Employment Opportunities for the International Students in California

Group 2 members:

Deepthimai Potla

Abda Fatima Syeda

Thulsi Buyyankar

Sharanya Chinnigari

Mohammed Shaik Afroz

Tanaya Dutt

Table of Contents

Executive Summary

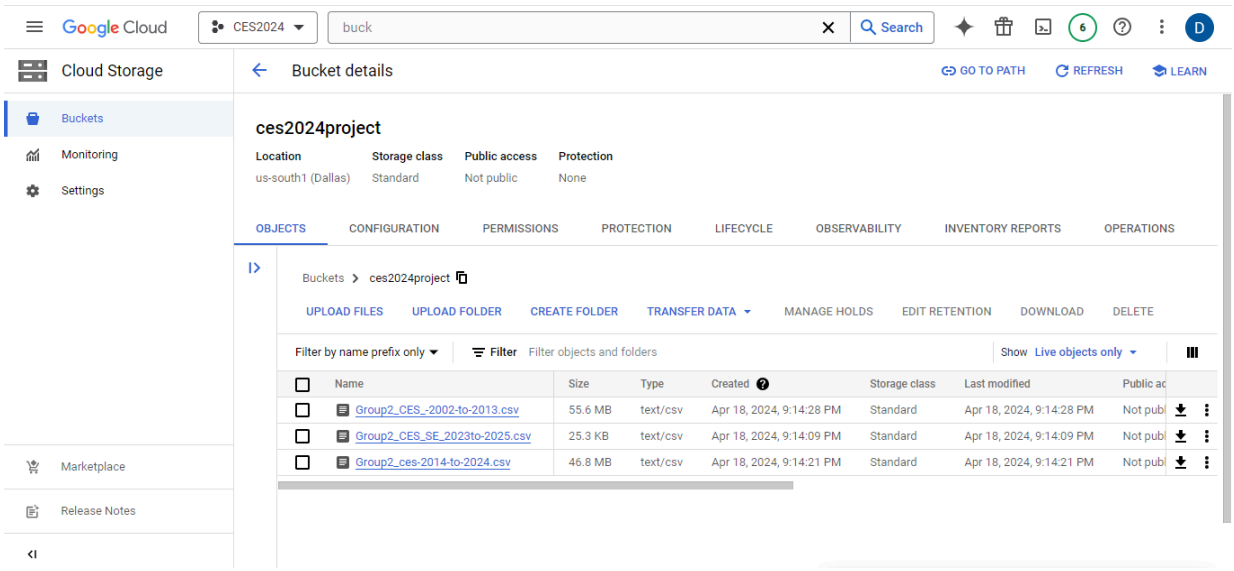
As a team of data engineers working for the International Student Employment Initiative, it is our job to collect and process the employment opportunities in California overtime so that our analysts can derive meaningful insights from the processed datasets. Our first goal was to find the right datasets in order to process, store and manage it. We found three datasets that contained data about the employment rates county wise over the state of california. We found these dataset via the California open data portal and data.gov. First we created a project on Google Cloud Platform (GCP) and then created storage buckets to store our three static datasets. After storing our data, we used tools such as compute engine and Dataproc to create clusters with one manager node and two worker nodes. Then we used OpenRefine to refine our data of any missing values or errors. After cleaning our data, we created a schema for all three datasets by ingesting our cleaned datasets using BigQuery Studio which is an extension of GCP and can be used for the data management and to analyze the datasets further we executed queries in BigQuery to ensure all our data loaded and analyzed using the SQL Queries to provide insights to our data analysts and data scientists team. Then we performed a few preliminary queries using Hive and Spark to observe which one would be more cost-effective and easy to use for our analysts.we built a system by using all the above mentioned tools helps our data analysts and scientists teams to perform deeper analysis of these cleaned datasets to meet the goal of this organization and provide the valuable insights to our International students.

Screenshots

We used Google cloud platform for setting up the project and storing our datasets.The below screenshot is the reference of how we created new GCP project for our employment opportunities.

Screenshot 1: Creation of the new GCP Project for our use case employment opportunities in California and creating the storage bucket to upload the three Static datasets.

Tools used : GCP



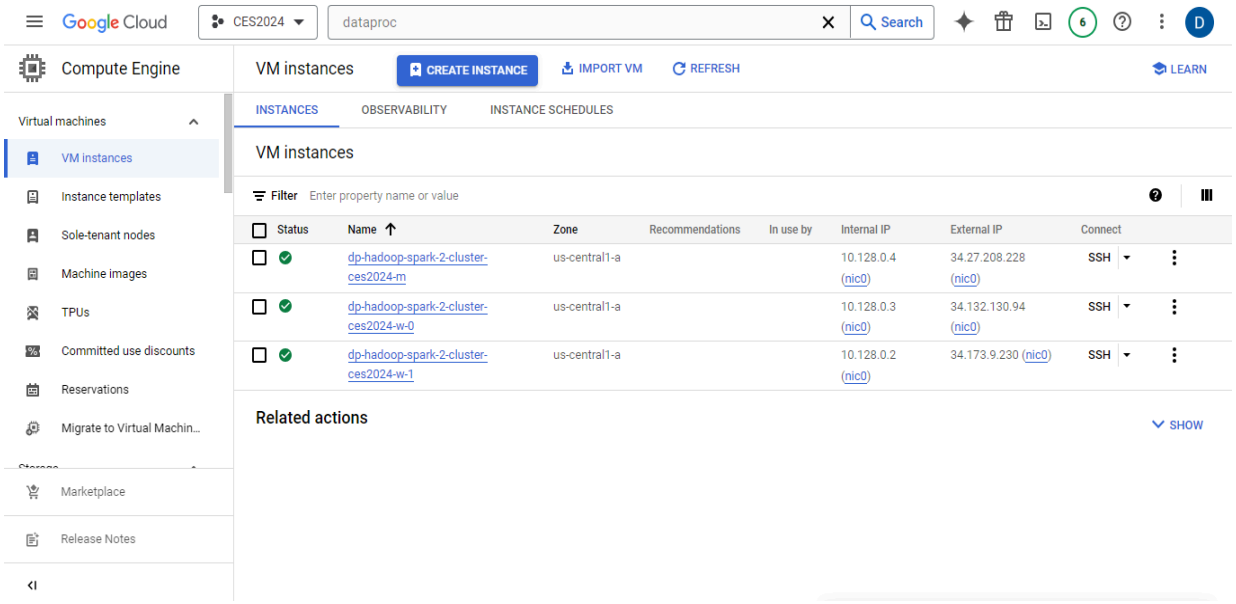
Explanation:

We have created the GCP project “CES2024” and storage bucket “CES2024project” and uploaded the three datasets Group2_CES_2002-to-2013, Group2_CES_2014-to-2024, Group2_CES_SE_2023-to-2025 respectively.

We setup a processing system for transforming our data for analysis by setting up the clusters and exploring the Hadoop and spark services which helps for parallel processing of the large datasets

we explored the services in Hadoop by using Bash shell command language by running the simple commands.VM instances provide an opportunity for parallel processing of our employment opportunities datasets.The below screenshots are how we set up VM Instances and explored hadoop services for our employment opportunities project.

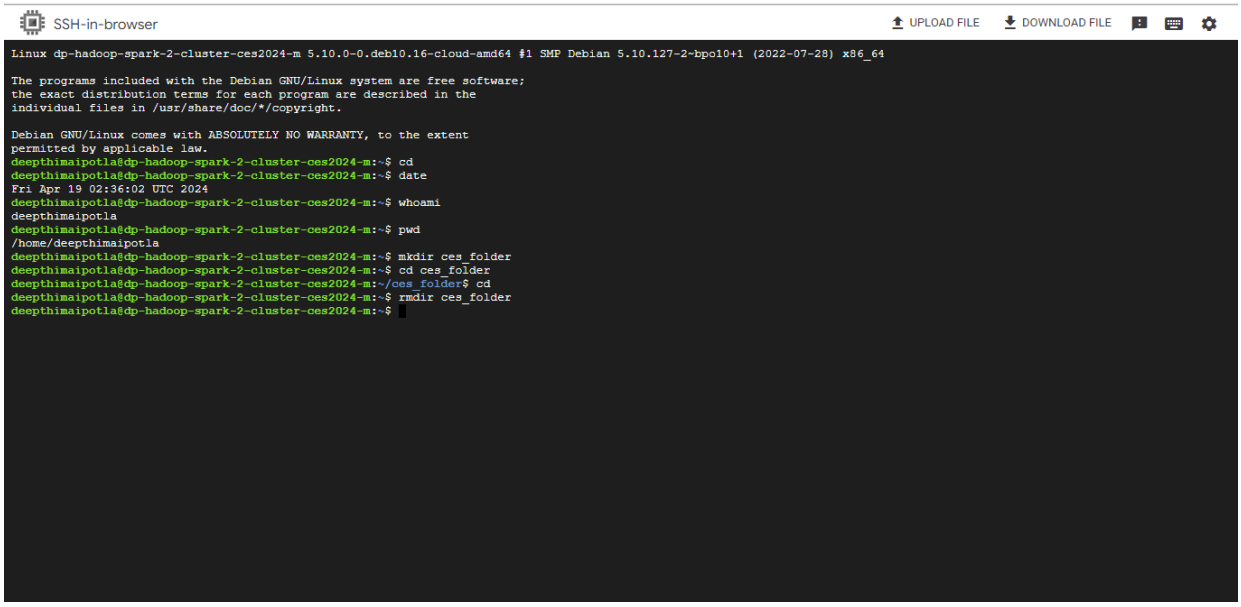
Screenshot 2 : VM instances running screenshot.



Explanation: Enabled both compute engine and Dataproc. Created clusters along with one manager node and two worker nodes in the following format:

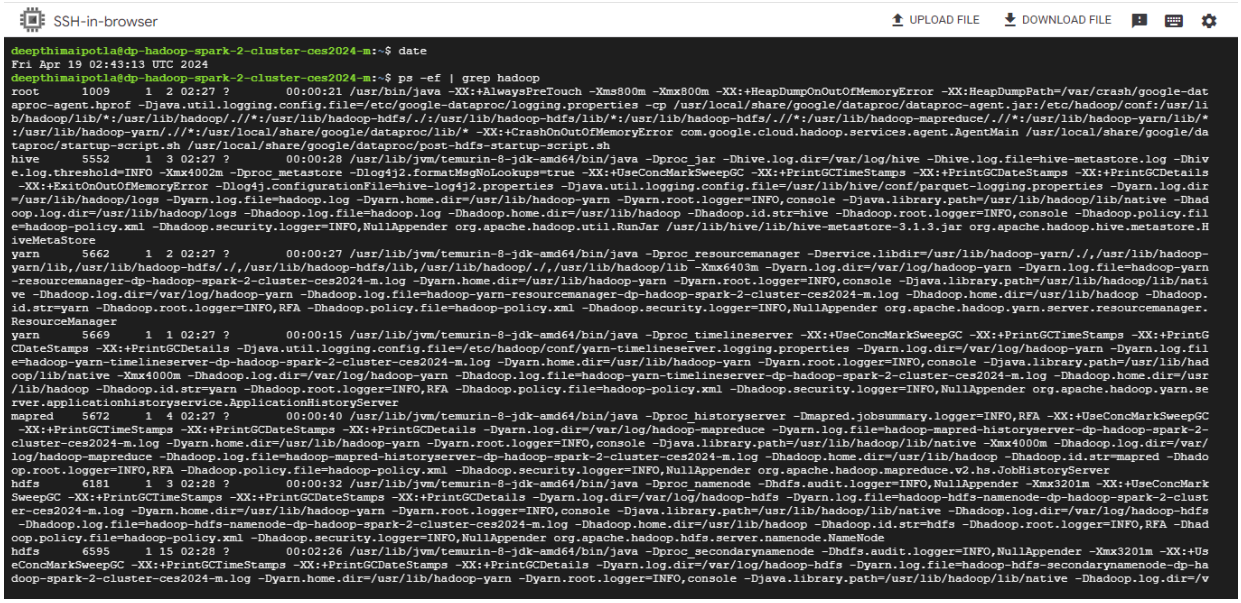
dp-hadoop-spark-2-cluster-ces2024-m, dp-hadoop-spark-2-cluster-ces2024.

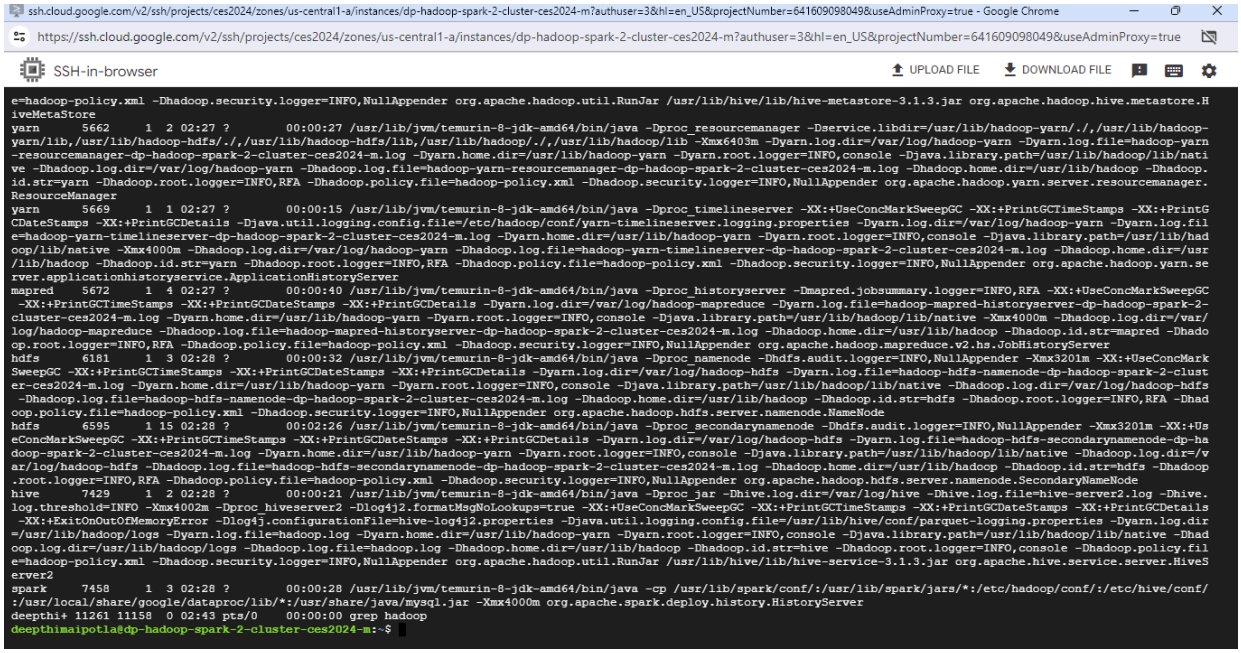
Screenshot 3 : Executing the linux commands in SSH terminal.



