Creating Date-Partitioned Tables in BigQuery

Task 1. Create a new dataset

1. First, you will create a dataset to store your tables.
2. Click the three dots next to your Qwiklabs project ID and select **Create dataset**:
3. Name your dataset **ecommerce**.
4. Click **Create dataset**.

Task 2. Creating tables with date partitions

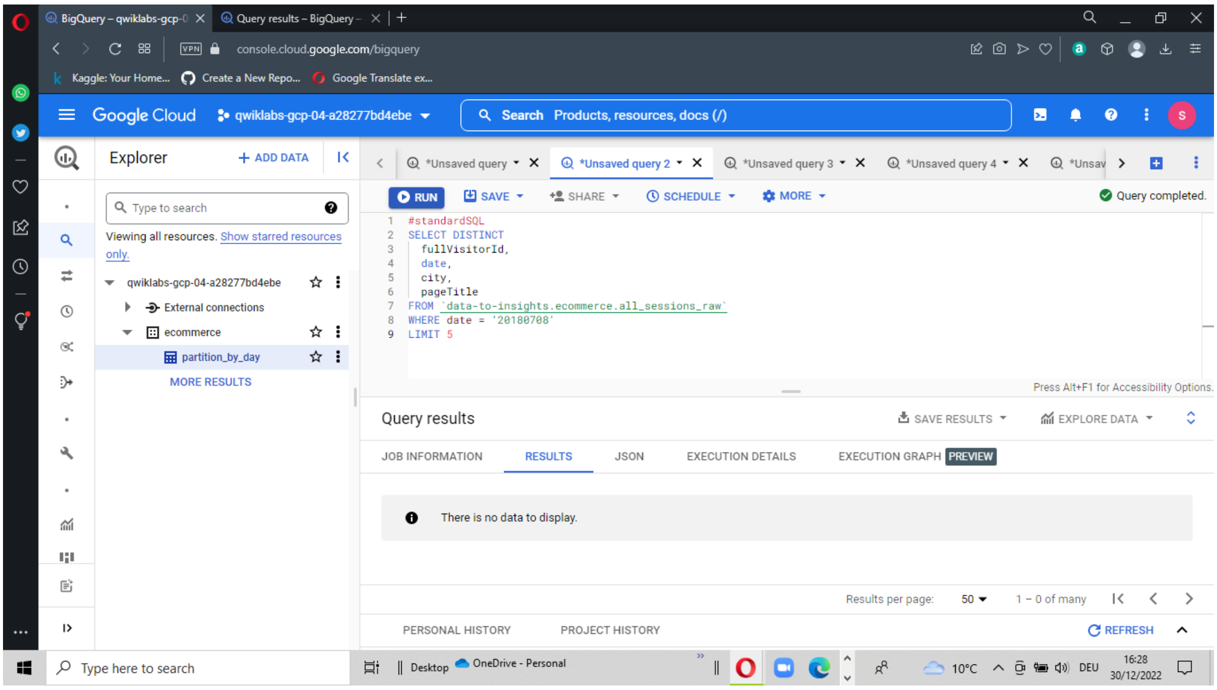
A partitioned table is a table that is divided into segments, called partitions, that make it easier to manage and query your data. By dividing a large table into smaller partitions, you can improve query performance, and control costs by reducing the number of bytes read by a query.

Now create a new table and bind a date or timestamp column as a partition. Before we do that, let's explore the data in the non-partitioned table first.

Query webpage analytics for a sample of visitors in 2017

1. In the **Query Editor**, add the below query:

Solution:



```sql
#standardSQL
SELECT DISTINCT
  fullVisitorId,
  date,
  city,
  pageTitle
FROM `data-to-insights.ecommerce.all_sessions_raw`
WHERE date = '20180708'
LIMIT 5
```

