

## **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**

1. Executive Summary
2. Screenshots:
  - GCP , Dataproc and Compute Engine(owners :Deepthi Mai and Mohammed)
  - Hadoop and Spark Clusters(owners :Deepthi Mai and Mohammed)
  - OpenRefine(owners : Sharanya & Tanaya)
  - BigQuery(owners :Sharanya & Tanaya)
  - Hive and Spark(owners :Abda Fatima Syeda & Thulsi Buuyankar)
3. Meeting notes
4. References

## **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**

The screenshot shows the Google Cloud Storage interface. The left sidebar has 'Cloud Storage' selected under 'Buckets'. The main area shows 'Bucket details' for 'ces2024project'. It lists the location as 'us-south1 (Dallas)', storage class as 'Standard', public access as 'Not public', and protection as 'None'. Below this is a table of objects in the bucket:

Name	Size	Type	Created	Storage class	Last modified	Public ac
Group2_CES_-2002-to-2013.csv	55.6 MB	text/csv	Apr 18, 2024, 9:14:28 PM	Standard	Apr 18, 2024, 9:14:28 PM	Not pub
Group2_CES_SE_2023to-2025.csv	25.3 KB	text/csv	Apr 18, 2024, 9:14:09 PM	Standard	Apr 18, 2024, 9:14:09 PM	Not pub
Group2_ces-2014-to-2024.csv	46.8 MB	text/csv	Apr 18, 2024, 9:14:21 PM	Standard	Apr 18, 2024, 9:14:21 PM	Not pub

### 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.

The screenshot shows the Google Cloud Compute Engine VM instances page. The left sidebar is collapsed. The main header has a dropdown for 'CES2024' and a search bar with 'datapro'. Below the header are buttons for 'CREATE INSTANCE', 'IMPORT VM', and 'REFRESH'. A 'LEARN' button is also present. The main area is titled 'INSTANCES' and shows a table of VM instances. The table columns are: Status, Name, Zone, Recommendations, In use by, Internal IP, External IP, and Connect. There are three entries:

Status	Name	Zone	Recommendations	In use by	Internal IP	External IP	Connect
<input type="checkbox"/>	<input checked="" type="checkbox"/> dp-hadoop-spark-2-cluster-ces2024-m	us-central1-a			10.128.0.4 (nic0)	34.27.208.228 (nic0)	SSH
<input type="checkbox"/>	<input checked="" type="checkbox"/> dp-hadoop-spark-2-cluster-ces2024-w_0	us-central1-a			10.128.0.3 (nic0)	34.132.130.94 (nic0)	SSH
<input type="checkbox"/>	<input checked="" type="checkbox"/> dp-hadoop-spark-2-cluster-ces2024-w_1	us-central1-a			10.128.0.2 (nic0)	34.173.9.230 (nic0)	SSH

Below the table is a 'Related actions' section with a 'SHOW' button.

**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.

```

Linux dp-hadoop-spark-2-cluster-ces2024-m 5.10.0-0.deb10.16-cloud-amd64 #1 SMP Debian 5.10.127-2-bpo10+1 (2022-07-28) x86_64
The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
deepthimaiptla@dp-hadoop-spark-2-cluster-ces2024-m:~$ cd
deepthimaiptla@dp-hadoop-spark-2-cluster-ces2024-m:~$ date
Fri Apr 19 02:36:02 UTC 2024
deepthimaiptla@dp-hadoop-spark-2-cluster-ces2024-m:~$ whoami
deepthimaiptla
deepthimaiptla@dp-hadoop-spark-2-cluster-ces2024-m:~$ pwd
/home/deepthimaiptla
deepthimaiptla@dp-hadoop-spark-2-cluster-ces2024-m:~$ mkdir ces_folder
deepthimaiptla@dp-hadoop-spark-2-cluster-ces2024-m:~$ cd ces_folder
deepthimaiptla@dp-hadoop-spark-2-cluster-ces2024-m:~/ces_folder$ cd
deepthimaiptla@dp-hadoop-spark-2-cluster-ces2024-m:~$ rmdir ces_folder
deepthimaiptla@dp-hadoop-spark-2-cluster-ces2024-m:~$ 

```

**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 successfully.

## Screenshot 4 : Exploring services using Hadoop and Spark clusters

```
ssh.cloud.google.com/v2/ssh/projects/ces2024/zones/us-central1-a/instances/dp-hadoop-spark-2-cluster-ces2024-m?authuser=3&hl=en_US&projectNumber=641609098049&useAdminProxy=true - Google Chrome
https://ssh.cloud.google.com/v2/ssh/projects/ces2024/zones/us-central1-a/instances/dp-hadoop-spark-2-cluster-ces2024-m?authuser=3&hl=en_US&projectNumber=641609098049&useAdminProxy=true

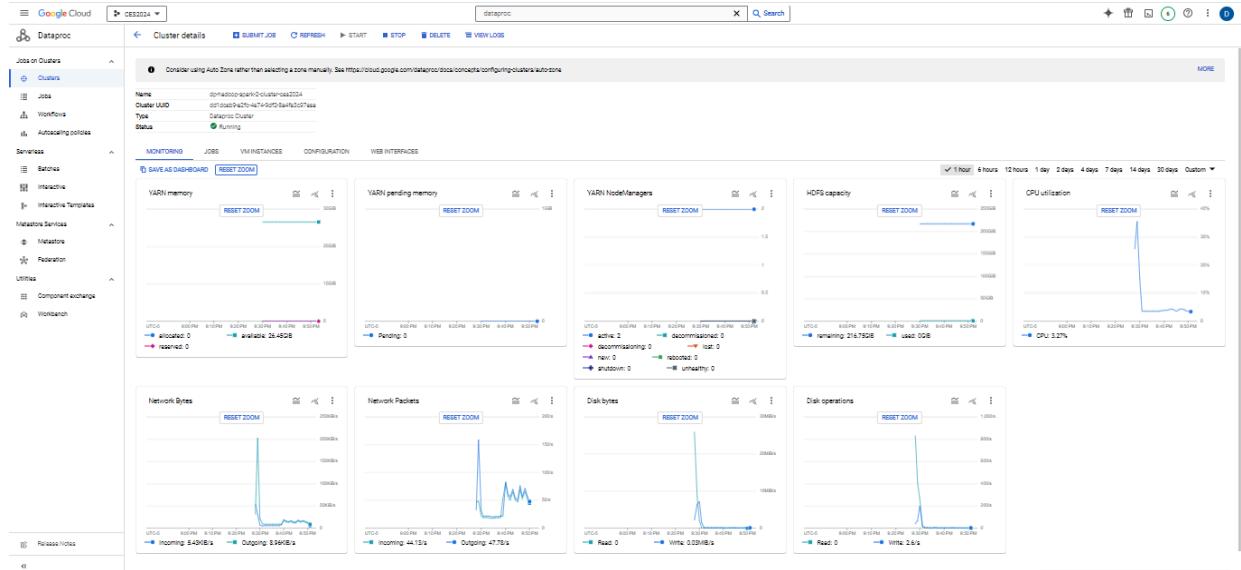
SSH-in-browser
↑ UPLOAD FILE ↓ DOWNLOAD FILE ⚙️ 🗑️

e-hadoop-policy.xml -Dhadoop.security.logger=INFO,NullAppender org.apache.hadoop.util.RunJar /usr/lib/hive/lib/hive-metastore-3.1.3.jar org.apache.hadoop.hive.metastore.HiveMetaStore
yarn      5662   1  2 02:27 ?          00:00:27 /usr/lib/jvm/temurin-8-jdk-amd64/bin/java -Dproc_resourcemanager -Dservice.libdir=/usr/lib/hadoop-yarn/..,/usr/lib/hadoop-yarn/lib/usr/lib/hadoop-hdfs/.,/usr/lib/hadoop-hdfs/lib/usr/lib/hadoop/. -Dxmx403m -Dyarn.log.dir=/var/log/hadoop-yarn -Dyarn.log.file=hadoop-yarn -resourcemanager-dp-hadoop-spark-2-cluster-ces2024-m.log -Dyarn.home.dir=/usr/lib/hadoop-yarn -Dyarn.root.logger=INFO,console -Djava.library.path=/usr/lib/hadoop/lib/native -Dhadoop.log.dir=/var/log/hadoop-yarn -Dhadoop.log.file=hadoop-yarn -resourcemanager-dp-hadoop-spark-2-cluster-ces2024-m.log -Dhadoop.home.dir=/usr/lib/hadoop -Dhadoop.id.str=yarn -Dhadoop.root.logger=INFO,RFA -Dhadoop.policy.file=hadoop-policy.xml -Dhadoop.security.logger=INFO,NullAppender org.apache.hadoop.yarn.server.resourcemanager.ResourceManager
yarn      5669   1  1 02:27 ?          00:00:15 /usr/lib/jvm/temurin-8-jdk-amd64/bin/java -Dproc_timelineserver -XX:+UseConcMarkSweepGC -XX:+PrintGCTimeStamps -XX:+PrintGCDateStamps -XX:+PrintGCDetails -Djava.util.logging.config.file=/etc/hadoop/conf/yarn-timelineserver.logging.properties -Dyarn.log.dir=/var/log/hadoop-yarn -Dyarn.log.fil
e=hadoop-yarn-timelineserver-dp-hadoop-spark-2-cluster-ces2024-m.log -Dyarn.home.dir=/usr/lib/hadoop-yarn -Dyarn.root.logger=INFO,console -Djava.library.path=/usr/lib/hadoop/lib/native -Xmx400m -Dhadoop.log.dir=/var/log/hadoop-yarn -Dhadoop.log.file=hadoop-yarn -Dhadoop-timelineserver-dp-hadoop-spark-2-cluster-ces2024-m.log -Dhadoop.home.dir=/usr/lib/hadoop -Dhadoop.id.str=yarn -Dhadoop.root.logger=INFO,RFA -Dhadoop.policy.file=hadoop-policy.xml -Dhadoop.security.logger=INFO,NullAppender org.apache.hadoop.yarn.server.rvmserviceapplication.HistoryServer -Djava.util.logging.config.file=/etc/hadoop/conf/yarn-timelineserver.logging.properties -Dyarn.log.dir=/var/log/hadoop-yarn -Dyarn.log.fil
e=hadoop-timelineserver-dp-hadoop-spark-2-cluster-ces2024-m.log -Dyarn.home.dir=/usr/lib/hadoop -Dhadoop.log.dir=/var/log/hadoop -Dhadoop.id.str=yarn -Dhadoop.root.logger=INFO,RFA -Dhadoop.policy.file=hadoop-policy.xml -Dhadoop.security.logger=INFO,NullAppender org.apache.hadoop.yarn.server.rvmserviceapplication.HistoryServer
yarn      5676   1  4 02:27 ?          00:00:40 /usr/lib/jvm/temurin-8-jdk-amd64/bin/java -Dproc_historyserver -Dmapred.jobsummary.logger=INFO,RFA -XX:+UseConcMarkSweepGC -XX:+PrintGCTimeStamps -XX:+PrintGCDateStamps -XX:+PrintGCDetails -Dyarn.log.dir=/var/log/hadoop-mapreduce -Dyarn.log.file=hadoop-mapreduce-historyserver-dp-hadoop-spark-2-cluster-ces2024-m.log -Dyarn.home.dir=/usr/lib/hadoop-yarn -Dyarn.root.logger=INFO,console -Djava.library.path=/usr/lib/hadoop/lib/native -Xmx400m -Dhadoop.log.dir=/var/log/hadoop-mapreduce -Dhadoop.log.file=hadoop-mapreduce-historyserver-dp-hadoop-spark-2-cluster-ces2024-m.log -Dhadoop.home.dir=/usr/lib/hadoop -Dhadoop.id.str=mapred -Dhadoop.root.logger=INFO,RFA -Dhadoop.policy.file=hadoop-policy.xml -Dhadoop.security.logger=INFO,NullAppender org.apache.hadoop.mapreduce.v2.hs.JobHistoryServer
hdfs      6181   1  3 02:28 ?          00:00:33 /usr/lib/jvm/temurin-8-jdk-amd64/bin/java -Dproc_namenode -Dhdfs.audit.logger=INFO,NullAppender org.apache.hadoop.hdfs -Dxmn3201m -XX:+UseConcMarkSweepGC -XX:+PrintGCTimeStamps -XX:+PrintGCDateStamps -XX:+PrintGCDetails -Dyarn.log.dir=/var/log/hadoop-hdfs -Dyarn.log.file=hadoop-hdfs-namenode-dp-hadoop-spark-2-cluster-ces2024-m.log -Dyarn.home.dir=/usr/lib/hadoop -Dyarn.root.logger=INFO,console -Djava.library.path=/usr/lib/hadoop/lib/native -Dhadoop.log.dir=/var/log/hadoop-hdfs -Dhadoop.log.file=hadoop-hdfs-namenode-dp-hadoop-spark-2-cluster-ces2024-m.log -Dhadoop.home.dir=/usr/lib/hadoop -Dhadoop.id.str=hdfs -Dhadoop.root.logger=INFO,RFA -Dhadoop.policy.file=hadoop-policy.xml -Dhadoop.security.logger=INFO,NullAppender org.apache.hadoop.hdfs.server.namenode.NameNode
hdfs      6595   1 15 02:28 ?          00:02:26 /usr/lib/jvm/temurin-8-jdk-amd64/bin/java -Dproc_secondarynamenode -Dhdfs.audit.logger=INFO,NullAppender -Xmx201m -XX:+Us
eConcMarkSweepGC -XX:+PrintGCTimeStamps -XX:+PrintGCDateStamps -XX:+PrintGCDetails -Dyarn.log.dir=/var/log/hadoop-hdfs -Dhdfs-secondarynamenode-dp-hadoop-spark-2-cluster-ces2024-m.log -Dyarn.home.dir=/usr/lib/hadoop -Dyarn.root.logger=INFO,console -Djava.library.path=/usr/lib/hadoop/lib/native -Dhadoop.log.dir=/var/log/hadoop-hdfs -Dhadoop.log.file=hadoop-hdfs -Dsecondarynamenode-dp-hadoop-spark-2-cluster-ces2024-m.log -Dyarn.home.dir=/usr/lib/hadoop -Dyarn.root.logger=INFO,NullAppender.org.apache.hadoop.hdfs.server.namenode.SecondaryNameNode
hive      7429   1  2 02:28 ?          00:00:21 /usr/lib/jvm/temurin-8-jdk-amd64/bin/java -Dproc_jar -Dhive.log.dir=/var/log/hive -Dhive.log.file=hive-service-3.1.3.jar.log -Dhive.log.threshold=INFO -Xmx102m -Dproc_hiveserver2 -Dlog4j.formatterMsgNoLookups=true -XX:+UseConcMarkSweepGC -XX:+PrintGCTimeStamps -XX:+PrintGCDateStamps -XX:+PrintGCDetails -XX:+ExitOnOutOfMemoryError -Dlog4j.configurationFile=hive-log4j2.properties -Djava.util.logging.config.file=/usr/lib/hadoop/logs/hive-log4j2.properties -Dyarn.log.dir=/var/log/hadoop/logs -Dyarn.root.logger=INFO,console -Djava.library.path=/usr/lib/hadoop/lib/native -Dhadoop.log.dir=/var/lib/hadoop/logs -Dhadoop.home.dir=/usr/lib/hadoop -Dhadoop.id.str=hive -Dhadoop.root.logger=INFO,console -Dhadoop.policy.file=hadoop-policy.xml -Dhadoop.security.logger=INFO,NullAppender.org.apache.hadoop.util.RunJar /usr/lib/hive/lib/hive-service-3.1.3.jar org.apache.hive.service.server.HiveService
spark      7458   1  3 02:28 ?          00:00:28 /usr/lib/jvm/temurin-8-jdk-amd64/bin/java -cp /usr/lib/spark/conf:/usr/lib/spark/jars/*:/etc/hadoop/conf:/etc/hive/conf:/usr/local/share/google/dataproxy/lib/*:/usr/share/java/mysql.jar -Xmx400m org.apache.spark.history.HistoryServer
deathplus 11261 11580  0  43 pts/0    00:00:00 grep hadoop
depthplus@ip-10-11-11-158:~$
```

