

MODUL 2

GOOGLE BIGQUERY ANALYSIS

THEME DESCRIPTION

Students understand and are able to perform managed big data analysis to provide new insights for business/organization progress using Google BigQuery cloud platform

WEEKLY LEARNING OUTCOMES (SUB-LESSONS)

CLO-2-Sub-CLO-2 Understand and be able to represent the concept of Data Science as a whole and manage it effectively in the Data Analysis process, C2.

Through the processing steps consisting of:

1. Query and load data into a table
2. Top 5 baby names for US males in 2014
3. London bicycles Public Data Analysis
 - I. How many unique bikes are rented for each month in 2015, ordered by month?
 - II. Start from which station has the most rental orders on the week the station was installed?
 - III. How many minutes is usually needed for the bike_id which ends with 9 to start and end the trip at the same station?
 - IV. What is the 3rd busiest starting station in 2015 (rental amount) and the cumulative weekly rental amount in 2016?

PRACTICUM SUPPORTS

1. Windows Operating System
2. (any) Browser Application

PRACTICUM STEPS

- 1) **Query and export a public dataset**
 - a. BigQuery public datasets are displayed by default in the GCP Console
To open the public datasets project manually, enter the following URL in your browser.
 - b. Go to the BigQuery web UI in the [GCP Console](#)
 - c. At the top right of the window, click **Compose new query**. If this text is dimmed, then the **Query editor** is already open.
 - d. Copy and paste the following query into the query text area :

```
SELECT
  name, gender,
  SUM(number) AS total
FROM
  `bigquery-public-data.usa_names.usa_1910_2013`
GROUP BY
  name, gender
ORDER BY
  total DESC
LIMIT
  10
```

- e. Click Run., The query results page displays below the query window. At the top of the query results page, the time elapsed and the data processed by the query are displayed.
- f. Below the Query complete... message, a table displays the query results with a header row containing the name of each column you selected in the query.

Query results				SAVE RESULTS	EXPLORE IN DATA STUDIO
Query complete (1.1 sec elapsed, 99.9 MB processed)					
Job information Results JSON Execution details					
Row	name	gender	total		
1	James	M	4924235		
2	John	M	4818746		
3	Robert	M	4703680		
4	Michael	M	4280040		
5	William	M	3811998		
6	Mary	F	3728041		
7	David	M	3541625		
8	Richard	M	2526927		
9	Joseph	M	2467298		
10	Charles	M	2237170		

- g. Create a dataset in the web UI to store the data, In the navigation panel, in the **Resources** section, click your project name.
- h. On the right side, in the details panel, click **Create dataset**.
- i. In the navigation panel, in the Resources section, click your project name.
- j. On the right side, in the details panel, click Create dataset :
 - o For Dataset ID, enter babynames.
 - o For Data location, choose United States (US). Currently, the public datasets are stored in the US multi-region location.
 - o For simplicity, you should place your dataset in the same location.

bigquery-umn-bda122

Run
Save query
Save view
Schedule query

bigquery-umn-bda12

CREATE DATASET

Create dataset

Dataset ID
babynames

Data location (Optional)
United States (US)

Default table expiration
☒ Never
☐ Number of days:

- k. Leave all of the other default settings in place and click **Create dataset**.

