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

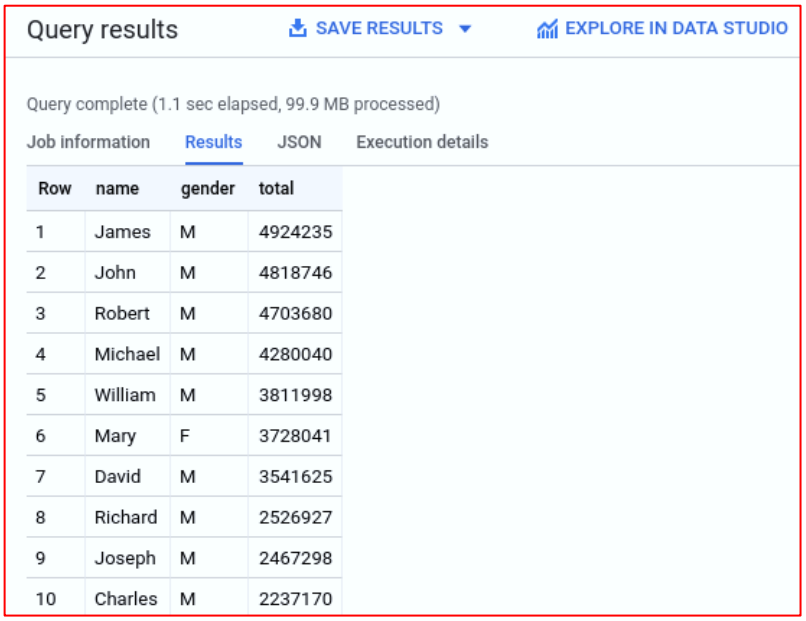
1) Query and export a public dataset

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.



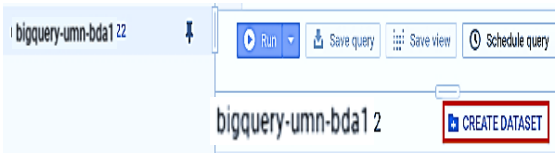
Row	name	gender	total
1	James	M	4924235
2	John	M	4818746
3	Robert	M	4703680
4	Michael	M	4280040
5	William	M	3811998
6	Mary	F	3728041
7	David	M	3541625
8	Richard	M	2526927
9	Joseph	M	2467298
10	Charles	M	2237170

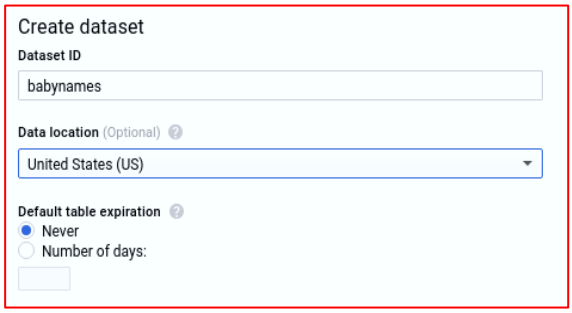
g. Create a dataset in the web UI to store the data, In the navigation panel, in the **Resources** section, click your project name.

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 :
   - For Dataset ID, enter babynames.
   - For Data location, choose United States (US). Currently,
   - the public datasets are stored in the US multi-region location.
   - For simplicity, you should place your dataset in the same location.





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:
   - For Source, click Empty table and choose Upload.
   - 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
```

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

```
SELECT
year_month,
count_bike
FROM(
  SELECT
  CONCAT(FORMAT_TIMESTAMP('%Y-%m',start_date)) as year_month,
  COUNT(DISTINCT bike_id) as count_bike
  FROM
  `bigquery-public-data.london_bicycles.cycle_hire`
  GROUP BY 1
  ORDER BY 1) AS Y
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:

```
SELECT
  b.year,
  b.week_number,
  b.start_station_name,
  b.count_bike
FROM (
  SELECT
    FORMAT_TIMESTAMP('%Y',end_date) AS year,
    EXTRACT(WEEK FROM DATE(end_date)) AS week_number,
    start_station_name,
    COUNT(DISTINCT bike_id) AS count_bike
  FROM
    `bigquery-public-data.london_bicycles.cycle_hire`
  GROUP BY
    1, 2, 3) AS b
WHERE
      count_bike = ( SELECT MAX(count_bike)FROM ( SELECT FORMAT_TIMESTAMP('%Y',end_date) AS year,
      EXTRACT(WEEK FROM DATE(end_date)) AS week_number,
      start_station_name,
      COUNT(DISTINCT bike_id) AS count_bike
    FROM
      `bigquery-public-data.london_bicycles.cycle_hire`
    GROUP BY
      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 :

