“Lack of Student Engagement in Group Projects”

Analysis Project

Tyler Rollinson & Gabrael Smallwood

University of North Texas

IPAC 4240: Principles of Data Structures, Harvesting and Wrangling

Dr. Schenita Floyd

March 3rd, 2024

Project Overview

This project aims to understand the issue of lack of student engagement in group projects. Leveraging Google Cloud Platform (GCP), the team meticulously navigated the data lifecycle.

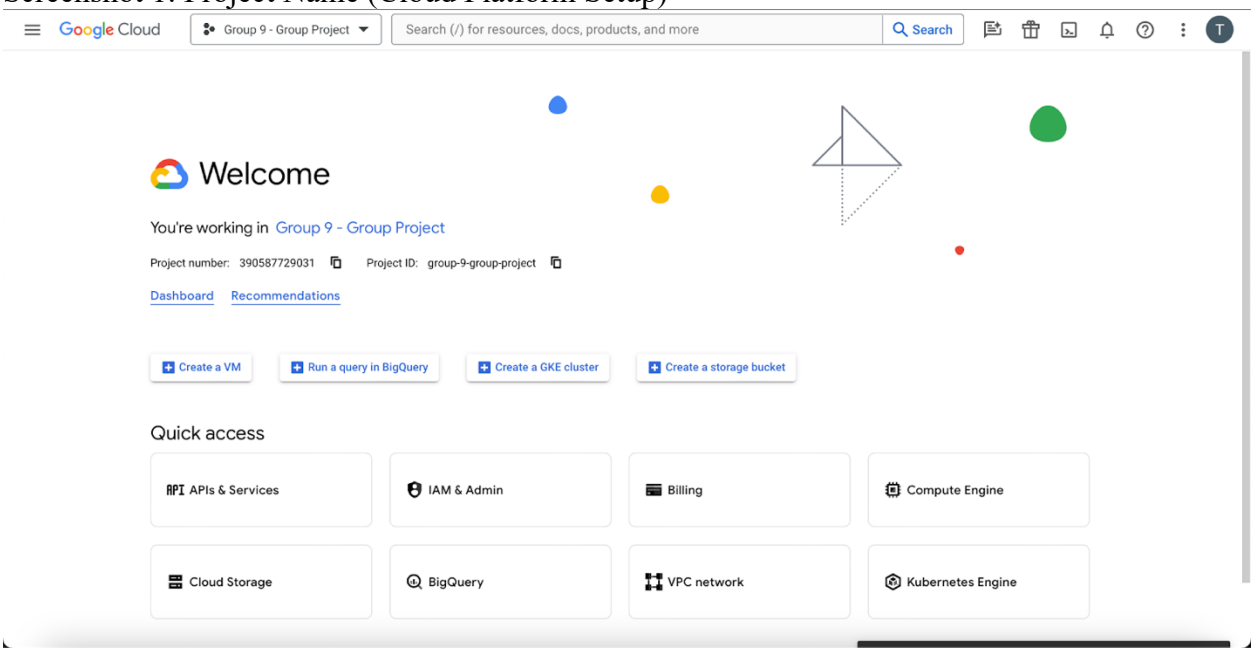
Initially, approved data sources from reputable platforms such as Data.gov and Reddit were identified and utilized. Data storage was established on GCP, with both static and streaming data sources uploaded for processing. Using GCP DataPrep by Trifacta, the team meticulously cleaned and preprocessed the data, ensuring its quality and relevance.

The processed data was then subjected to analysis using tools like Google Dataproc and BigQuery. Queries were executed to extract meaningful insights into student engagement patterns, shedding light on potential factors contributing to the lack of engagement.