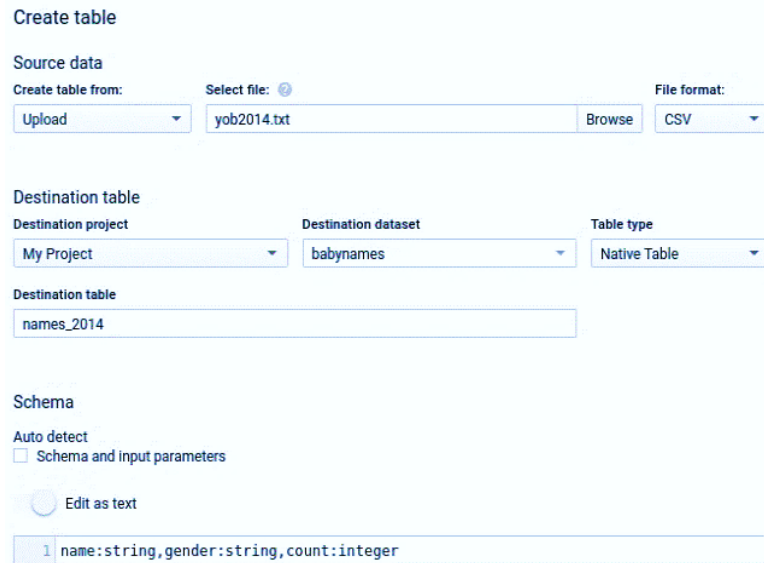
Load the data into a new table

- l. In the navigation panel, in the Resources section, click the babynames dataset you have created.
- m. On the right side, in the details panel, click Create table.
- n. Use the default values for all settings unless otherwise indicated.
- o. On the Create table page:
 - o For Source, click Empty table and choose Upload.
 - o For Select file, click Browse, navigate to the IS429-Lab-Week#2-yob2014.txt file (downloaded from your UMN e-learning) and click Open.

- For File format, click Avro and choose CSV.
- For Destination, enter babynames_2014.
- In the Schema section, click the Edit as text toggle and paste the following schema definition in the box: *name:string,gender:string,count:integer*

1. Click **Create table**.

p. Wait for BigQuery to create the table and load the data. While BigQuery loads the data, a **(1 running)** string displays beside the job history in the navigation panel. The string disappears after the data is loaded.



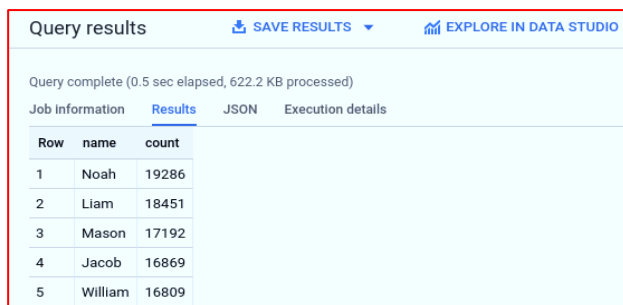
Preview the table

- q. After the **(1 running)** string disappears, you can access the table.
- r. To preview the first few rows of the data:
- s. Select babynames > names_2014 in the navigation panel.
- t. In the details panel, click the Preview tab as a result shown on Figure 2.1

2) Top 5 baby names for US males in 2014

- ▶ Now that you've loaded data into a table, you can query it.
- ▶ The process is identical to the previous example, except that this time, you're querying your table instead of a public table.
- a. If necessary, click the Compose new query button. Unless you hid the query window previously, it should still be visible.
- b. Copy and paste the following query into the query text area. This query retrieves the top 5 baby names for US males in 2014.
- c. Click Run. The results are displayed right-side the query window.

```
SELECT
  name,
  count
FROM
  `babynames.names_2014`
WHERE
  gender = 'M'
ORDER BY
  count DESC
LIMIT
  5
```



Row	name	count
1	Noah	19286
2	Liam	18451
3	Mason	17192
4	Jacob	16869
5	William	16809

- d. Save your query to be able to run again later, screenshot of your job should be as shown on Figure 2.2

3) Public Data Analysis

Utilise BigQuery to learn to do data analysis using London bicycles public data

- a. How many unique bikes are rented for each month in 2015, ordered by month?

To perform this analysis, type in your query as follows:

Unsaved query Edited

```

1 SELECT
2   year_month,
3   count_bike
4 FROM(
5   SELECT
6     CONCAT(FORMAT_TIMESTAMP('%Y-%m',start_date)) as year_month,
7     COUNT(DISTINCT bike_id) as count_bike
8   FROM
9     `bigquery-public-data.london_bicycles.cycle_hire`
10  GROUP BY 1
11  ORDER BY 1) AS Y
12 WHERE year_month LIKE '2015%'

```

The result of your query will be shown as Figure 2.3, save as your job result in screenshot.

- b. Start from which station has the most rental orders on the week the station was installed?

