.68999999...
5	2024	5	8455.489999999...
6	2024	6	7384.549999999...
7	2024	7	6797.079999999...
8	2024	8	7278.109999999...

EXECUTE QUERIES ON BIG QUERY UI (WEB CONSOLE) (5/5)

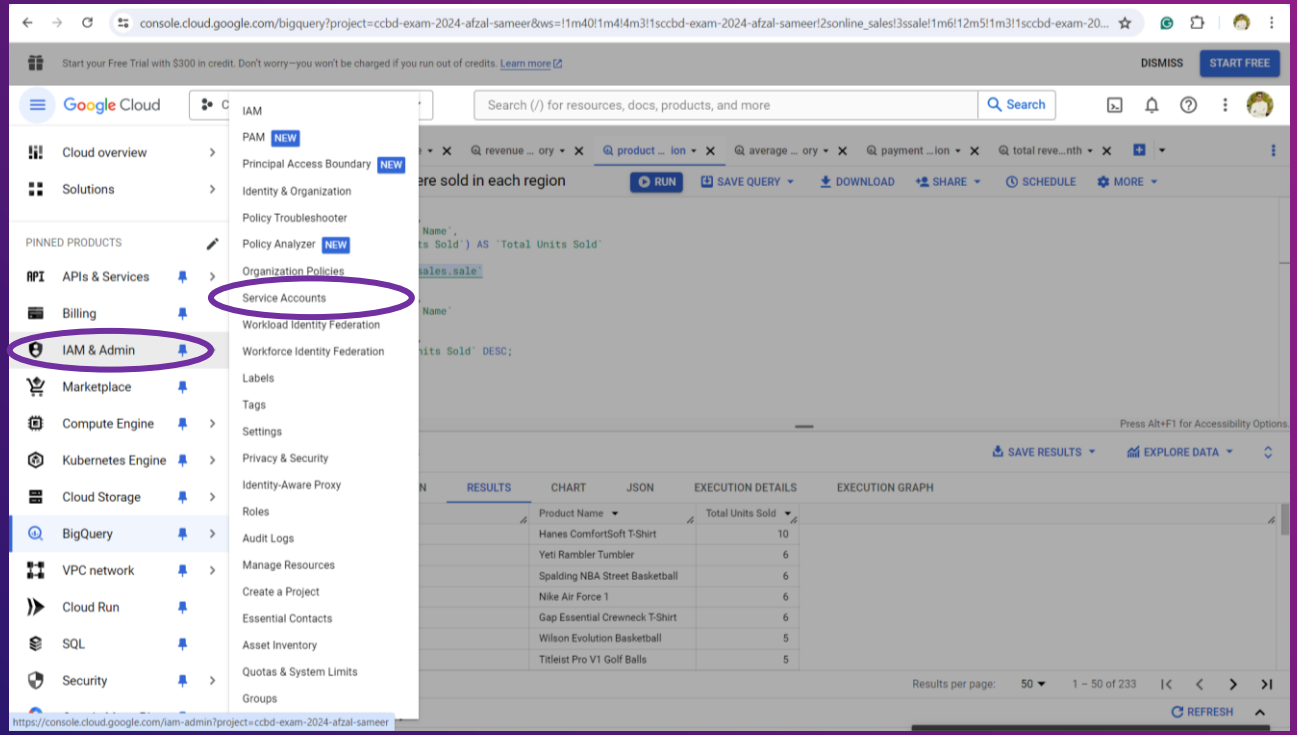
AVERAGE UNIT PRICE BY PRODUCT CATEGORY

```
SELECT
  `Product Category`,
  AVG(`Unit Price`) AS `Average Unit Price`
FROM
  `online_sales.sale`
GROUP BY
  `Product Category`
ORDER BY
  `Average Unit Price` DESC;
```

Row	Product Category	Average Unit Price
1	Electronics	691.5915000000...
2	Home Appliances	320.1855000000...
3	Sports	261.2839999999...
4	Clothing	67.53649999999...
5	Beauty Products	61.62300000000...
6	Books	16.15299999999...