```
SELECT
  rental_id,
  bike_id,
  duration/60 AS duration_minute,
  start_station_name,
  end_station_name
FROM
  `bigquery-public-data.london_bicycles.cycle_hire`
WHERE
  bike_id = 9
  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 :

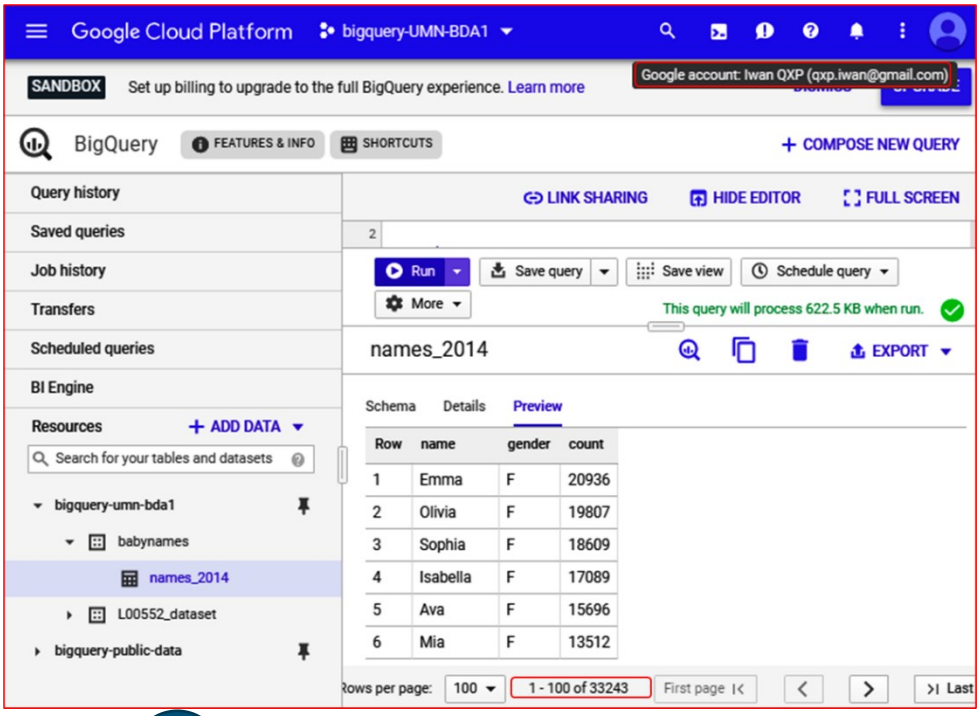
New Query

```
SELECT
  id,
  week_number,
  rental_count,
  SUM(rental_count) OVER(ORDER BY week_number)cumulative_sum
FROM (
  SELECT
    id,
    EXTRACT(week
    FROM
      ch.start_date)week_number,
    COUNT(rental_id)rental_count
  FROM (
    SELECT
      cs.id,
      COUNT(ch.rental_id)rental_count,
      RANK() OVER(ORDER BY COUNT(ch.rental_id) DESC)rank
    FROM
      `bigquery-public-data.london_bicycles.cycle_stations` cs
    JOIN
      `bigquery-public-data.london_bicycles.cycle_hire` ch
    ON
      cs.id = ch.start_station_id
    WHERE
      start_date BETWEEN '2015-01-01 00:00:00' AND '2015-12-31 23:59:59'
    GROUP BY
      1)t1
  JOIN
    `bigquery-public-data.london_bicycles.cycle_hire` ch
  ON
    ch.start_station_id=t1.id
    AND ch.start_date BETWEEN '2016-01-01 00:00:00' AND '2016-12-31 23:59:59'
  WHERE
    rank=3
  GROUP BY
    1,2)t2
```

