

# DATA ANALYSIS FOR ONLINE SALE

CLOUD COMPUTING AND BIG DATA



# TABLE OF CONTENT

---

DATA ANALYSIS

01

CLOUD SETUP

---

04

DATA ANALYSIS  
USING SPARK NOTEBOOKS

---

02

DATA INGESTION

---

05

DATA ENRICHMENT

---

03

DATA MANIPULATION  
USING DATABASES

---

06

DATA VISUALIZATION

---

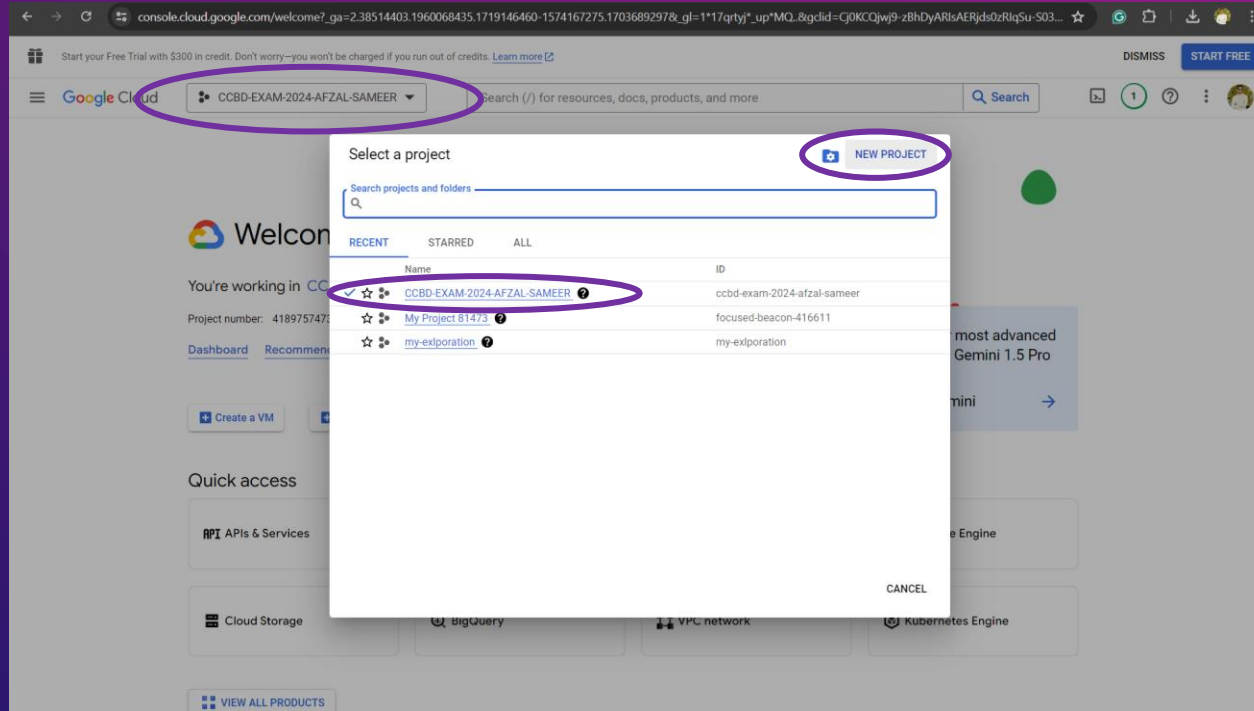




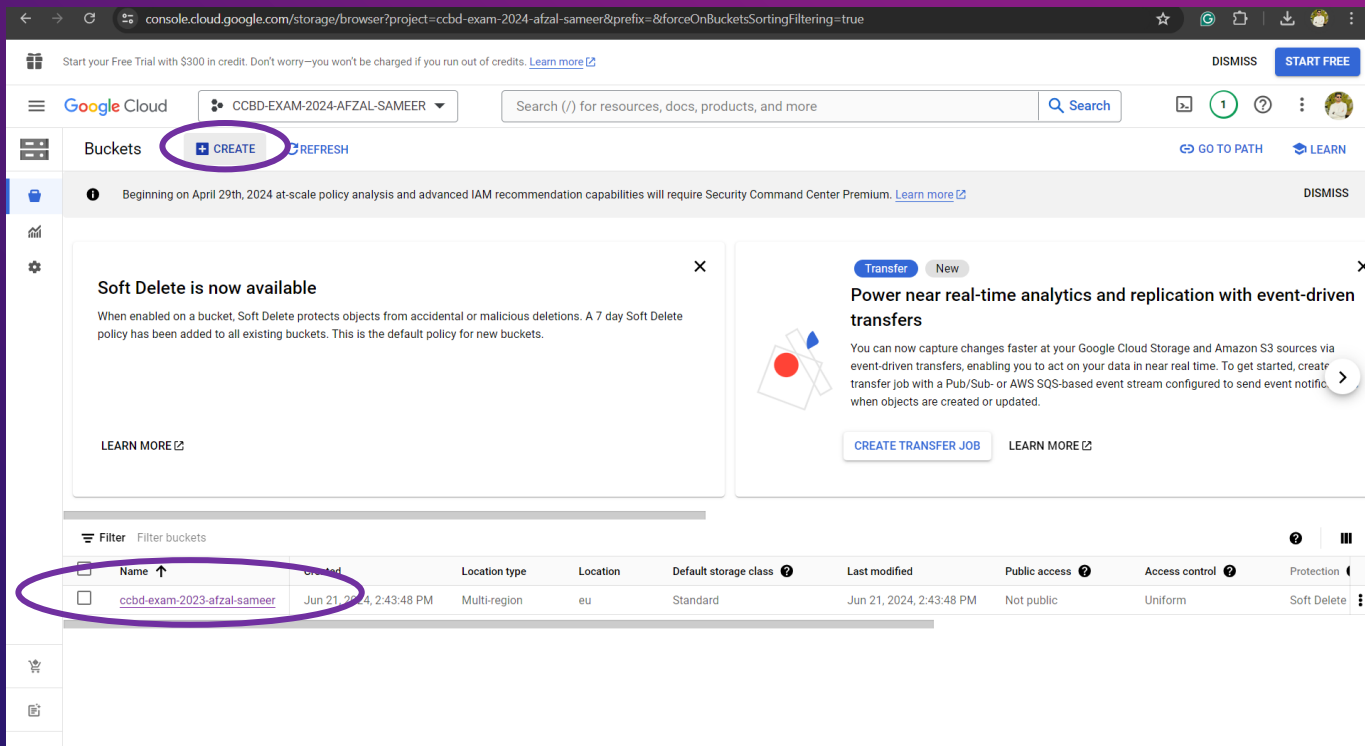
# CLOUD SETUP

---

# CREATE A GOOGLE CLOUD PROJECT



# CREATE A CLOUD STORAGE BUCKET



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](#)