**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

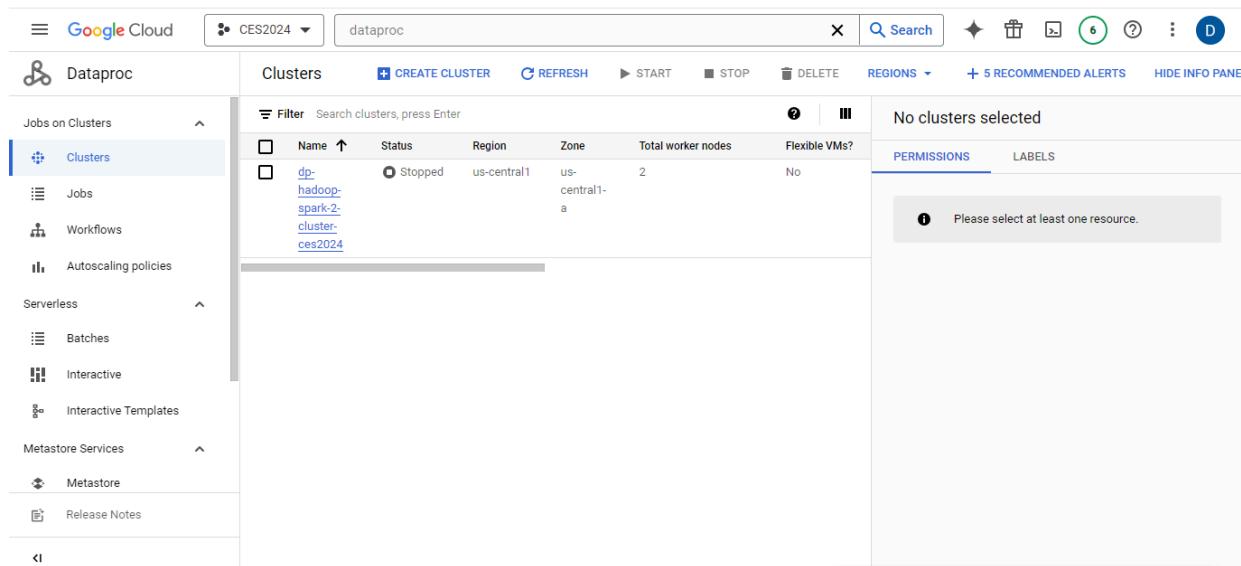
window, and find all the core subsystems that are currently running in the manager node of the cluster.

## Screenshot 5 : Monitoring the Hadoop and Spark clusters.

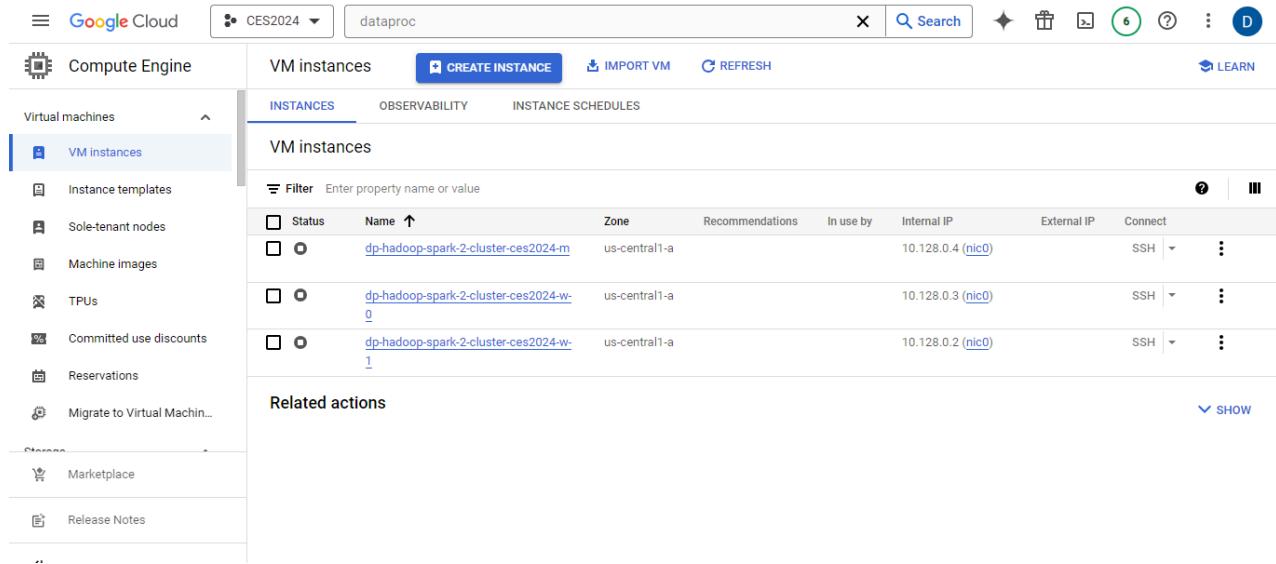


**Explanation:** Here when we click on the manager node we will get the panel like this which contains a monitoring tab, click on it. In this monitoring tab we can see the monitoring panels of the cluster and gather information about YARN memory, YARN pending memory, YARN NodeManagers, HDFS Capacity,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

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



Here, we stopped the cluster.



The screenshot shows the Google Cloud Compute Engine VM instances page. The left sidebar is titled 'Compute Engine' and has a 'Virtual machines' section with 'VM instances' selected. The main area is titled 'VM instances' and shows a table of three stopped instances:

Status	Name	Zone	Recommendations	In use by	Internal IP	External IP	Connect
stopped	dp-hadoop-spark-2-cluster-ces2024-m	us-central1-a			10.128.0.4 (nic0)		SSH
stopped	dp-hadoop-spark-2-cluster-ces2024-w-0	us-central1-a			10.128.0.3 (nic0)		SSH
stopped	dp-hadoop-spark-2-cluster-ces2024-w-1	us-central1-a			10.128.0.2 (nic0)		SSH

Below the table, there is a 'Related actions' section with a 'SHOW' button.

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.