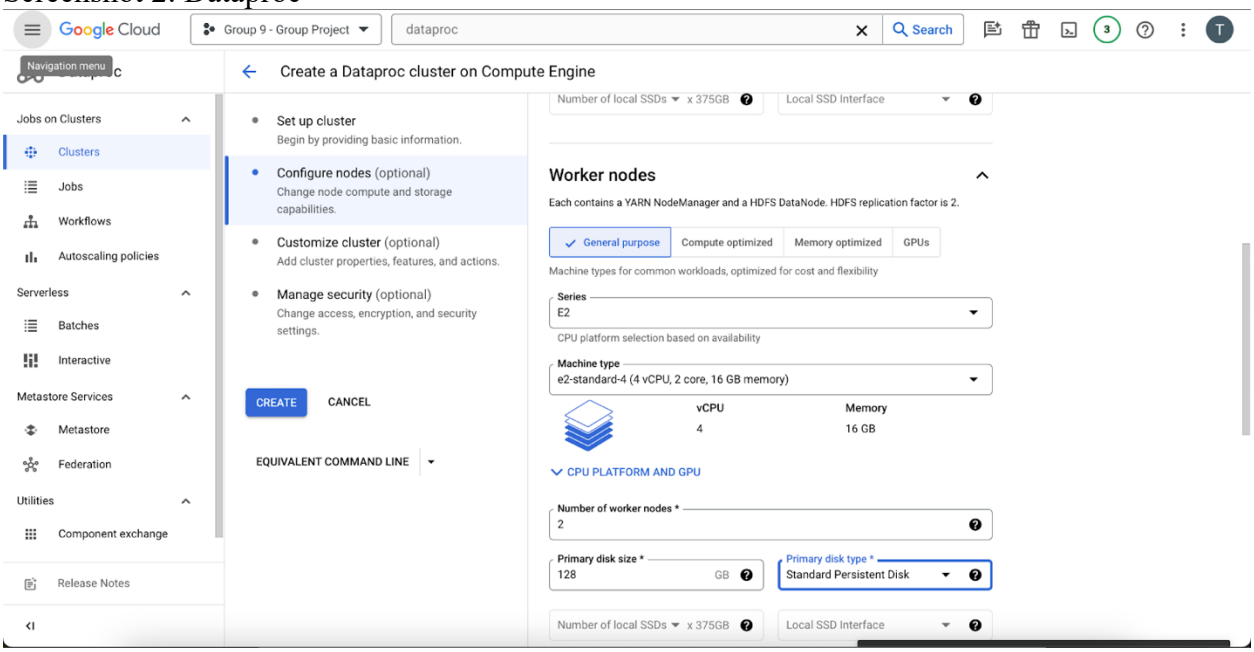
Our Cloud Provider

Our team has utilized Google Cloud Platform (GCP) to address the lack of student engagement n group projects, leveraging various tools provided by GCP. We've established a project and set up a Hadoop infrastructure using Google Dataproc. Below are screenshots for your reference, showcasing the setup:

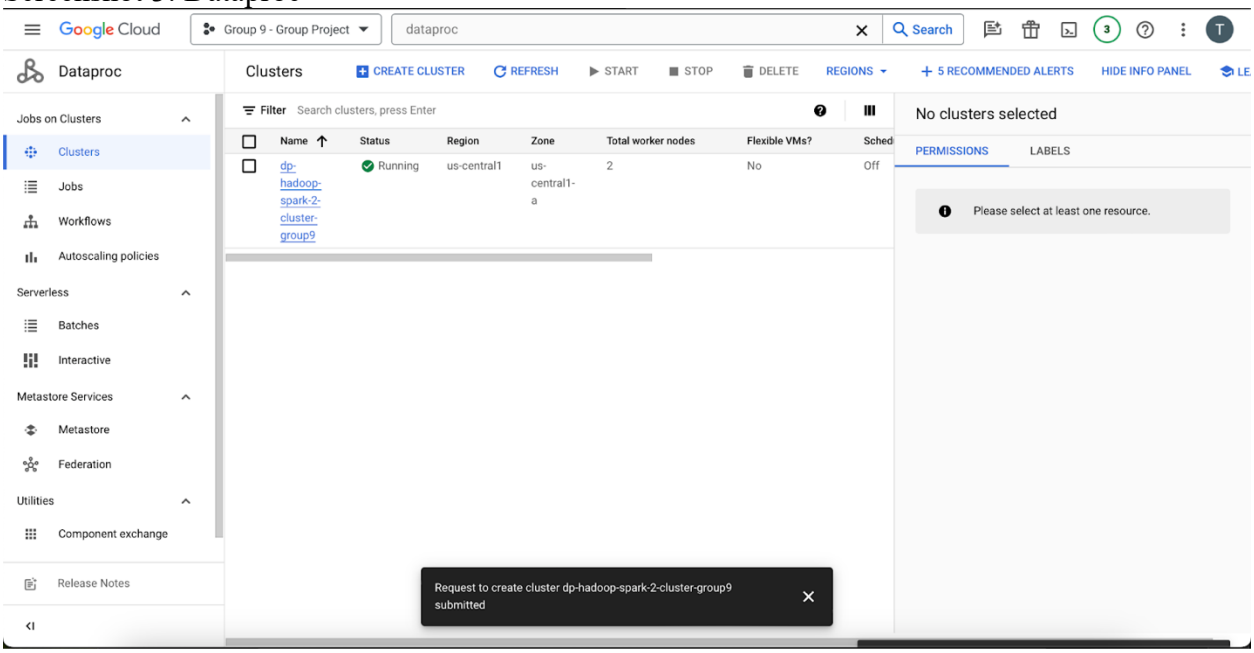
Screenshot 1: Project Name (Cloud Platform Setup)