**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](#)

**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 a transfer job with a Pub/Sub- or AWS SQS-based event stream configured to send event notifications when objects are created or updated.  
[CREATE TRANSFER JOB](#) [LEARN MORE](#)

**Filter** Filter buckets

<input type="checkbox"/>	Name ↑	Created	Location type	Location	Default storage class ?	Last modified	Public access ?	Access control ?	Protection ?
<input type="checkbox"/>	<a href="#">ccbd-exam-2023-afzal-sameer</a>	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

# UPLOAD DATA TO G.C.S PREVIOUSLY DEFINED

The screenshot displays the Google Cloud Storage console interface. The top navigation bar includes the Google Cloud logo, a project selector set to 'CCBD-EXAM-2024-AFZAL-SAMEER', a search bar, and a 'START FREE' button. The left sidebar shows navigation options: Buckets, Monitoring, and Settings. The main content area is titled 'Bucket details' for the bucket 'ccbd-exam-2023-afzal-sameer'. It lists properties: Location (eu), Storage class (Standard), Public access (Not public), and Protection (Soft Delete). Below this, a tabbed interface shows 'OBJECTS' selected. The 'Folder browser' section shows the bucket's contents. A table lists objects, with 'OnlineSalesData.csv' highlighted. Above the table, buttons for 'UPLOAD FILES', 'UPLOAD FOLDER', 'CREATE FOLDER', 'TRANSFER DATA', 'MANAGE HOLDS', 'EDIT RETENTION', and 'DOWNLOAD' are visible. At the bottom, a notification states '1 file successfully uploaded' and a progress bar shows 'OnlineSalesData.csv' as 'Complete'.

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 Bucket details GO TO PATH REFRESH LEARN

Buckets Monitoring Settings

ccbd-exam-2023-afzal-sameer

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

OBJECTS CONFIGURATION PERMISSIONS PROTECTION LIFECYCLE OBSERVABILITY INVENTORY REPORTS OPERATIONS

Folder browser Buckets > ccbd-exam-2023-afzal-sameer

ccbd-exam-2023-afzal-sameer

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

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

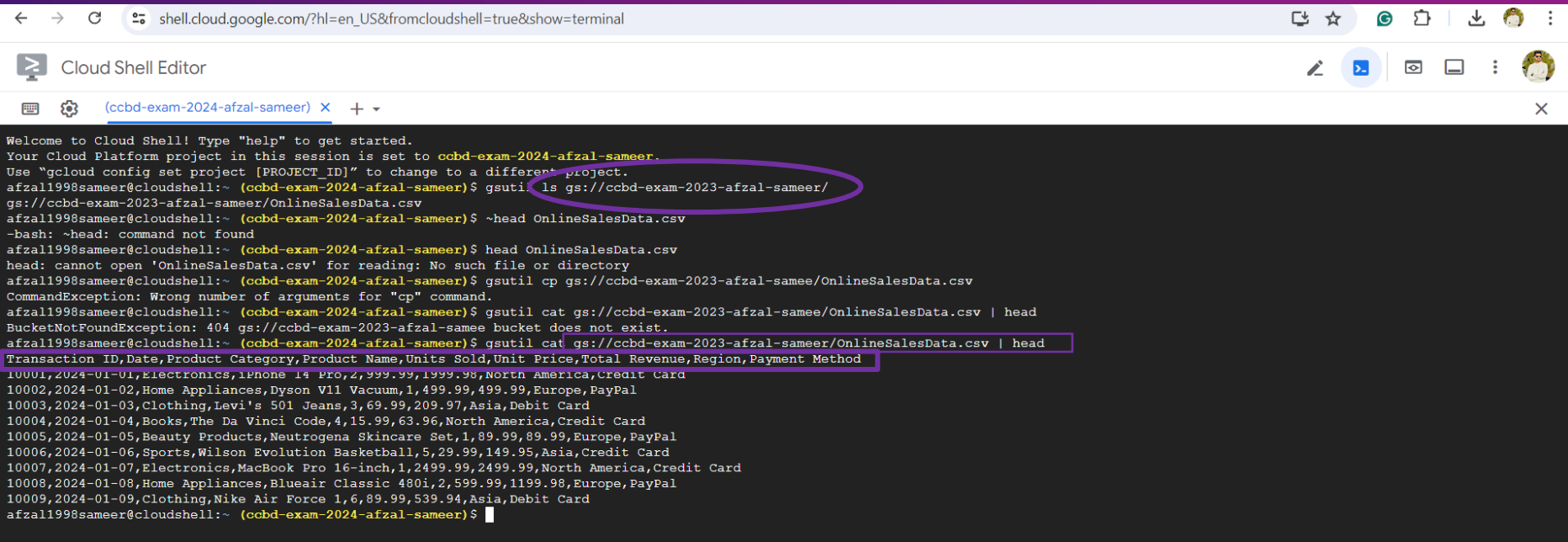
Name	Size	Type	Created	Storage class	Last modified
OnlineSalesData.csv	21 KB	text/csv	Jun 21, 2024, 3:14:48 PM	Standard	Jun 21, 2024, 3:14:48