**OpenRefine Group2\_ces 2002 to 2013** Permalink

Facet / Filter Undo / Redo 3 / 3

581880 rows

Show as: rows records Show: 5 10 25 50 100 500 1000 rows

	Area Type	Area Name	Year	Month	Date	Series Code	Industry Title	Current Employment
1.	County	Alameda County	2002	January	01/1/2002	40000000	Trade, Transportation, and Utilities	1429000
2.	County	Alameda County	2002	January	01/1/2002	31000000	Durable Goods	554000
3.	County	Alameda County	2002	January	01/1/2002	01000000	Total Wage and Salary	8964000
4.	County	Alameda County	2002	January	01/1/2002	00000000	Total Nonfarm	6955000
5.	County	Alameda County	2002	January	01/1/2002	15000000	Mining, Logging and Construction	363000
6.	County	Alameda County	2002	January	01/1/2002	00500000	Administrative and Support and Waste Management and Remediation Services	371000
7.	County	Alameda County	2002	January	01/1/2002	32000000	Non-Durable Goods	240000
8.	County	Alameda County	2002	January	01/1/2002	59530000	Real Estate and Rental and Leasing	180000
9.	County	Alameda County	2002	January	01/1/2002	59000000	Financial Activities	260000
10.	County	Alameda County	2002	January	01/1/2002	60000000	Professional and Business Services	1025000
11.	County	Alameda County	2002	January	01/1/2002	11000000	Total Farm	800
12.	County	Alameda County	2002	January	01/1/2002	65610000	Private Educational Services	102000
13.	County	Alameda County	2002	January	01/1/2002	68540000	Professional, Scientific, and Technical Services	473000
14.	County	Alameda County	2002	January	01/1/2002	07000000	Service-Providing	579300
15.	County	Alameda County	2002	January	01/1/2002	70000000	Leisure and Hospitality	494000
16.	County	Alameda County	2002	January	01/1/2002	06000000	Goods Producing	1153000
17.	County	Alameda County	2002	January	01/1/2002	42000000	Retail Trade	682000
18.	County	Alameda County	2002	January	01/1/2002	59520000	Finance and Insurance	154000
19.	County	Alameda County	2002	January	01/1/2002	41000000	Wholesale Trade	434000
20.	County	Alameda County	2002	January	01/1/2002	65000000	Private Education and Health Services	820000
21.	County	Alameda County	2002	January	01/1/2002	50000000	Information	200000
22.	County	Alameda County	2002	January	01/1/2002	65420000	Health Care and Social Assistance	720000
23.	County	Alameda County	2002	January	01/1/2002	30000000	Manufacturing	800000
24.	County	Alameda County	2002	January	01/1/2002	76720000	Accommodation and Food Services	454000

**OpenRefine Group2\_ces 2014 to 2024** Permalink

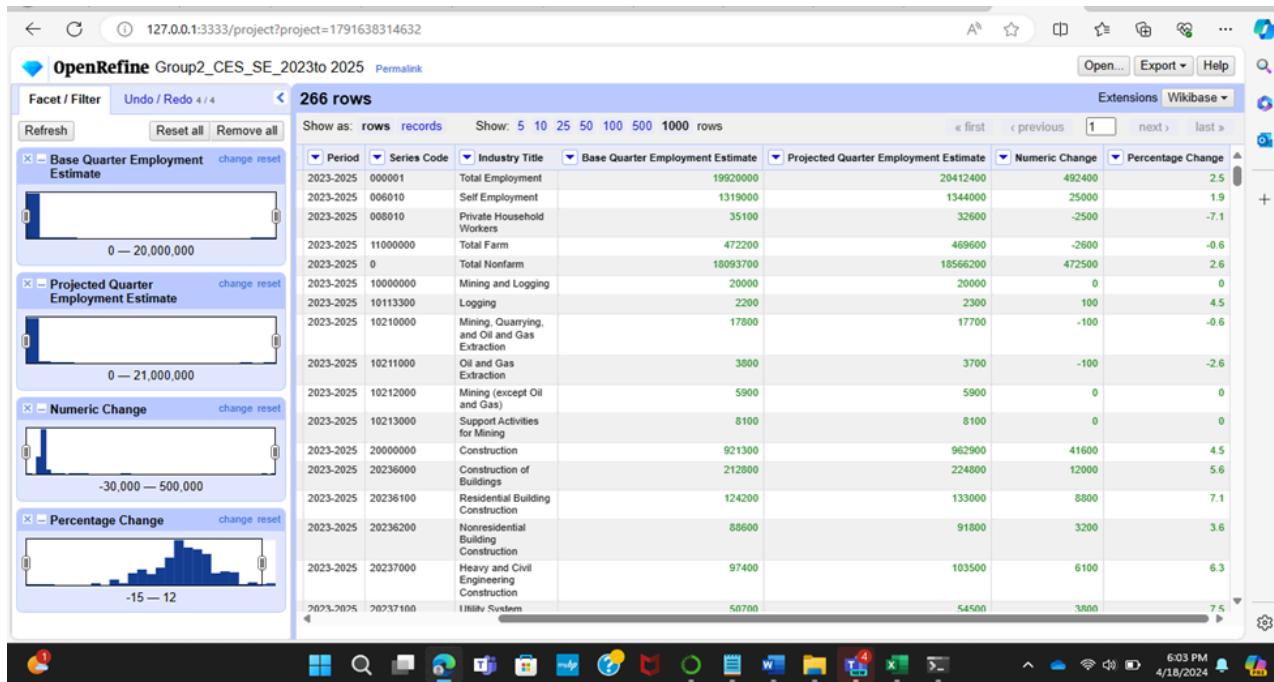
Facet / Filter Undo / Redo 3 / 3

488817 rows

Show as: rows records Show: 5 10 25 50 100 500 1000 rows

	Area Type	Area Name	Year	Month	Date	Series Code	Industry Title	Current Employment
1.	County	Alameda County	2014	January	01/1/2014	31000000	Durable Goods	442000
2.	County	Alameda County	2014	January	01/1/2014	06000000	Goods Producing	1000000
3.	County	Alameda County	2014	January	01/1/2014	00000000	Total Nonfarm	7007000
4.	County	Alameda County	2014	January	01/1/2014	42000000	Retail Trade	685000
5.	County	Alameda County	2014	January	01/1/2014	70710000	Arts, Entertainment, and Recreation	8800
6.	County	Alameda County	2014	January	01/1/2014	11000000	Total Farm	500
7.	County	Alameda County	2014	January	01/1/2014	07000000	Service-Providing	6007000
8.	County	Alameda County	2014	January	01/1/2014	40000000	Trade, Transportation, and Utilities	1296000
9.	County	Alameda County	2014	January	01/1/2014	01000000	Total Wage and Salary	7012000
10.	County	Alameda County	2014	January	01/1/2014	60000000	Professional and Business Services	1172000
11.	County	Alameda County	2014	January	01/1/2014	60540000	Professional, Scientific, and Technical Services	636000
12.	County	Alameda County	2014	January	01/1/2014	65610000	Private Educational Services	154000
13.	County	Alameda County	2014	January	01/1/2014	70720000	Accommodation and Food Services	530000
14.	County	Alameda County	2014	January	01/1/2014	70000000	Leisure and Hospitality	618000
15.	County	Alameda County	2014	January	01/1/2014	80000000	Other Services	244000
16.	County	Alameda County	2014	January	01/1/2014	90910000	Federal Government	9300
17.	County	Alameda County	2014	January	01/1/2014	90000000	Government	1148000
18.	County	Alameda County	2014	January	01/1/2014	90930000	Local Government	690000
19.	County	Alameda County	2014	January	01/1/2014	32000000	Non-Durable Goods	21100
20.	County	Alameda County	2014	January	01/1/2014	15000000	Mining, Logging and Construction	34700
21.	County	Alameda County	2014	January	01/1/2014	55530000	Real Estate and Rental and Leasing	9500
22.	County	Alameda County	2014	January	01/1/2014	55000000	Financial Activities	28300
23.	County	Alameda County	2014	January	01/1/2014	60560000	Administrative and Support and Waste Management and Remediation Services	35200
24.	County	Alameda County	2014	January	01/1/2014	30000000	Manufacturing	65300

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.

The screenshot shows the Google Cloud BigQuery console interface. The left sidebar displays the project structure under 'ces2024group2'. The main area shows the schema for the dataset 'ces2024table1'. The schema includes columns: Area\_Type (STRING, NULLABLE), Area\_Name (STRING, NULLABLE), Year (INTEGER, NULLABLE), Month (STRING, NULLABLE), Date (DATE, NULLABLE), Series\_Code (INTEGER, NULLABLE), Industry\_Title (STRING, NULLABLE), and Current\_Employment (INTEGER, NULLABLE). Below the schema, there are buttons for 'EDIT SCHEMA' and 'VIEW ROW ACCESS POLICIES'.

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

The screenshot shows the Google Cloud BigQuery console interface. The left sidebar displays the project structure under 'ces2024group2'. The main area shows the schema for the dataset 'ces2023toces2025'. The schema includes columns: Area\_Type (STRING, NULLABLE), Area\_Name (STRING, NULLABLE), Period (STRING, NULLABLE), Series\_Code (INTEGER, NULLABLE), Industry\_Title (STRING, NULLABLE), Base\_Quarter\_Employment\_Estimate (INTEGER, NULLABLE), Projected\_Quarter\_Employment\_Estimate (INTEGER, NULLABLE), Numeric\_Change (INTEGER, NULLABLE), and Percentage\_Change (FLOAT, NULLABLE). Below the schema, there are buttons for 'EDIT SCHEMA' and 'VIEW ROW ACCESS POLICIES'.

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;

```

The screenshot shows the Google Cloud BigQuery interface. The top navigation bar includes the URL <https://console.cloud.google.com/bigquery?hl=en&project=ces2024group2&ws=!1m1!1m4!4m3!1scs2024group2!2scs202401!3sces2024table1>, a 'START FREE' button, and a search bar. The left sidebar shows project structure under 'ces2024group2'. The main area displays an 'Untitled query' with the following SQL code:

```

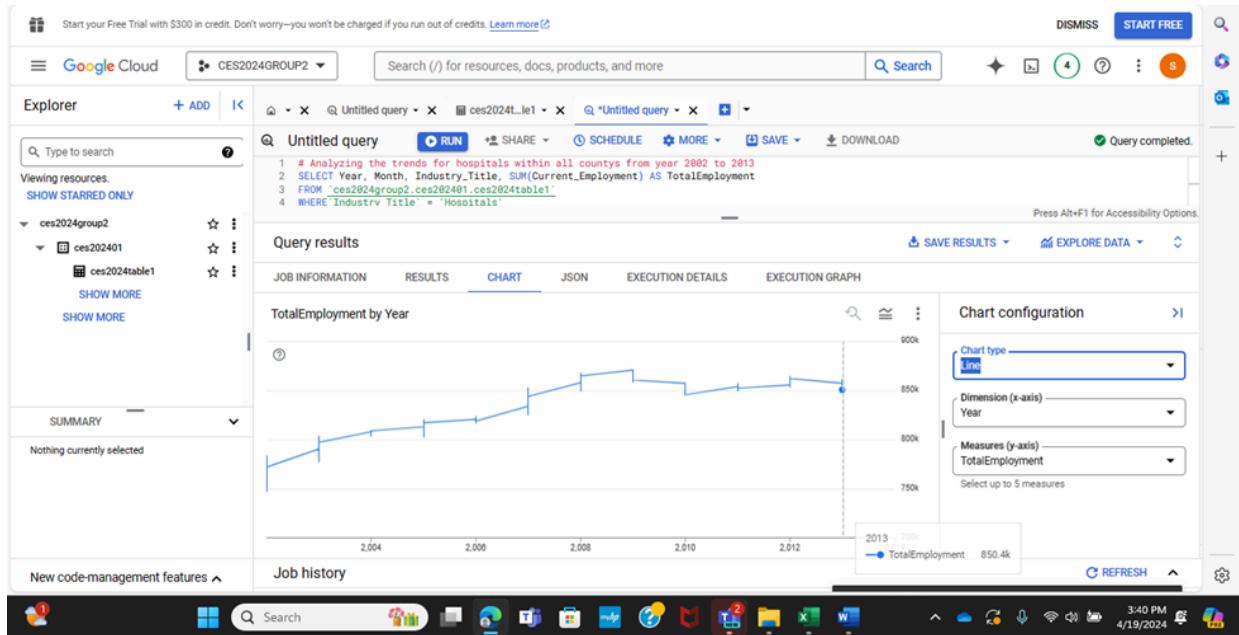
# Analyzing the trends for hospitals within all counties
SELECT Year, Month, Industry_Title, SUM(Current_Employment) AS TotalEmployment
FROM `ces2024group2.ces202401.ces2024table1`
WHERE `Industry_Title` = 'Hospitals'
GROUP BY Year, Month, Industry_Title
ORDER BY Year, Month
Limit 50;

```

The 'Query results' section shows a table with the following data:

Row	Year	Month	Industry_Title	TotalEmployment
1	2002	April	Hospitals	746900
2	2002	August	Hospitals	767400
3	2002	December	Hospitals	783400
4	2002	February	Hospitals	753900
5	2002	January	Hospitals	751600

The bottom status bar indicates the time as 3:35 PM and the date as 4/19/2024.



\*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 screenshot shows the Google Cloud BigQuery interface. In the top navigation bar, there are tabs for 'Untitled query', 'ces2024...\_le1', 'ces2024...\_le2', and 'ces2024...\_le3'. Below the tabs, the main area displays an 'Untitled query' with the following SQL code:

```

1 # Analyzing the trends for hospitals within all counties from year 2014 to 2023
2 SELECT Year, Month, Industry_Title, SUM(Current_Employment) AS TotalEmployment
3 FROM `ces2024group2.ces202401.ces2024table2`
4 WHERE `Industry_Title` = 'Hospitals'
5 GROUP BY Year, Month, Industry_Title
6 ORDER BY Year, Month
7 LIMIT 50;

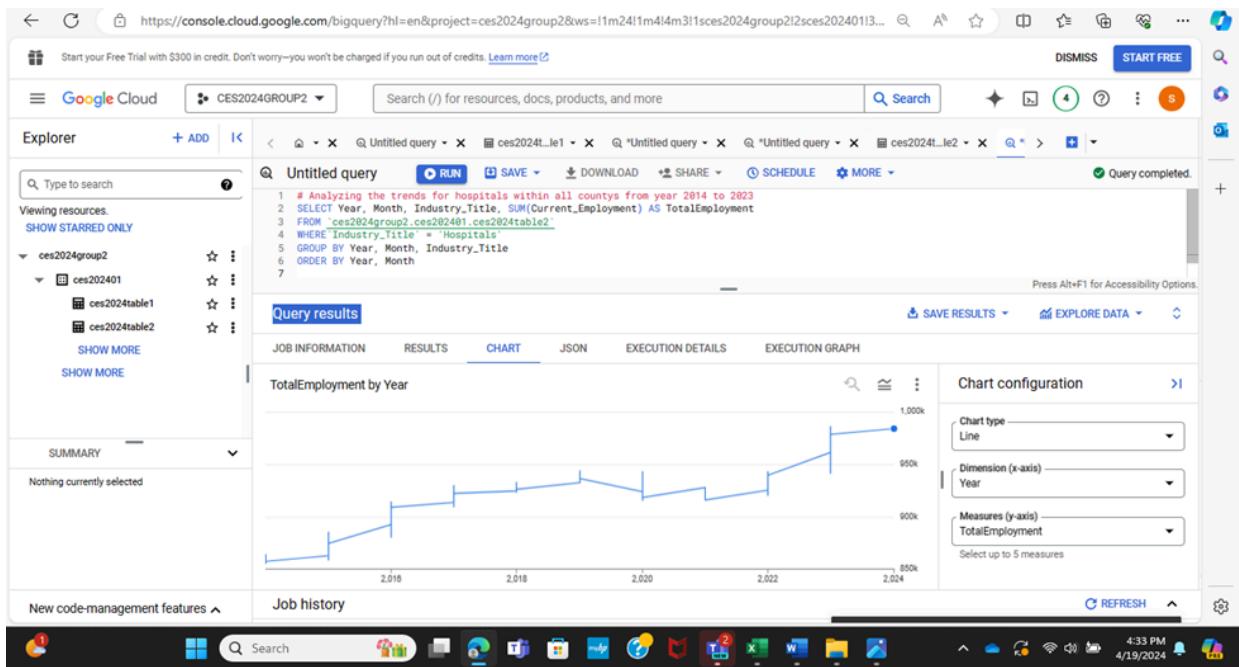
```

The 'RESULTS' tab is selected, showing a table with the following data:

Row	Year	Month	Industry_Title	TotalEmployment
1	2014	April	Hospitals	858000
2	2014	August	Hospitals	858900
3	2014	December	Hospitals	860100
4	2014	February	Hospitals	860900
5	2014	January	Hospitals	863100
6	2014	July	Hospitals	854800
7	2014	June	Hospitals	856100
8	2014	March	Hospitals	864500

At the bottom of the interface, there is a 'Job history' section and a Windows taskbar.

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.

```

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

```

The screenshot shows the Google Cloud BigQuery interface. On the left, there's a sidebar with various services like Analytics Hub, Dataform, Partner Center, Migration, Assessment, SQL translation, Administration, Monitoring, Capacity management, BI Engine, Policy tags, and Release Notes. The main area has tabs for 'Explorer' and 'Untitled query'. The 'Untitled query' tab contains the following SQL code:

```

1 SELECT Industry_Title, Projected_Quarter_Employment_Estimate, Base_Quarter_Employment_Estimate
2 FROM `ces2024group2.ces202401.ces2023toces2025`
3 ORDER BY Base_Quarter_Employment_Estimate DESC
4 LIMIT 10;

```

The 'Query results' section displays the output of the query in a table format. The columns are 'Row', 'Industry\_Title', 'Projected\_Quarter\_E', and 'Base\_Quarter\_Emp'. The data includes:

Row	Industry_Title	Projected_Quarter_E	Base_Quarter_Emp
1	Total Employment	20412400	19920000
2	Total Nonfarm	18566200	18093700
3	Trade, Transportation, and Utilit...	3147000	31191100
4	Educational Services (Private), He...	3280000	3091600
5	Professional and Business Ser...	2943100	2900100
6	Health Care and Social Assista...	2856500	2682100
7	Government	2657100	2621700
8	State and Local Government	2406600	2373300
9	Leisure and Hospitality	2154300	2057500
10	Local Government	1836800	1812100

This screenshot is identical to the one above, showing the same BigQuery interface and query results. However, the 'RESULTS' tab has been replaced by a 'CHART' tab, which displays a bar chart of the data. The x-axis represents the industry categories, and the y-axis represents the projected quarter employment estimate. The chart shows that 'Total Employment' is the highest category, followed by 'Total Nonfarm'.

- \* 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)

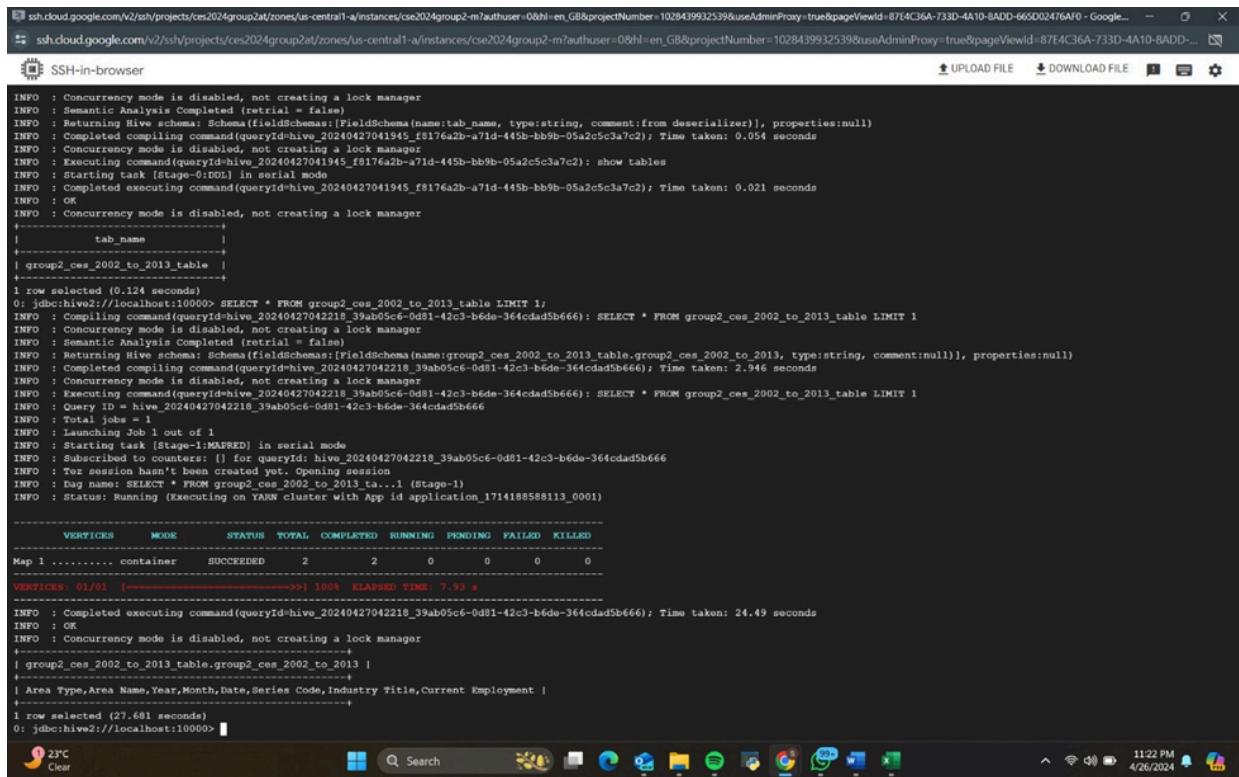
## First dataset: 2002\_to\_2013

### Hive Screenshot: Creating the table.

```
ssh.cloud.google.com/v2/ssh/projects/ces2024group2/instances/us-central1-a/instances/ces2024group2-m?authUser=0&hl=en_GB&projectNumber=1028439932539&useAdminProxy=true&pageViewId=874C36A-733D-4A10-8ADD-665D02476AF0 - Google...
ssh.cloud.google.com/v2/ssh/projects/ces2024group2/instances/us-central1-a/instances/ces2024group2-m?authUser=0&hl=en_GB&projectNumber=1028439932539&useAdminProxy=true&pageViewId=874C36A-733D-4A10-8ADD-665D02476AF0 - Google...
SSH-in-browser
 UPLOAD FILE DOWNLOAD FILE
No rows selected (0.408 seconds)
0: jdbc:hive2://localhost:10000> CREATE EXTERNAL TABLE IF NOT EXISTS Group2_CES_2002_to_2013_table
+-----+
|          .-> (Group2_CES_2002_to_2013_string)
|          .-> ROW FORMAT DELIMITED
|          .-> STORED AS TEXTFILE
|          .-> Location '/user/syedaabdafatima/data/Group2_CES_2002_to_2013/';
Error: Error while compiling statement: FAILED: ParseException line 1:29 missing KW_EXISTS at 'EXISTSGroup2_CES_2002_to_2013_table' near '<EOF>' (state=42000,code=40000)
0: jdbc:hive2://localhost:10000> CREATE EXTERNAL TABLE IF NOT EXISTS Group2_CES_2002_to_2013_table
+-----+
|          .-> (Group2_CES_2002_to_2013_string)
|          .-> ROW FORMAT DELIMITED
|          .-> STORED AS TEXTFILE
|          .-> Location '/user/syedaabdafatima/data/Group2_CES_2002_to_2013/';
INFO : Compiling command(queryId=hive_20240427041936_12fa5713-7dea-4cc2-9691-867ce5b18470): CREATE EXTERNAL TABLE IF NOT EXISTS Group2_CES_2002_to_2013_table
(Group2_CES_2002_to_2013_string)
STORED AS TEXTFILE
Location '/user/syedaabdafatima/data/Group2_CES_2002_to_2013/'
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis completed (retryal = false)
INFO : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO : Completed compiling command(queryId=hive_20240427041936_12fa5713-7dea-4cc2-9691-867ce5b18470); Time taken: 0.104 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20240427041936_12fa5713-7dea-4cc2-9691-867ce5b18470): CREATE EXTERNAL TABLE IF NOT EXISTS Group2_CES_2002_to_2013_table
(Group2_CES_2002_to_2013_string)
STORED AS TEXTFILE
Location '/user/syedaabdafatima/data/Group2_CES_2002_to_2013/'
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20240427041936_12fa5713-7dea-4cc2-9691-867ce5b18470); Time taken: 0.479 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
No rows affected (0.605 seconds)
0: jdbc:hive2://localhost:10000> show tables;
INFO : Compiling command(queryId=hive_20240427041945_f8176a2b-a71d-445b-bb9b-05a2c5c3a7c2): show tables
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retryal = false)
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, type:string, comment:from deserializer)], properties:null)
INFO : Completed compiling command(queryId=hive_20240427041945_f8176a2b-a71d-445b-bb9b-05a2c5c3a7c2); Time taken: 0.054 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20240427041945_f8176a2b-a71d-445b-bb9b-05a2c5c3a7c2): show tables
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20240427041945_f8176a2b-a71d-445b-bb9b-05a2c5c3a7c2); Time taken: 0.021 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
|      tab_name      |
+-----+
| group2_ces_2002_to_2013_table |
+-----+
1 row selected (0.124 seconds)
0: jdbc:hive2://localhost:10000>
```

**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



The screenshot shows an SSH-in-browser session on a Windows desktop. The terminal window displays the output of a Hive query. The query selected all columns from the 'group2\_ces\_2002\_to\_2013\_table' limit 1. The execution took 24.49 seconds. The session also shows the status of the task and the DAG.