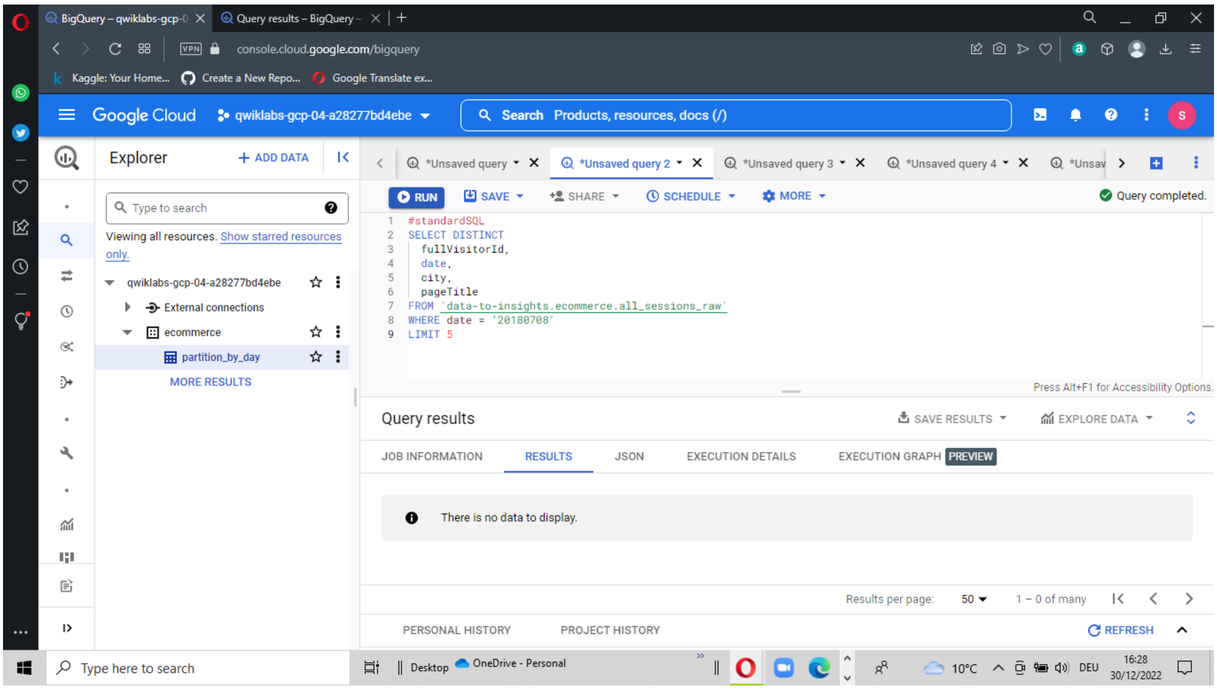
There is no data to display.

Query webpage analytics for a sample of visitors in 2018

Let's modify the query to look at visitors for 2018 now.

1. Click **COMPOSE NEW QUERY** to clear the **Query Editor**, then add this new query. Note the WHERE date parameter is changed to 20180708:

solution:



```
#standardSQL
SELECT DISTINCT
  fullVisitorId,
  date,
  city,
  pageTitle
FROM `data-to-insights.ecommerce.all_sessions_raw`
WHERE date = '20180708'
LIMIT 5
```

Query completed.

There is no data to display.

Common use-cases for date-partitioned tables

Scanning through the entire dataset everytime to compare rows against a WHERE condition is wasteful. This is especially true if you only really care about records for a specific period of time like:

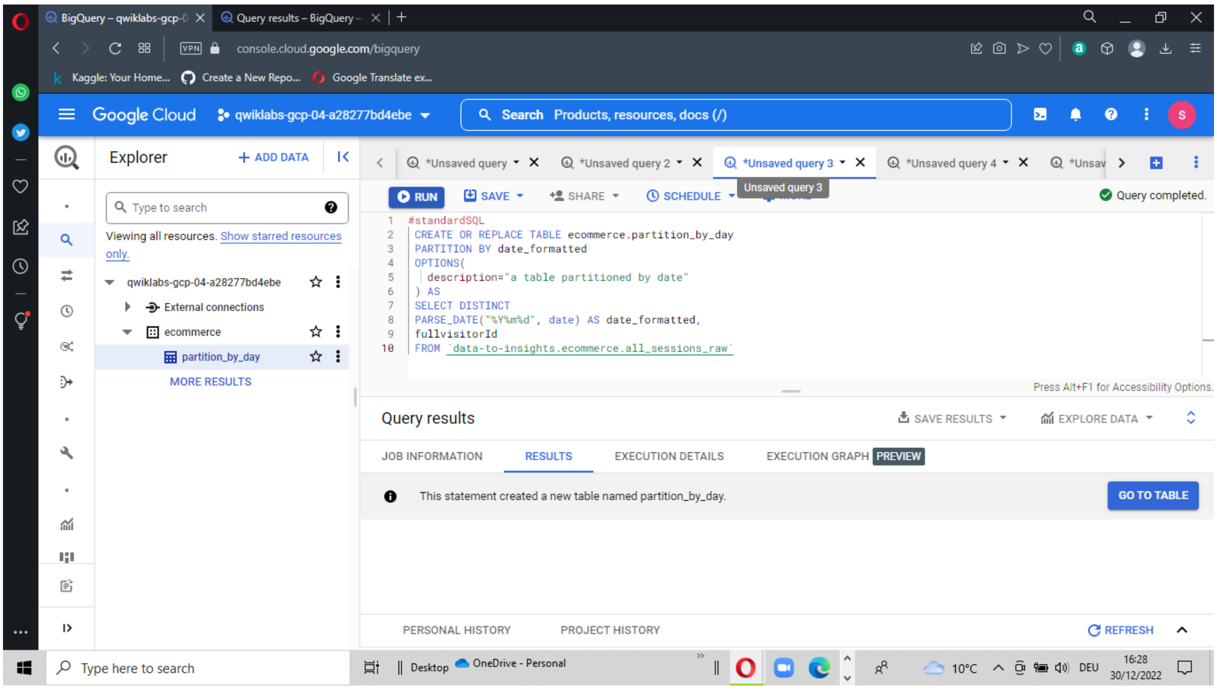
- All transactions for the last year
- All visitor interactions within the last 7 days
- All products sold in the last month

Instead of scanning the entire dataset and filtering on a date field like we did in the earlier queries, Now set up a date-partitioned table. This allows you to completely ignore scanning records in certain partitions if they are irrelevant to our query.

Create a new partitioned table based on date

1. Click **COMPOSE NEW QUERY** , add the query, then click **Run**:

solution:



```
#standardSQL
CREATE OR REPLACE TABLE ecommerce.partition_by_day
PARTITION BY date_formatted
OPTIONS(
  description="a table partitioned by date"
) AS
SELECT DISTINCT
PARSE_DATE("%Y%m%d", date) AS date_formatted,
fullvisitorId
FROM `data-to-insights.ecommerce.all_sessions_raw`
```

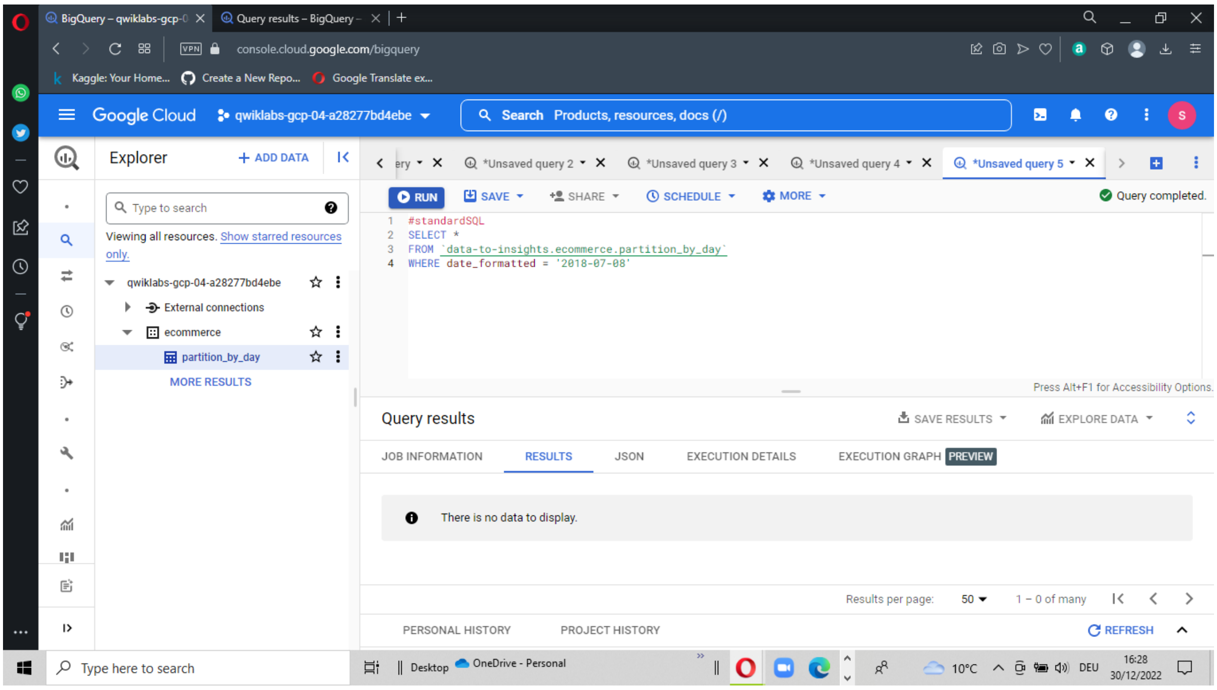
Query completed.