Explanation: Here we have connected to the manager node of the cluster and started the cluster. In the SSH terminal, we accessed the home directory of the account and executed the linux commands succesfully**.**

Screenshot 4 : Exploring services using Hadoop and Spark clusters

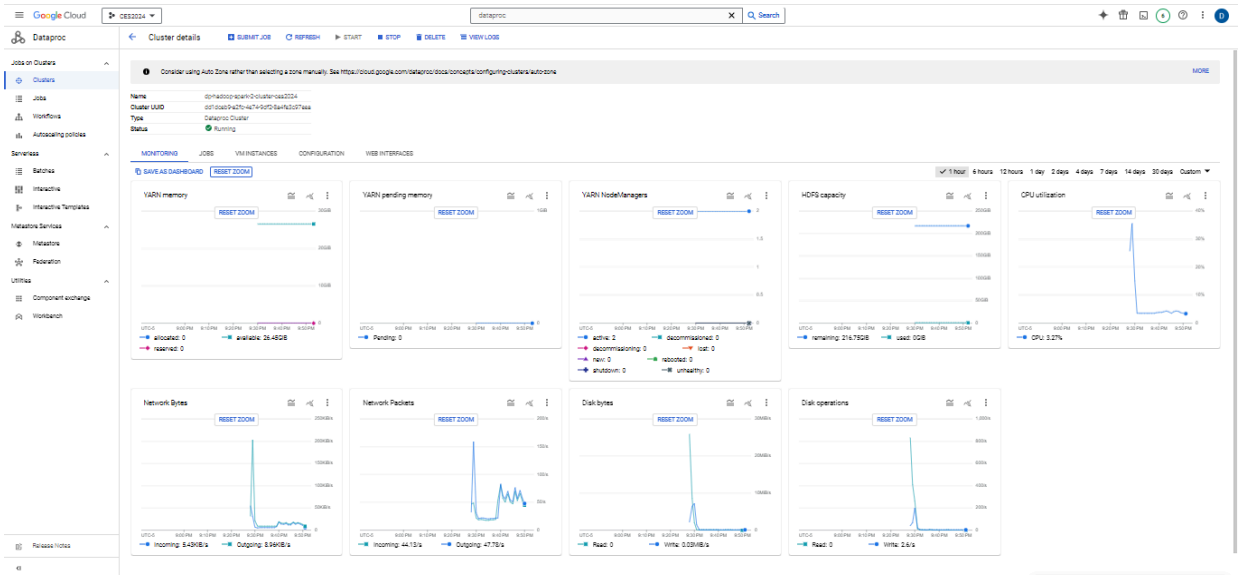




Explanation: Here we have entered the following command to the services that are running inside it ps -ef | grep hadoop. Where ps stands for process status -e flag to print all the processes within the system, -f to see a more detailed output. Here we can view the logs shown in the terminal

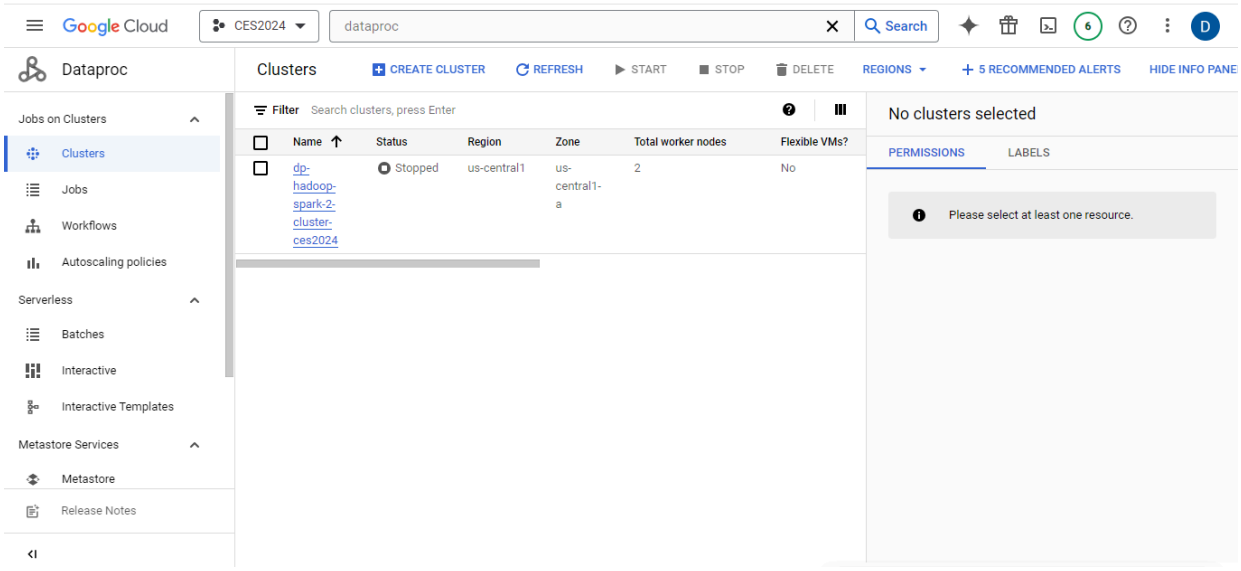
window, and find all the core subsystems that are currently running in the manager node of the cluster.

Screenshot 5 : Monitoring the Hadoop and Spark clusters.

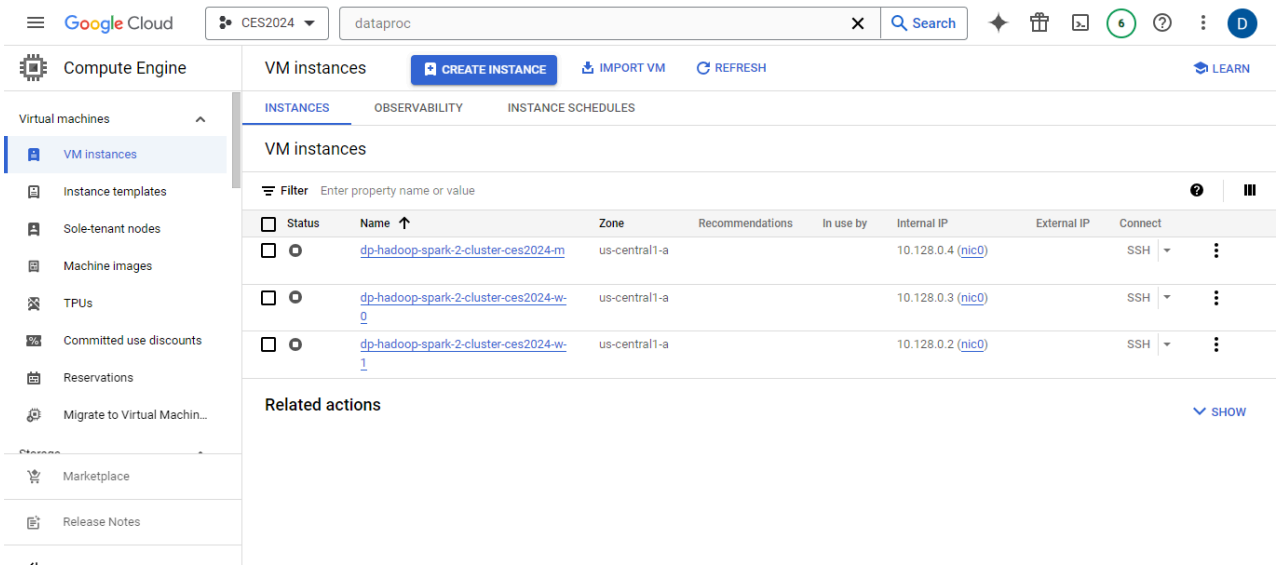


Explanation: Here when we click on the manager node we will get the panel like this which contains a monitoring tab, click on it. In this monitoring tab we can see the monitoring panels of the cluster and gather information about YARN memory. YARN pending memory, YARN NodeManagers, HDFS Capacity, etc.

Screenshot 6 : Stopping the Dataproc cluster, manager and worker nodes.



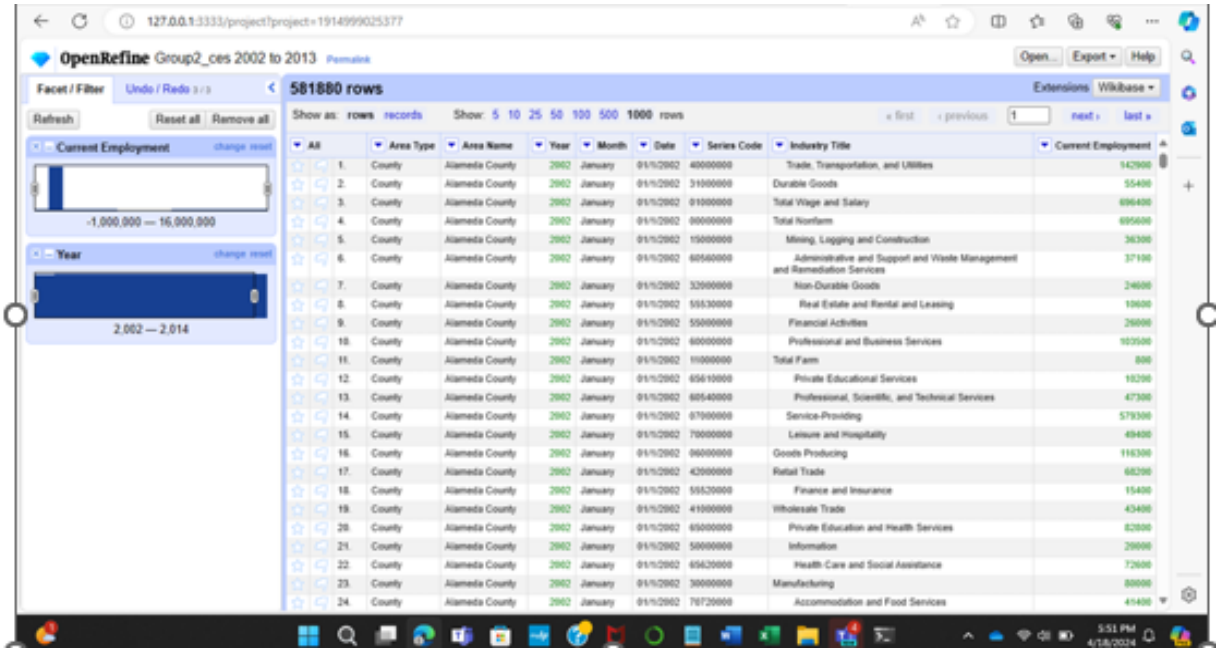
Here, we stopped the cluster.

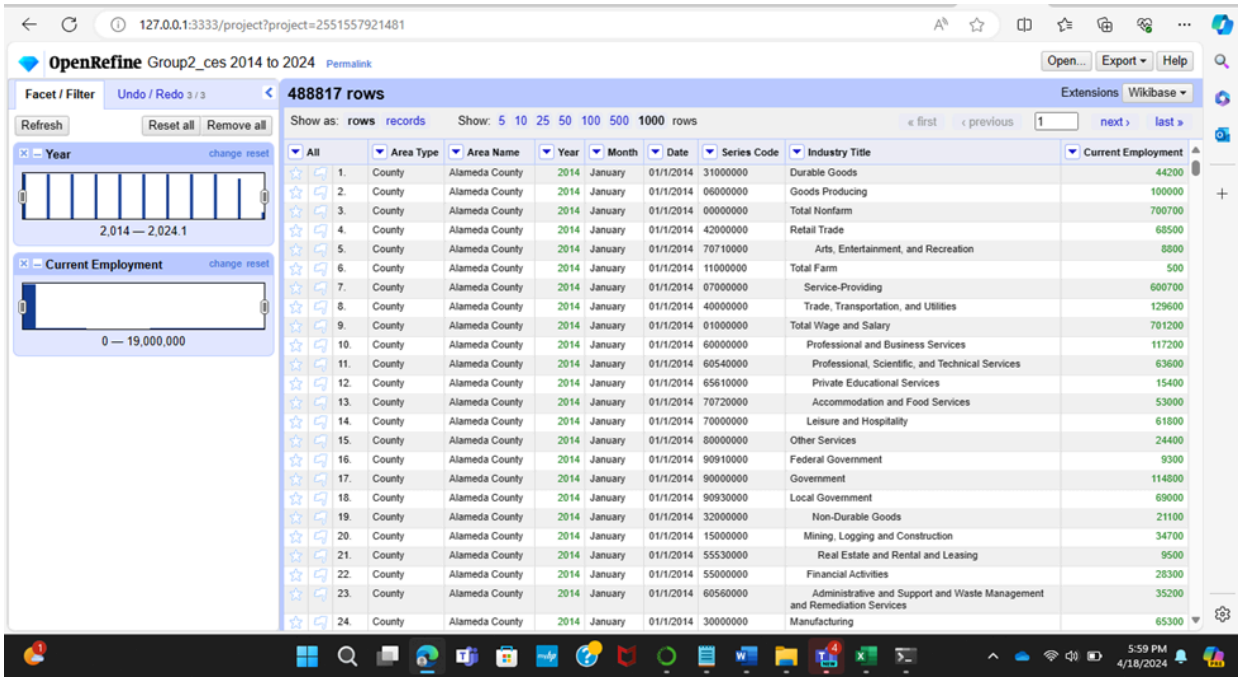


In the above screenshot all the three cluster nodes, manager and worker nodes are stopped.

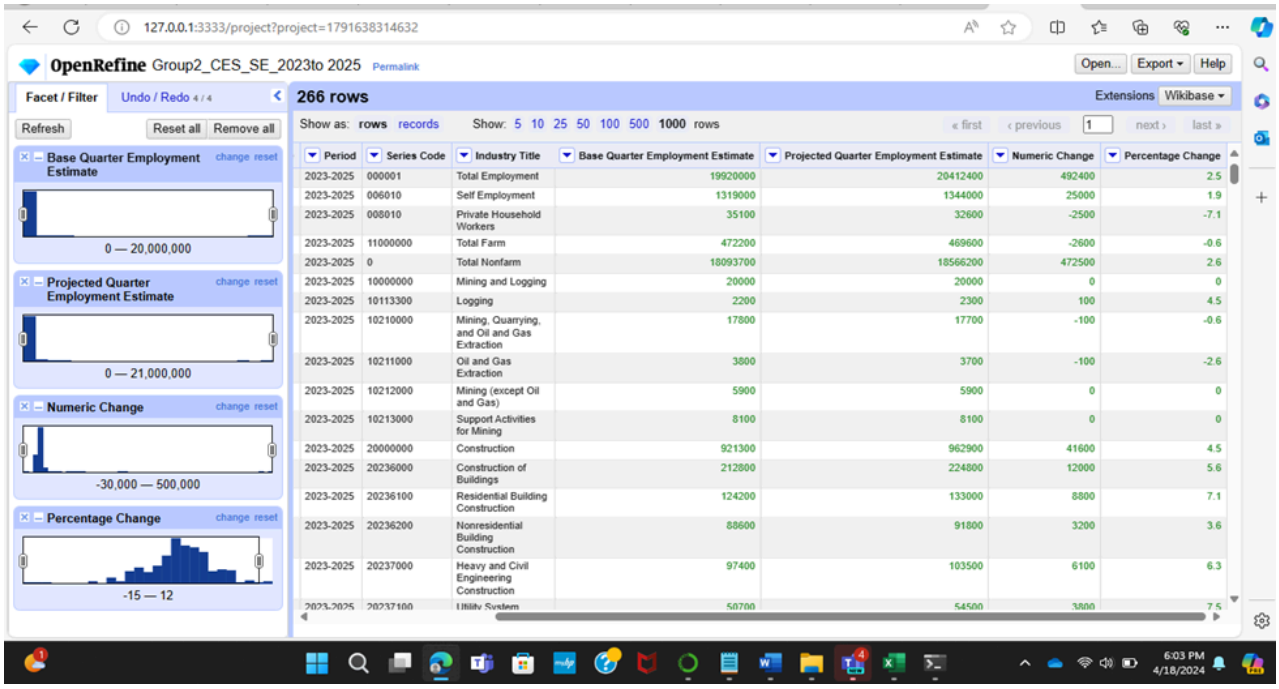
Screenshot 5: Preprocessing the raw data using openrefine(Sharanya & Tanaya)

We preprocessed our raw data using the openrefine and deleted the unnecessary column which is the seasonal variety. We also checked for missing values in each 3 static datasets using the text facet feature of all the columns in the datasets ces2002 to 2013 and ces2014 to ces2024 .we did not find any missing values in the both datasets.The year column and employment type column were converted to numeric values from their original text formats during the initial phase.The screenshots that follow are the results of preprocessing the data in both datasets.