1 file successfully uploaded

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

OnlineSalesData.csv

# 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) x

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-sameer/OnlineSalesData.csv
CommandException: Wrong number of arguments for "cp" command.
afzal1998sameer@cloudshell:~ (ccbd-exam-2024-afzal-sameer)$ gsutil cat gs://ccbd-exam-2023-afzal-sameer/OnlineSalesData.csv | head
BucketNotFoundException: 404 gs://ccbd-exam-2023-afzal-sameer 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)$
```

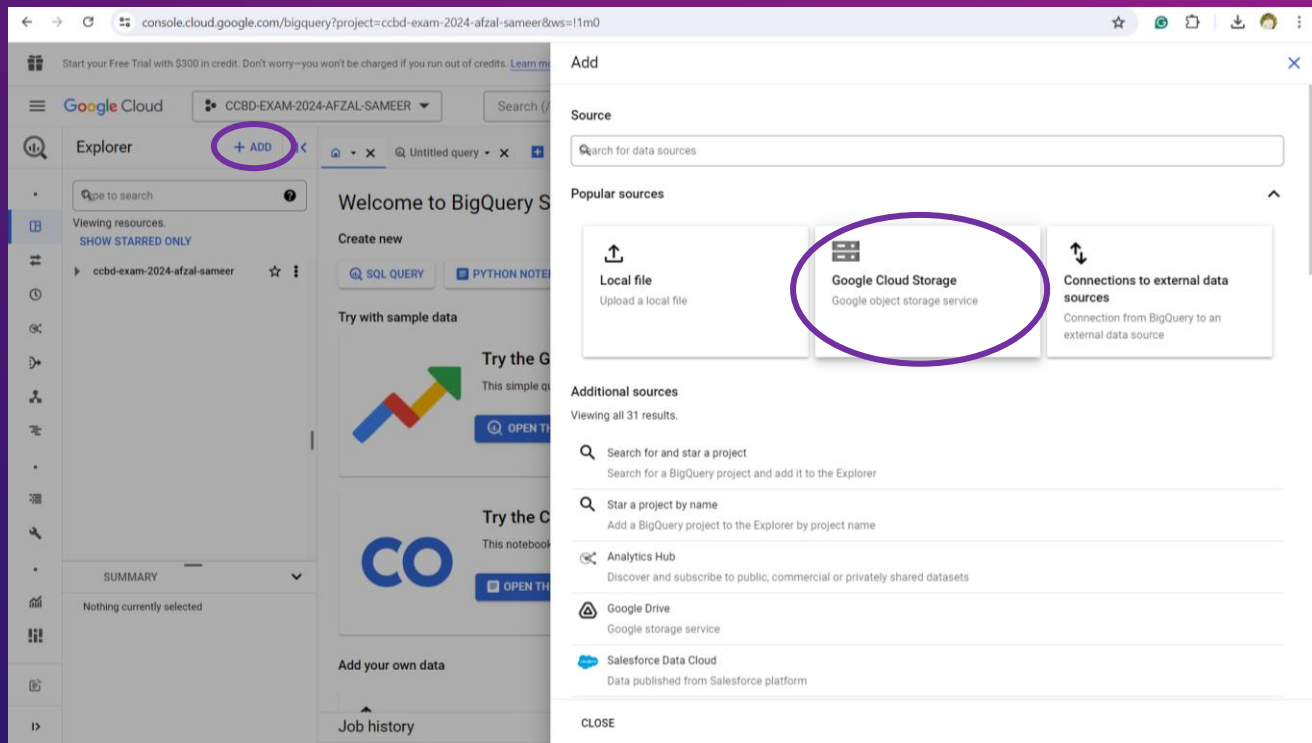




# DATA MANIPULATION USING DATABASES

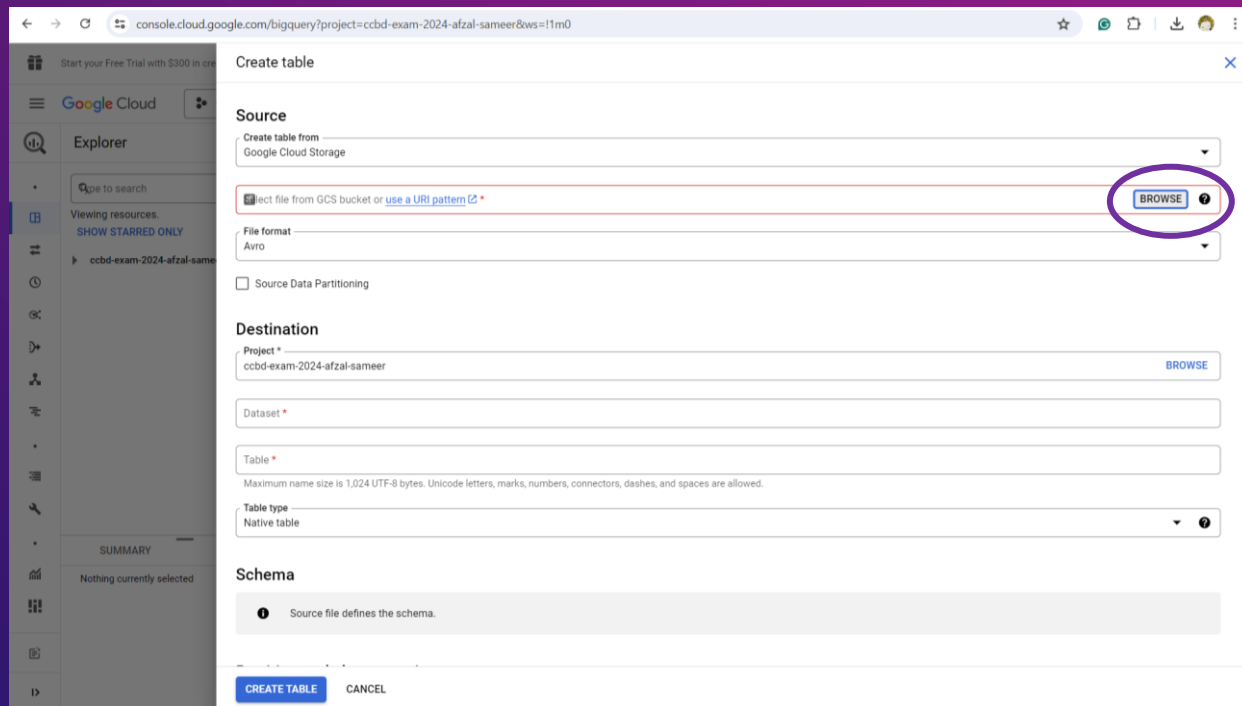
---

# BIG QUERY WEB CONSOLE



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

(1/5)



console.cloud.google.com/bigquery?project=ccbd-exam-2024-afzal-sameer&ws=11m0

### Create table

Source

Create table from Google Cloud Storage

Select file from GCS bucket or [use a URI pattern](#)

**BROWSE**

File format Avro

☐ Source Data Partitioning

Destination

Project \* ccbd-exam-2024-afzal-sameer **BROWSE**

Dataset \*

Table \*

Maximum name size is 1,024 UTF-8 bytes. Unicode letters, marks, numbers, connectors, dashes, and spaces are allowed.

Table type Native table

Schema

1 Source file defines the schema.

**CREATE TABLE** CANCEL

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

(2/5)

The screenshot shows the Google Cloud console interface for creating a new table in BigQuery. The main window is titled 'Create table' and has a sidebar on the left with the Google Cloud logo and an 'Explorer' section. The 'Source' section is highlighted with a red box, and the 'Destination' section is highlighted with a blue box. A modal window titled 'Choose a file' is open, showing a list of buckets with 'ccbd-exam-2023-afzal-sameer' selected and circled in purple.

**Create table**

**Source**

Create table from Google Cloud Storage

Select file from GCS bucket or [use a URI pattern](#)

File format: Avro

☐ Source Data Partitioning

**Destination**

Project: ccbd-exam-2024-afzal-sameer

Dataset:

Table:

Maximum name size is 1,024 UTF-8 bytes. Unicode letters, marks, numbers, connectors, dashes, and spaces are allowed.

Table type: Native table

**Schema**

Source file defines the schema.

**Choose a file**

Buckets

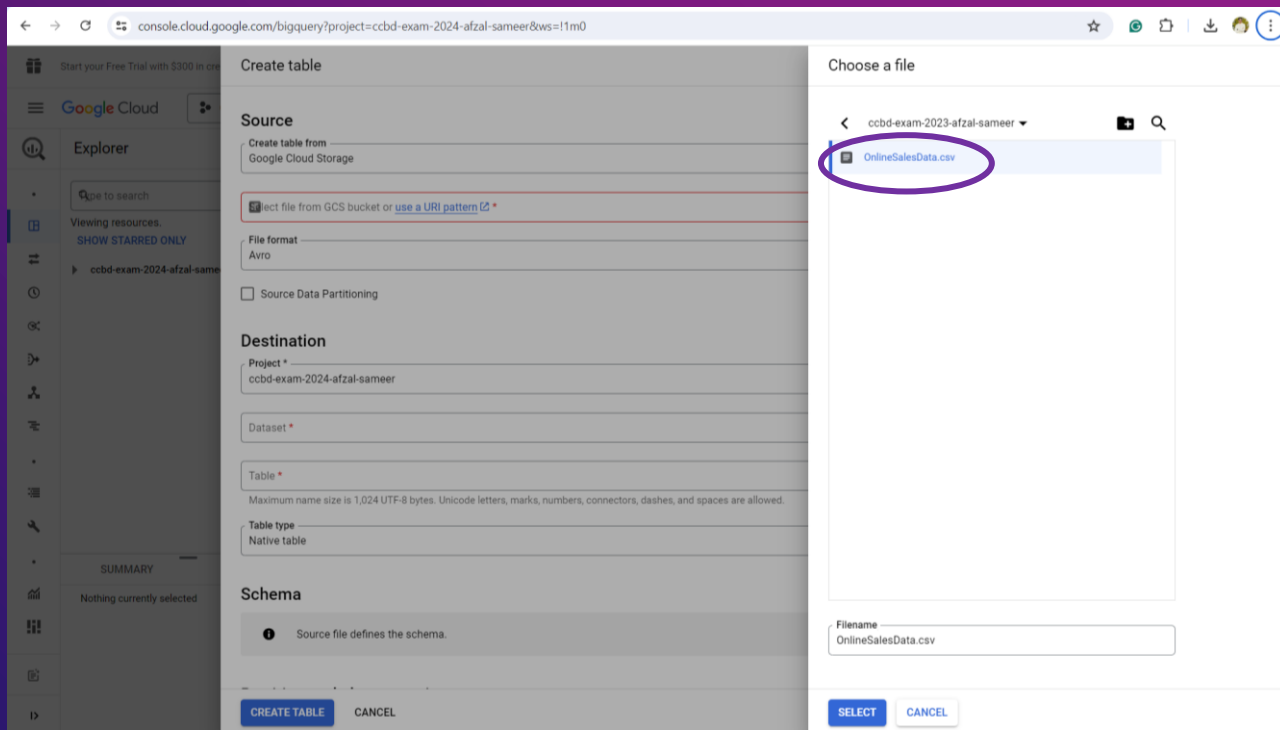
ccbd-exam-2023-afzal-sameer

Filename

SELECT CANCEL

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

(3/5)



The screenshot displays the Google Cloud BigQuery console interface. The main panel is titled 'Create table' and contains the following sections:

- Source:** 'Create table from Google Cloud Storage'. Below this, there is a red-bordered box containing the text 'Select file from GCS bucket or [use a URI pattern](#)'. The 'File format' is set to 'Avro', and 'Source Data Partitioning' is unchecked.
- Destination:** 'Project \*' is 'ccbd-exam-2024-afzal-sameer'. 'Dataset \*' and 'Table \*' are empty. A note states: 'Maximum name size is 1,024 UTF-8 bytes. Unicode letters, marks, numbers, connectors, dashes, and spaces are allowed.' The 'Table type' is 'Native table'.
- Schema:** A message indicates 'Source file defines the schema.'

At the bottom of the 'Create table' panel are 'CREATE TABLE' and 'CANCEL' buttons.

A 'Choose a file' dialog is overlaid on the right side of the console. It shows a file list for the project 'ccbd-exam-2023-afzal-sameer'. The file 'OnlineSalesData.csv' is highlighted and circled in purple. Below the file list, the 'Filename' field contains 'OnlineSalesData.csv'. At the bottom of the dialog are 'SELECT' and 'CANCEL' buttons.

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

(4/5)

console.cloud.google.com/bigquery?project=ccbd-exam-2024-afzal-sameer&ws=1m0

Start your Free Trial with \$300 in credit

Google Cloud

Explorer

Viewing resources. SHOW STARRED ONLY

ccbd-exam-2024-afzal-sameer

Nothing currently selected

### Create table

Create table from  
Google Cloud Storage

Select file from GCS bucket or use a URI pattern [?](#)  
ccbd-exam-2023-afzal-sameer/OnlineSalesData.csv BROWSE

File format  
CSV

☐ Source Data Partitioning

### Destination

Project \*  
ccbd-exam-2024-afzal-sameer BROWSE

Dataset \*  
online\_sales

CREATE NEW DATASET  
Launches a form to create a new dataset before continuing

Loaded datasets  
online\_sales ✓

### Schema

☒ Auto detect

Schema will be automatically generated.

### Partition and cluster settings

CREATE TABLE CANCEL

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

(5/5)

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

Explorer

+ ADD

IK

🔍

Type to search

?

Viewing resources.

SHOW STARRED ONLY

ccbd-exam-2024-afzal-sameer

🔍 Queries

📓 Notebooks

🗂 Data canvases

🔗 External connections

📊 online\_sales

📊 sale

SUMMARY

sale

ccbd-exam-2024-afzal-sameer.online\_sales

Last modified Jun 21, 2024, 3:33:21 PM UTC+2

Data EU

location

Description

sale

QUERY

SHARE

COPY

SNAPSHOT

DELETE

EXPORT

REFRESH

SCHEMA

DETAILS

PREVIEW

LINEAGE

DATA PROFILE

DATA QUALITY

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

Results per page: 50

1 - 50 of 240

IK < > I>

REFRESH

Job history

REFRESH

# BUSINESS QUESTIONS

---

## TOTAL REVENUE BY PRODUCT CATEGORY

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

01

02

## MOST POPULAR PAYMENT METHOD BY REGION

What is the most popular payment method in each region?

## 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?

04

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?

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

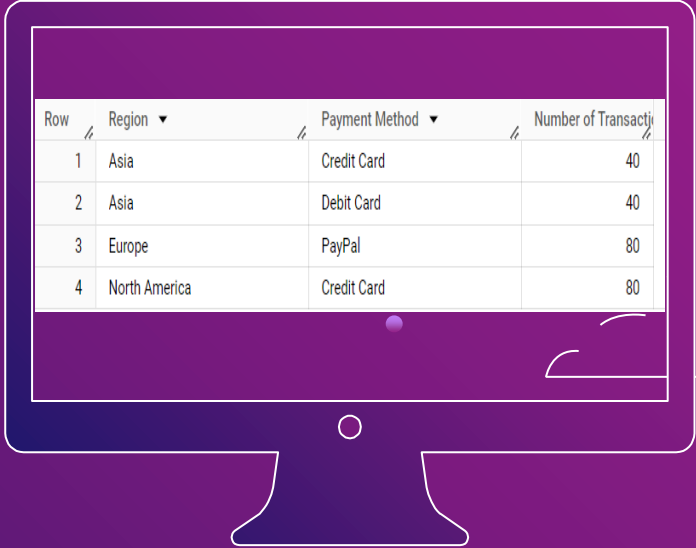
```
1 SELECT
2   `Product Category`,
3   SUM(`Total Revenue`) AS `Total Revenue`
4 FROM
5   `online_sales.sale`
6 GROUP BY
7   `Product Category`
8 ORDER BY
9   `Total Revenue` DESC;
```

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;
```