This statement created a new table named partition_by_day.

In the above query, note the new option - PARTITION BY a field. The two options available to partition are DATE and TIMESTAMP. The PARSE_DATE function is used on the date field (stored as a string) to get it into the proper DATE type for partitioning.

Task 3. View data processed with a partitioned table

1. Run the query, and note the total bytes to be processed:



```
#standardSQL
SELECT *
FROM `data-to-insights.ecommerce.partition_by_day`
WHERE date_formatted = '2018-07-08'
```

Query completed.

There is no data to display.

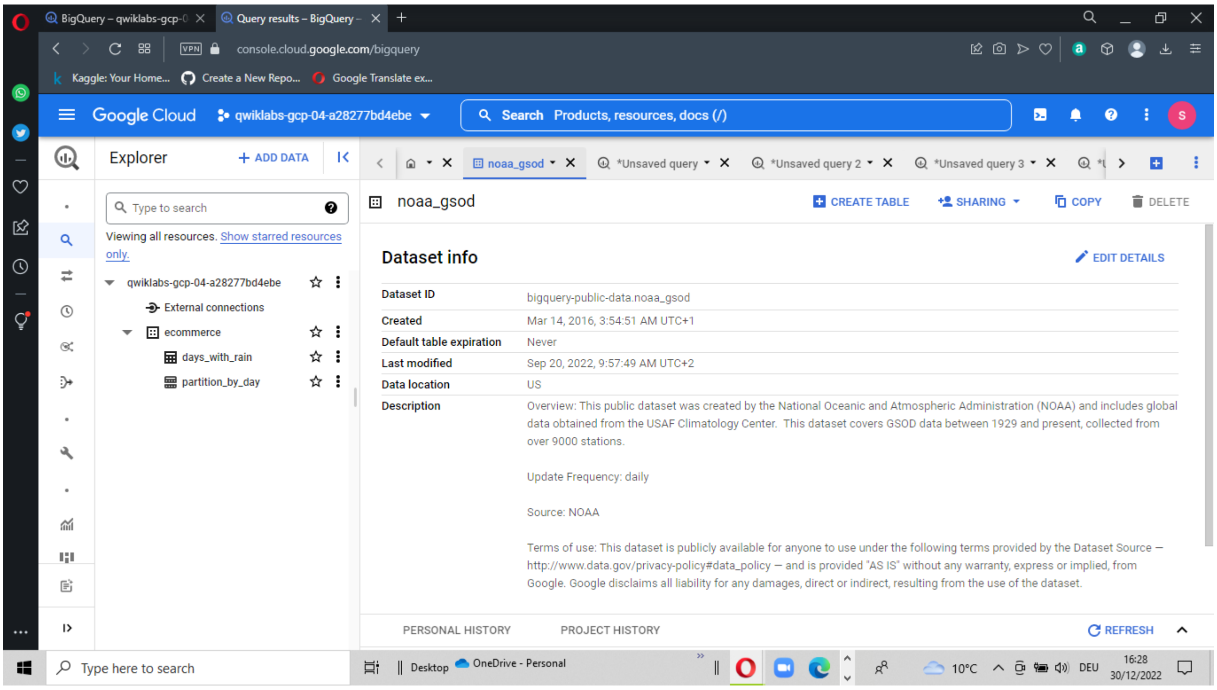
Task 4. Creating an auto-expiring partitioned table

Auto-expiring partitioned tables are used to comply with data privacy statutes and can be used to avoid unnecessary storage (which you'll be charged for in a production environment). If you want to create a rolling window of data, add an expiration date so the partition disappears after you're finished using it.