Screenshot 2: Dataproc



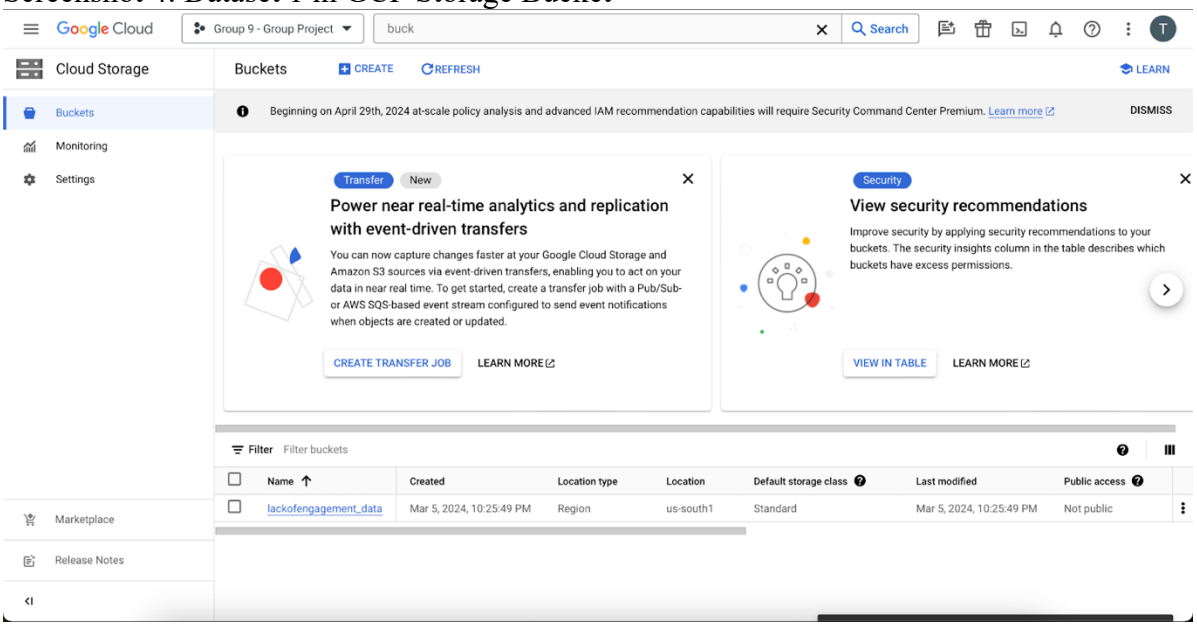
Screenshot 3: Dataproc



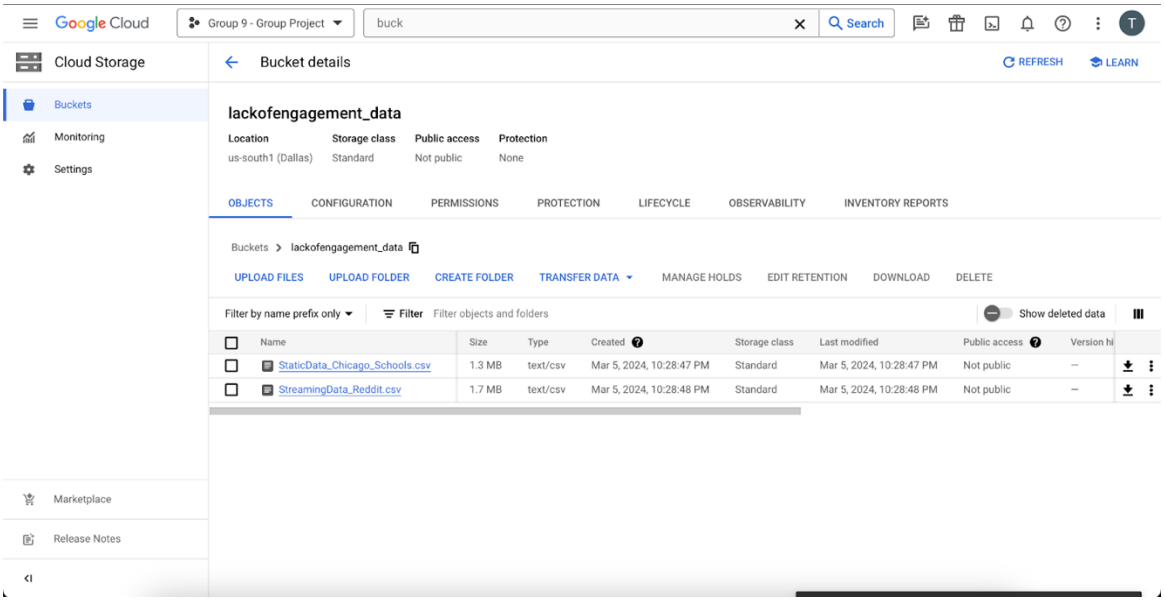
Our Data Storage

We've established a dedicated storage area to manage our data sources. Our efforts include acquiring and uploading two types of