To perform this analysis, type in your query as follows:

```

1 SELECT
2   b.year,
3   b.week_number,
4   b.start_station_name,
5   b.count_bike
6 FROM (
7   SELECT
8     FORMAT_TIMESTAMP('%Y',end_date) AS year,
9     EXTRACT(WEEK FROM DATE(end_date)) AS week_number,
10    start_station_name,
11    COUNT(DISTINCT bike_id) AS count_bike
12  FROM
13    `bigquery-public-data.london_bicycles.cycle_hire`
14  GROUP BY
15    1, 2, 3) AS b
16 WHERE
17   count_bike = ( SELECT MAX(count_bike)FROM ( SELECT FORMAT_TIMESTAMP('%Y',end_date) AS year,
18     EXTRACT(WEEK FROM DATE(end_date)) AS week_number,
19     start_station_name,
20     COUNT(DISTINCT bike_id) AS count_bike
21  FROM
22    `bigquery-public-data.london_bicycles.cycle_hire`
23  GROUP BY
24    1, 2, 3))

```

Valid: This query will process 1.02 GB when run.

The result of your query will be shown as Figure 2.4, save as your job result in screenshot.

- c. How many minutes is usually needed for the bike_id which ends with 9 to start and end the trip at the same station? To perform this analysis, type in your query as follows :

```

1 SELECT
2   rental_id,
3   bike_id,
4   duration/60 AS duration_minute,
5   start_station_name,
6   end_station_name
7 FROM
8   `bigquery-public-data.london_bicycles.cycle_hire`
9 WHERE
10  bike_id = 9
11  AND end_station_name = start_station_name

```

The result of your query will be shown as Figure 2.5, save as your job result in screenshot.

- d. What is the 3rd busiest starting station in 2015 (rental amount) and the cumulative weekly rental amount in 2016 ? To perform this analysis, type in your query as follows :

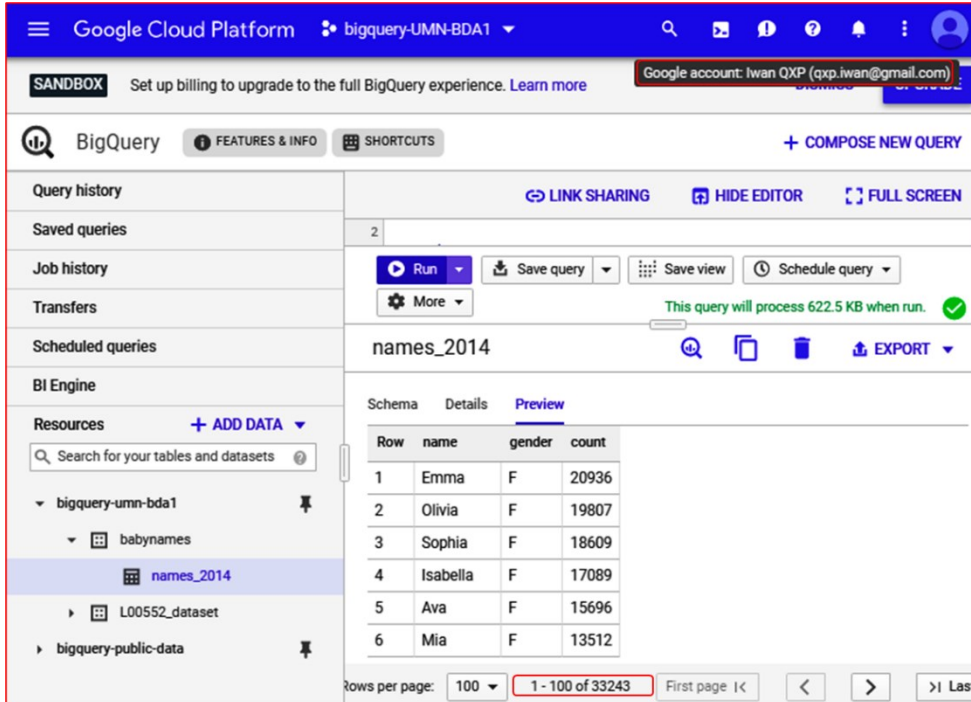
```

New Query ?
1 SELECT
2   id,
3   week_number,
4   rental_count,
5   SUM(rental_count) OVER(ORDER BY week_number) cumulative_sum
6 FROM (
7   SELECT
8     id,
9     EXTRACT(week
10    FROM
11      ch.start_date) week_number,
12     COUNT(rental_id) rental_count
13   FROM (
14     SELECT
15       cs.id,
16       COUNT(ch.rental_id) rental_count,
17       RANK() OVER(ORDER BY COUNT(ch.rental_id) DESC) rank
18     FROM
19       `bigquery-public-data.london_bicycles.cycle_stations` cs
20     JOIN
21       `bigquery-public-data.london_bicycles.cycle_hire` ch
22     ON
23       cs.id = ch.start_station_id
24   WHERE
25     start_date BETWEEN '2015-01-01 00:00:00' AND '2015-12-31 23:59:59'
26   GROUP BY
27     1) t1
28   JOIN
29     `bigquery-public-data.london_bicycles.cycle_hire` ch
30   ON
31     ch.start_station_id=t1.id
32   AND ch.start_date BETWEEN '2016-01-01 00:00:00' AND '2016-12-31 23:59:59'
33   WHERE
34     rank=3
35   GROUP BY
36     1,2) t2
  