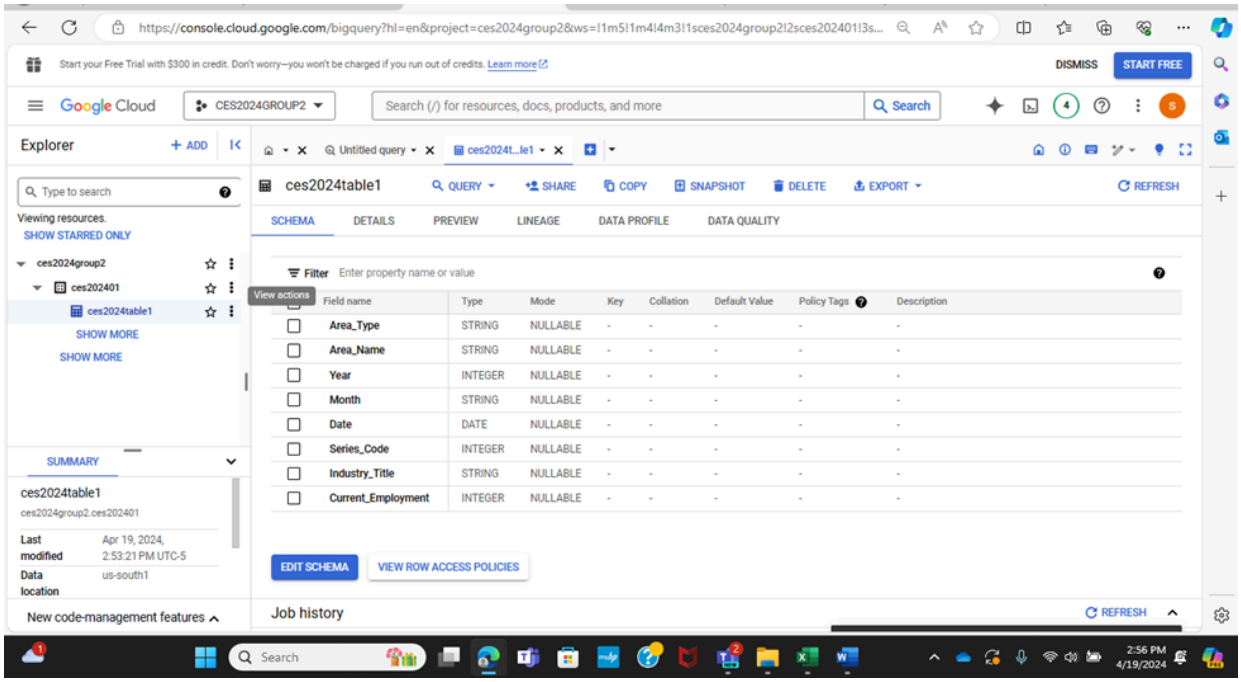
For the standard estimators dataset we checked for the missing values using the text facet feature of openrefine and we have many numeric columns like base Quarter employment estimate,projected employment estimate,numeric changes,percentage change.These cleaned data helps in deriving the insights.



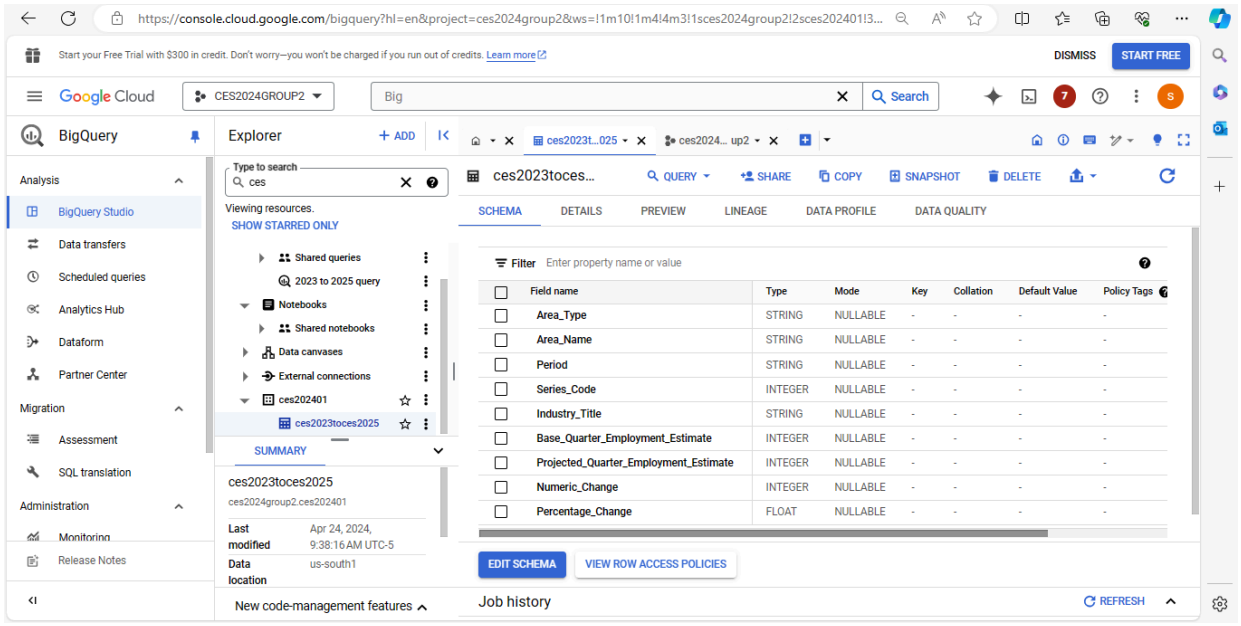
BigQuery : It is a Data Warehouse tool provided by the GCP that helps in managing our data encrypting our data and performing the analytics and visualizations.The below screenshots are the reference of how we setup our data in Bigquery and helping our analysts to perform the analytics and provide more insights to our data analysts and data scientists team.

Screenshot 6: BigQuery Screenshots (Sharanya & Tanaya)

After cleaning the three static datasets we uploaded each dataset into our project and were able to see the column names in the big Query studio.This helps in understanding different columns present in each dataset. We have the same column names in the two datasets.



Below are the column names for the ces 2023 to ces2025 datasets.



We chose a specific job category, 'Hospitals', and created a query to show the total employment for hospitals from 2002 to 2013. We also generated a time series graph to assist our analysts and scientists in understanding trends and delving deeper into analysis utilizing these queries for different types of employment. The Query which we selected is by using

SELECT Year,Month,Industry_Title,SUM(Current_Employment) As Total Employment

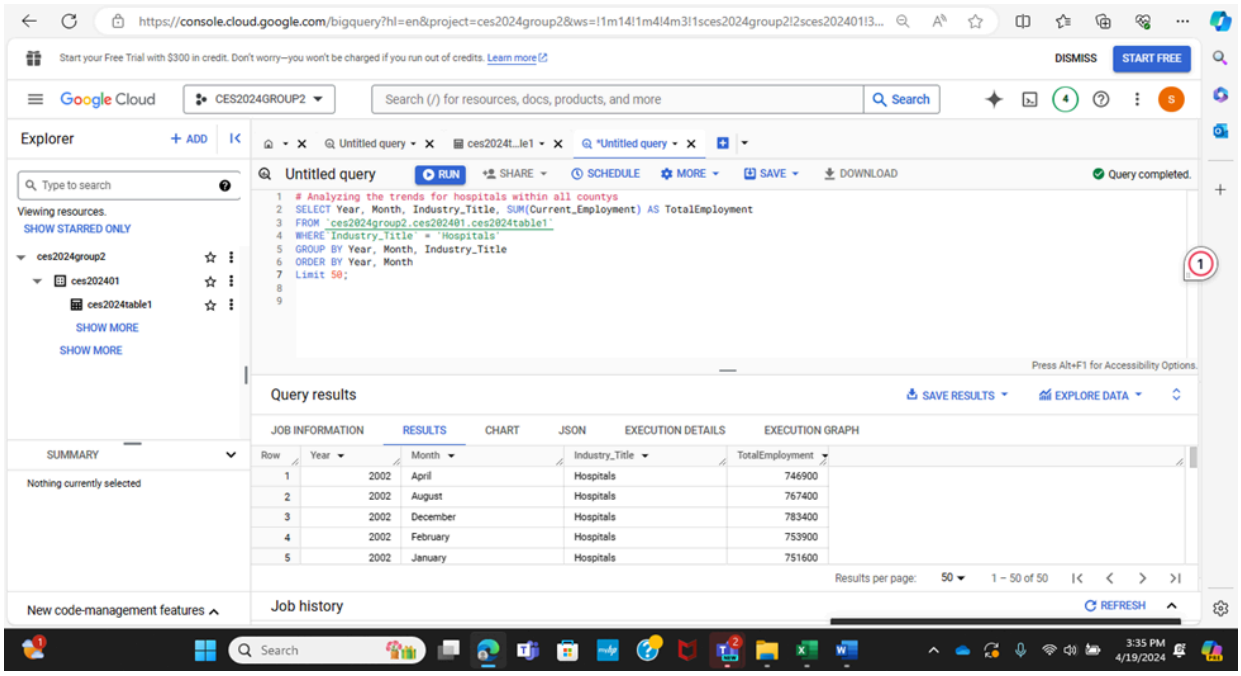
FROM ces2024group2.ces202401.ces2024table1

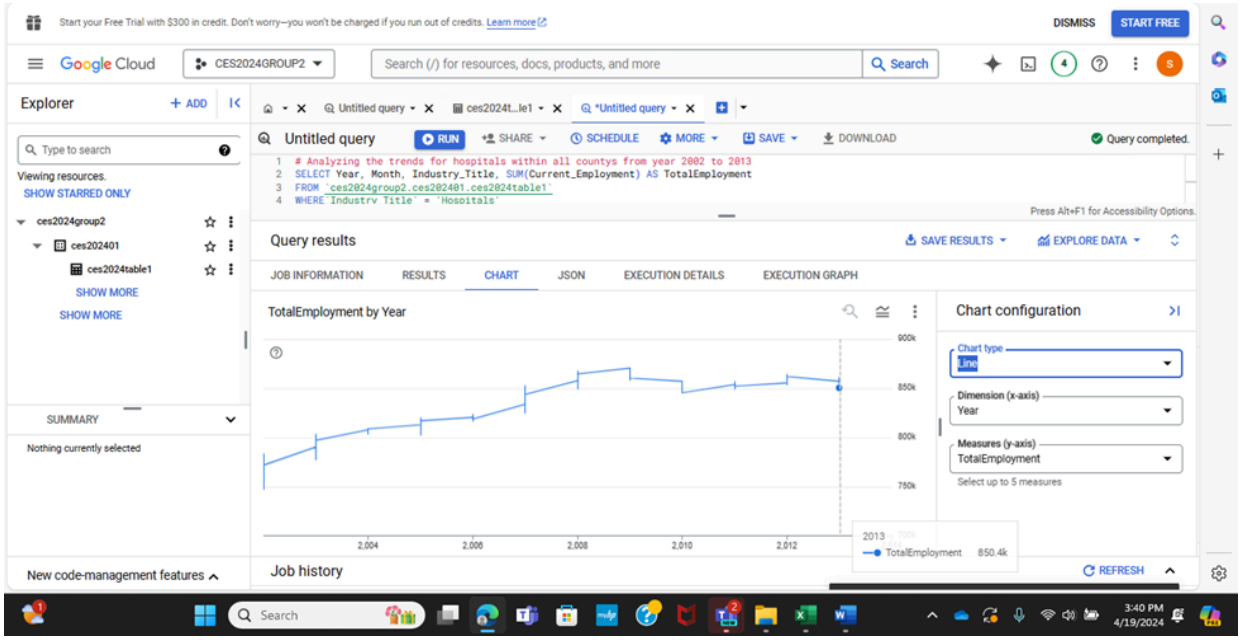
WHERE Industry_Title` = 'Hospitals'

GROUP BY Year, Month, Industry_Title

ORDER BY Year, Month

Limit 50;





*We used the same query for analyzing the hospital employment trends from 2014 to 2024.

SELECT Year,Month,Industry_Title,SUM(Current_Employment) As Total Employment

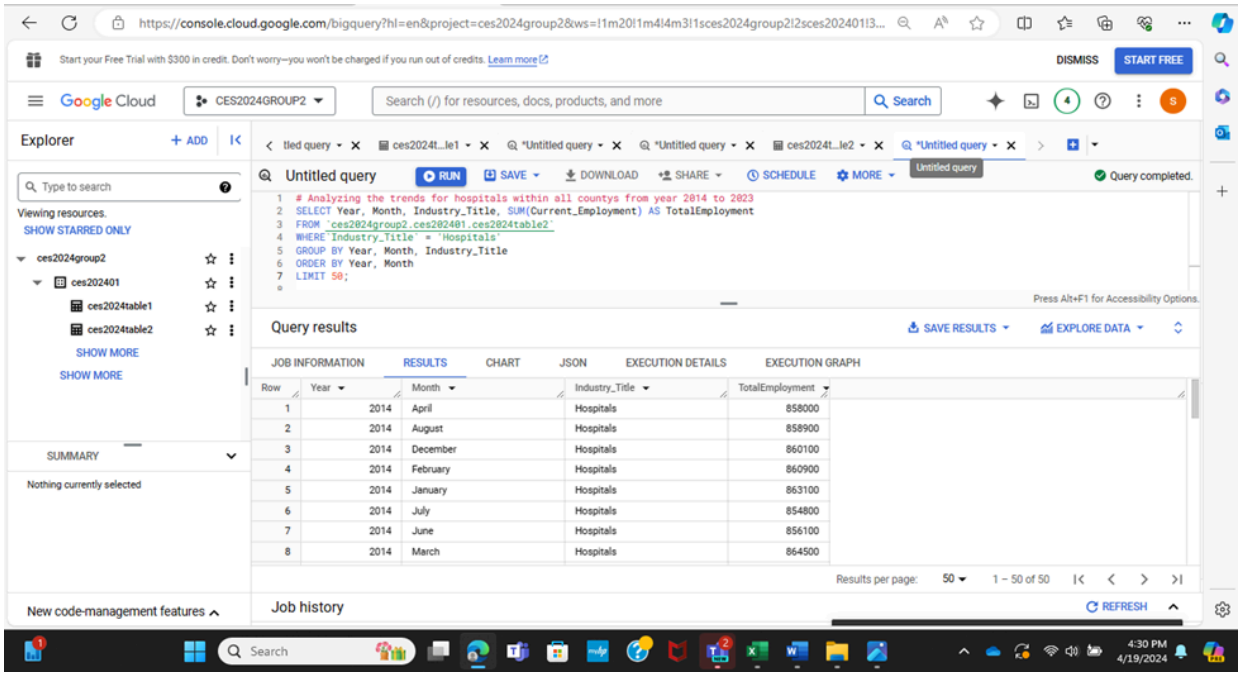
FROM ces2024group2.ces202401.ces2024table1

WHERE Industry_Title` = 'Hospitals'

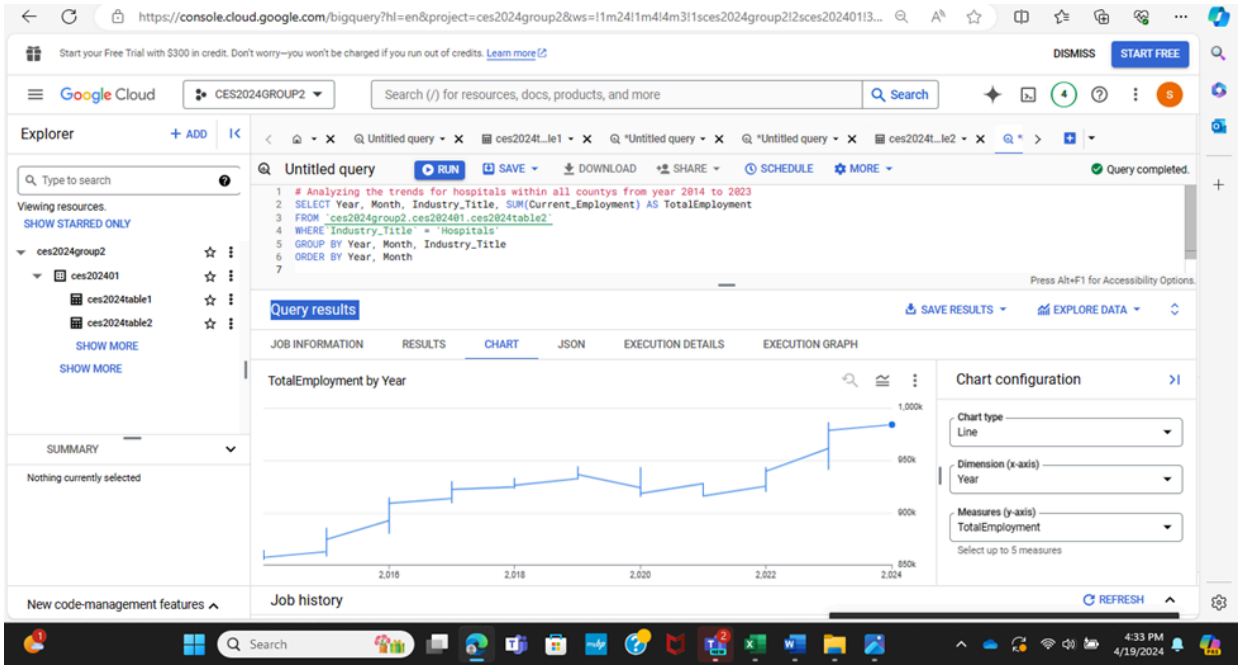
GROUP BY Year, Month, Industry_Title

ORDER BY Year, Month

Limit 50;