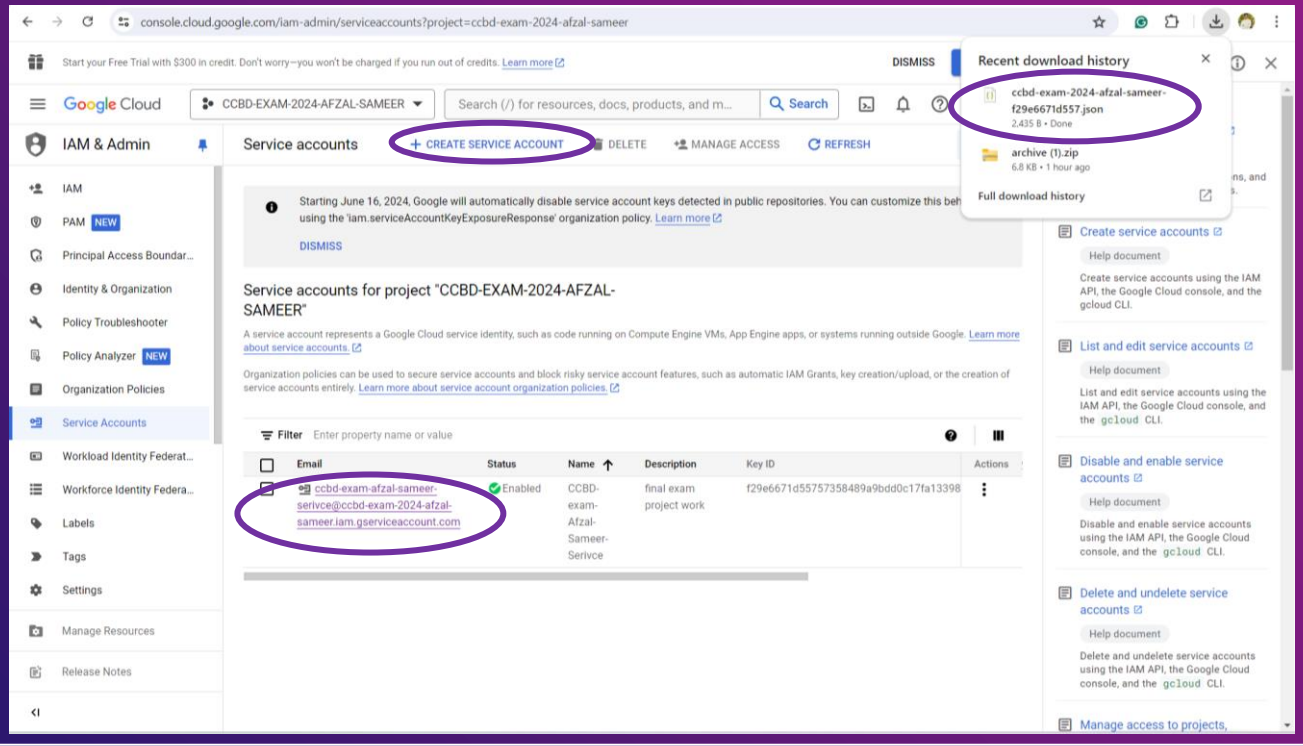
EXECUTE QUERIES ON JUPYTER NOTEBOOK (1/3)

CONNECTING BIG QUERY TO JUPYTER NOTEBOOK LOCALLY (CREATE JSON)



EXECUTE QUERIES ON JUPYTER NOTEBOOK (2/3)

CONNECTING BIG QUERY TO JUPYTER NOTEBOOK LOCALLY (CREATE AND DOWNLOAD JSON)