data: one static file containing information on Chicago schools and one streaming data source from Reddit. Attached are screenshots showcasing our storage location and the uploaded data.

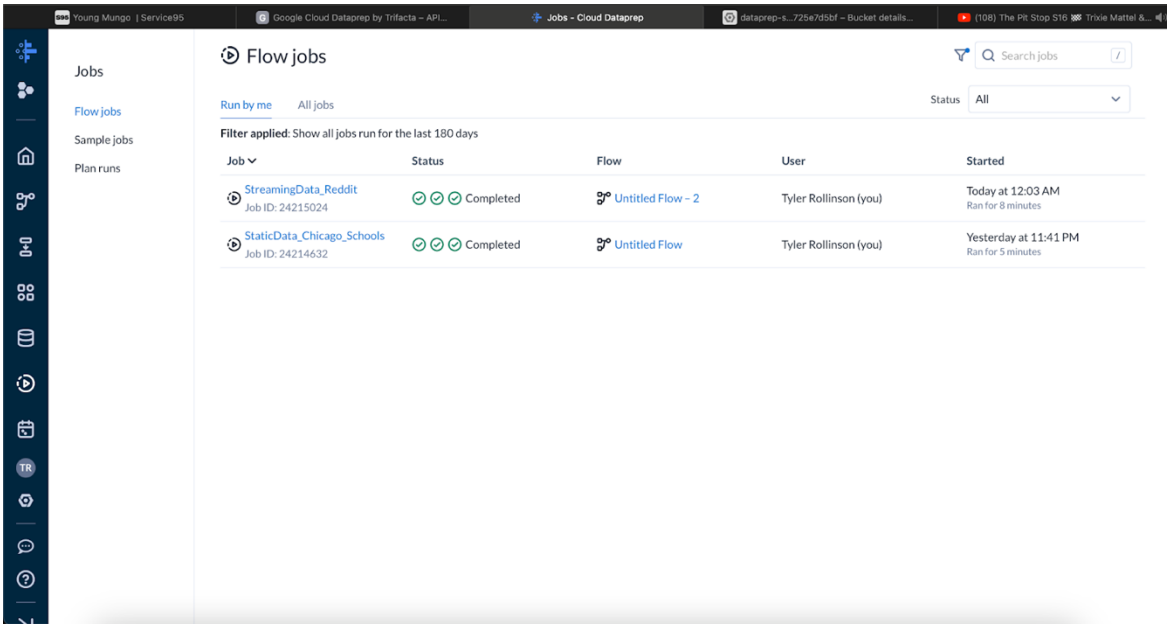
Screenshot 4: Dataset 1 in GCP Storage Bucket