```

Standard SQL Dialect X

- ❖ Save all your screenshot(s) result as Figure 2.1 until 2.6 into 2.Tugas IS-429 Lab BDA-1 GBQ ANALYSIS yourname-NIM.doc and submit to e-Learning IS-429 Lab Week#2.

OUTPUT/ RESULTS

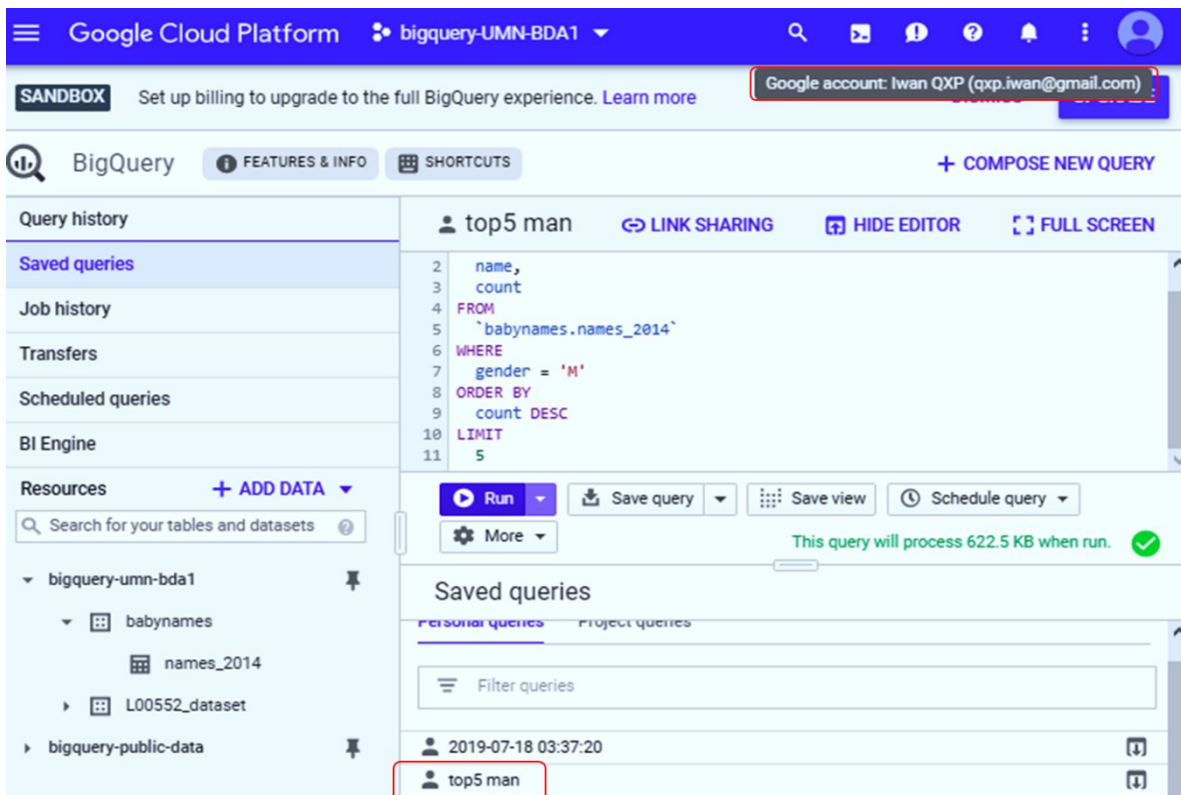


The screenshot shows the Google Cloud Platform BigQuery interface. The left sidebar contains a 'Resources' section with a search bar and a list of datasets. The 'babynames' dataset is expanded, showing 'names_2014' selected. The main area displays the 'names_2014' dataset with a 'Preview' tab active. The preview shows a table with columns 'Row', 'name', 'gender', and 'count'. The first six rows are visible, showing names like Emma, Olivia, Sophia, Isabella, Ava, and Mia. The bottom of the preview indicates '1 - 100 of 33243' rows.