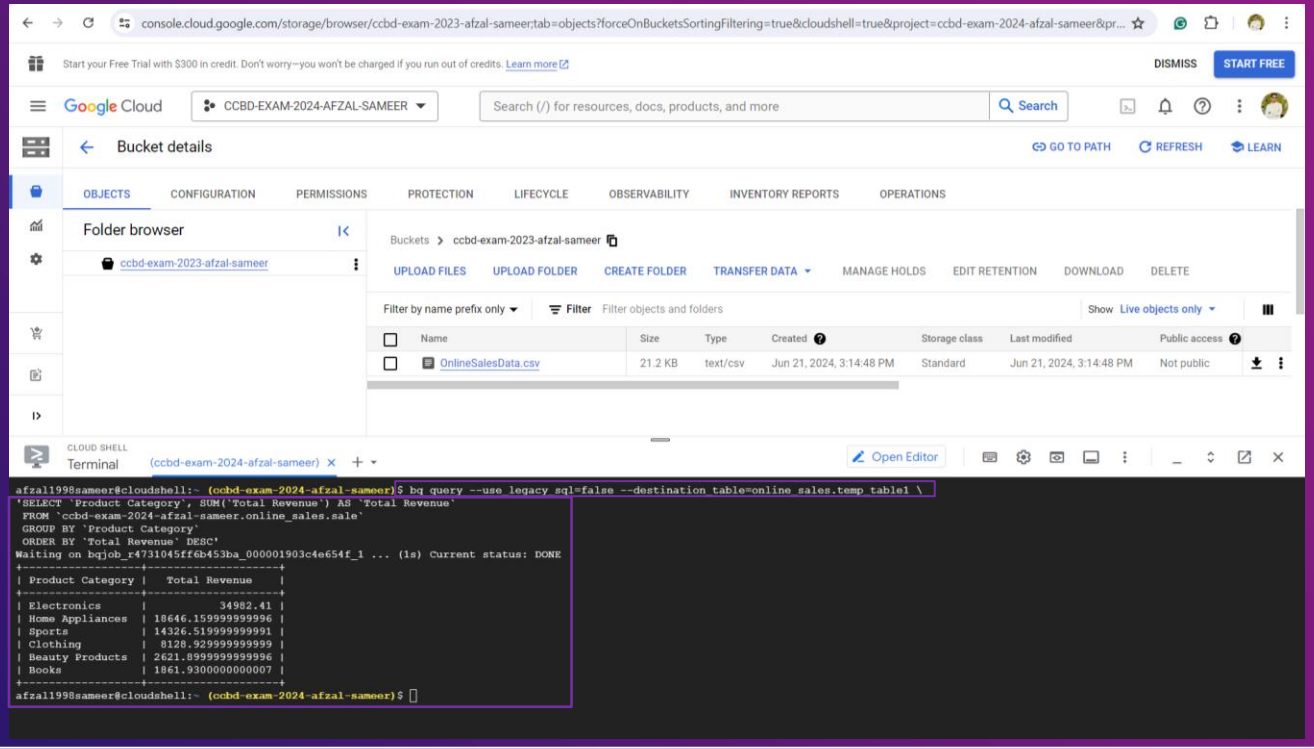


EXECUTE QUERIES ON JUPYTER NOTEBOOK (3/3)