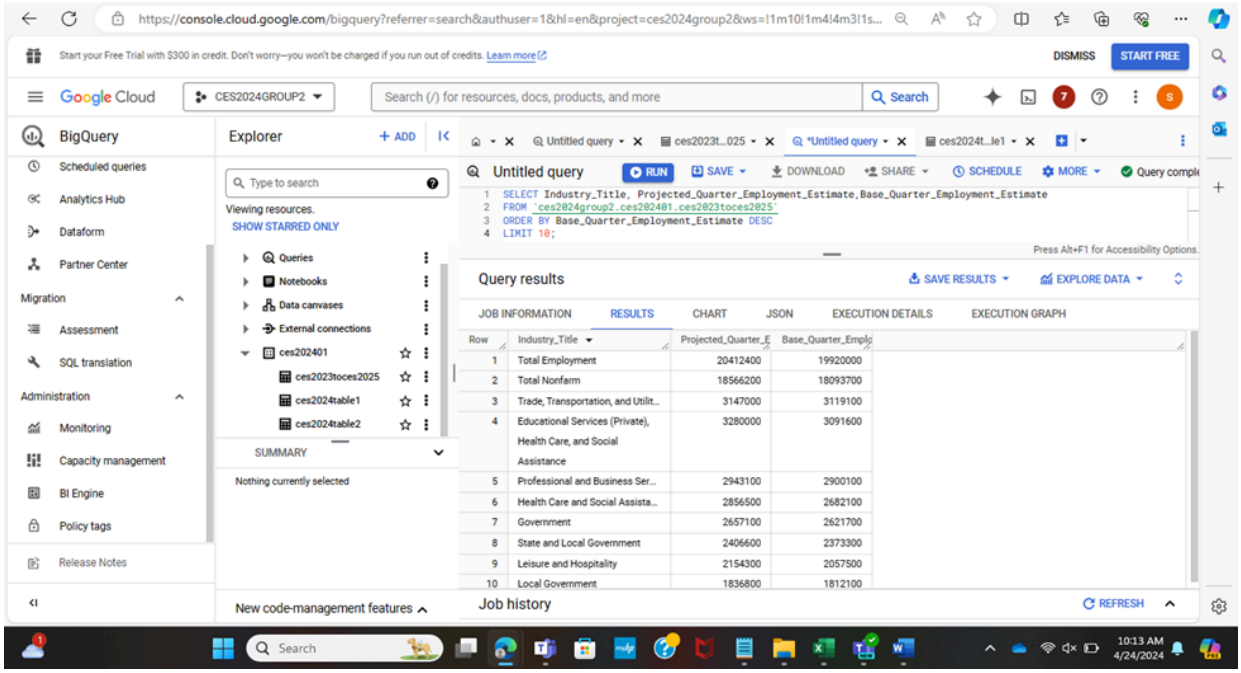
The visualization of the above query with the trends using the line graph.

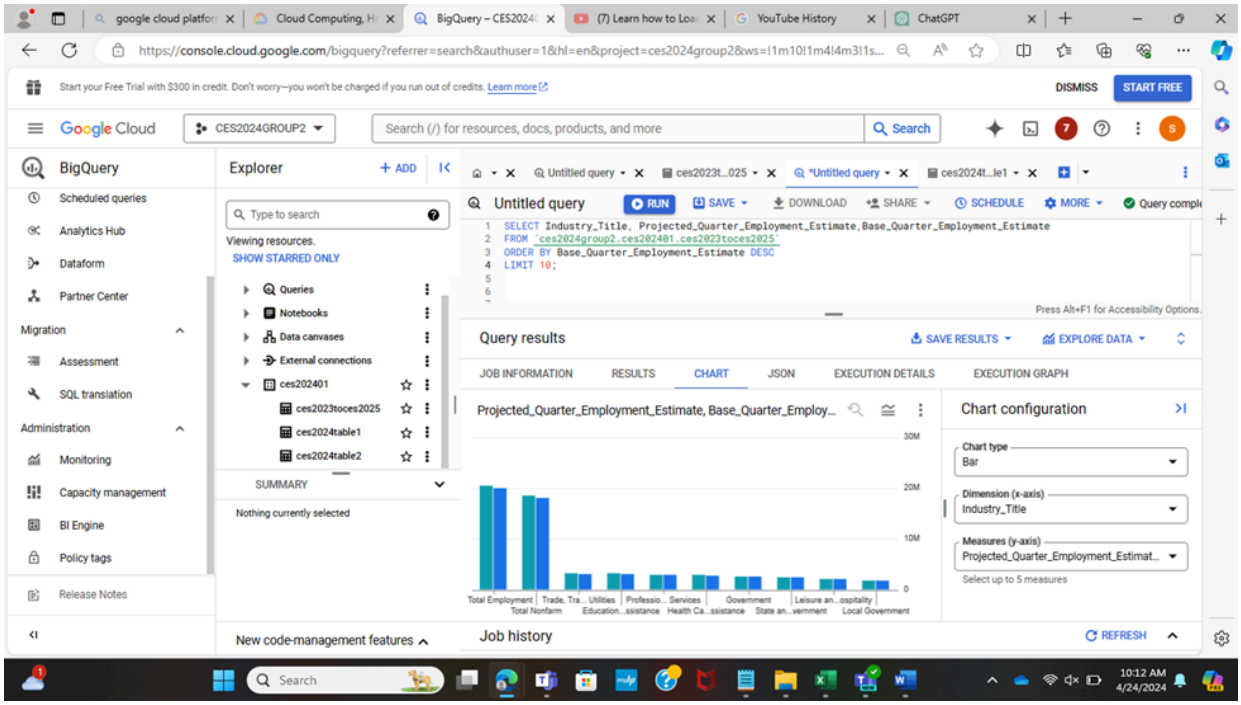


We are selecting the top 10 industries with the base employment and Quarter employment using the queries in the below screenshot and visualizing it by comparing the base employment.

We executed below screenshot Query that helps in providing the Top 10 of both projected and Base Quarter employment by the Industry.

```sql
SELECT Industry_Title, Projected_Quarter_Employment_Estimate,Base_Quarter_Employment_Estimate
FROM `ces2024group2.ces202401.ces2023toces2025`
ORDER BY Base_Quarter_Employment_Estimate DESC
LIMIT 10;
```





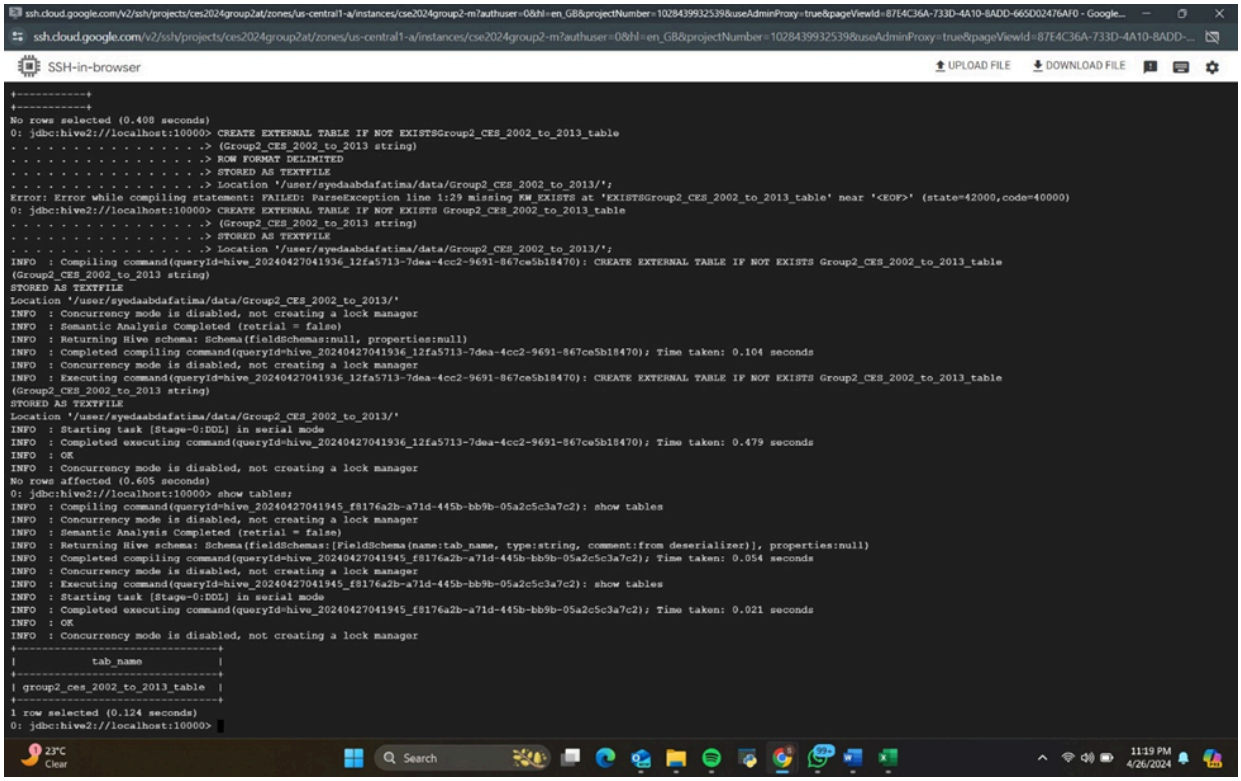
* By setting up these data management using Bigquery and encrypting them helps the analysts to perform the analytic techniques for the deeper analysis on the data for various other categories of the employment and give insights.

Screenshot 7 : Hive and Spark

Wrangling and Querying Data (Hive & Spark SQL)

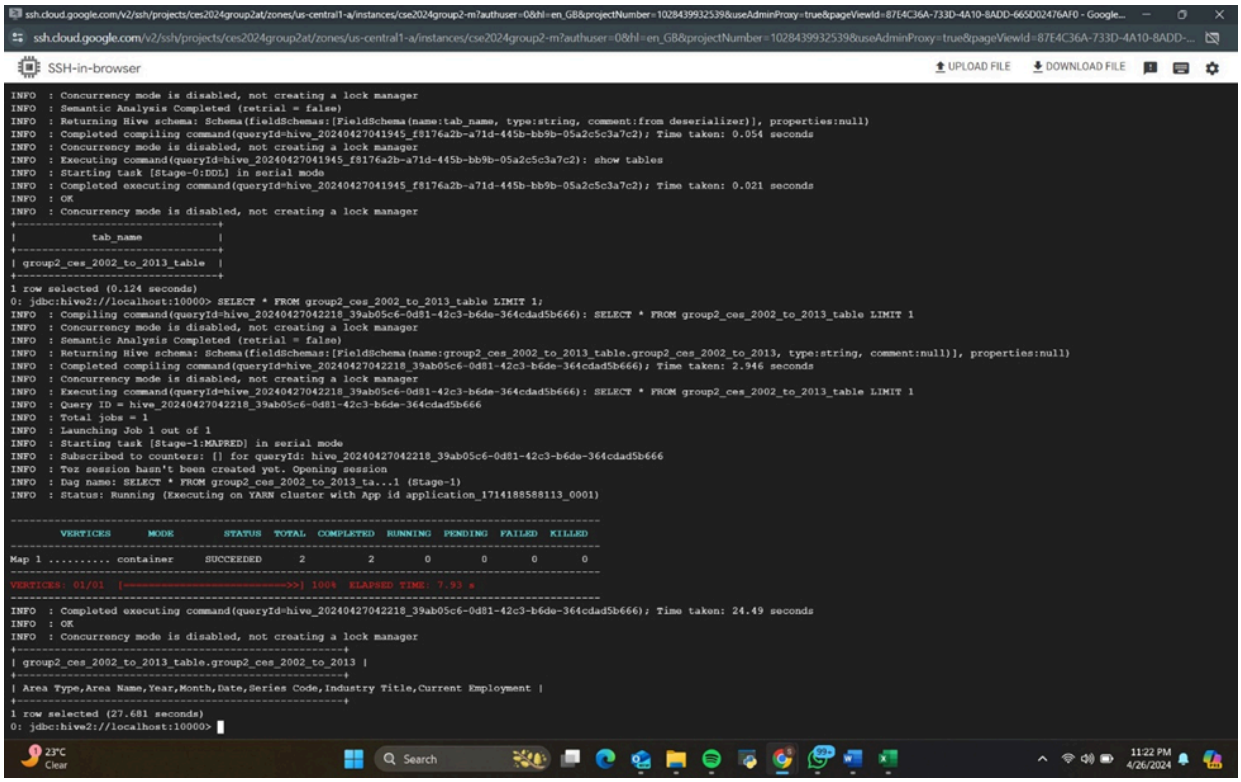
First dataset: 2002_to_2013

Hive Screenshot: Creating the table.



Tools used :creating the external table is Apache Hive. This is done in order to import the data from google cloud platform's storage bucket and transfer into a Hive metastore. Tables are created in Hive and stored for further processing of the data. The create table will be stored in the sub directory of the database where the table is stored

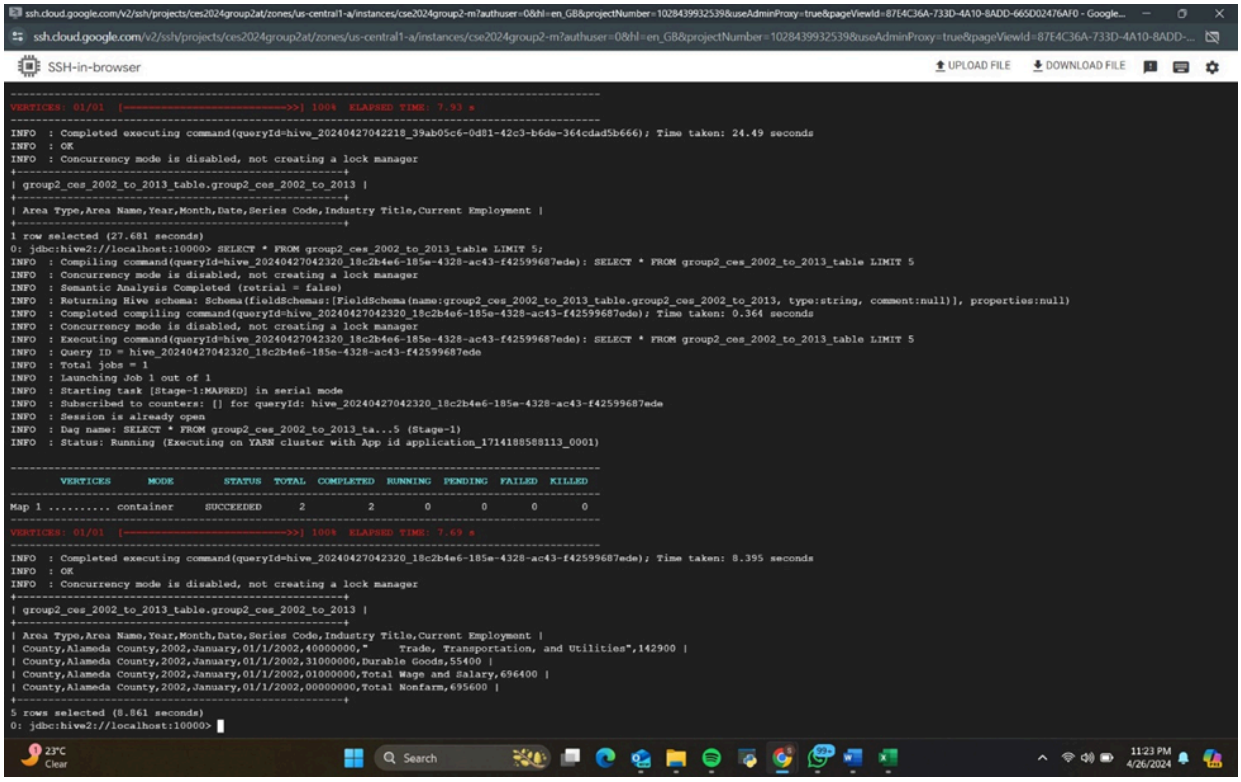
Hive Query Results



Query 1 in Hive: SELECT * FROM Group2_CES_2002_to_2013_table LIMIT 1;

Time: 24.49 seconds.