Screenshot 5: Dataset 2



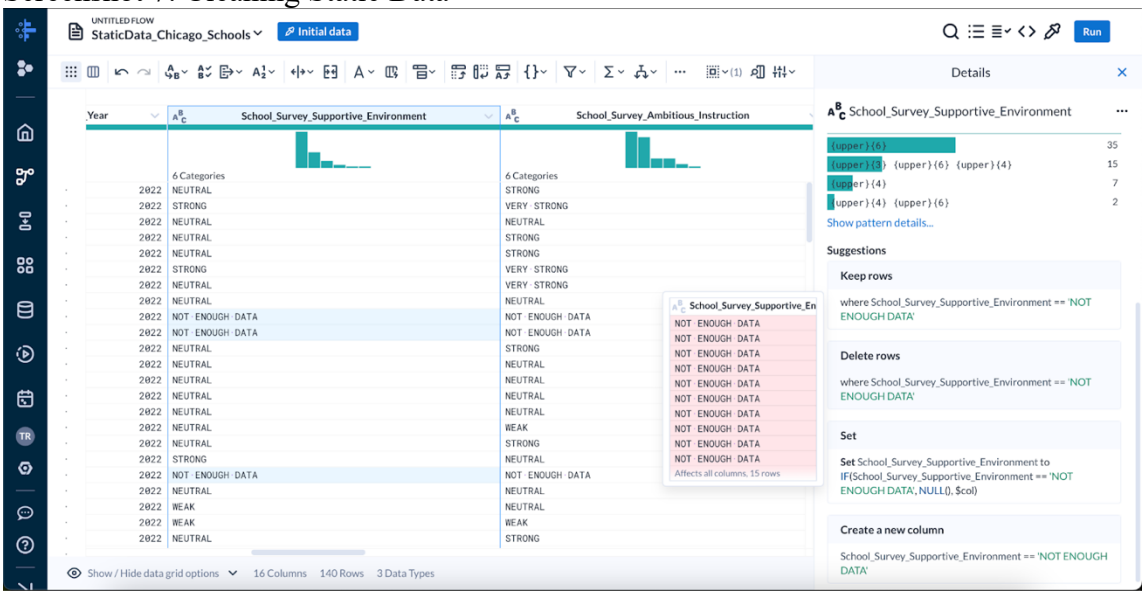
Screenshot 6: Folders (Data Cleaning and Pre processing)