```
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retryal = false)
INFO : Returning Hive schema: Schema[fieldschemas:[FieldSchema(name:tab_name, type:string, comment:from deserializer)], properties:null]
INFO : Completed compiling command(queryId=hive_20240427041945_f8176a2b-a71d-445b-bb9b-05a2c5c3a7c2); Time taken: 0.054 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20240427041945_f8176a2b-a71d-445b-bb9b-05a2c5c3a7c2): show tables
INFO : Starting task [Stage-0:IDLE] in serial mode
INFO : completed executing command(queryId=hive_20240427041945_f8176a2b-a71d-445b-bb9b-05a2c5c3a7c2); Time taken: 0.021 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| tab_name |
+-----+
| group2_ces_2002_to_2013_table |
+-----+
1 row selected (0.124 seconds)
0: jdbc:hive2://localhost:10000> SELECT * FROM group2_ces_2002_to_2013_table LIMIT 1;
INFO : Compiling command(queryId=hive_20240427042218_39ab05c6-0d81-42c3-b6de-364cdad5b666): SELECT * FROM group2_ces_2002_to_2013_table LIMIT 1
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retryal = false)
INFO : Returning Hive schema: Schema[fieldschemas:[FieldSchema(name:group2_ces_2002_to_2013_table.group2_ces_2002_to_2013, type:string, comment:null)], properties:null]
INFO : Completed compiling command(queryId=hive_20240427042218_39ab05c6-0d81-42c3-b6de-364cdad5b666); Time taken: 2.946 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20240427042218_39ab05c6-0d81-42c3-b6de-364cdad5b666): SELECT * FROM group2_ces_2002_to_2013_table LIMIT 1
INFO : Query ID = hive_20240427042218_39ab05c6-0d81-42c3-b6de-364cdad5b666
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Subscribed to counters: [] for queryId: hive_20240427042218_39ab05c6-0d81-42c3-b6de-364cdad5b666
INFO : Ter session hasn't been created yet. Opening session
INFO : Dag name: SELECT * FROM group2_ces_2002_to_2013_table..1 (Stage-1)
INFO : Status: Running (Executing on YARN cluster with App id application_1714188588113_0001)

-----+
VERTICES    MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----+
Map 1 ..... container    SUCCEEDED   2       2       0       0       0       0       0
-----+
VERTICES: 01/01 [=====>>>] 100% ELAPSED TIME: 24.49 s
-----+
INFO : Completed executing command(queryId=hive_20240427042218_39ab05c6-0d81-42c3-b6de-364cdad5b666); Time taken: 24.49 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| group2_ces_2002_to_2013_table.group2_ces_2002_to_2013 |
+-----+
| Area Type,Area Name,Year,Month,Date,Series Code,Industry Title,Current Employment |
+-----+
1 row selected (27.681 seconds)
0: jdbc:hive2://localhost:10000>
```

**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

```

VERTICES: 01/01 [----->] 100% ELAPSED TIME: 7.93 s
INFO : Completed executing command(queryId=hive_20240427042218_39ab05c6-0d81-42c3-b6de-364cdad5b666); Time taken: 24.49 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| group2_ces_2002_to_2013_table.group2_ces_2002_to_2013 |
+-----+
| Area_Type,Area_Name,Year,Month,Date,Series_Code,Industry_Title,Current_Employment |
+-----+
1 row selected (0.861 seconds)
0: jdbc:hive2://localhost:10000> SELECT * FROM group2_ces_2002_to_2013_table LIMIT 5;
INFO : Compiling command(queryId=hive_20240427042320_18c2b4e6-185e-4328-ac43-f42599687ede): SELECT * FROM group2_ces_2002_to_2013_table LIMIT 5
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retail = false)
INFO : Returning Hive schema: Schema(fieldsSchemas:[FieldSchema(name:group2_ces_2002_to_2013_table, type:string, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20240427042320_18c2b4e6-185e-4328-ac43-f42599687ede); Time taken: 0.364 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20240427042320_18c2b4e6-185e-4328-ac43-f42599687ede): SELECT * FROM group2_ces_2002_to_2013_table LIMIT 5
INFO : Query ID = hive_20240427042320_18c2b4e6-185e-4328-ac43-f42599687ede
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting Job = [Stage-1] (PENDING) in serial mode
INFO : Session is already open
INFO : Dag name: SELECT * FROM group2_ces_2002_to_2013_table.5 (Stage-1)
INFO : Status: Running (Executing on YARN cluster with App id application_171418850113_0001)

VERTICES      MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED
Map 1 ..... container SUCCEEDED   2      2      0      0      0      0      0
VERTICES: 01/01 [----->] 100% ELAPSED TIME: 7.69 s
INFO : Completed executing command(queryId=hive_20240427042320_18c2b4e6-185e-4328-ac43-f42599687ede); Time taken: 8.395 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| group2_ces_2002_to_2013_table.group2_ces_2002_to_2013 |
+-----+
| Area_Type,Area_Name,Year,Month,Date,Series_Code,Industry_Title,Current_Employment |
| County,Alameda County,2002,January,01/1/2002,40000000,"Trade, Transportation, and Utilities",142900 |
| County,Alameda County,2002,January,01/1/2002,31000000,Durable Goods,55400 |
| County,Alameda County,2002,January,01/1/2002,01000000,Total Wage and Salary,696400 |
| County,Alameda County,2002,January,01/1/2002,00000000,Total Nonfarm,695600 |
+-----+
5 rows selected (8.861 seconds)
0: jdbc:hive2://localhost:10000> 
```

**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.

```
ssh.cloud.google.com/v2/ssh/projects/ces2024group2at/zones/us-central1-a/instances/cse2024group2-m?authuser=0&hl=en_GB&projectNumber=1028439932539&useAdminProxy=true&pageViewId=87E4C36A-733D-4A10-8ADD-665D02476A0 - Google... - ○ X
ssh.cloud.google.com/v2/ssh/projects/ces2024group2at/zones/us-central1-a/instances/cse2024group2-m?authuser=0&hl=en_GB&projectNumber=1028439932539&useAdminProxy=true&pageViewId=87E4C36A-733D-4A10-8ADD-665D02476A0 - Google... - ○ X

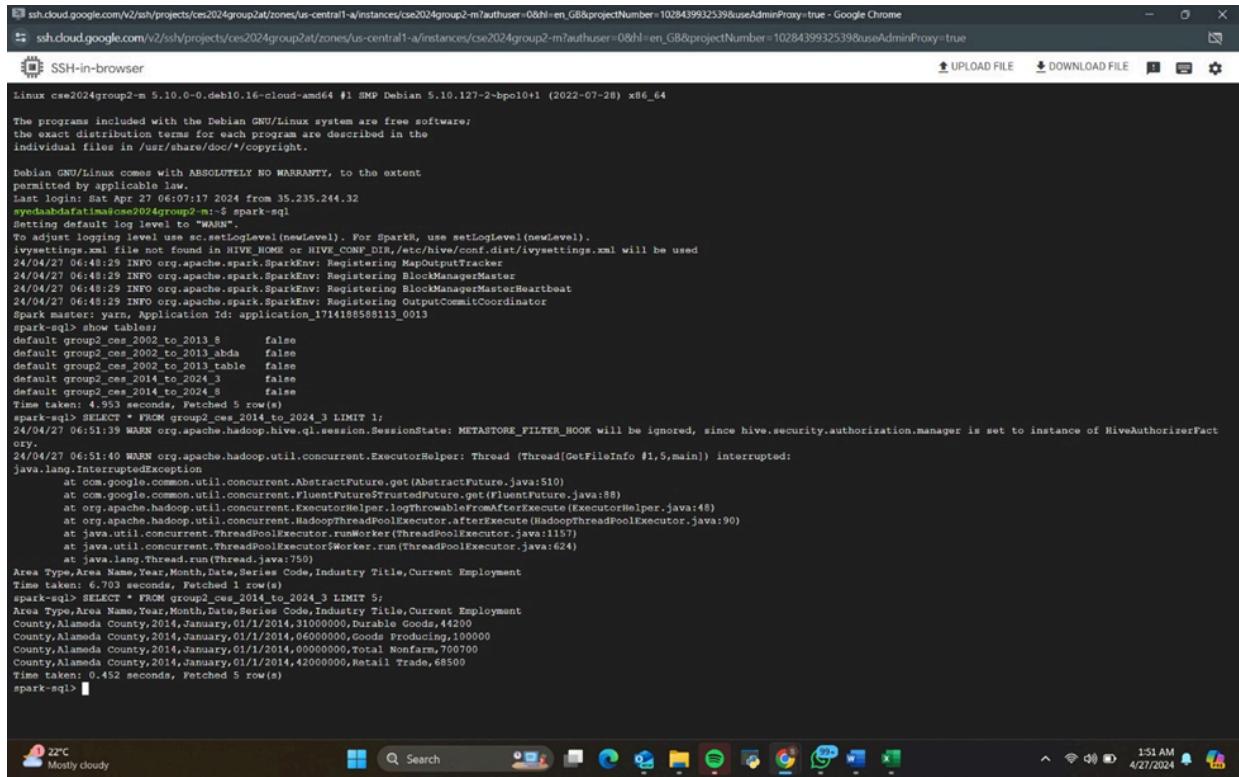
SSH-in-browser
UPLOAD FILE DOWNLOAD FILE
SELECT
    'Industry Title',
    SUM('Current Employment') AS total_employment
FROM
    group2_ces_2002_to_2013_abda
WHERE
    'Area Type' = 'State'
    AND 'Area Name' = 'California'
GROUP BY
    'Industry Title'
ORDER BY
    total_employment DESC
LIMIT 5;
INFO : Compiling command(queryId=hive_20240427054538_ea7f9392-2779-4532-b7dc-b3c61fb34e2): SELECT
'Industry Title',
SUM('Current Employment') AS total_employment
FROM
group2_ces_2002_to_2013_abda
WHERE
    'Area Type' = 'State'
    AND 'Area Name' = 'California'
GROUP BY
    'Industry Title'
ORDER BY
total_employment DESC
LIMIT 5
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retryable = false)
INFO : Returning Hive schema: Schema(fieldsSchemas:[FieldSchema(name:industry title, type:string, comment:null), FieldSchema(name:total_employment, type:double, comment:null)]}, properties=null
INFO : Compiled compiling command(queryId=hive_20240427054538_ea7f9392-2779-4532-b7dc-b3c61fb34e2); time taken: 0.346 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20240427054538_ea7f9392-2779-4532-b7dc-b3c61fb34e2): SELECT
'Industry Title',
SUM('Current Employment') AS total_employment
FROM
group2_ces_2002_to_2013_abda
WHERE
    'Area Type' = 'State'
    AND 'Area Name' = 'California'
GROUP BY
    'Industry Title'
ORDER BY
total_employment DESC
LIMIT 5
INFO : Query ID = hive_20240427054538_ea7f9392-2779-4532-b7dc-b3c61fb34e2
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Subscribed to counters: {} for queryId: hive_20240427054538_ea7f9392-2779-4532-b7dc-b3c61fb34e2
ZIM Clear
Search
12:46 AM 4/27/2024
```

**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