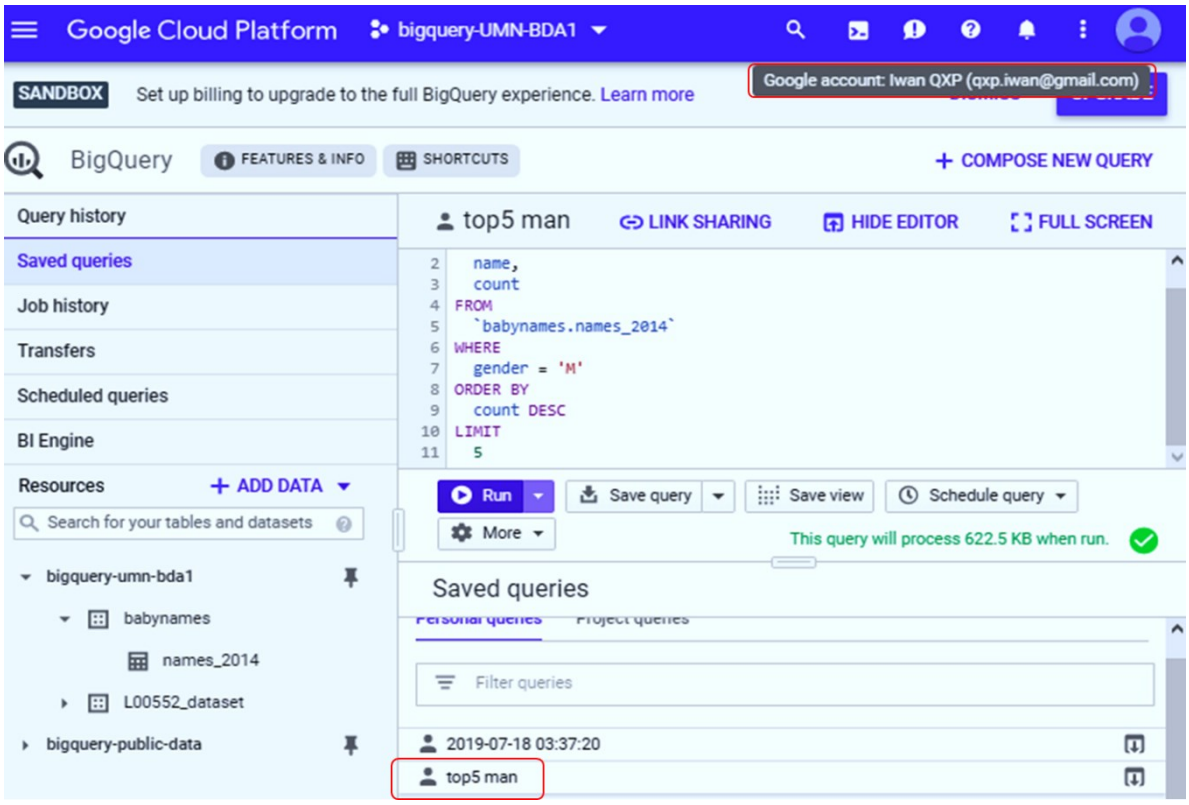
Standard SQL Dialect

- 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.