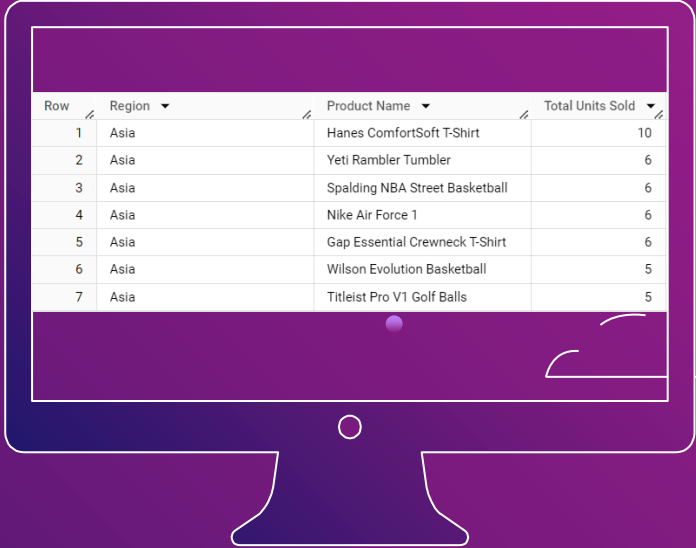
A computer monitor is shown on the right side of the slide, displaying a table with the following data:

Row	Region	Payment Method	Number of Transactions
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

```
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	8455.48999999...
6	2024	6	7384.54999999...
7	2024	7	6797.07999999...
8	2024	8	7278.10999999...

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

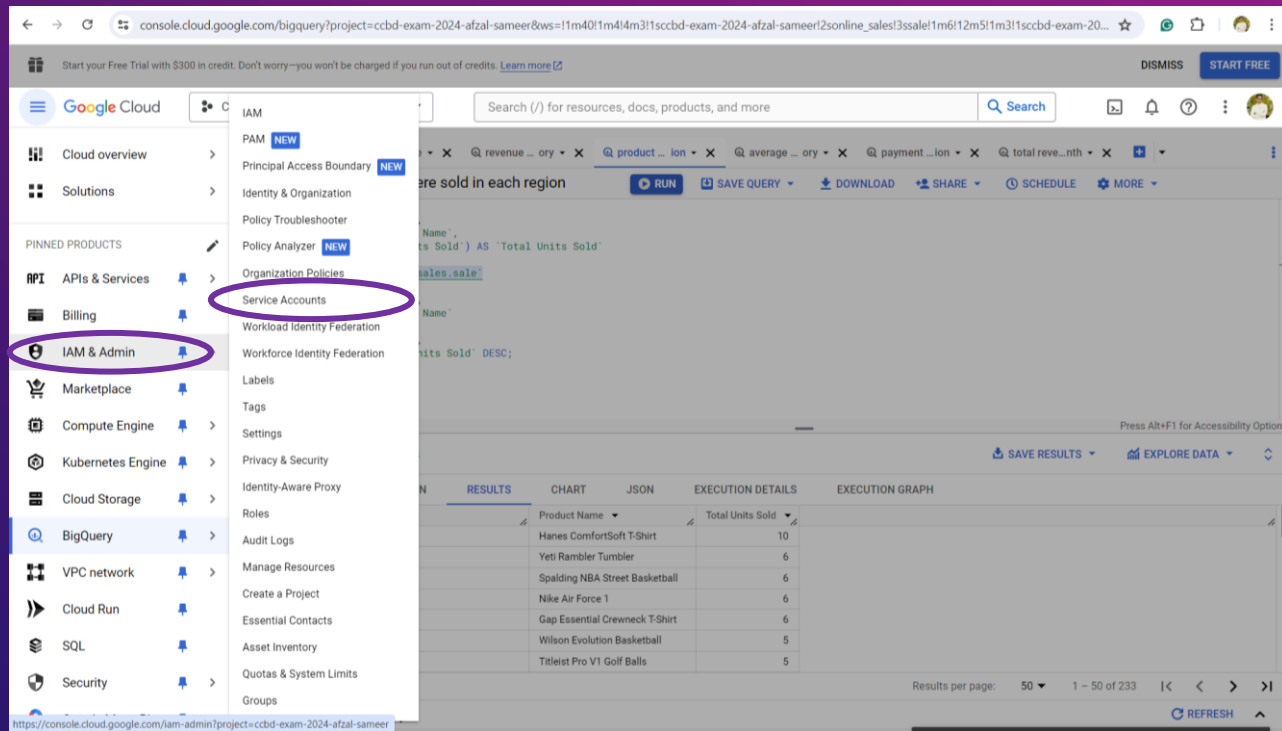
## AVERAGE UNIT PRICE BY PRODUCT CATEGORY

```
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	67.5364999999...
5	Beauty Products	61.623000000000...
6	Books	16.1529999999...

# EXECUTE QUERIES ON JUPYTER NOTEBOOK (1/3)

## CONNECTING BIG QUERY TO JUPYTER NOTEBOOK LOCALLY ( CREATE JSON)



The screenshot shows the Google Cloud IAM & Admin console. The left sidebar is open, and the 'Service Accounts' option is highlighted with a red circle. The main content area displays a list of service accounts, including 'Hanes ComfortSoft T-Shirt', 'Yeti Rambler Tumbler', 'Spalding NBA Street Basketball', 'Nike Air Force 1', 'Gap Essential Crewneck T-Shirt', 'Wilson Evolution Basketball', and 'Titleist Pro V1 Golf Balls'. The 'Total Units Sold' column shows values for each account.

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

# EXECUTE QUERIES ON JUPYTER NOTEBOOK (2/3)

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

The screenshot shows the Google Cloud IAM & Admin console for project "CCBD-EXAM-2024-AFZAL-SAMEER". The "Service accounts" section is active, and the "CREATE SERVICE ACCOUNT" button is circled in purple. Below the button, a table lists the service accounts for the project. The first service account is circled in purple. A "Recent download history" popup is visible in the top right corner, showing a download of "ccbd-exam-2024-afzal-sameer-f29e6671d557.json".