The Screenshot includes the querying data set Group2_2002_to_2013 which in the storage bucket and processing the results. The time taken to execute the query is 24.49 seconds



Query 2 in Hive: SELECT * FROM Group2_CES_2002_to_2013_table LIMIT 5;

Time: 8.861 seconds.

HiveQL is been used to query the data stored in Apache Hive. The query is run on the data of group2_ces_2002_to_2013 limit 5 this implies that the first 5 rows of the table are retrieved from the table.The time taken by Hive to execute this query is 8.861 seconds

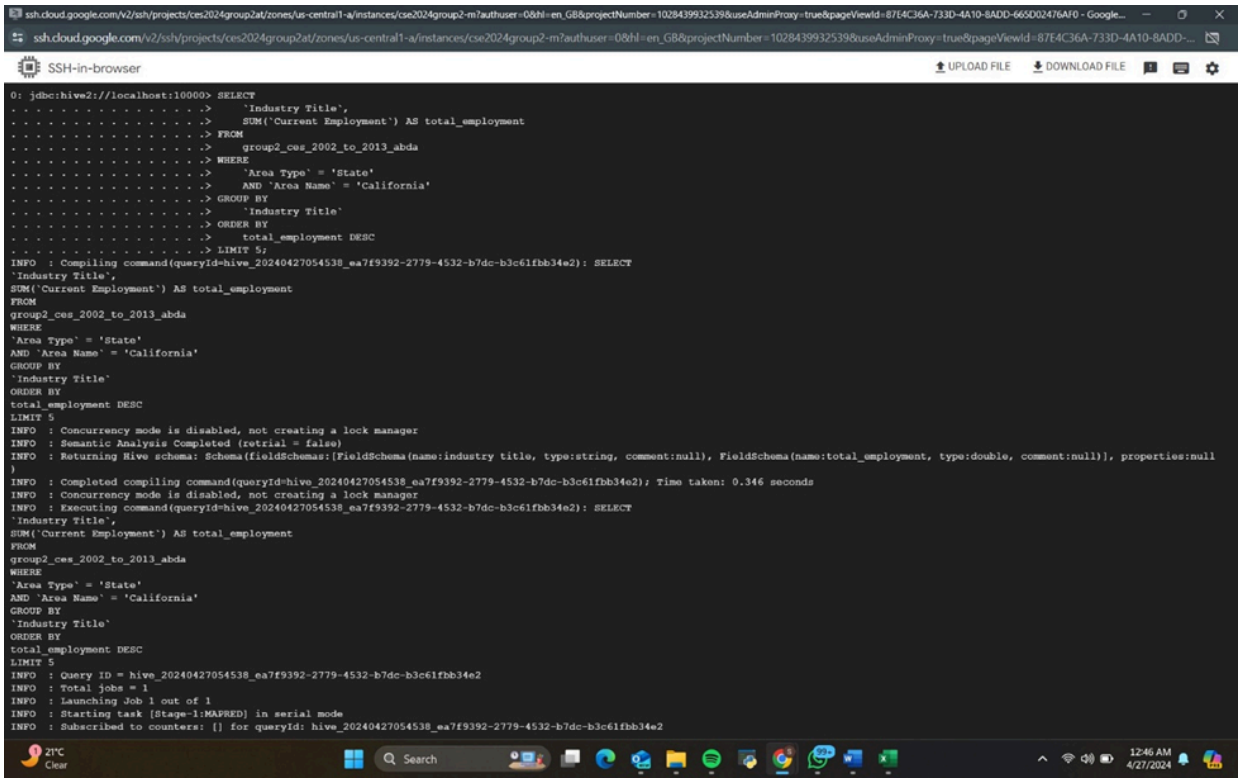
Running a more complex query.

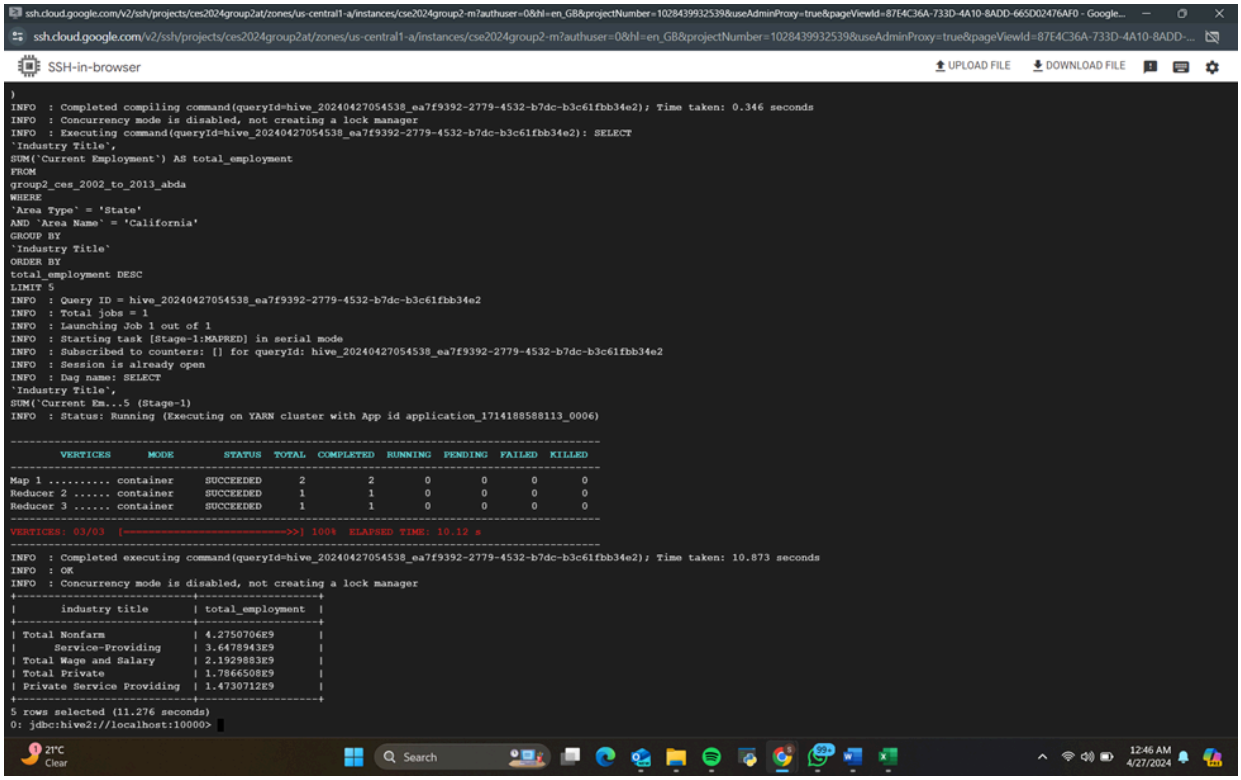
Query Question:

What are the top 5 industries with the highest average employment in California from 2002 to 2013?

Explanation: Running a more complex query.

This query analyzes employment data from the years 2002 to 2013 in California. It calculates the average employment for each industry and selects the top 5 industries with the highest average employment. The result provides insight into which industries had the most significant average employment in California during the specified period.



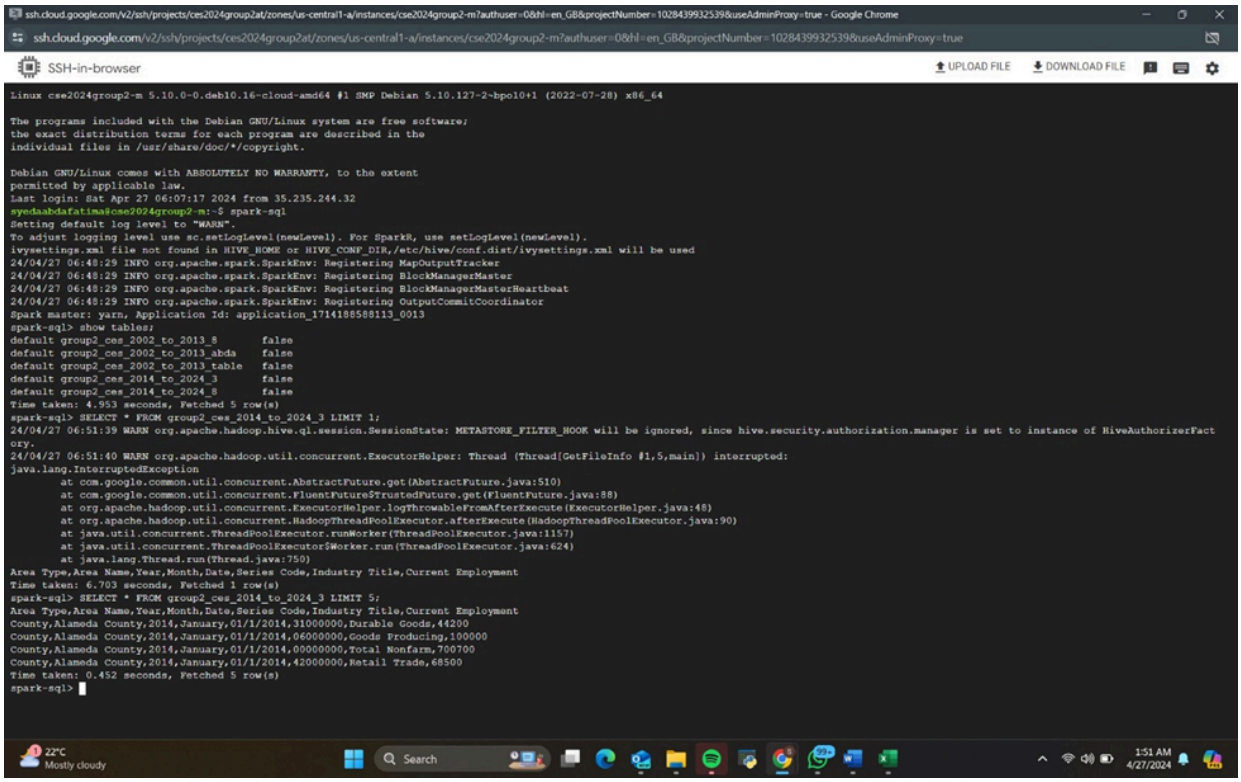


Query 3 in Hive: top 5 industries with the highest average employment in California from 2002 to 2013

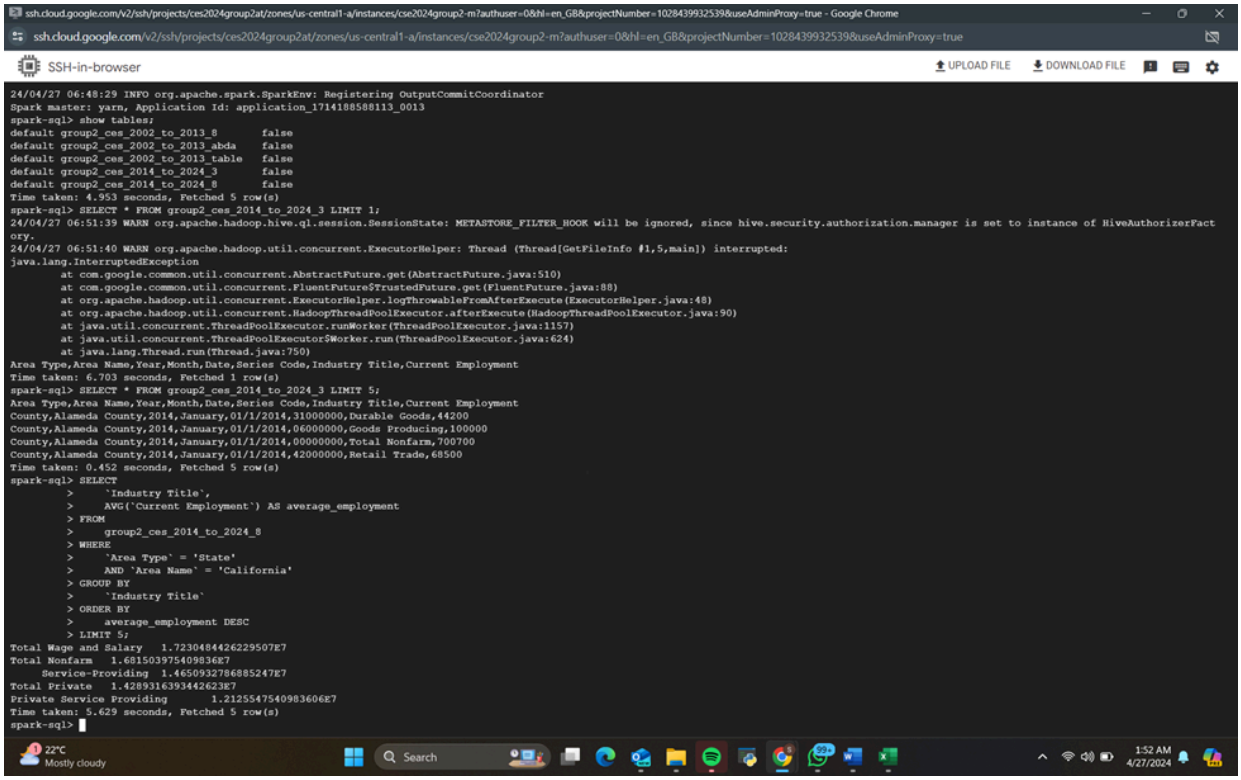
Time: 11.276 seconds.

Several data columns, including "Industry Title," "SUM(Current Employment) AS total_employment," "Area Type," "Area Name," and "State," are specified in the query to be returned. This implies that the inquiry is focused on total employment by area type, location (state), and industry. The top 5 industries with highest average employment are retrieved the time taken by to retrieve this complex query is 11.27 seconds.

Spark Screenshot: running queries in spark



Spark SQL acts as distributed query engine using its JDBC/ODBC or command-line interface. once this is executed, one can interact with Spark SQL directly to run SQL queries. The same queries are run in spark and the time taken by spark to execute the queries is compared.



Spark Results

Query 1 in Spark: SELECT * FROM Group2_CES_2002_to_2013_3 LIMIT 1;

Time: 6.703 seconds.

Query 2 in Spark: SELECT * FROM Group2_CES_2002_to_2013_3 LIMIT 5;

Time: 0.452 seconds.

Query 3 in Spark: top 5 industries with the highest average employment in California from 2002 to 2013

Time: 5.629 seconds.