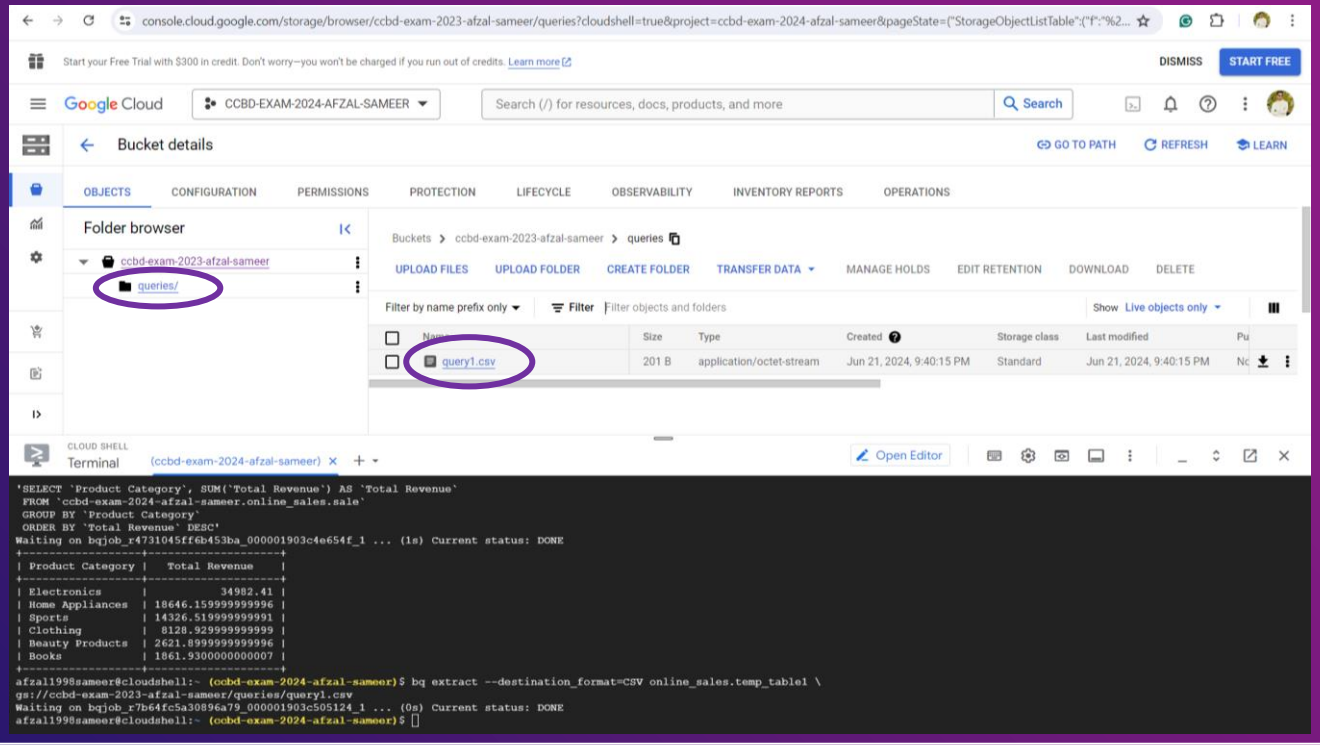
QUERIES ON JUPYTER NOTEBOOK

CCDB-EXAM-SAMEER-AFZAL-1000053143.IPYNB

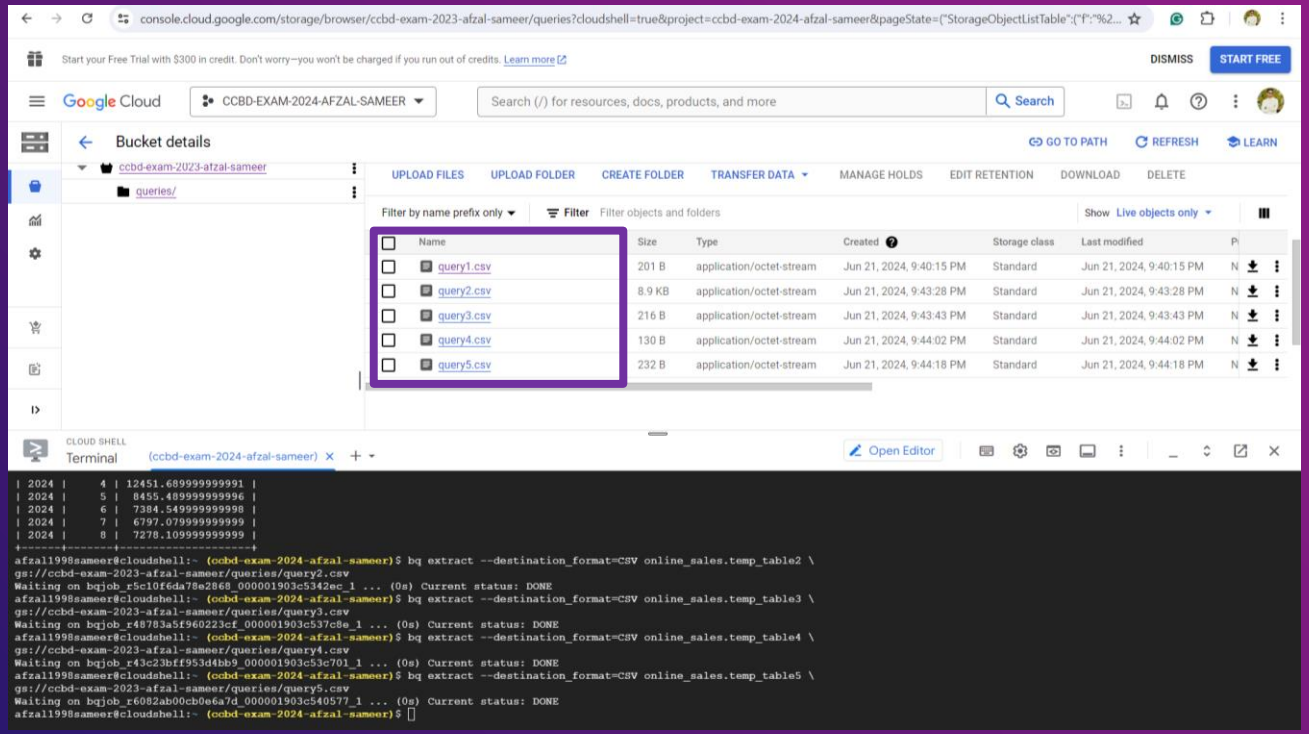