Our Initial Tools

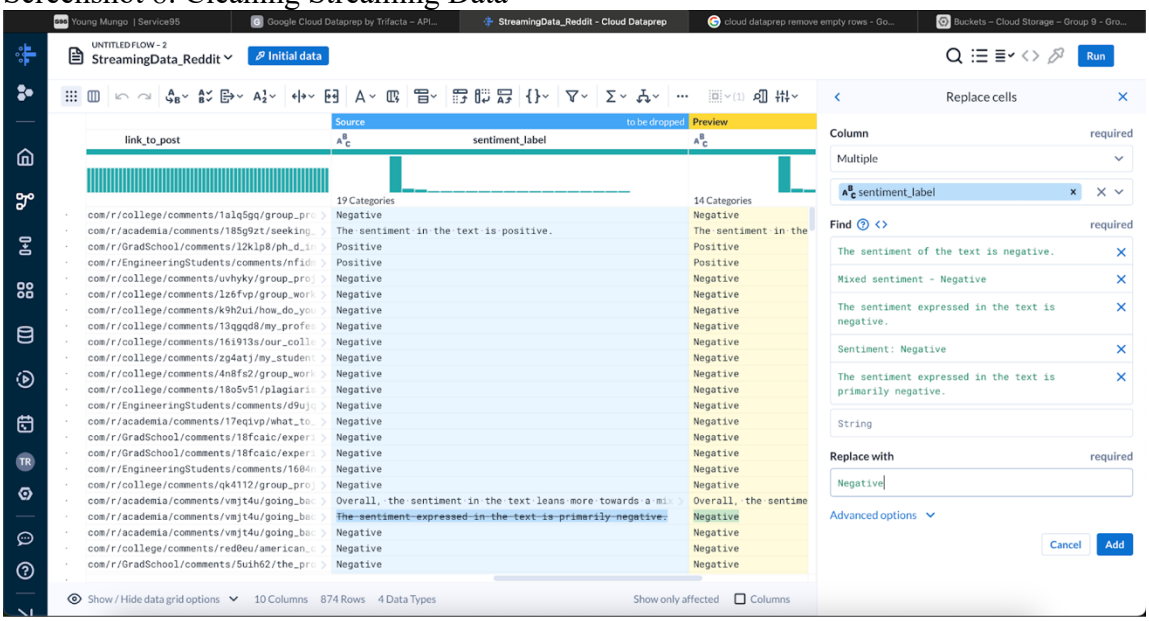
We employed GCP DataPrep by Trifacta to preprocess our data. The screenshots below illustrate the cleaning procedures applied to both the Chicago schools file and the Reddit file. Our initial data processing steps are detailed in the provided screenshots.

Screenshot 7: Cleaning Static Data



Removed rows with missing data for graduation and college enrollment rates.

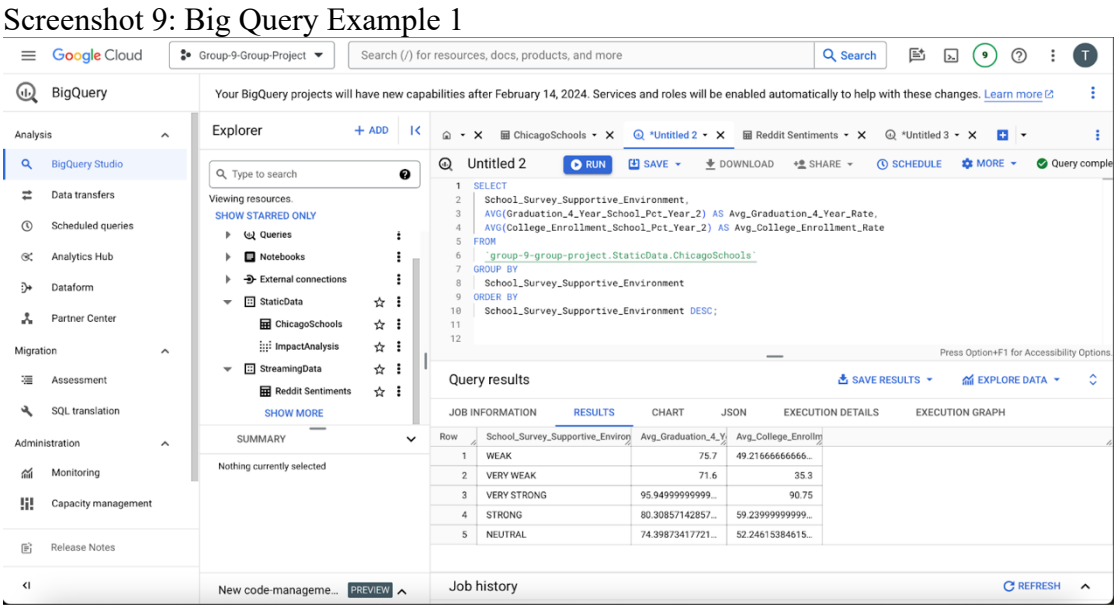
Screenshot 8: Cleaning Streaming Data



Standardized data field for the sentiment_label to three data points; Negative, Positive, and Neutral. Removed all rows of data that had missing data in the "text" field.