RESULTS: CES_2002_to_2013

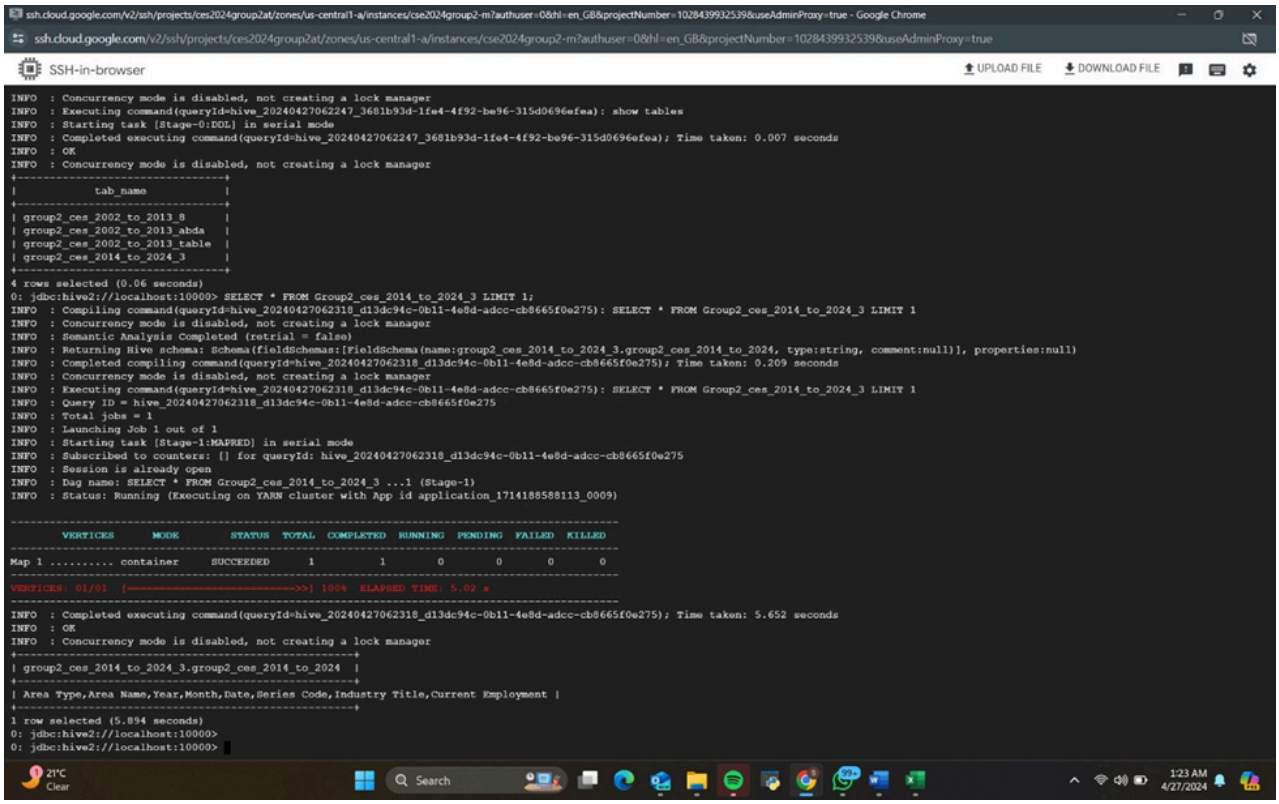
Query 1: It was **24.49 seconds** in Hive. In Spark, it was **6.703 seconds.**

Query 2: It was **8.861 seconds** in Hive. In Spark, it was **0.452 seconds.**

Query 3: It was **11.276 seconds** in Hive. In Spark, it was **5.629 seconds.**

Second Dataset:CES_2014_to_2023

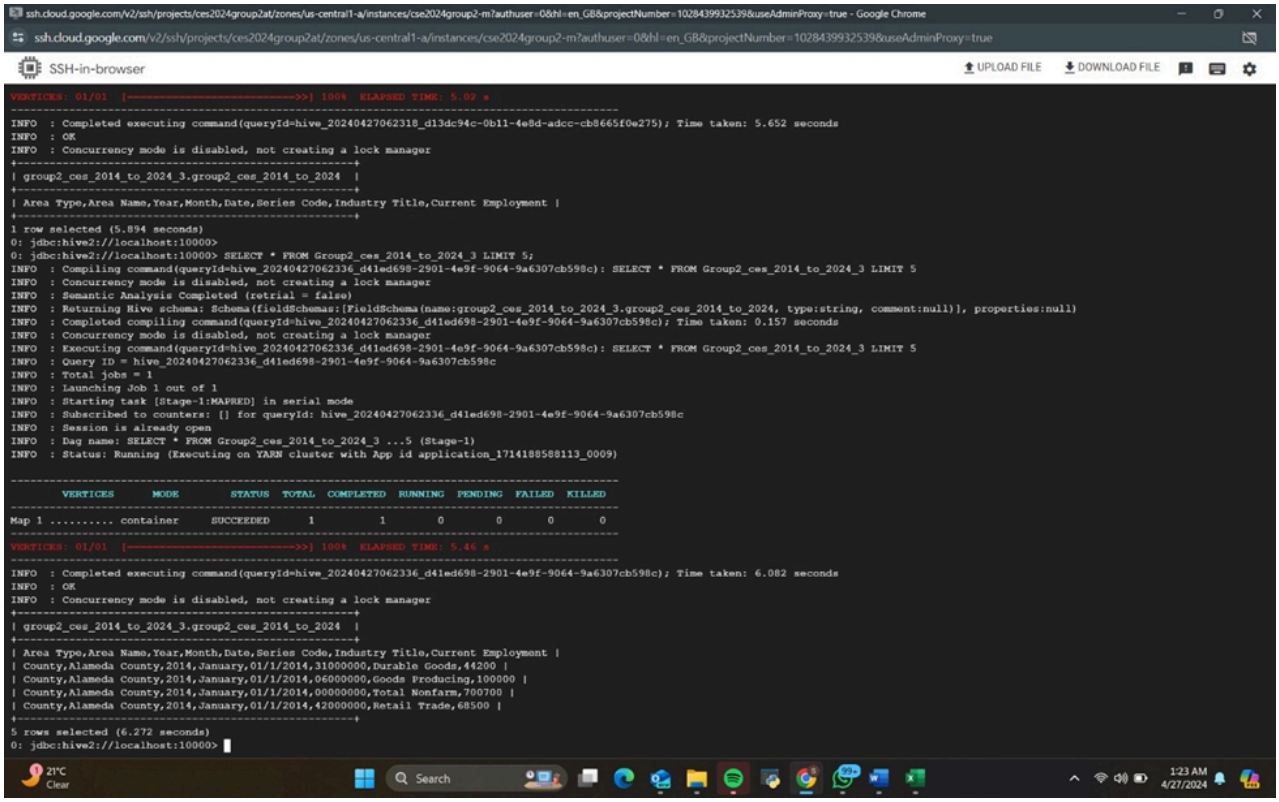
Hive Screenshots



HIVE RESULTS

Query 1 in Hive: SELECT * FROM Group2_ces_2014_to_2024_3 LIMIT 1;

Time: 5.894 seconds.



Query 2 in Hive: SELECT * FROM Group2_ces_2014_to_2024_3 LIMIT 5;

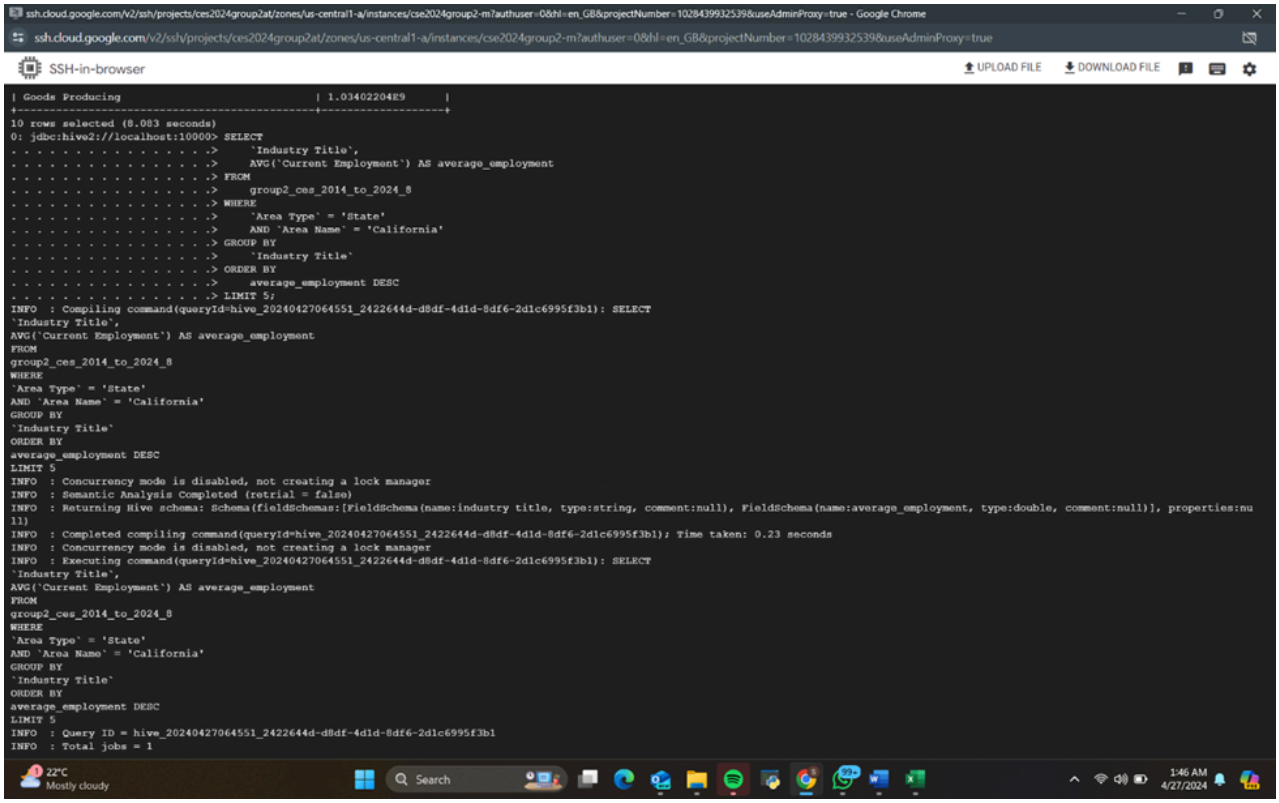
Time: 6.272 seconds.

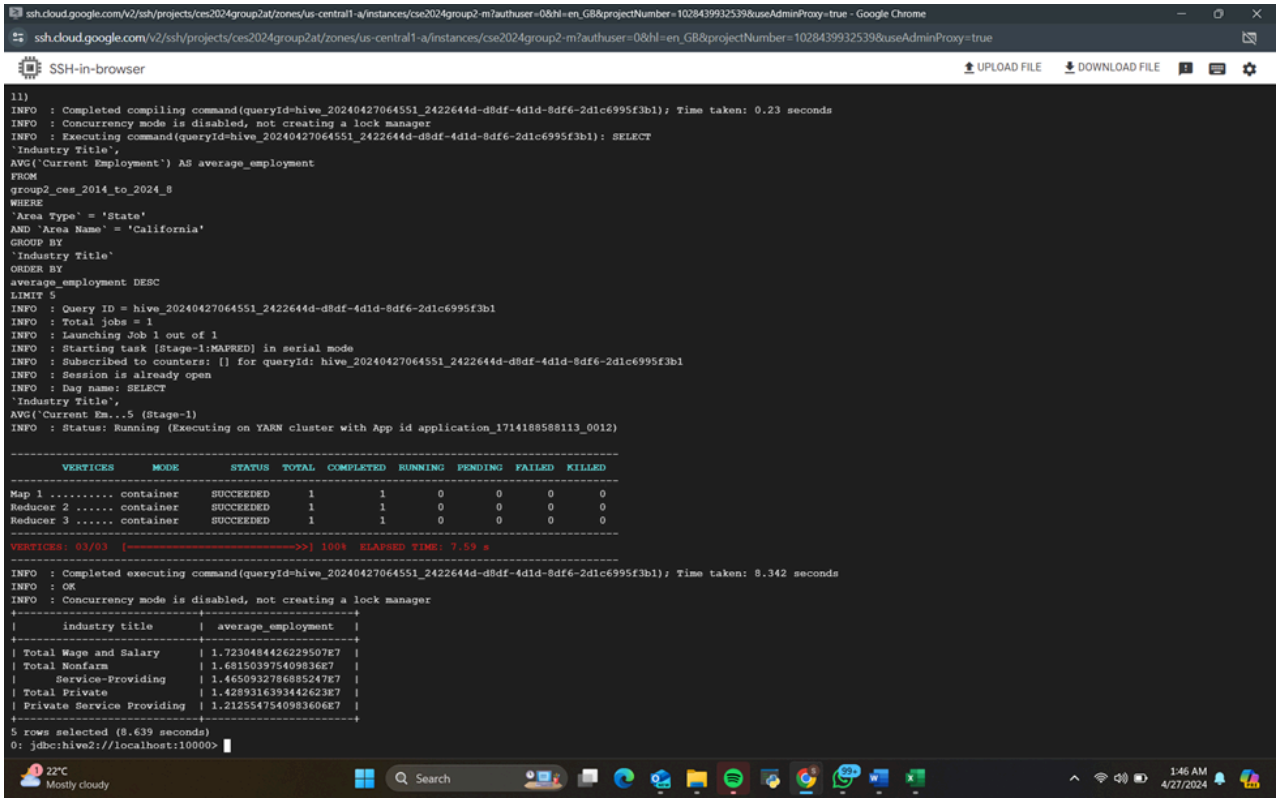
Running a more complex query.

Query 2: What are the top 5 industries with the highest average employment in California from 2014 to 2024?

Explanation:

This query analyzes employment data from the years 2014 to 2024 in California. It calculates the average employment for each industry and selects the top 5 industries with the highest average employment. The result provides insight into which industries have the most significant contribution to employment in California during the specified period.