EXPORT QUERY RESULTS INTO THE G.C.S BUCKET (1/3)



EXPORT QUERY RESULTS INTO THE G.C.S BUCKET (2/3)



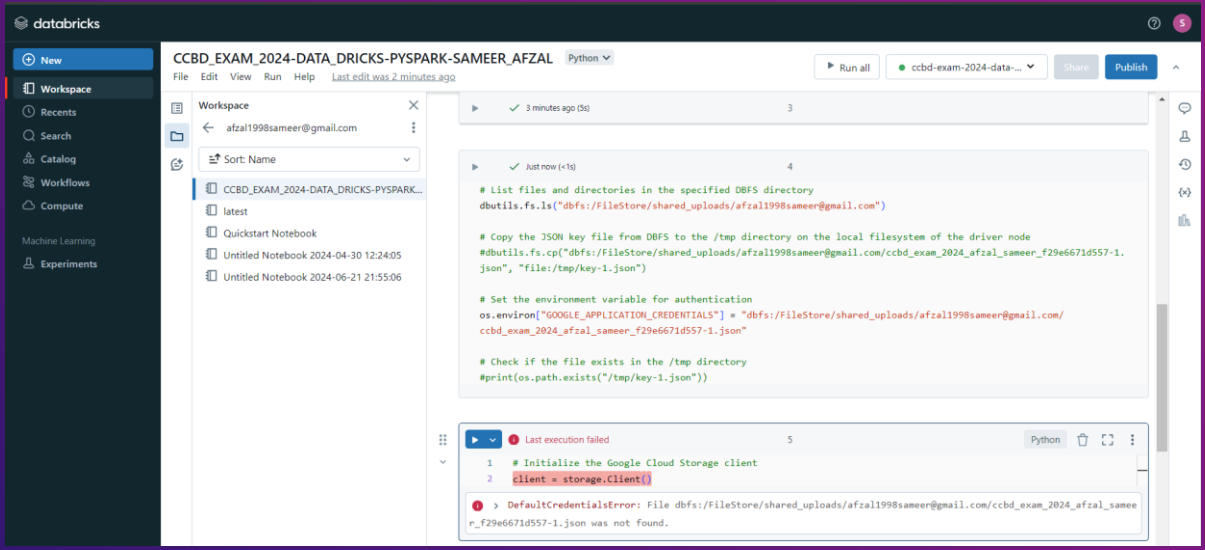
EXPORT QUERY RESULTS INTO THE G.C.S BUCKET (3/3)