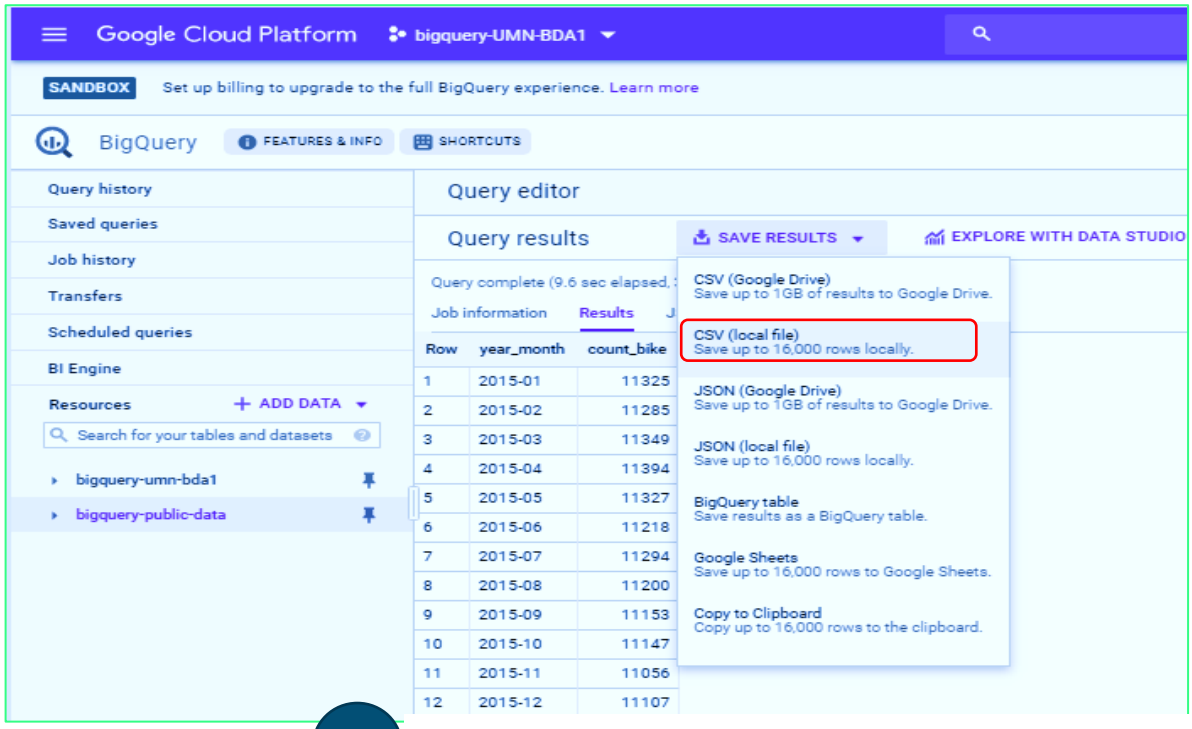
OUTPUT/ RESULTS



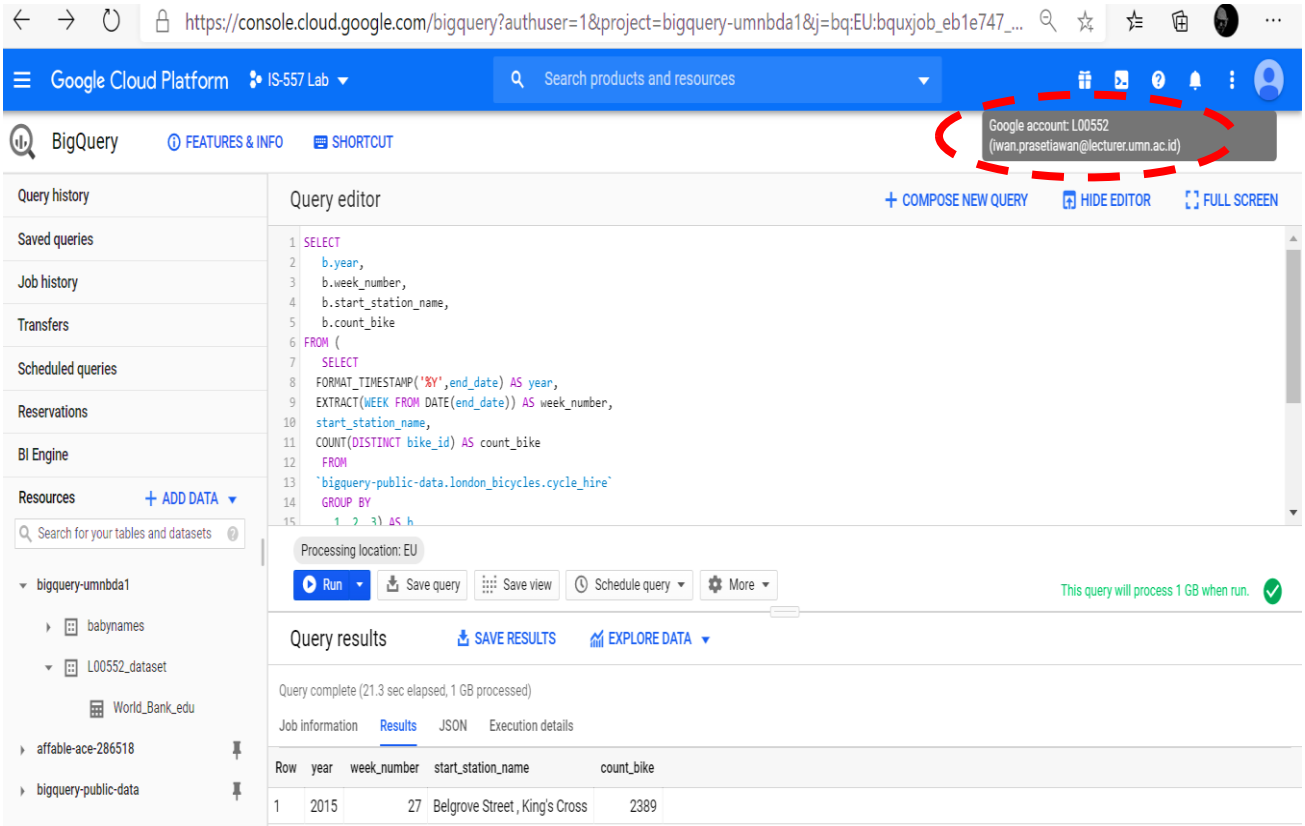
1 Figure 2.1 Load the data into a new table