Service accounts for project "CCBD-EXAM-2024-AFZAL-SAMEER"

Filter	Email	Status	Name	Description	Key ID	Actions
	<a href="mailto:ccbd-exam-afzal-sameer-service@ccbd-exam-2024-afzal-sameer.iam.gserviceaccount.com">ccbd-exam-afzal-sameer-service@ccbd-exam-2024-afzal-sameer.iam.gserviceaccount.com</a>	Enabled	CCBD-exam-Afzal-Sameer-Service	final exam project work	f29e6671d55757358489a9bdd0c17fa13398	

# EXECUTE QUERIES ON JUPYTER NOTEBOOK <sup>(3/3)</sup>

---



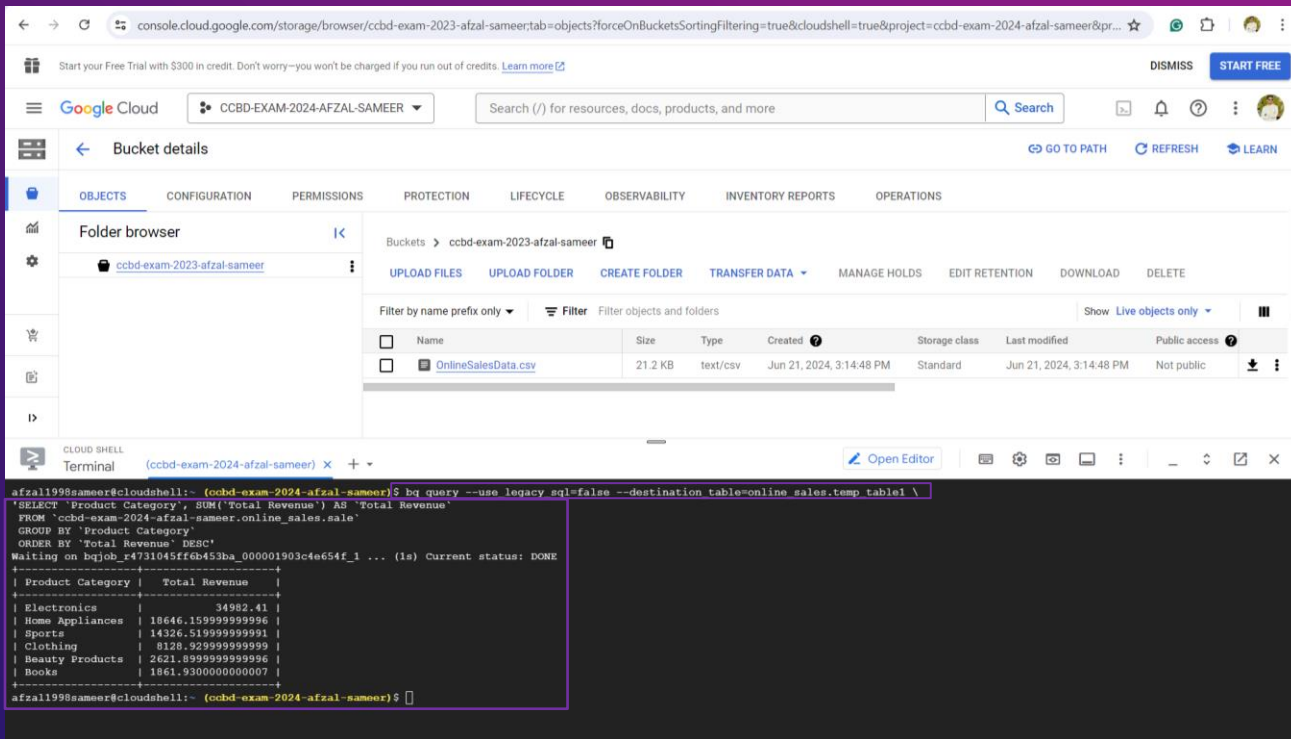
## QUERIES **ON** JUPYTER NOTEBOOK

CCDB-EXAM-SAMEER-AFZAL-1000053143.IPYNB





# EXPORT QUERY RESULTS INTO THE G.C.S BUCKET [1/3]



The screenshot displays the Google Cloud Storage console for the bucket 'ccbd-exam-2023-afzal-sameer'. The 'Folder browser' tab is active, showing a list of objects. A file named 'OnlineSalesData.csv' is visible, with a size of 21.2 KB, created on Jun 21, 2024, at 3:14:48 PM, and stored in the Standard storage class. Below the console, a Cloud Shell terminal window is open, showing a BigQuery query being executed. The query selects the product category and total revenue from the 'ccbd-exam-2024-afzal-sameer.online\_sales.sale' table, grouped by product category and ordered by total revenue in descending order. The query results are displayed in a table format.

Bucket details

OBJECTS CONFIGURATION PERMISSIONS PROTECTION LIFECYCLE OBSERVABILITY INVENTORY REPORTS OPERATIONS

Folder browser

ccbd-exam-2023-afzal-sameer

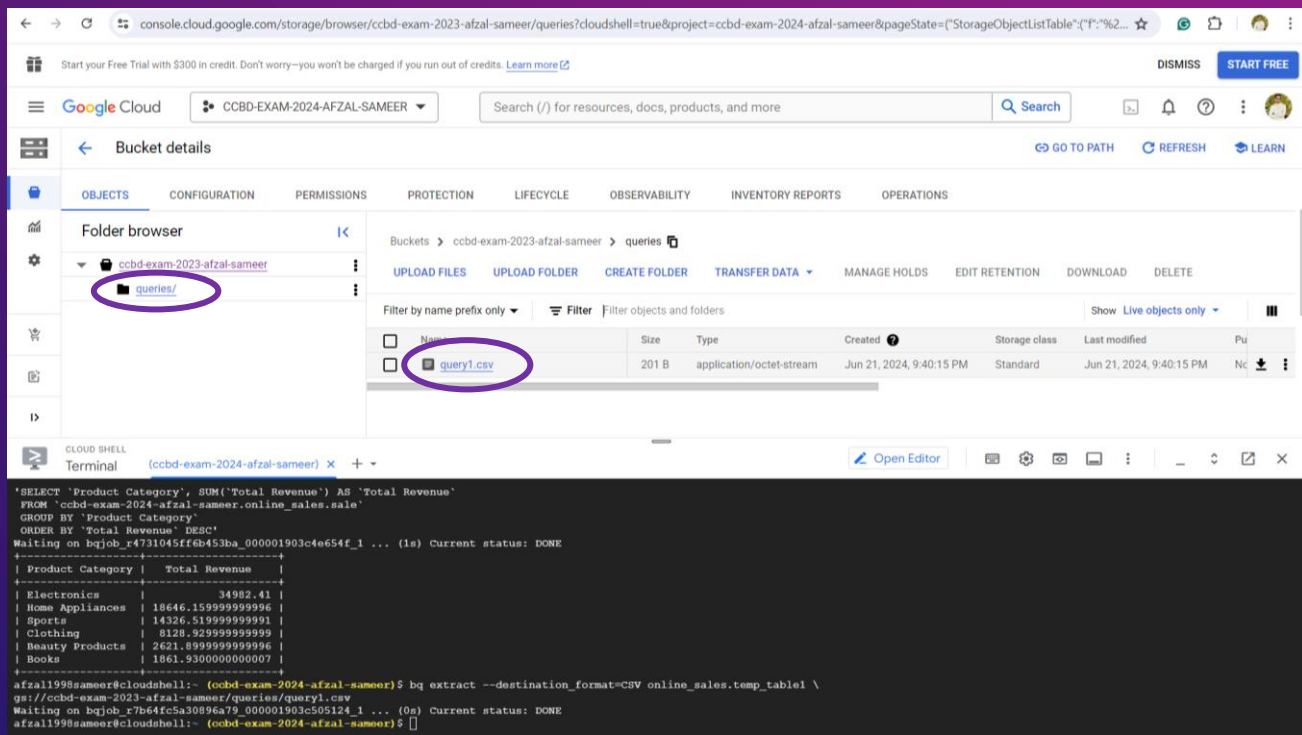
Upload FILES Upload FOLDER Create FOLDER Transfer DATA Manage Holds Edit Retention Download Delete