```
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
Linux cse2024group2-m 5.10.0-0.deb10.16-cloud-amd64 #1 SMP Debian 5.10.127-2-bpo1+1 (2022-07-28) x86_64
The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Sat Apr 27 06:07:17 2024 from 35.235.244.32
sysadmin@ces2024group2-m:~$ spark-sql
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
hive-site.xml not found in HIVE_HOME or HIVE_CONF_DIR, /etc/hive/conf.dist/hive-site.xml will be used
24/04/27 06:48:29 INFO org.apache.hadoop.mapred.lib.MmapOutputFormat
24/04/27 06:48:29 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
24/04/27 06:48:29 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
24/04/27 06:48:29 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
Spark master: yarn, Application Id: application_1714188580113_0013
spark-sql> show tables;
default group2_ces_2002_to_2013_8      false
default group2_ces_2002_to_2013_abda    false
default group2_ces_2002_to_2013_table   false
default group2_ces_2014_to_2024_3      false
default group2_ces_2014_to_2024_8      false
Time taken: 4.953 seconds. Fetched 5 row(s)
spark-sql> select * FROM group2_ces_2014_to_2024_3 LIMIT 1;
24/04/27 06:51:39 WARN org.apache.hive.qs.SessionState: METASTORE_FILTER_HOOK will be ignored, since hive.security.authorization.manager is set to instance of HiveAuthorizerFact
ory.
24/04/27 06:51:40 WARN org.apache.hadoop.util.concurrent.ExecutorHelper: Thread (Thread[GetFileInfo #1,5,main]) interrupted:
java.lang.InterruptedException
    at com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)
    at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(FluentFuture.java:98)
    at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfterExecute(ExecutorHelper.java:48)
    at org.apache.hadoop.util.concurrent.HadoopThreadReadPoolExecutor.afterExecute(HadoopThreadReadPoolExecutor.java:90)
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1157)
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
    at java.lang.Thread.run(Thread.java:750)
Area_Type,Area_Name,Year,Month,Date,Serial_Code,Industry,Title,Current_Employment
Time taken: 6.703 seconds. Fetched 1 row(s)
spark-sql> SELECT * FROM group2_ces_2014_to_2024_3 LIMIT 5;
Area_Type,Area_Name,Year,Month,Date,Serial_Code,Industry,Title,Current_Employment
County,Alameda County,2014,January,01/1/2014,31000000,Durable Goods,44200
County,Alameda County,2014,January,01/1/2014,06000000,Goods Producing,100000
County,Alameda County,2014,January,01/1/2014,00000000,Total Nonfarm,700700
County,Alameda County,2014,January,01/1/2014,42000000,Retail Trade,68500
Time taken: 0.452 seconds. Fetched 5 row(s)
spark-sql>
```

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.

```
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/sshy/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
24/04/27 06:48:29 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
Spark master: yarn, Application Id: application_171186588113_0013
spark-sql> show tables;
default group2_ces_2002_to_2013_8 false
default group2_ces_2002_to_2013_abda false
default group2_ces_2002_to_2013_table false
default group2_ces_2014_to_2024_3 false
default group2_ces_2014_to_2024_8 false
Time taken: 0.503 seconds, Fetched 5 row(s)
spark-sql> SELECT * FROM group2_ces_2014_to_2024_3 LIMIT 1;
24/04/27 06:51:39 WARN org.apache.hadoop.hive.ql.session.SessionState: METASTORE_FILTER_HOOK will be ignored, since hive.security.authorization.manager is set to instance of HiveAuthorizerFact
ory.
24/04/27 06:51:40 WARN org.apache.hadoop.util.concurrent.ExecutorHelper: Thread (Thread[GetFileInfo #1,5,main]) interrupted:
java.lang.InterruptedIOException
    at com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)
    at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(FluentFuture.java:88)
    at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfterExecute(ExecutorHelper.java:48)
    at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.afterExecute(HadoopThreadPoolExecutor.java:90)
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1157)
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
    at java.lang.Thread.run(Thread.java:750)
Area Type,Area Name,Year,Month,Date,Series Code,Industry Title,Current Employment
Time taken: 6.703 seconds, Fetched 1 row(s)
spark-sql> SELECT * FROM group2_ces_2014_to_2024_3 LIMIT 5;
+-----+-----+-----+-----+-----+
|Area Type|Area Name|Year|Month|Date|Series Code|Industry Title|Current Employment|
+-----+-----+-----+-----+-----+
|County,Alameda County,2014,January,01/1/2014,31000000,Durable Goods,44200|
|County,Alameda County,2014,January,01/1/2014,06000000,Goods Producing,100000|
|County,Alameda County,2014,January,01/1/2014,00000000,Total Nonfarm,790700|
|County,Alameda County,2014,January,01/1/2014,42000000,Retail Trade,68500|
+-----+-----+-----+-----+-----+
Time taken: 0.452 seconds, Fetched 5 row(s)
spark-sql> SELECT
    >     'Industry Title',
    >     AVG('Current Employment') AS average_employment
    > FROM
    >     group2_ces_2014_to_2024_8
    > WHERE
    >     'Area Type' = 'State'
    >     AND 'Area Name' = 'California'
    > GROUP BY
    >     'Industry Title'
    > ORDER BY
    >     average_employment DESC
    > LIMIT 5;
Total Wage and Salary 1.7230484426229507E7
Total Nonfarm 1.681503975409836E7
Service-Providing 1.4650932786885247E7
Total Private 1.4289316393442628E7
Private Service Providing 1.2125547540983606E7
Time taken: 5.629 seconds, Fetched 5 row(s)
spark-sql>
```

## 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

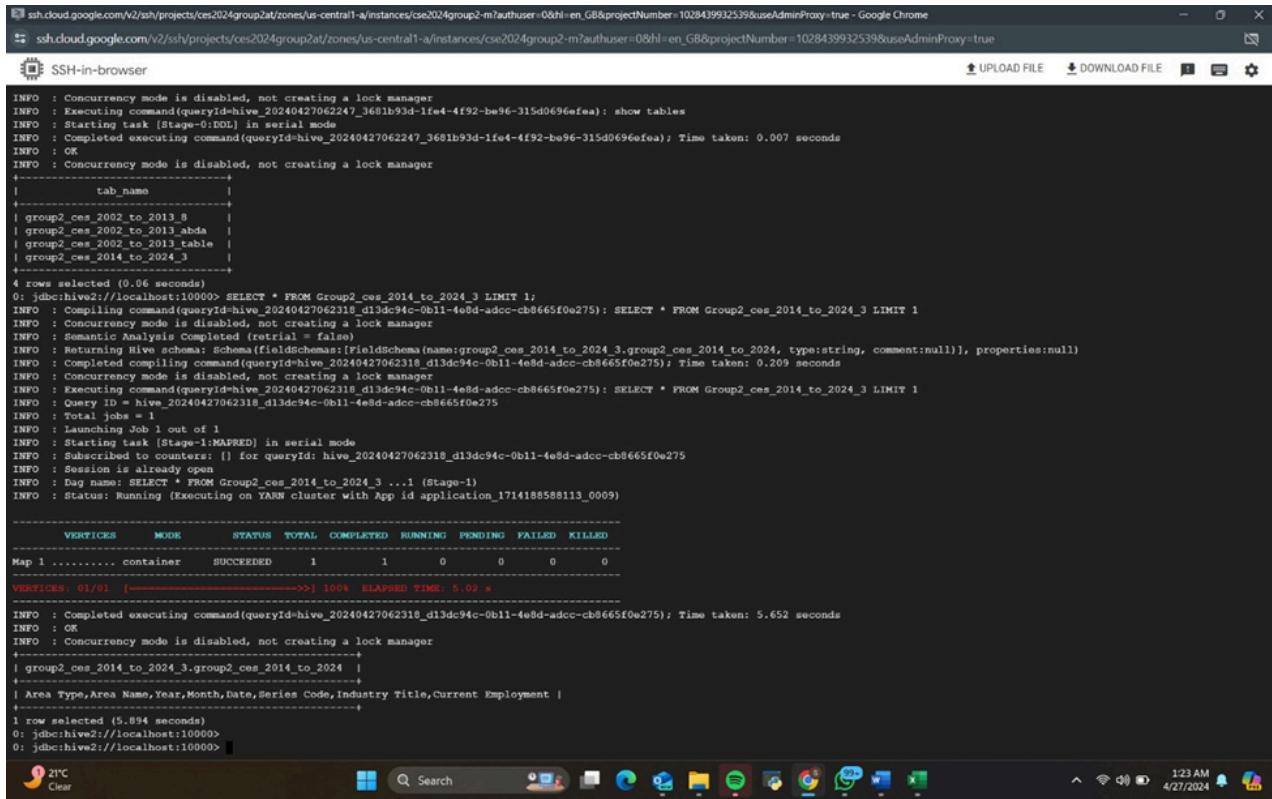
**Query 1:** It was **24.49 seconds** in Hive. In Spark, it was **6.703 seconds**.

**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



The screenshot shows a Google Chrome browser window with two tabs open, both displaying the same Hive query results. The title bar for both tabs reads "ssh.cloud.google.com/v2/ssh/projects/cse2024group2at/zones/us-central1-a/instances/cse2024group2-m?authuser=0&hl=en\_GB&projectNumber=1028439932539&useAdminProxy=true - Google Chrome". The main content area shows the output of a Hive query:

```
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20240427062247_3601b93d-1fe4-4f92-be96-315d0696e0ea): show tables
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20240427062247_3601b93d-1fe4-4f92-be96-315d0696e0ea); Time taken: 0.007 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| tab_name          |
+-----+
| group2_ces_2002_to_2013_8   |
| group2_ces_2002_to_2013_abda |
| group2_ces_2002_to_2013_table |
| group2_ces_2014_to_2024_3   |
+-----+
4 rows selected (0.06 seconds)
0: jdbc:hive2://localhost:10000> SELECT * FROM Group2_ces_2014_to_2024_3 LIMIT 1;
INFO : Compiling command(queryId=hive_20240427062318_d13dc94c-0b11-4e8d-adcc-cb8665f0e275): SELECT * FROM Group2_ces_2014_to_2024_3 LIMIT 1
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (trialist = false)
INFO : Returning Hive schema: Schema(fieldschemas:[FieldSchema(name:group2_ces_2014_to_2024_3.group2_ces_2014_to_2024, type:string, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20240427062318_d13dc94c-0b11-4e8d-adcc-cb8665f0e275); Time taken: 0.209 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20240427062318_d13dc94c-0b11-4e8d-adcc-cb8665f0e275): SELECT * FROM Group2_ces_2014_to_2024_3 LIMIT 1
INFO : Job ID: hive_20240427062318_d13dc94c-0b11-4e8d-adcc-cb8665f0e275
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Subscribed to counters: [] for queryId: hive_20240427062318_d13dc94c-0b11-4e8d-adcc-cb8665f0e275
INFO : Session is already open
INFO : Dag name: SELECT * FROM Group2_ces_2014_to_2024_3 ...1 (Stage-1)
INFO : Status: Running (Executing on YARN cluster with App id application_1714188588113_0009)