Row	name	gender	count
1	Emma	F	20936
2	Olivia	F	19807
3	Sophia	F	18609
4	Isabella	F	17089
5	Ava	F	15696
6	Mia	F	13512

1

Figure 2.1 Load the data into a new table



The screenshot shows the Google Cloud Platform BigQuery interface. The left sidebar is the same as in Figure 2.1. The main area displays a query editor with a query titled 'top5 man'. The query is a SQL statement that filters for male names ('gender = 'M'') and orders them by count in descending order, limiting the results to 5. The query is saved as a query named 'top5 man' in the 'Saved queries' section. The 'top5 man' query is highlighted in the 'Saved queries' list.

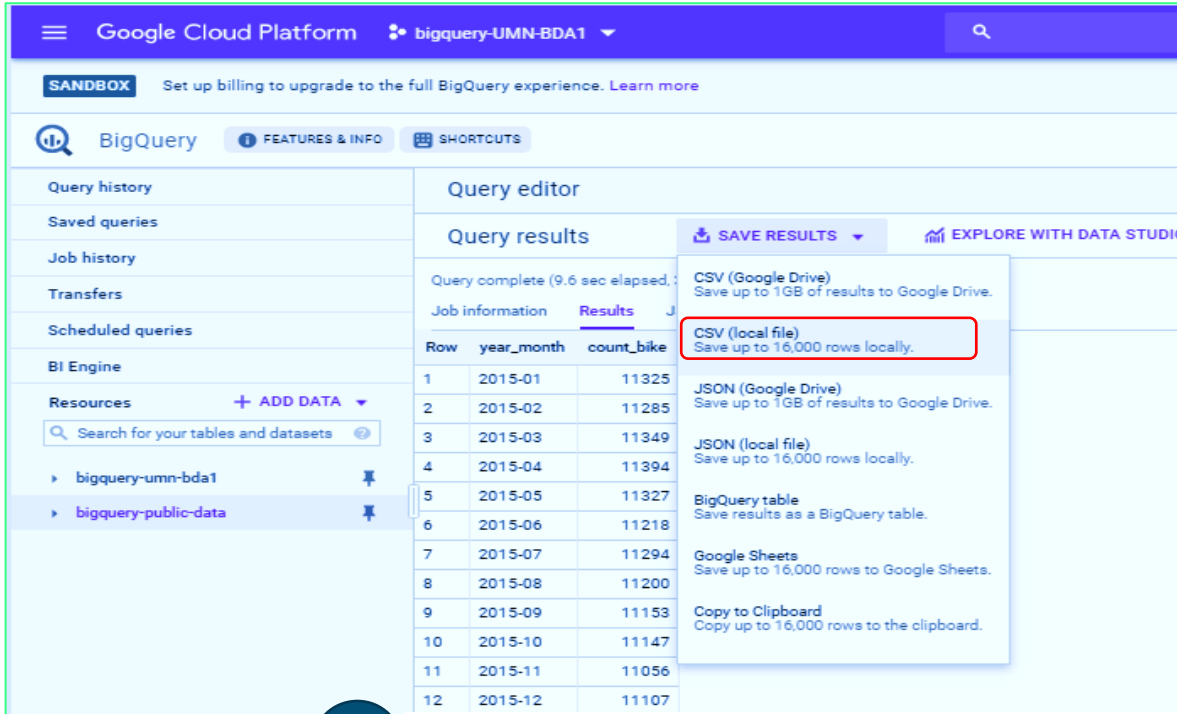
```

2  name,
3  count
4  FROM
5  `babynames.names_2014`
6  WHERE
7  gender = 'M'
8  ORDER BY
9  count DESC
10 LIMIT
11  5