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

Name	Size	Type	Created	Storage class	Last modified	Public access
OnlineSalesData.csv	21.2 KB	text/csv	Jun 21, 2024, 3:14:48 PM	Standard	Jun 21, 2024, 3:14:48 PM	Not public

```
afzal1998sameer@cloudshell:~ (ccbd-exam-2024-afzal-sameer)$ bq query --use legacy sql=false --destination table=online_sales.temp_table1 \
'SELECT `Product Category`, SUM(`Total Revenue`) AS `Total Revenue`
FROM `ccbd-exam-2024-afzal-sameer.online_sales.sale`
GROUP BY `Product Category`
ORDER BY `Total Revenue` DESC'
Waiting on bqjob_r4731045ff6b453ba_000001903c4e654f_1 ... (1s) Current status: DONE
+-----+
| Product Category | Total Revenue |
+-----+
| Electronics      | 34982.41      |
| Home Appliances  | 18646.159999999996 |
| Sports           | 14326.519999999991 |
| Clothing         | 8128.929999999999 |
| Beauty Products  | 2621.8999999999996 |
| Books            | 1861.9300000000007 |
+-----+
afzal1998sameer@cloudshell:~ (ccbd-exam-2024-afzal-sameer)$
```

# EXPORT QUERY RESULTS INTO THE G.C.S BUCKET [2/3]



The screenshot shows the Google Cloud Storage console interface. The top navigation bar includes the Google Cloud logo, a project selector set to 'CCBD-EXAM-2024-AFZAL-SAMEER', a search bar, and a 'START FREE' button. The main content area is titled 'Bucket details' and shows the 'OBJECTS' tab for the bucket 'ccbd-exam-2023-afzal-sameer'. The 'queries' folder is highlighted in the left sidebar. The main view shows a table of objects with columns: Name, Size, Type, Created, Storage class, Last modified, and Permissions. A file named 'query1.csv' is listed with a size of 201 B and a type of 'application/octet-stream'. Below the table, a terminal window shows a SQL query and its results.

Bucket details

OBJECTS CONFIGURATION PERMISSIONS PROTECTION LIFECYCLE OBSERVABILITY INVENTORY REPORTS OPERATIONS

Folder browser

ccbd-exam-2023-afzal-sameer

queries

query1.csv

Filter by name prefix only

Name	Size	Type	Created	Storage class	Last modified	Permissions
query1.csv	201 B	application/octet-stream	Jun 21, 2024, 9:40:15 PM	Standard	Jun 21, 2024, 9:40:15 PM	Permissions

```
'SELECT 'Product Category', SUM('Total Revenue') AS 'Total Revenue'
FROM 'ccbd-exam-2024-afzal-sameer.online_sales.sale'
GROUP BY 'Product Category'
ORDER BY 'Total Revenue' DESC'
Waiting on bqjob_r4731045f6b453ba_000001903c4e654f_1 ... (1s) Current status: DONE
+-----+-----+
| Product Category | Total Revenue |
+-----+-----+
| Electronics      | 34982.41      |
| Home Appliances  | 18646.159999999996 |
| Sports           | 14326.519999999991 |
| Clothing         | 8128.929999999999 |
| Beauty Products | 2621.8999999999996 |
| Books            | 1861.9300000000007 |
+-----+-----+
afzal1998sameer@cloudshell:~ (ccbd-exam-2024-afzal-sameer) $ bq extract --destination_format=CSV online_sales.temp_table1 \
gs://ccbd-exam-2023-afzal-sameer/queries/query1.csv
Waiting on bqjob_r7b64fc5a30896a79_000001903c505124_1 ... (0s) Current status: DONE
afzal1998sameer@cloudshell:~ (ccbd-exam-2024-afzal-sameer) $
```

# EXPORT QUERY RESULTS INTO THE G.C.S BUCKET [3/3]

The screenshot displays the Google Cloud Storage console interface. At the top, the breadcrumb navigation shows the path: `ccbd-exam-2023-afzal-sameer/queries/`. Below this, a table lists five CSV files: `query1.csv`, `query2.csv`, `query3.csv`, `query4.csv`, and `query5.csv`. A purple box highlights the 'Name' column of this table. Below the table, a terminal window is open, showing the execution of the `bq extract` command to export query results to the GCS bucket. The terminal output shows the command being executed and the status 'DONE'.