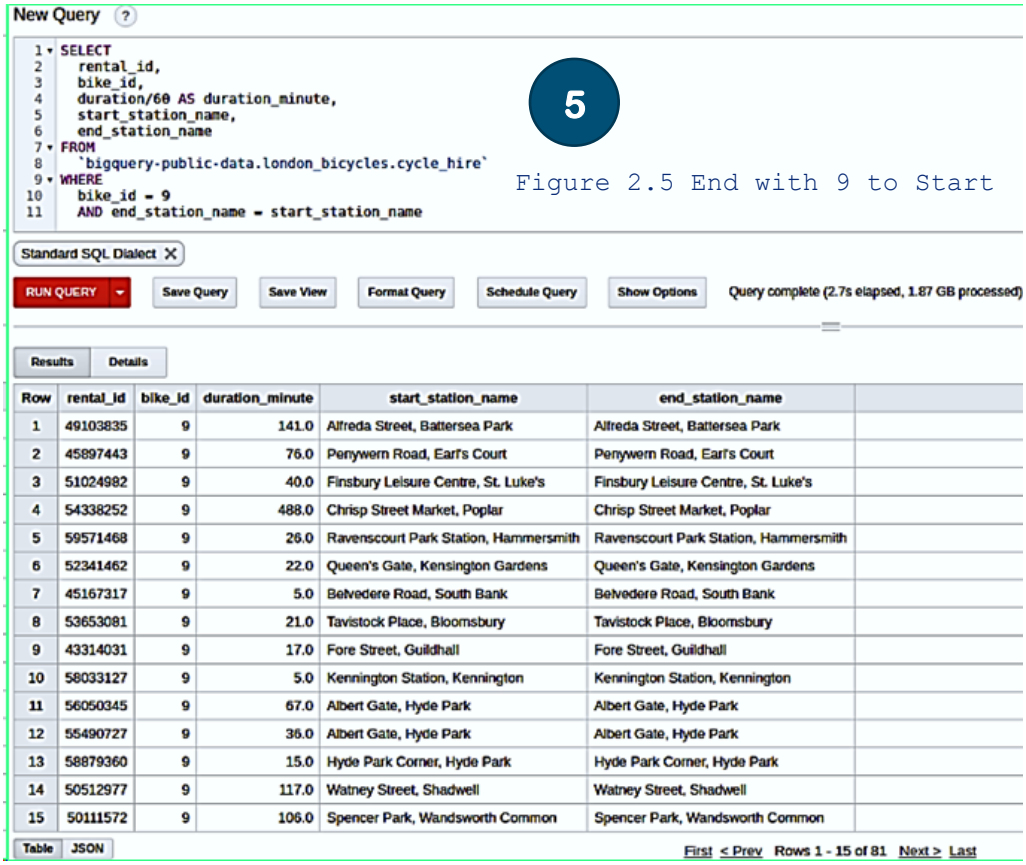
2 Figure 2.2 Top 5 Man Baby



3 Figure 2.3 Orders by Month in 2015



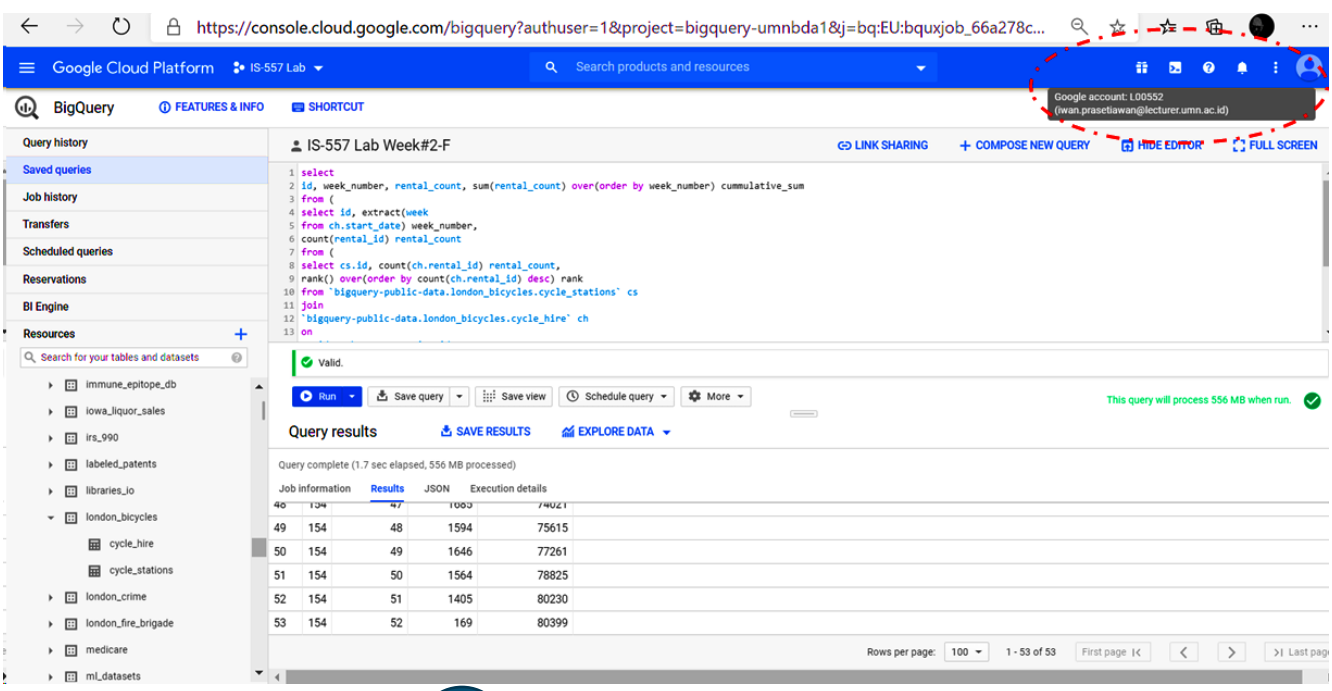
4 Figure 2.4 On the Week Station was Installed



Row	rental_id	bike_id	duration_minute	start_station_name	end_station_name
1	49103835	9	141.0	Alfreda Street, Battersea Park	Alfreda Street, Battersea Park
2	45897443	9	76.0	Penywern Road, Earl's Court	Penywern Road, Earl's Court
3	51024982	9	40.0	Finsbury Leisure Centre, St. Luke's	Finsbury Leisure Centre, St. Luke's
4	54338252	9	488.0	Chrisp Street Market, Poplar	Chrisp Street Market, Poplar
5	59571468	9	26.0	Ravenscourt Park Station, Hammersmith	Ravenscourt Park Station, Hammersmith
6	52341462	9	22.0	Queen's Gate, Kensington Gardens	Queen's Gate, Kensington Gardens
7	45167317	9	5.0	Belvedere Road, South Bank	Belvedere Road, South Bank
8	53653081	9	21.0	Tavistock Place, Bloomsbury	Tavistock Place, Bloomsbury
9	43314031	9	17.0	Fore Street, Guildhall	Fore Street, Guildhall
10	58033127	9	5.0	Kennington Station, Kennington	Kennington Station, Kennington
11	56050345	9	67.0	Albert Gate, Hyde Park	Albert Gate, Hyde Park
12	55490727	9	36.0	Albert Gate, Hyde Park	Albert Gate, Hyde Park
13	58879360	9	15.0	Hyde Park Corner, Hyde Park	Hyde Park Corner, Hyde Park
14	50512977	9	117.0	Watney Street, Shadwell	Watney Street, Shadwell
15	50111572	9	106.0	Spencer Park, Wandsworth Common	Spencer Park, Wandsworth Common

5

Figure 2.5 End with 9 to Start



6

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.