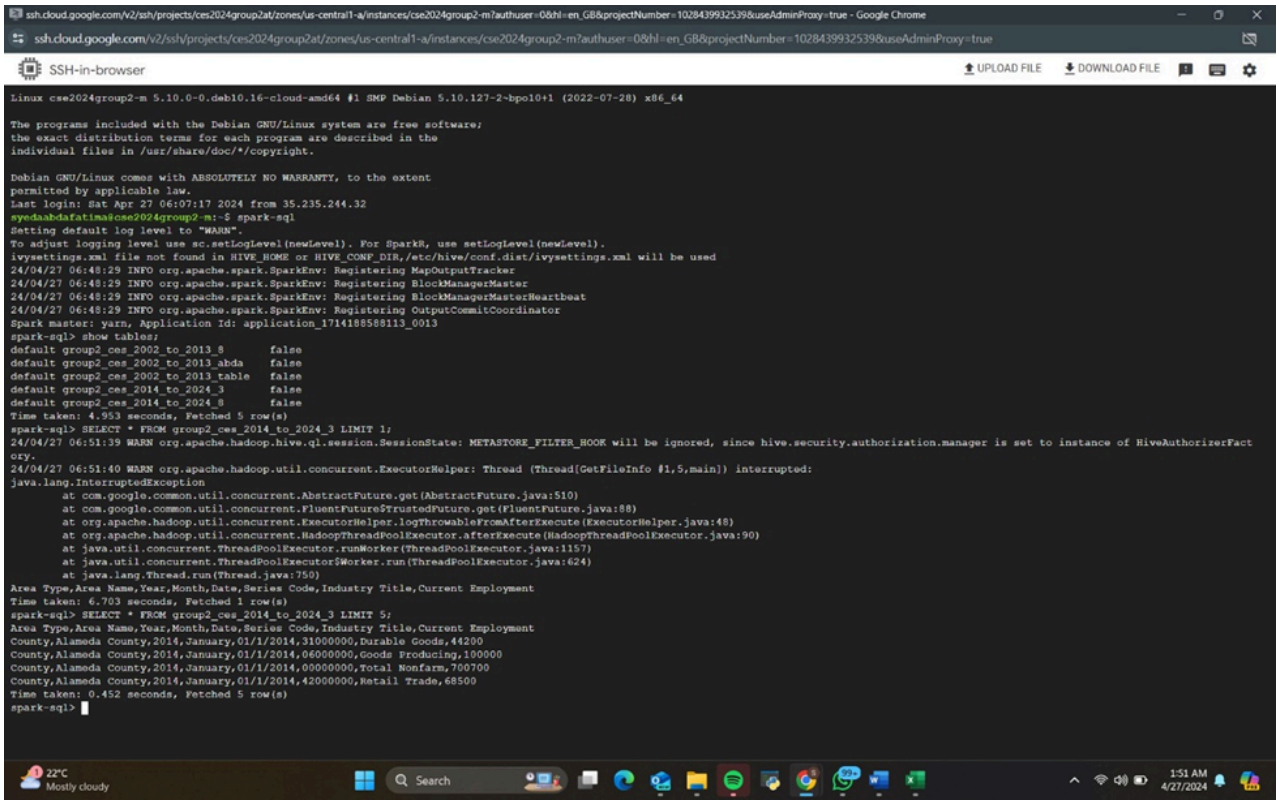


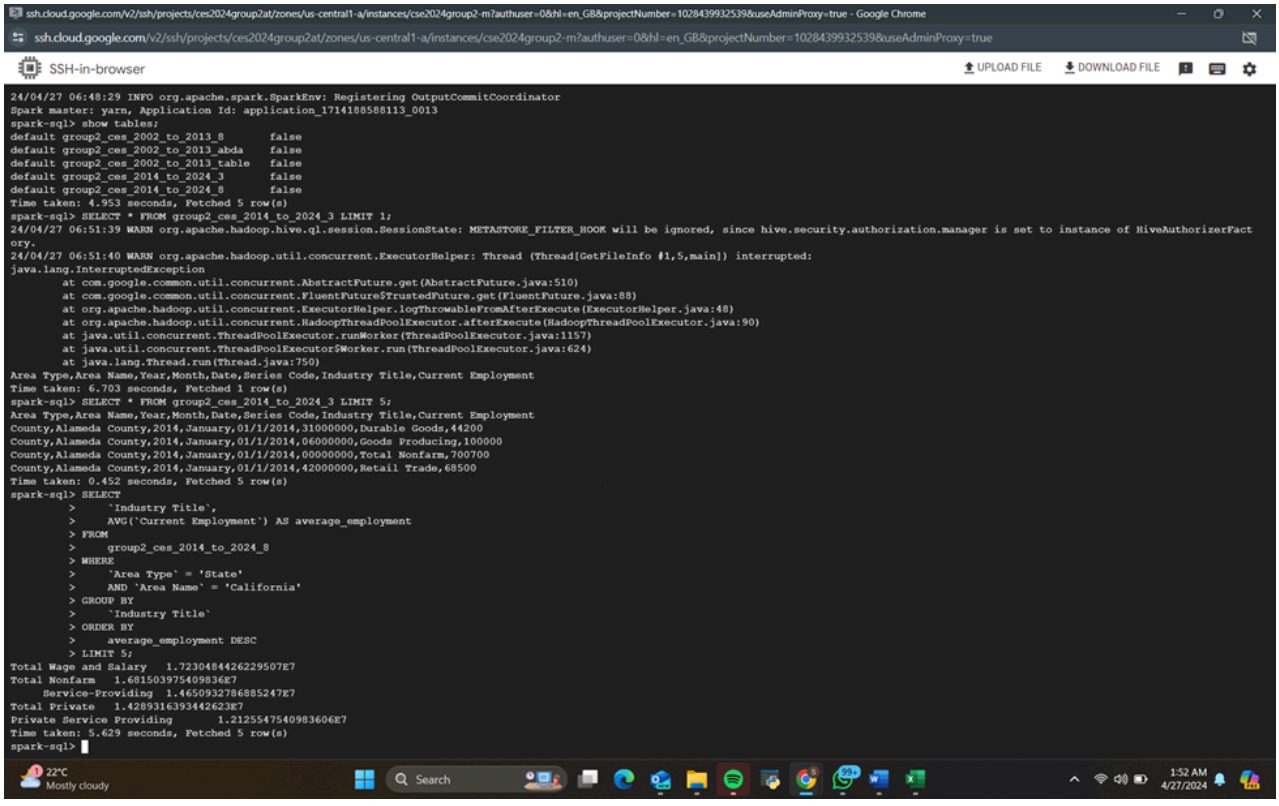


Query 3 in Hive: top 5 industries with the highest average employment in California from 2014 to 2024.

Time: 8.639 seconds.

Screenshot 8: Spark Screenshot





Spark API Results

Query 1 in Spark: SELECT * FROM Group2_ces_2014_to_2024_3 LIMIT 1;

Time: 6.703 seconds.

Query 2 in Spark: SELECT * FROM Group2_ces_2014_to_2024_3 LIMIT 5;

Time: 0.452 seconds.

Query 3 in Spark: top 5 industries with the highest average employment in California from 2014 to 2024.

Time: 5.629 seconds.

RESULT - CES_2014_to_2024

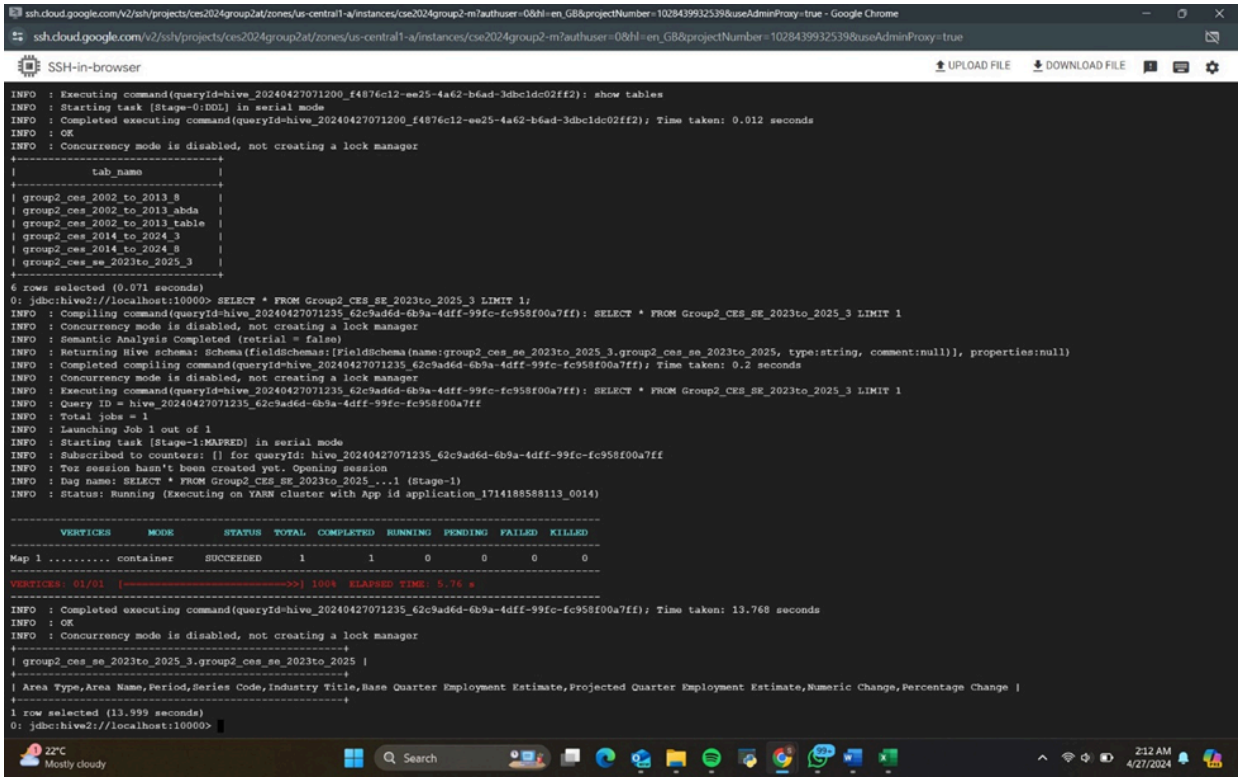
Query 1: It was **5.894 seconds** in Hive. In Spark, it was **6.703 seconds.**

Query 2: It was **6.272 seconds** in Hive. In Spark, it was **0.452 seconds.**

Query 3: It was **8.639 seconds** in Hive. In Spark, it was **5.629 seconds.**

CES_2023_to_2025

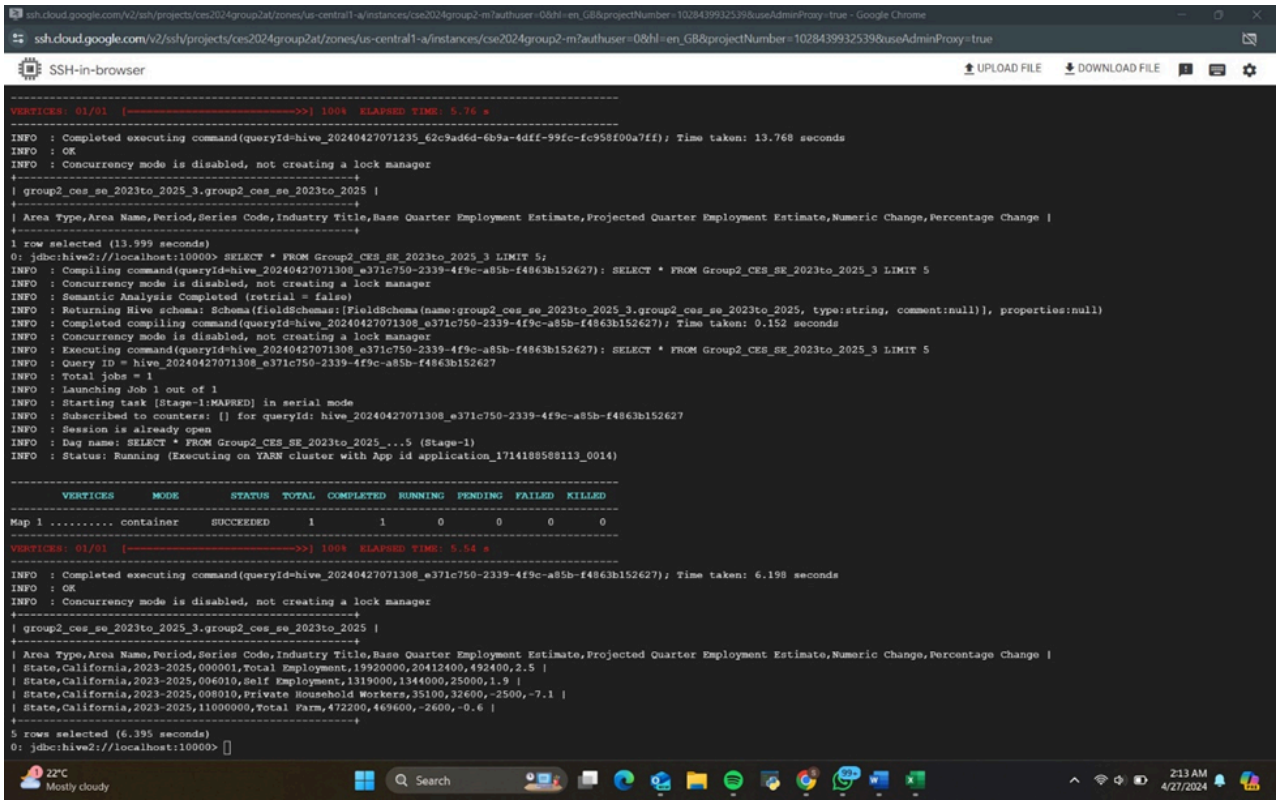
SCREEN SHOT #1- Hive Screenshot



Query 1 in Hive: SELECT * FROM Group2_CES_SE_2023_to_2025_3 LIMIT 1;

Time: 13.999 seconds.

Hive Screenshot



Query 2 in Hive: SELECT * FROM Group2_CES_SE_2023to-2025_3 LIMIT 5;

Time: 6.395 seconds

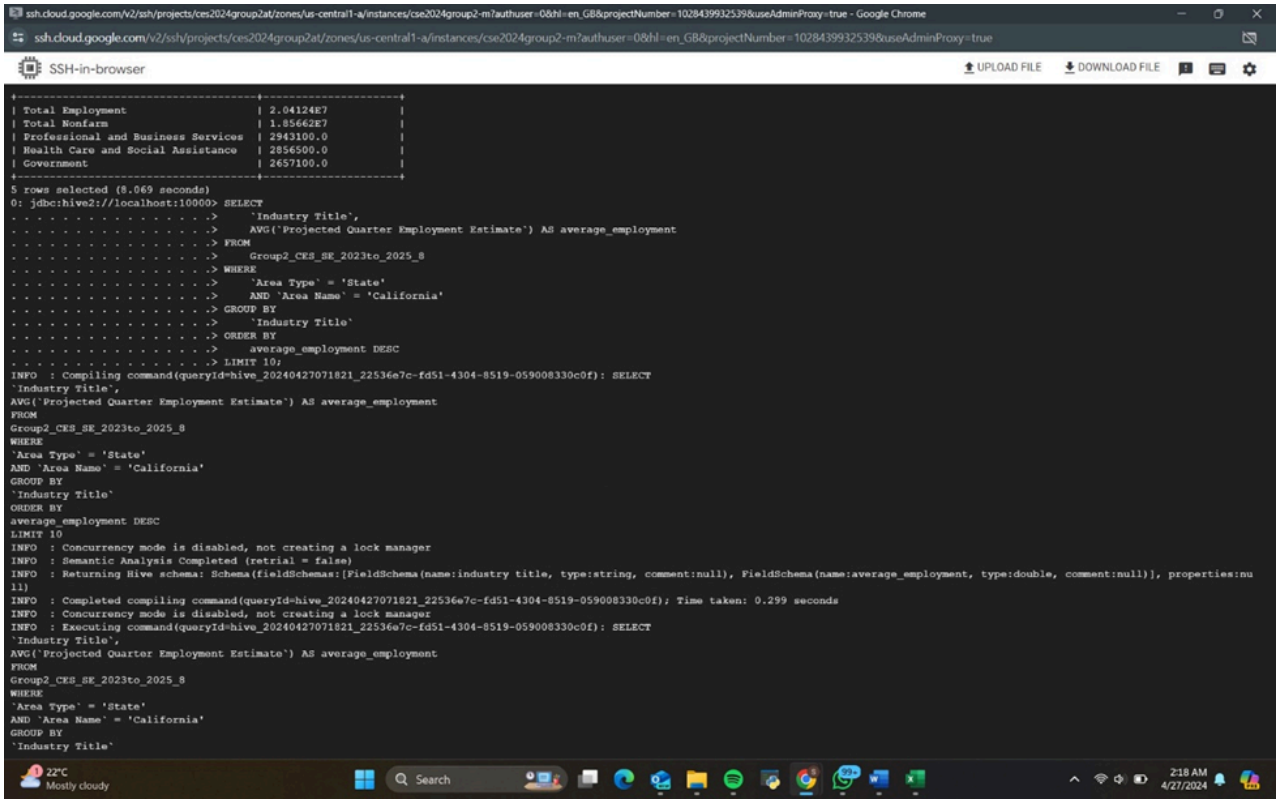
Hive Screenshot

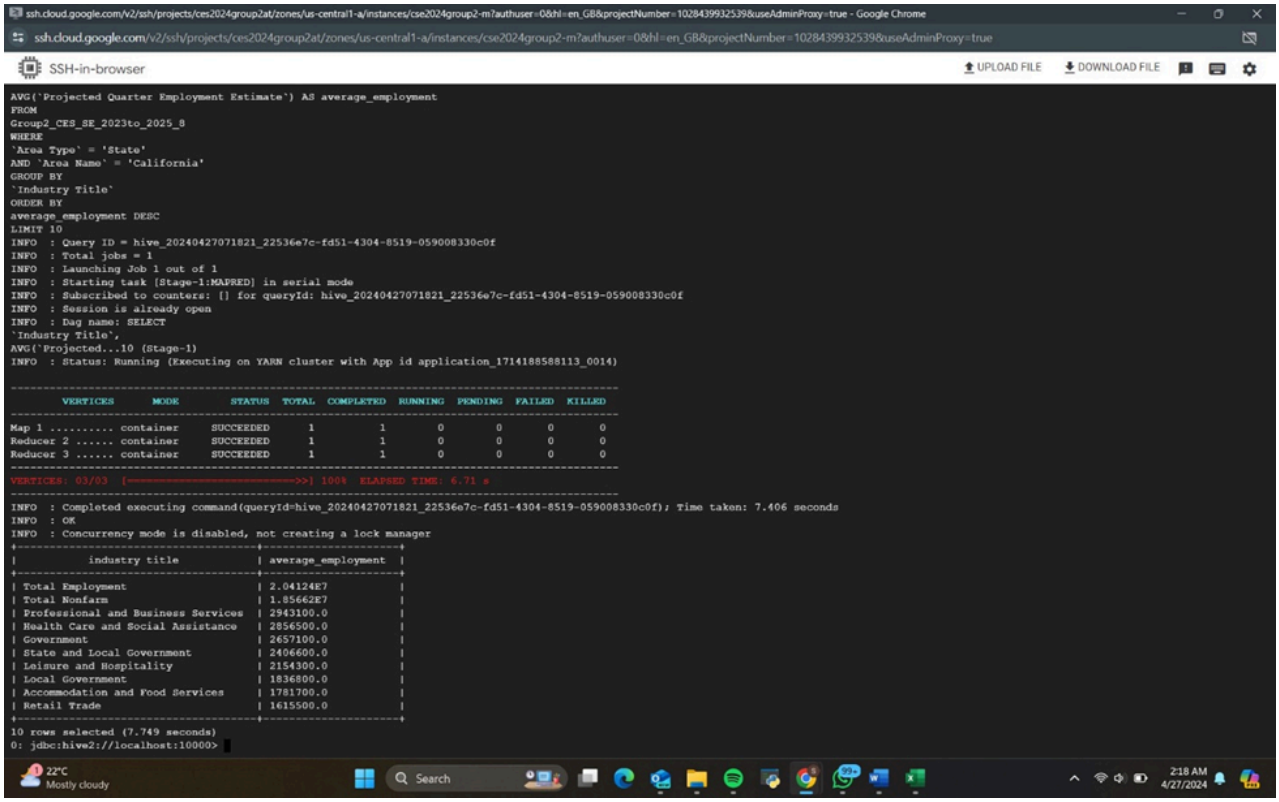
Running a more complex query.