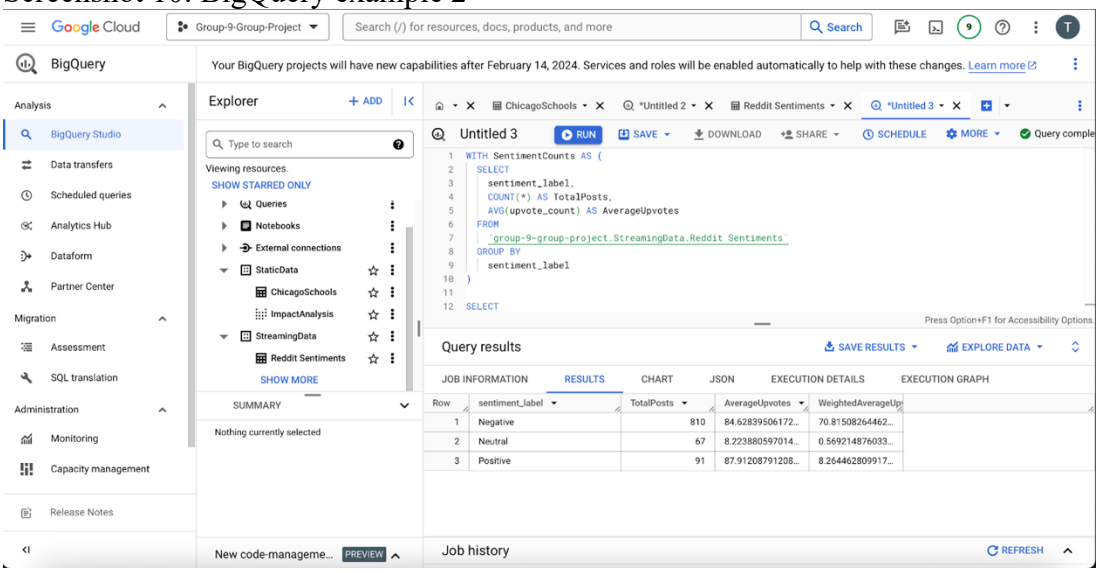
After completing the data cleaning process, we uploaded the files and executed two queries. Attached are screenshots depicting the queries we performed.

Screenshot 9: Big Query Example 1



We ran a query to see if there is any correlation between a schools "supportive enviroment" rank and the graduation or college enrollment rate. The data shows that there is a strong correlation between a higher (Very Strong) supportive ranking and a higher graduation and college enrollment rate.

Screenshot 10: BigQuery example 2



We ran a query on the data from Reddit to understand the popularity of sentiments towards group projects. We first retrieved a total count of all sentiment ratings in the dataset. We then calculated the average upvotes for each rating; Negative, Positive, Neutral. Given that the dataset was largely populated with “Negative” sentiments, we then weighed the average upvote to more fairly represent the popularity of opinion. The data shows that from our data set “Negative” sentiments are 8x’s more frequent than positive sentiments.

Screenshot 11: Hive Query

```
ssh.cloud.google.com
SSH-in-browser                                                    UPLOAD FILE   DOWNLOAD FILE
INFO  : Completed compiling command(queryId=hive_20240307210719_330d71b0-80f0-4405-a71e-73e2fa2b796c); Time taken: 0.206 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240307210719_330d71b0-80f0-4405-a71e-73e2fa2b796c): SELECT sentiment_label, COUNT(*) AS count
FROM sentiment_analysis
GROUP BY sentiment_label
WARN  : Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X rele
ases.
INFO  : Query ID = hive_20240307210719_330d71b0-80f0-4405-a71e-73e2fa2b796c
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Number of reduce tasks not specified. Estimated from input data size: 1
INFO  : In order to change the average load for a reducer (in bytes):
INFO  :   set hive.exec.reducers.bytes.per.reducer=<number>
INFO  : In order to limit the maximum number of reducers:
INFO  :   set hive.exec.reducers.max=<number>
INFO  : In order to set a constant number of reducers:
INFO  :   set mapreduce.job.reduces=<number>
INFO  : number of splits:1
INFO  : Submitting tokens for job: job_1709844499804_0004
INFO  : Executing with tokens: []
INFO  : The url to track the job: http://dp-hadoop-spark-2-cluster-group9-m:8088/proxy/application_1709844499804_0004/
INFO  : Starting Job = job_1709844499804_0004, Tracking URL = http://dp-hadoop-spark-2-cluster-group9-m:8088/proxy/application_1709844499804_0004/
INFO  : Kill Command = /usr/lib/hadoop/bin/mapred job  -kill job_1709844499804_0004
INFO  : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
INFO  : 2024-03-07 21:07:32,189 Stage-1 map = 0%,  reduce = 0%
INFO  : 2024-03-07 21:07:40,450 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.98 sec
INFO  : 2024-03-07 21:07:49,770 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.09 sec
INFO  : MapReduce Total cumulative CPU time: 8 seconds 90 msec
INFO  : Ended Job = job_1709844499804_0004
INFO  : MapReduce Jobs Launched:
INFO  : Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.09 sec   HDFS Read: 26382 HDFS Write: 189 SUCCESS
INFO  : Total MapReduce CPU Time Spent: 8 seconds 90 msec
INFO  : Completed executing command(queryId=hive_20240307210719_330d71b0-80f0-4405-a71e-73e2fa2b796c); Time taken: 31.698 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+------------------+--------+
| sentiment_label  | count  |
+------------------+--------+
| Negative         | 810    |
| Neutral          | 67     |
| Positive         | 91     |
| sentiment_label  | 1      |
+------------------+--------+
4 rows selected (31.959 seconds)
0: jdbc:hive2://localhost:10000>
```