DATA ANALYSIS USING SPARK NOTEBOOKS

CCBD_EXAM_2024-DATA_DRICKS-PYSPARK-SAMEER_AFZAL.IPYNB

LOAD YOUR DATASET FROM G.C.S



Yesterday (<1s) 4 Python

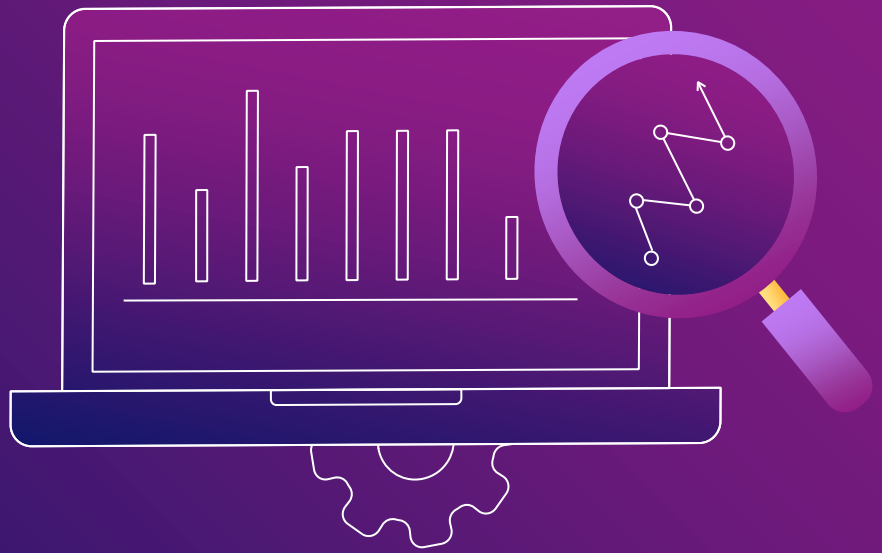
```python
# List files and directories in the specified DBFS directory
#dbutils.fs.ls("dbfs:/FileStore/shared_uploads/afzal1998sameer@gmail.com")

# Copy the JSON key file from DBFS to the /tmp directory on the local filesystem of the driver node
dbutils.fs.cp("dbfs:/FileStore/shared_uploads/afzal1998sameer@gmail.com/ccbd_exam_2024_afzal_sameer_f29e6671d557-1.json", "file:/tmp/
ccbd_exam_2024_afzal_sameer_f29e6671d557-1.json")

# Set the environment variable for authentication
os.environ["GOOGLE_APPLICATION_CREDENTIALS"] = "/tmp/ccbd_exam_2024_afzal_sameer_f29e6671d557-1.json"
```

DATA ENRICHMENT

EXECUTE A MACHINE LEARNING ALGORITHM ON THE DATASET USING SPARK MLIB IN DATA BRICKS



CCBD_EXAM_2024-DATA_DRICKS-PYSPARK-SAMEER_AFZAL.IPYNB

DATA VISUALIZATION

Looker Studio

databricks

CCBD-EXAM-2024-GLS-SAMEER-AFZAL

CCBD_EXAM_2024-DATA_DRICKS-PYSPARK-SAMEER_AFZAL.IPYNB

THANKS!

SAMEER AFZAL

DO YOU HAVE ANY QUESTIONS?