Query: What are the top 10 industries with the highest average projected quarter employment in California?

Explanation:

This query retrieves the top 10 industries with the highest average projected quarter employment in California for the period 2023-2025.

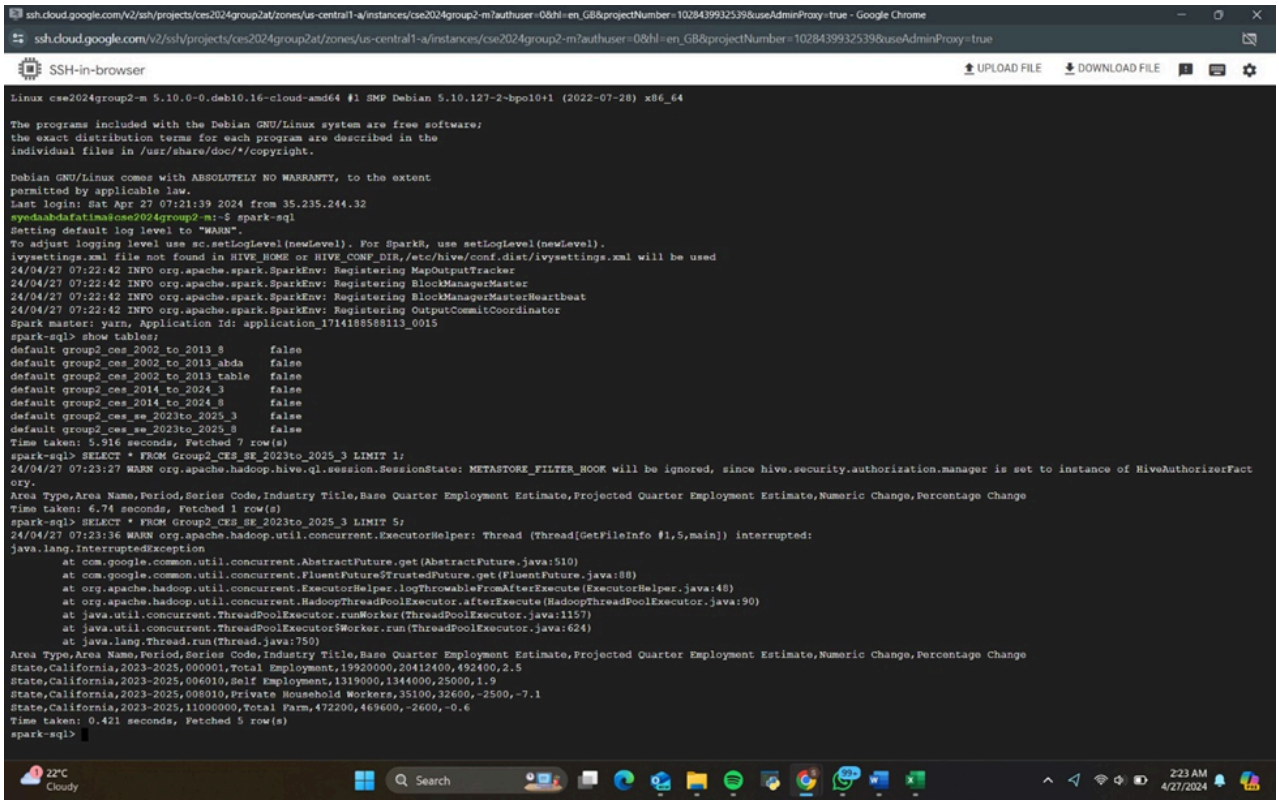

ssh.cloud.google.com/v2/ssh/projects/ces2024group2at/zones/us-central1-a/instances/cse2024group2-m?authuser=0&hl=en_GB&projectNumber=1028439932539&useAdminProxy=true - Google Chrome
ssh.cloud.google.com/v2/ssh/projects/ces2024group2at/zones/us-central1-a/instances/cse2024group2-m?authuser=0&hl=en_GB&projectNumber=1028439932539&useAdminProxy=true
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE
+-----------------------------------+---------------------+
Total Employment	2.04124E7
Total Nonfarm	1.85662E7
Professional and Business Services	2943100.0
Health Care and Social Assistance	2856500.0
Government	2657100.0
+-----------------------------------+---------------------+
5 rows selected (8.069 seconds)
0: jdbc:hive2://localhost:10000> SELECT
. . . . . .> `Industry Title`,
. . . . . .> AVG(`Projected Quarter Employment Estimate`) AS average_employment
. . . . . .> FROM
. . . . . .> Group2_CES_SE_2023to_2025_8
. . . . . .> WHERE
. . . . . .> `Area Type` = 'State'
. . . . . .> AND `Area Name` = 'California'
. . . . . .> GROUP BY
. . . . . .> `Industry Title`
. . . . . .> ORDER BY
. . . . . .> average_employment DESC
. . . . . .> LIMIT 10;
INFO : Compiling command(queryId=hive_20240427071821_22536e7c-fd51-4304-8519-059008330c0f): SELECT
`Industry Title`,
AVG(`Projected Quarter Employment Estimate`) AS average_employment
FROM
Group2_CES_SE_2023to_2025_8
WHERE
`Area Type` = 'State'
AND `Area Name` = 'California'
GROUP BY
`Industry Title`
ORDER BY
average_employment DESC
LIMIT 10
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:industry title, type:string, comment:null), FieldSchema(name:average_employment, type:double, comment:null)], properties:nu
ll)
INFO : Completed compiling command(queryId=hive_20240427071821_22536e7c-fd51-4304-8519-059008330c0f); Time taken: 0.299 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20240427071821_22536e7c-fd51-4304-8519-059008330c0f): SELECT
`Industry Title`,
AVG(`Projected Quarter Employment Estimate`) AS average_employment
FROM
Group2_CES_SE_2023to_2025_8
WHERE
`Area Type` = 'State'
AND `Area Name` = 'California'
GROUP BY
`Industry Title`
22°C
Mostly cloudy
Search
2:18 AM
4/27/2024

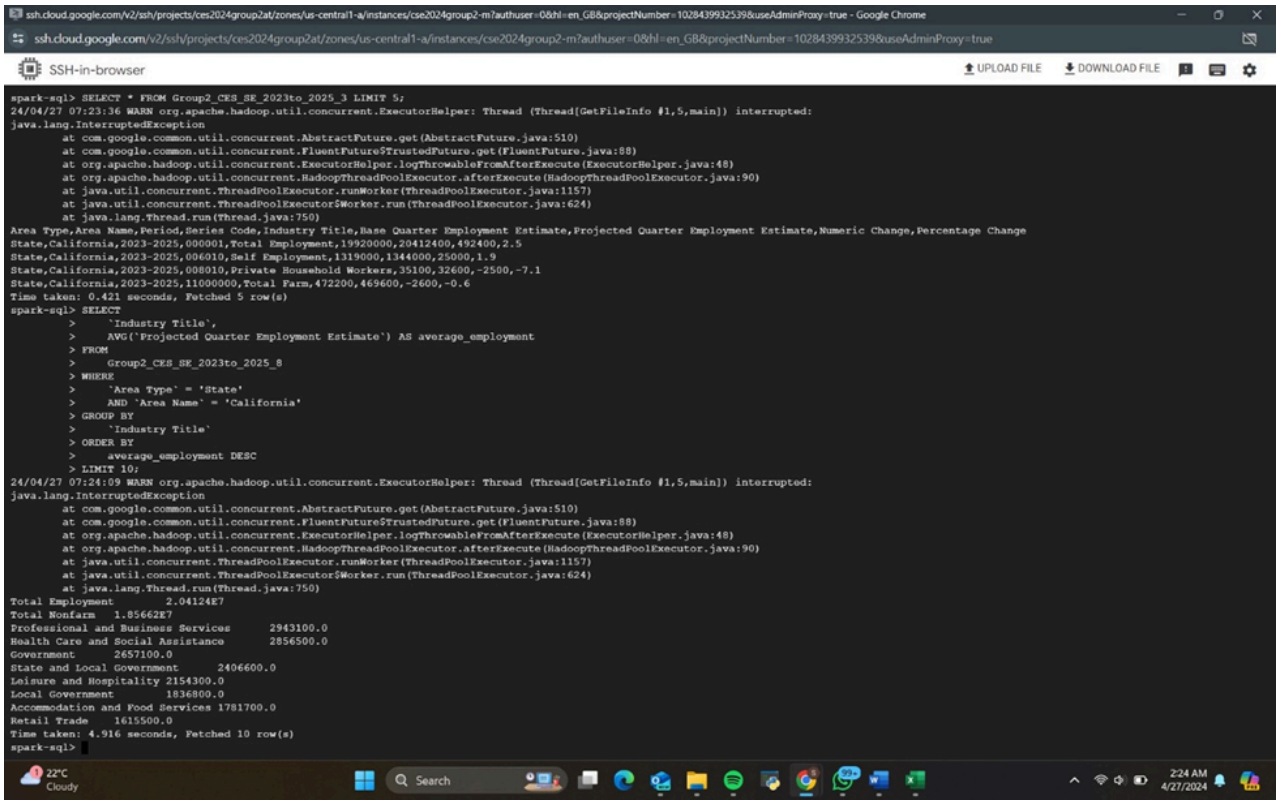


Query 3 in Hive: top 10 industries with the highest average projected quarter employment in California

Time: 7.749 seconds.

Screen Shot #2. Spark Screenshot #Syeda Abda Fatima





Spark API Results

Query 1 in Spark: SELECT * FROM Group2_CES_SE_2023to-2025_3 LIMIT 1;

Time: 6.74 seconds.

Query 2 in Spark: SELECT * FROM Group2_CES_SE_2023to-2025_3 LIMIT 5;

Time: 0.421 seconds.

Query 3 in Spark: top 10 industries with the highest average projected quarter employment in California

Time: 4.916 seconds.

RESULT - user/syedaabdafatima/data/Group2_CES_SE_2023to-2025

Query 1: It was **13.999 seconds** in Hive. In Spark, it was **6.74 seconds.**

Query 2: It was **6.395 seconds** in Hive. In Spark, it was **0.421 seconds.**

Query 3: It was **7.749 seconds** in Hive. In Spark, it was **4.916 seconds.**

Group 2 Project Meeting Notes-1

Date: April 13th, 2024

Start Time: 3:00 PM

End Time: 5:00 PM

Note-taker: Tanaya Dutt

Attendees: Sharanya Chinnigari, Tanaya Dutt, Abda Fatima Syed, Deepthimai Potla, Thulsi Buyyankar

Notes:

- Discussed a rough layout of what steps we need to take for data cleaning and creating storage buckets and clusters.

- Discussed the datasets for the project.

- Decided on the logo and name for our committee

Decisions:

1. Assigned tasks to each group member

2. Set a deadline for those tasks

3.Set the next meeting date

4.Selected datasets

Action Items:

Action Item #	Description	Owner	Due Date	Status
1	GCP: creating the project and storage bucket	Deepthimai and Mohammed	April 20th, 2024,	Pending
2	Dataproc: creating clusters	Deepthimai and Mohammed	April 20th, 2024	Pending

3	Open Refine: data cleaning	Sharanya and Tanaya	April 20th, 2024	Pending
4	BigQuery: Extracting data and SQL code	Sharanya and Tanaya	April 20th, 2024	Pending
5	Hive and Spark: SQL	Thulsi and Fatima	April 20th, 2024	Pending

Group 2 Project Meeting Notes-2

Attendees: Sharanya Chinnigari, Abda Fatima Syed, Deepthimai Potla, Thulsi Buyyankar

Date: April 20th, 2024

Start Time: 2:00 PM

End Time: 3:00 PM

Note-taker: Abda Fatima Syeda

Notes:

- Discussed a rough layout of what steps we need to take for data cleaning and creating storage buckets and clusters.
- Discussed the datasets for the project.
- Decided on the logo and name for our committee.

Decisions:

1. Assigned tasks to each group member

2. Set a deadline for those tasks

3.Set the next meeting date.

Action Items:

Action Item #	Description	Owner	Due Date	Status
1	Project Report Discussion	All	April 27th, 2024,	Completed
2	Assigned tasks regarding the report	All	April 27th, 2024	Completed
3	Team Name and Logo	Tanaya Dutt	April 27th, 2024	Completed

Group 2 Project Meeting Notes-3

Date: April 27th, 2024

Start Time: 1:00 PM

End Time: 5:00 PM

Note-taker: Tanaya Dutt

Attendees: Sharanya Chinnigari, Abda Fatima Syed, Deepthimai Potla, Thulsi Buyyankar

Notes:

- Working and completing the report
- Finishing the presentation

Decisions:

1. Complete the presentation.
2. Assign slides to each member.

Action Items:

Action Item #	Description	Owner	Due Date	Status
1	Project Report completion	All	April 27th, 2024,	Completed
2	Presentation completion	All	April 27th, 2024	completed

References:

1.Dutt, T. (2024). *Logo*.
https://www.canva.com/design/DAGDdDYwIHU/sw2QQb7FvtMuEdoV39Y96A/edit
2.Links to the three static datasets:

- https://data.ca.gov/dataset/current-employment-statistics-ces-2
- https://data.ca.gov/dataset/current-employment-statistics-ces-2
- https://catalog.data.gov/dataset/short-term-industry-employment-projections

3.computing, hosting services, and APIs. (n.d.). Google Cloud.

https://cloud.google.com/gcp?utm_source=bing&utm_medium=cpc&utm_campaign=na-US-all-en-dr-bkws-all-all-trial-e-dr-1707554&utm_content=text-ad-none-any-DEV_c-CRE_-ADGP_Desk+%7C+BKWS+-+EXA+%7C+Txt-Core-General+GCP-KWID_43700063341843290-kwd-77240896075639:loc-190&utm_term=KW_gcp-ST_gcp&gclid=ed32af75a6891d8a6336c944a1e5b2ee&gclsrc=3p.ds&msclkid=ed32af75a6891d8a6336c944a1e5b2ee&hl=en