----- VERTICES      MODE      STATUS      TOTAL      COMPLETED      RUNNING      PENDING      FAILED      KILLED -----
Map 1 ..... container      SUCCEEDED      1      1      0      0      0      0
----- VERTICES: 0/1 [----->] 100% ELAPSED TIME: 5.02 s
INFO : Completed executing command(queryId=hive_20240427062318_d13dc94c-0b11-4e8d-adcc-cb8665f0e275); Time taken: 5.652 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| group2_ces_2014_to_2024_3_group2_ces_2014_to_2024 |
+-----+
| Area Type,Area Name,Year,Month,date,Series Code,Industry Title,Current Employment |
+-----+
1 row selected (5.894 seconds)
0: jdbc:hive2://localhost:10000>
0: jdbc:hive2://localhost:10000>
```

## HIVE RESULTS

**Query 1 in Hive:** SELECT \* FROM Group2\_ces\_2014\_to\_2024\_3 LIMIT 1;

**Time: 5.894 seconds.**

```

ssh.cloud.google.com/v2/ssh/projects/ces2024group2at/zones/us-central1-a/instances/ces2024group2-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/ces2024group2-m?authuser=0&hl=en_GB&projectNumber=1028439932539&useAdminProxy=true

SSH-in-browser
UPLOAD FILE DOWNLOAD FILE
VERITIES: 01/01 [----->] 100% ELAPSED TIME: 5.02 s
INFO : Completed executing command(queryId=hive_20240427062318_d13dc94c-0b11-4e8d-adcc-cb8665f0e275); Time taken: 5.652 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| group2_ces_2014_to_2024_3.group2_ces_2014_to_2024 |
+-----+
| Area_Type,Area_Name,Year,Month,date,Series_Code,Industry_Title,Current_Employment |
+-----+
1 row selected (5.894 seconds)
0: jdbc:hive2://localhost:10000>
0: jdbc:hive2://localhost:10000> SELECT * FROM Group2_ces_2014_to_2024_3 LIMIT 5;
INFO : Compiling command(queryId=hive_20240427062336_d41ed698-2901-4e9f-9064-9a6307cb598c): SELECT * FROM Group2_ces_2014_to_2024_3 LIMIT 5
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retryal = false)
INFO : Returning Hive schema: Schema(fieldsSchemas:[FieldSchema(name:group2_ces_2014_to_2024_3.group2_ces_2014_to_2024, type:string, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20240427062336_d41ed698-2901-4e9f-9064-9a6307cb598c); Time taken: 0.157 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20240427062336_d41ed698-2901-4e9f-9064-9a6307cb598c): SELECT * FROM Group2_ces_2014_to_2024_3 LIMIT 5
INFO : Query ID = hive_20240427062336_d41ed698-2901-4e9f-9064-9a6307cb598c
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Subscribed to counters [] for queryId: hive_20240427062336_d41ed698-2901-4e9f-9064-9a6307cb598c
INFO : Session is already open
INFO : Dag name: SELECT * FROM Group2_ces_2014_to_2024_3 ...5 (Stage-1)
INFO : Status: Running (Executing on YARN cluster with App id application_1714180580113_0009)

VERITIES: 01/01 [----->] 100% ELAPSED TIME: 5.46 s
INFO : Completed executing command(queryId=hive_20240427062336_d41ed698-2901-4e9f-9064-9a6307cb598c); Time taken: 6.082 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| group2_ces_2014_to_2024_3.group2_ces_2014_to_2024 |
+-----+
| Area_Type,Area_Name,Year,Month,date,Series_Code,Industry_Title,Current_Employment |
| County,Alameda County,2014,January,01/1/2014,31000000,Durable Goods,44200 |
| County,Alameda County,2014,January,01/1/2014,06000000,Goods Producing,100000 |
| County,Alameda County,2014,January,01/1/2014,00000000,Total Nonfarm,700700 |
| County,Alameda County,2014,January,01/1/2014,42000000,Retail Trade,68500 |
+-----+
5 rows selected (6.272 seconds)
0: jdbc:hive2://localhost:10000>

```

**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.

```

11)
INFO : Completed compiling command(queryId=hive_20240427064551_2422644d-d8df-4d1d-8df6-2d1c6995f3b1); Time taken: 0.23 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20240427064551_2422644d-d8df-4d1d-8df6-2d1c6995f3b1): SELECT
`Industry title`,
AVG(`Current Employment`) AS average_employment
FROM
group2_ces_2014_to_2024_0
WHERE
`Area Type` = 'State'
AND `Area Name` = 'California'
GROUP BY
`Industry title`
ORDER BY
average_employment DESC
LIMIT 5
INFO : Query ID = hive_20240427064551_2422644d-d8df-4d1d-8df6-2d1c6995f3b1
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Subscribed to counters: [] for queryId: hive_20240427064551_2422644d-d8df-4d1d-8df6-2d1c6995f3b1
INFO : Session is already open
INFO : Dag name: SELECT
`Industry title`,
AVG(`Current Em..5 (Stage-1)
INFO : Status: Running (Executing on YARN cluster with App id application_171418858813_0012)

-----  

VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

-----  

Map 1 ..... container SUCCEEDED 1 1 0 0 0 0  

Reducer 2 ..... container SUCCEEDED 1 1 0 0 0 0  

Reducer 3 ..... container SUCCEEDED 1 1 0 0 0 0  

-----  

VERTICES: 03/03 [=====>] 100% ELAPSED TIME: 7.59 s  

-----  

INFO : Completed executing command(queryId=hive_20240427064551_2422644d-d8df-4d1d-8df6-2d1c6995f3b1); Time taken: 8.342 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+-----+
| industry title | average_employment |
+-----+-----+
| Total Wage and Salary | 1.7230484426229507E7 |
| Total Nonfarm | 1.681503975409836E7 |
| Service-Providing | 1.4650932786885247E7 |
| Total Private | 1.4289316393442623E7 |
| Private Service Providing | 1.2125547540983606E7 |
+-----+
5 rows selected (8.639 seconds)
0: jdbc:hive2://localhost:10000>

```

The screenshot shows a terminal window within a browser tab titled 'SSH-in-browser'. The terminal output displays the execution of a Hive query. The query selects the top 5 industries with the highest average employment in California from 2014 to 2024. The results are shown in a table with columns 'industry title' and 'average\_employment'. The industries listed are Total Wage and Salary, Total Nonfarm, Service-Providing, Total Private, and Private Service Providing, each with an average employment value.

**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

```
ssh.cloud.google.com:v2/ssh/projects/ces2024group2a1/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/ces2024group2a1/zones/us-central1-a/instances/cse2024group2-m?authuser=0&hl=en_GB&projectNumber=1028439932539&useAdminProxy=true

SSH-in-browser
 UPLOAD FILE DOWNLOAD FILE
Linux cse2024group2-m 5.10.0-0.deb10.16-cloud-amd64 #1 SMP Debian 5.10.127-2-bpo10+1 (2022-07-28) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.

Last login: Sat Apr 27 06:07:17 2024 from 35.235.244.32
syndication@ces2024group2-m:~$ spark-sql
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel).
ivysettings.xml file not found in HIVE_HOME or HIVE_CONF_DIR, /etc/hive/conf.dist/ivysettings.xml will be used
24/04/27 06:48:29 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
24/04/27 06:48:29 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
24/04/27 06:48:29 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
24/04/27 06:48:29 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
Spark master: yarn, Application Id: application_1714188589113_0013
spark-sql> show tables;
default group2_ces_2002_to_2013_8      false
default group2_ces_2002_to_2013_abda    false
default group2_ces_2002_to_2013_table   false
default group2_ces_2014_to_2024_3       false
default group2_ces_2014_to_2024_3       false
Time taken: 0.19 seconds, Fetched 5 row(s)
spark-sql> SELECT * FROM group2_ces_2014_to_2024_3 LIMIT 1;
24/04/27 06:51:39 WARN org.apache.hadoop.hive.ql.session.SessionState: METASTORE_FILTER_HOOK will be ignored, since hive.security.authorization.manager is set to instance of HiveAuthorizerFactory.
24/04/27 06:51:40 WARN org.apache.hadoop.util.concurrent.ExecutorHelper: Thread (Thread[GetFileInfo #1,main]) interrupted: java.lang.InterruptedIOException
at com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)
at com.google.common.util.concurrent.FluentFuture$trustedFuture.get(FluentFuture.java:88)
at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfterExecute(ExecutorHelper.java:46)
at org.apache.hadoop.util.concurrent.HadoopThreadpoolExecutor.afterExecute(HadoopThreadpoolExecutor.java:90)
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:117)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
at java.lang.Thread.run(Thread.java:750)
Area Type,Area,Name,Year,Month,Date,Series,Code,Industry,Title,Current Employment
Time taken: 6.703 seconds, Fetched 1 row(s)
spark-sql> SELECT * FROM group2_ces_2014_to_2024_3 LIMIT 5;
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|Area Type|Area|Name|Year|Month|Date|Series|Code|Industry|Title|Current Employment|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|County|Alameda County, 2014, January, 01/1/2014, 31000000,Durable Goods,44200|County|Alameda County, 2014, January, 01/1/2014, 31000000,Durable Goods,44200|County|Alameda County, 2014, January, 01/1/2014, 60000000,Goods Producing,100000|County|Alameda County, 2014, January, 01/1/2014, 00000000,Total Nonfarm,700700|County|Alameda County, 2014, January, 01/1/2014, 42000000,Retail Trade,68500
Time taken: 0.452 seconds, Fetched 5 row(s)
spark-sql> 
```

## 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

```
ssh.cloud.google.com/v2/ssh/projects/cse2024group2at/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/cse2024group2at/zones/us-central1-a/instances/cse2024group2-m?authuser=0&hl=en_GB&projectNumber=1028439932539&useAdminProxy=true
SSH-in-browser
INFO : Executing command(queryId=hive_20240427071200_f4876c12-ee25-4a62-b6ad-3dbc1dc02ff2): show tables
INFO : Starting task [Stage-0:EDD] in serial mode
INFO : Completed executing command(queryId=hive_20240427071200_f4876c12-ee25-4a62-b6ad-3dbc1dc02ff2); Time taken: 0.012 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| tab_name |
+-----+
| group2_ces_2002_to_2013_8 |
| group2_ces_2002_to_2013_abda |
| group2_ces_2002_to_2013_table |
| group2_ces_2014_to_2024_3 |
| group2_ces_2014_to_2024_8 |
| group2_ces_se_2023to_2025_3 |
+-----+
6 rows selected (0.071 seconds)
0:jdbc:hive2://localhost:10000> SELECT * FROM Group2_CES_SE_2023to_2025_3 LIMIT 1;
INFO : Compiling command(queryId=hive_20240427071235_62c9ad6d-6b9a-4dff-99fc-fc958f00a7ff): SELECT * FROM Group2_CES_SE_2023to_2025_3 LIMIT 1
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:group2_ces_se_2023to_2025_3.group2_ces_se_2023to_2025, type:string, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20240427071235_62c9ad6d-6b9a-4dff-99fc-fc958f00a7ff); Time taken: 0.2 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20240427071235_62c9ad6d-6b9a-4dff-99fc-fc958f00a7ff): SELECT * FROM Group2_CES_SE_2023to_2025_3 LIMIT 1
INFO : Query ID = hive_20240427071235_62c9ad6d-6b9a-4dff-99fc-fc958f00a7ff
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Starting task [Stage-1:MAPRED] for queryId: hive_20240427071235_62c9ad6d-6b9a-4dff-99fc-fc958f00a7ff
INFO : No session has been created yet. Opening session
INFO : Dag name: SELECT * FROM Group2_CES_SE_2023to_2025...1 (Stage-1)
INFO : Status: Running (Executing on YARN cluster with App id application_1714188588113_0014)

----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
Map 1 ..... container SUCCEEDED 1 1 0 0 0 0
----- VERTICES: 01/01 [=====>>>] 100% ELAPSED TIME: 5.76 s
----- INFO : Completed executing command(queryId=hive_20240427071235_62c9ad6d-6b9a-4dff-99fc-fc958f00a7ff); Time taken: 13.768 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| group2_ces_se_2023to_2025_3.group2_ces_se_2023to_2025 |
+-----+
| Area Type,Area Name,Period,Series Code,Industry Title,Base Quarter Employment Estimate,Projected Quarter Employment Estimate,Numeric Change,Percentage Change |
+-----+
1 row selected (13.999 seconds)
0:jdbc:hive2://localhost:10000>

22°C Mostly cloudy
Search
212 AM 4/27/2024
```

Query 1 in Hive: SELECT \* FROM Group2\_CES\_SE\_2023\_to\_2025\_3 LIMIT 1;

Time: 13.999 seconds.

### Hive Screenshot

The screenshot shows an SSH session in Google Cloud Shell. The terminal window displays the output of a Hive query. The query is: SELECT \* FROM Group2\_CES\_SE\_2023to\_2025\_3 LIMIT 1;. The output shows the execution details, including command completion, compilation, and execution time (13.999 seconds). It also shows the resulting data, which is a single row of employment statistics for California from 2023 to 2025. The terminal window has a header with tabs for 'SSH-in-browser' and 'File', and a toolbar with icons for upload, download, and settings. The system tray at the bottom shows network status, battery level, and the date/time (4/27/2024 2:13 AM).