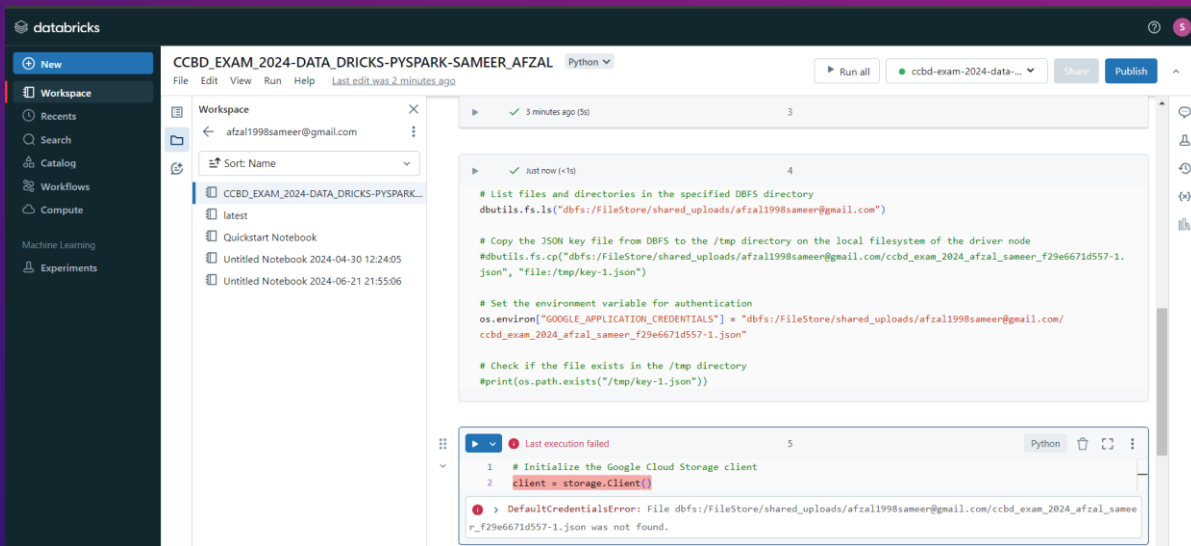
Name	Size	Type	Created	Storage class	Last modified
<a href="#">query1.csv</a>	201 B	application/octet-stream	Jun 21, 2024, 9:40:15 PM	Standard	Jun 21, 2024, 9:40:15 PM
<a href="#">query2.csv</a>	8.9 KB	application/octet-stream	Jun 21, 2024, 9:43:28 PM	Standard	Jun 21, 2024, 9:43:28 PM
<a href="#">query3.csv</a>	216 B	application/octet-stream	Jun 21, 2024, 9:43:43 PM	Standard	Jun 21, 2024, 9:43:43 PM
<a href="#">query4.csv</a>	130 B	application/octet-stream	Jun 21, 2024, 9:44:02 PM	Standard	Jun 21, 2024, 9:44:02 PM
<a href="#">query5.csv</a>	232 B	application/octet-stream	Jun 21, 2024, 9:44:18 PM	Standard	Jun 21, 2024, 9:44:18 PM

```
afzal1998sameer@cloudshell:~ (ccbd-exam-2024-afzal-sameer) $ bq extract --destination_format=CSV online_sales.temp_table2 \
gs://ccbd-exam-2023-afzal-sameer/queries/query2.csv
Waiting on bqjob_r5c10f6da70e2869_000001903c5342ec_1 ... (0s) Current status: DONE
afzal1998sameer@cloudshell:~ (ccbd-exam-2024-afzal-sameer) $ bq extract --destination_format=CSV online_sales.temp_table3 \
gs://ccbd-exam-2023-afzal-sameer/queries/query3.csv
Waiting on bqjob_r48783a5f60223cf_000001903c537c8e_1 ... (0s) Current status: DONE
afzal1998sameer@cloudshell:~ (ccbd-exam-2024-afzal-sameer) $ bq extract --destination_format=CSV online_sales.temp_table4 \
gs://ccbd-exam-2023-afzal-sameer/queries/query4.csv
Waiting on bqjob_r43c2bfbf953d4bb9_000001903c53c701_1 ... (0s) Current status: DONE
afzal1998sameer@cloudshell:~ (ccbd-exam-2024-afzal-sameer) $ bq extract --destination_format=CSV online_sales.temp_table5 \
gs://ccbd-exam-2023-afzal-sameer/queries/query5.csv
Waiting on bqjob_r6082ab0cb0e6a7d_000001903c540577_1 ... (0s) Current status: DONE
afzal1998sameer@cloudshell:~ (ccbd-exam-2024-afzal-sameer) $
```



# DATA ANALYSIS USING SPARK NOTEBOOKS

# LOAD YOUR DATASET FROM G.C.S



**CCBD\_EXAM\_2024-DATA\_DRICKS-PYSPARK-SAMEER\_AFZAL** Python

File Edit View Run Help Last edit was 2 minutes ago

Workspace

afzal1998sameer@gmail.com

Sort Name

- CCBD\_EXAM\_2024-DATA\_DRICKS-PYSPARK...
- latest
- Quickstart Notebook
- Untitled Notebook 2024-04-30 12:24:05
- Untitled Notebook 2024-06-21 21:55:06

3 minutes ago (5s) 3

```
# 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/key-1.json")

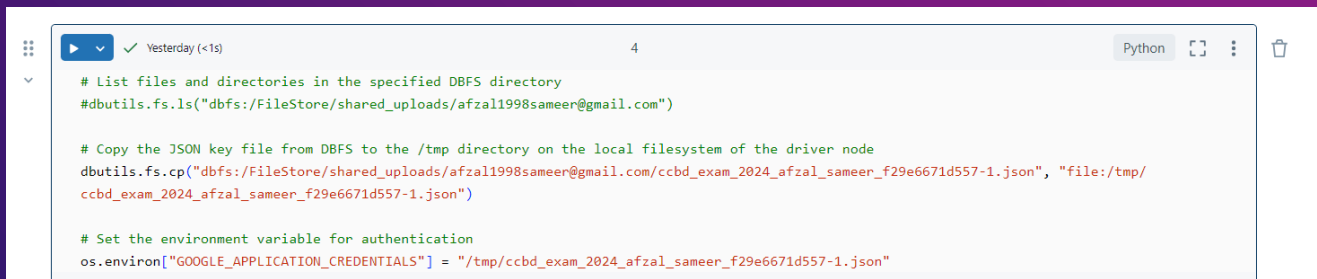
# Set the environment variable for authentication
os.environ["GOOGLE_APPLICATION_CREDENTIALS"] = "dbfs:/FileStore/shared_uploads/afzal1998sameer@gmail.com/ccbd_exam_2024_afzal_sameer_f29e6671d557-1.json"

# Check if the file exists in the /tmp directory
#print(os.path.exists("/tmp/key-1.json"))
```

Last execution failed 5

```
1 # Initialize the Google Cloud Storage client
2 client = storage.Client()
```

DefaultCredentialsError: File dbfs:/FileStore/shared\_uploads/afzal1998sameer@gmail.com/ccbd\_exam\_2024\_afzal\_sameer\_f29e6671d557-1.json was not found.



Yesterday (<1s) 4 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



# DATA VISUALIZATION

---



Looker Studio

CCBD-EXAM-2024-GLS-SAMEER-AFZAL



databricks

CCBD\_EXAM\_2024-DATA\_DRICKS-PYSPARK-SAMEER\_AFZAL.IPYNB

# THANKS!

---

DATA ANALYSIS



**SAMEER AFZAL**

---

---

**DO YOU HAVE ANY QUESTIONS?**