```

2

Figure 2.2 Top 5 Man Baby



Google Cloud Platform bigquery-UMN-BDA1

SANDBOX Set up billing to upgrade to the full BigQuery experience. [Learn more](#)

BigQuery FEATURES & INFO SHORTCUTS

Query history
Saved queries
Job history
Transfers
Scheduled queries
BI Engine
Resources + ADD DATA

Search for your tables and datasets

bigquery-umn-bda1
bigquery-public-data

Query editor

Query results

Query complete (9.6 sec elapsed)

Job information Results

Row year_month count_bike

1 2015-01 11325
2 2015-02 11285
3 2015-03 11349
4 2015-04 11394
5 2015-05 11327
6 2015-06 11218
7 2015-07 11294
8 2015-08 11200
9 2015-09 11153
10 2015-10 11147
11 2015-11 11056
12 2015-12 11107

SAVE RESULTS EXPLORE WITH DATA STUDIO

CSV (Google Drive)
Save up to 1GB of results to Google Drive.

CSV (local file)
Save up to 16,000 rows locally.

JSON (Google Drive)
Save up to 1GB of results to Google Drive.

JSON (local file)
Save up to 16,000 rows locally.

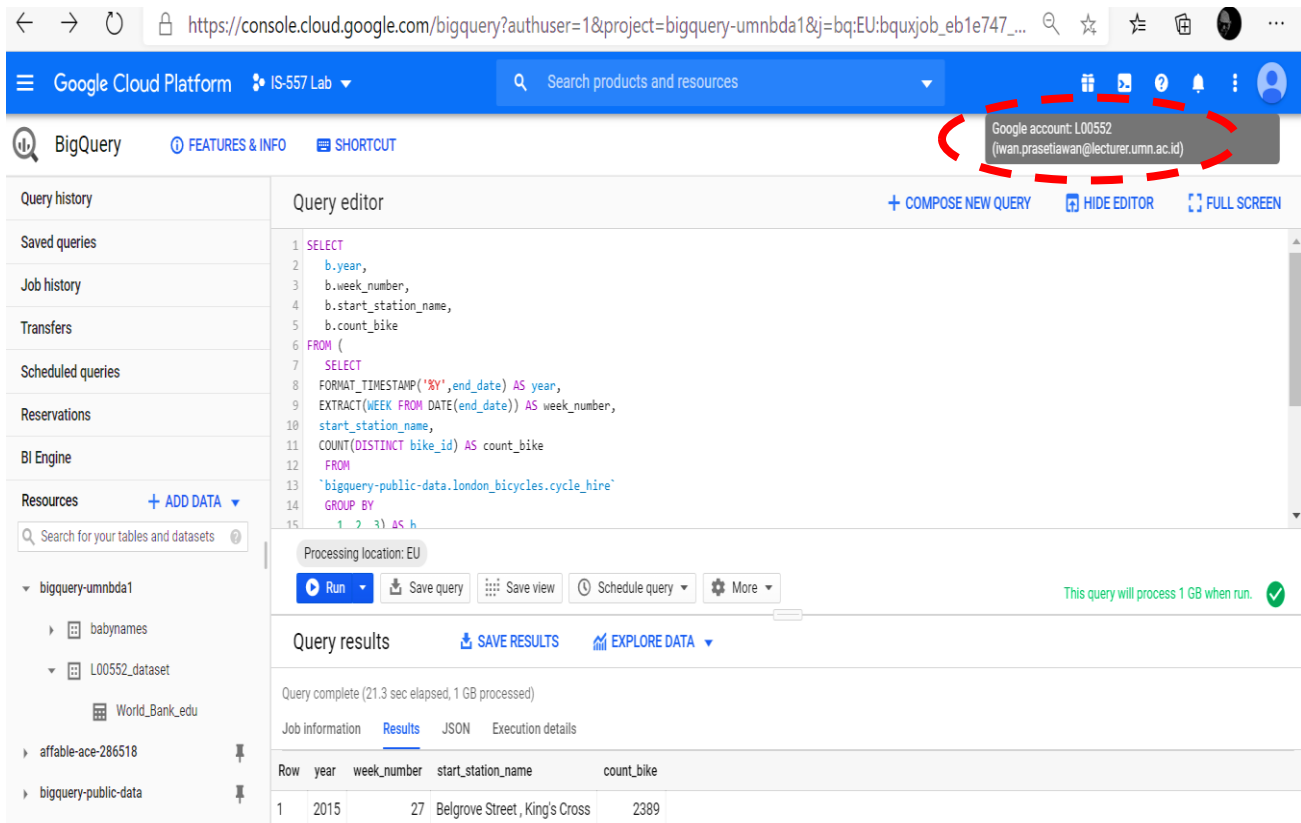
BigQuery table
Save results as a BigQuery table.