```
VERTICES 01/01 [----->>>] 100% ELAPSED TIME: 5.76 s
INFO : Completed executing command(queryId=hive_20240427071235_62c9ad6d-6b9a-4dff-99fc-fc958f00a7ff); Time taken: 13.768 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| group2_ces_se_2023to_2025_3.group2_ces_se_2023to_2025 |
+-----+
| Area Type,Area Name,Period,Series Code,Industry Title,Base Quarter Employment Estimate,Projected Quarter Employment Estimate,Numeric Change,Percentage Change |
+-----+
1 row selected (13.999 seconds)
0: jdbc:hive2://localhost:10000> SELECT * FROM Group2_CES_SE_2023to_2025_3 LIMIT 1;
INFO : Compiling command(queryId=hive_20240427071308_e371c750-2339-4f9c-a85b-f4863b152627): SELECT * FROM Group2_CES_SE_2023to_2025_3 LIMIT 5
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (trial= false)
INFO : Returning Hive schema: Schema(fieldschemas:[FieldSchema(name:group2_ces_se_2023to_2025_3.group2_ces_se_2023to_2025, type:string, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20240427071308_e371c750-2339-4f9c-a85b-f4863b152627); Time taken: 0.152 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20240427071308_e371c750-2339-4f9c-a85b-f4863b152627): SELECT * FROM Group2_CES_SE_2023to_2025_3 LIMIT 5
INFO : Query ID = hive_20240427071308_e371c750-2339-4f9c-a85b-f4863b152627
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage=1:MAPRED] in serial mode
INFO : Subscribed to counters: [] for queryId: hive_20240427071308_e371c750-2339-4f9c-a85b-f4863b152627
INFO : Session is already open
INFO : Dag name: SELECT * FROM Group2_CES_SE_2023to_2025...5 (Stage=1)
INFO : Status: Running (Executing on YARN cluster with App id application_1714188580113_0014)

----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
Map 1 ..... container SUCCEEDED 1 1 0 0 0 0
----- VERTICES 01/01 [----->>>] 100% ELAPSED TIME: 5.54 s
INFO : Completed executing command(queryId=hive_20240427071308_e371c750-2339-4f9c-a85b-f4863b152627); Time taken: 6.198 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| group2_ces_se_2023to_2025_3.group2_ces_se_2023to_2025 |
+-----+
| Area Type,Area Name,Period,Series Code,Industry Title,Base Quarter Employment Estimate,Projected Quarter Employment Estimate,Numeric Change,Percentage Change |
| state,California,2023-2025,000001,Total Employment,19920000,20412400,492400,2.5 |
| state,California,2023-2025,006010,Self Employment,13190000,13440000,25000,1.9 |
| state,California,2023-2025,008010,Private Household Workers,35100,32600,-2500,-7.1 |
| state,California,2023-2025,11000000,Total Farm,472200,469600,-2600,-0.6 |
+-----+
5 rows selected (6.395 seconds)
0: jdbc:hive2://localhost:10000> 
```

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/ces2024group2-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/ces2024group2-m?authuser=0&hl=en_GB&projectNumber=1028439932539&useAdminProxy=true

SSH-in-browser
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 : Query ID = hive_20240427071821_22536e7c-fd51-4304-8519-059008330c0f
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Subscribed to counters: [] for queryId: hive_20240427071821_22536e7c-fd51-4304-8519-059008330c0f
INFO : Session is already open
INFO : Dag name: SELECT
`Industry Title`,
AVG('Projected...10 (Stage-1)
INFO : Status: Running (Executing on YARN cluster with App id application_1714188588113_0014)

----- VERTICES MODE STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED -----
Map 1 ..... container SUCCEEDED 1 1 0 0 0 0
Reducer 2 ..... container SUCCEEDED 1 1 0 0 0 0
Reducer 3 ..... container SUCCEEDED 1 1 0 0 0 0
VERTICES: 03/03 [=====>>>] 100% ELAPSED TIME: 6.71 s
INFO : Completed executing command(queryId=hive_20240427071821_22536e7c-fd51-4304-8519-059008330c0f); time taken: 7.406 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| industry title | average_employment |
+-----+
| Total Employment | 2.0412487 |
| Total Nonfarm | 1.8566287 |
| Professional and Business Services | 2000000.0 |
| Health Care and Social Assistance | 2856500.0 |
| Government | 2657100.0 |
| State and Local Government | 2406600.0 |
| Leisure and Hospitality | 2154300.0 |
| Local Government | 1836800.0 |
| Accommodation and Food Services | 1781700.0 |
| Retail Trade | 1615500.0 |
+-----+
10 rows selected (7.749 seconds)
0: jdbc:hive2://localhost:10000>

```

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

```
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

Linux cse2024group2-m 5.10.0-0.deb10.16-cloud-amd64 #1 SMP Debian 5.10.127-2+bp10+1 (2022-07-28) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Sat Apr 27 07:21:39 2024 from 35.235.244.32
syednabdafatimae2024group2-m:~$ spark-sql
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkSQL, use setLogLevel(newLevel).
ivysettings.xml file not found in HIVE_HOME or HIVE_CONF_DIR, /etc/hive/conf.dist/ivysettings.xml will be used
24/04/27 07:22:42 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
24/04/27 07:22:42 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
24/04/27 07:22:42 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
24/04/27 07:22:42 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
Spark Master: yarn Application Id: application_171418558813_0015
Spark User: syednabdafatimae2024group2-m
default group2_ces_2002_to_2013_8 false
default group2_ces_2002_to_2013_shda false
default group2_ces_2002_to_2013_table false
default group2_ces_2014_to_2024_3 false
default group2_ces_2014_to_2024_8 false
default group2_ces_2023to2025_3 false
default group2_ces_2023to2025_8 false
Time taken: 5.916 seconds, Fetched 7 row(s)
spark-sql> SELECT * FROM Group2_CES_SE_2023to_2025_3 LIMIT 1;
24/04/27 07:23:27 WARN org.apache.hadoop.hive.session.SessionState: METASTORE_FILTER_HOOK will be ignored, since hive.security.authorization.manager is set to instance of HiveAuthorizerFactory
Area Type,Area Name,Period,Series Code,Industry Title,Base Quarter Employment Estimate,Projected Quarter Employment Estimate,Numeric Change,Percentage Change
Time taken: 6.74 seconds, Fetched 1 row(s)
spark-sql> SELECT * FROM Group2_CES_SE_2023to_2025_3 LIMIT 5;
24/04/27 07:23:36 WARN org.apache.hadoop.util.concurrent.ExecutorHelper: Thread (Thread[GetFileInfo #1,5,main]) interrupted:
java.lang.InterruptedException
    at com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)
    at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(FluentFuture.java:88)
    at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfterExecute(ExecutorHelper.java:48)
    at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.afterExecute(HadoopThreadPoolExecutor.java:90)
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1157)
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
    at java.lang.Thread.run(Thread.java:750)
Area Type,Area Name,Period,Series Code,Industry Title,Base Quarter Employment Estimate,Projected Quarter Employment Estimate,Numeric Change,Percentage Change
State,California,2023-2025,,000001>Total Employment,19920000,20412400,492400,2.5
State,California,2023-2025,006010,Self Employment,1319000,1344000,25000,1.9
State,California,2023-2025,,008010,Private Household Workers,35100,32600,-2500,-7.1
State,California,2023-2025,11000000>Total Farm,472200,469600,-2600,-0.6
Time taken: 0.421 seconds, Fetched 5 row(s)
spark-sql>

22°C Cloudy
 Search
 223 AM 4/27/2024

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

spark-sql> SELECT * FROM Group2_CES_SE_2023to_2025_3 LIMIT 5;
24/04/27 07:23:36 WARN org.apache.hadoop.util.concurrent.ExecutorHelper: Thread (Thread[GetFileInfo #1,5,main]) interrupted:
java.lang.InterruptedException
    at com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)
    at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(FluentFuture.java:88)
    at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfterExecute(ExecutorHelper.java:48)
    at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.afterExecute(HadoopThreadPoolExecutor.java:90)
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1157)
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
    at java.lang.Thread.run(Thread.java:750)
Area Type,Area Name,Period,Series Code,Industry Title,Base Quarter Employment Estimate,Projected Quarter Employment Estimate,Numeric Change,Percentage Change
State,California,2023-2025,,000001>Total Employment,19920000,20412400,492400,2.5
State,California,2023-2025,006010,Self Employment,1319000,1344000,25000,1.9
State,California,2023-2025,,008010,Private Household Workers,35100,32600,-2500,-7.1
State,California,2023-2025,11000000>Total Farm,472200,469600,-2600,-0.6
Time taken: 0.421 seconds, Fetched 5 row(s)
spark-sql> 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;
24/04/27 07:24:09 WARN org.apache.hadoop.util.concurrent.ExecutorHelper: Thread (Thread[GetFileInfo #1,5,main]) interrupted:
java.lang.InterruptedException
    at com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)
    at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(FluentFuture.java:88)
    at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfterExecute(ExecutorHelper.java:48)
    at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.afterExecute(HadoopThreadPoolExecutor.java:90)
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1157)
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
    at java.lang.Thread.run(Thread.java:750)
Total Employment 2.04124E7
Total Nonfarm 1.85662E7
Professional and Business Services 2943100.0
Health Care and Social Assistance 2856500.0
Government 2657100.0
State and Local Government 2406600.0
Arts, Entertainment and Hospitality 2154300.0
Local Government 1836800.0
Accommodation and Food Services 1781700.0
Retail Trade 1615500.0
Time taken: 4.916 seconds, Fetched 10 row(s)
spark-sql>

22°C Cloudy
 Search
 224 AM 4/27/2024
```

## **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 13<sup>th</sup>, 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 20 <sup>th</sup> , 2024,	Pending
2	Dataproc: creating clusters	Deepthimai and Mohammed	April 20 <sup>th</sup> , 2024	Pending

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

## **Group 2 Project Meeting Notes-2**

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

Date: April 20<sup>th</sup>, 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 27 <sup>th</sup> , 2024,	Completed
2	Assigned tasks regarding the report	All	April 27 <sup>th</sup> , 2024	Completed
3	Team Name and Logo	Tanaya Dutt	April 27 <sup>th</sup> , 2024	Completed

### **Group 2 Project Meeting Notes-3**

Date: April 27<sup>th</sup>, 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 27 <sup>th</sup> , 2024,	Completed
2	Presentation completion	All	April 27 <sup>th</sup> , 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&gclidcsrc=3p.ds&msclkid=ed32af75a6891d8a6336c944a1e5b2ee&hl=en](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&gclidcsrc=3p.ds&msclkid=ed32af75a6891d8a6336c944a1e5b2ee&hl=en)