Screenshot 12: Spark Query

```
ssh.cloud.google.com
SSH-in-browser                                                    UPLOAD FILE   DOWNLOAD FILE
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: Thu Mar  7 20:50:01 2024 from 35.235.244.33
tyler_rollinson@dp-hadoop-spark-2-cluster-group9-m:~$ spark-sql
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
ivysettings.xml file not found in HIVE_HOME or HIVE_CONF_DIR,/etc/hive/conf.dist/ivysettings.xml will be used
24/03/07 21:10:28 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
24/03/07 21:10:28 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
24/03/07 21:10:28 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
24/03/07 21:10:28 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
Spark master: yarn, Application Id: application_1709844499804_0005
spark-sql> show tables;
default reddit_sentiment        false
default reddit_sentiments       false
default sentiment_analysis      false
default sentiment_data  false
Time taken: 4.31 seconds, Fetched 4 row(s)
spark-sql> SELECT * FROM sentiment_analysis LIMIT 5;
24/03/07 21:11:18 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.
24/03/07 21:11:19 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)
NULL    sentiment_label
667     Negative
667     Negative
3       Negative
1       Negative
Time taken: 6.105 seconds, Fetched 5 row(s)
spark-sql> SELECT sentiment_label, COUNT(*) AS count
         > FROM sentiment_analysis
         > GROUP BY sentiment_label;
sentiment_label 1
Neutral 67
Positive        91
Negative        810
Time taken: 3.073 seconds, Fetched 4 row(s)
spark-sql>
```

Here is the time summary test comparison between hive and spark:

Hive Query Times:

Query 1: SELECT * FROM sentiment_analysis LIMIT 5; Time: 24.685 seconds

Query 2: SELECT sentiment_label, COUNT(*) AS count FROM sentiment_analysis GROUP BY sentiment_label; Time: 31.959 seconds

Spark SQL Query Times:

Query 1: SELECT * FROM sentiment_analysis LIMIT 5; Time: 6.105 seconds

Query 2: SELECT sentiment_label, COUNT(*) AS count FROM sentiment_analysis GROUP BY sentiment_label; Time: 3.073 seconds

Supporting Documents

Chicago Schools Survey

Dates: 03/01/2024
Website: https://catalog.data.gov/dataset/chicago-public-schools-school-progress-reports-sy2324
Quality: Data is in a consistent format but will employ a more abstract approach to our research as the data does not direclty reference student engagement in group projects. We found this to be difficult subject to research with limited data sources available to us. We will be comparing schools "supportive enviroment" ranking with the graduation rate to understand if there is any correlation between the two.

Reddit API "Group Projects"

Subreddits: r/college; r/education; r/StudentLife
Dates: 03/05/2024
Website: www.reddit.com/dev/api/
Quality: Data is more relavant to our research topic, however, there are numerous fields with missing data and inconsistent data format. Data requires more transformation to be useful in analysis.