Google Sheets
Save up to 16,000 rows to Google Sheets.

Copy to Clipboard
Copy up to 16,000 rows to the clipboard.

3

Figure 2.3 Orders by Month in 2015



Google Cloud Platform IS-557 Lab

Search products and resources

BigQuery FEATURES & INFO SHORTCUT

Query history
Saved queries
Job history
Transfers
Scheduled queries
Reservations
BI Engine
Resources + ADD DATA

Search for your tables and datasets

bigquery-umn-bda1
babynames
L00552_dataset
World_Bank_edu
affable-ace-286518
bigquery-public-data

Query editor

+ COMPOSE NEW QUERY HIDE EDITOR FULL SCREEN

```
1 SELECT
2   b.year,
3   b.week_number,
4   b.start_station_name,
5   b.count_bike
6 FROM (
7   SELECT
8     FORMAT_TIMESTAMP('%Y', end_date) AS year,
9     EXTRACT(WEEK FROM DATE(end_date)) AS week_number,
10    start_station_name,
11    COUNT(DISTINCT bike_id) AS count_bike
12 FROM
13   `bigquery-public-data.london_bicycles.cycle_hire`
14 GROUP BY
15   1, 2, 3) AS h
```

Processing location: EU

Run Save query Save view Schedule query More

This query will process 1 GB when run.

Query results

SAVE RESULTS EXPLORE DATA

Query complete (21.3 sec elapsed, 1 GB processed)

Job information Results JSON Execution details

Row year week_number start_station_name count_bike

1 2015 27 Belgrove Street, King's Cross 2389

4

Figure 2.4 On the Week Station was Installed

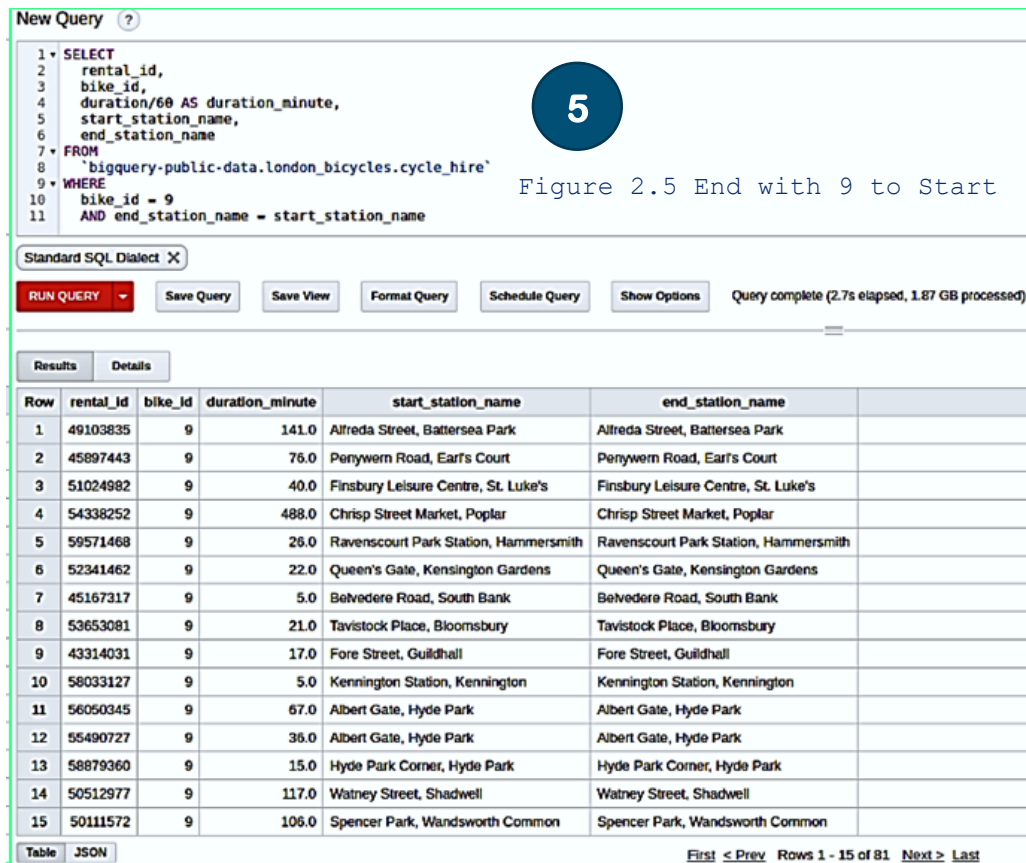


Figure 2.5 End with 9 to Start

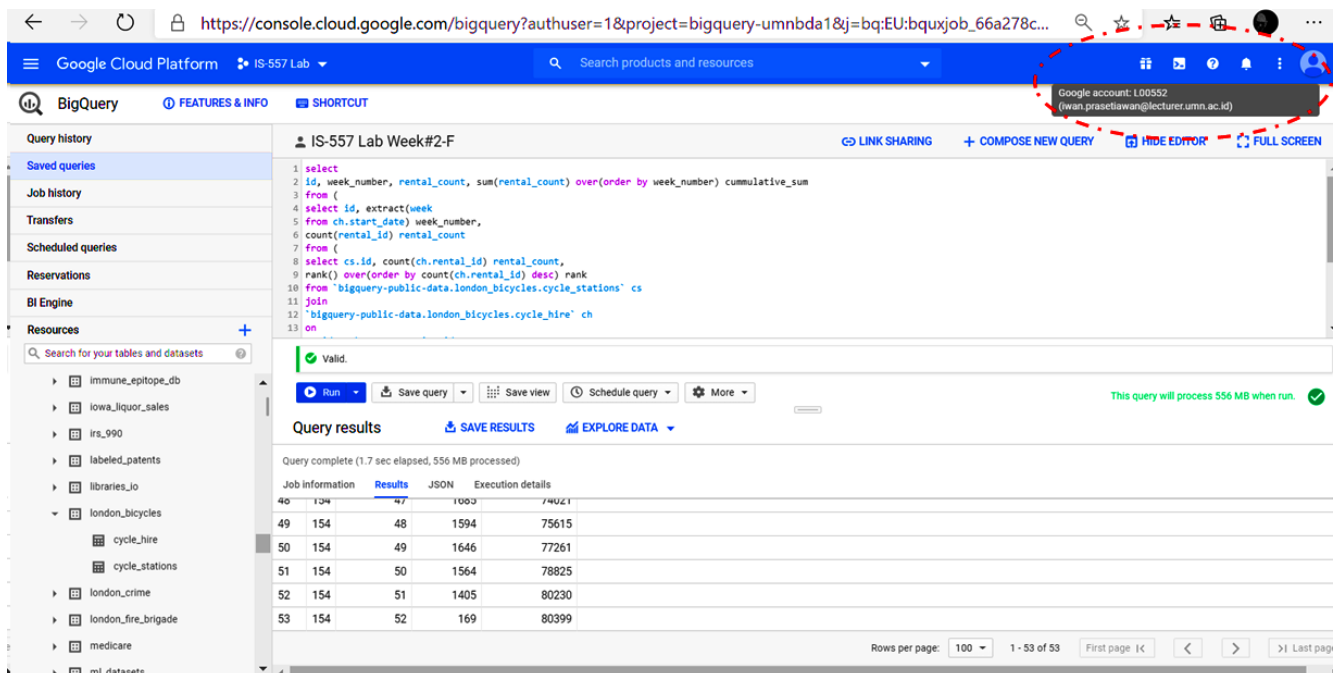


Figure 2.6 3rd Busiest Station 2015-2016

REFERENCE

1. [Documentation | Google Cloud](#)
2. [Cloud Google | BigQuery](#)
3. Other additional references are excerpts from various Online Learning/websites.

The End