Explore the available NOAA weather data tables

1. In the left menu, in Explorer, click on **Add Data** and select **Explore public datasets**.

solution:



noaa_gsod

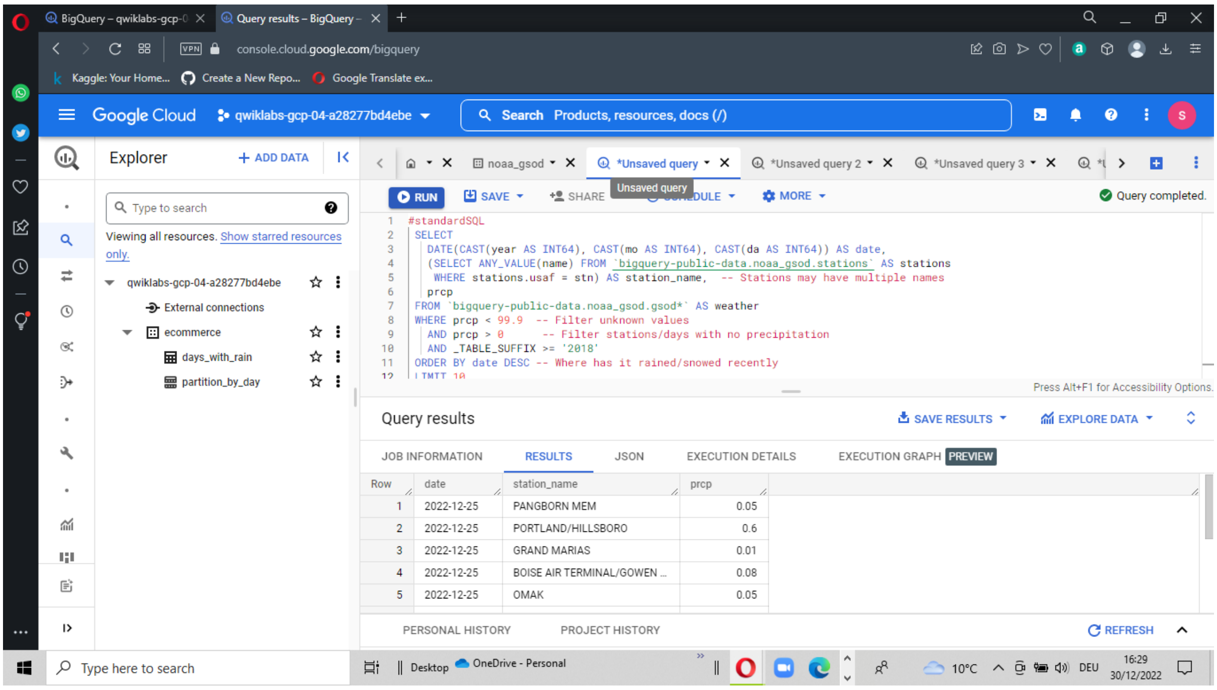
Dataset info

Dataset ID	bigquery-public-data.noaa_gsod
Created	Mar 14, 2016, 3:54:51 AM UTC+1
Default table expiration	Never
Last modified	Sep 20, 2022, 9:57:49 AM UTC+2
Data location	US
Description	Overview: This public dataset was created by the National Oceanic and Atmospheric Administration (NOAA) and includes global data obtained from the USAF Climatology Center. This dataset covers GSOD data between 1929 and present, collected from over 9000 stations. Update Frequency: daily Source: NOAA Terms of use: This dataset is publicly available for anyone to use under the following terms provided by the Dataset Source — http://www.data.gov/privacy-policy#data_policy — and is provided "AS IS" without any warranty, express or implied, from Google. Google disclaims all liability for any damages, direct or indirect, resulting from the use of the dataset.

Your goal is to create a table that:

- Queries on weather data from 2018 onward
- Filters to only include days that have had some precipitation (rain, snow, etc.)
- Only stores each partition of data for 90 days from that partition's date (rolling window)

solution:



```sql
#standardSQL
 SELECT
   DATE(CAST(year AS INT64), CAST(mo AS INT64), CAST(da AS INT64)) AS date,
   (SELECT ANY_VALUE(name) FROM `bigquery-public-data.noaa_gsod.stations` AS stations
    WHERE stations.usaf = stn) AS station_name,  -- Stations may have multiple names
   prcp
 FROM `bigquery-public-data.noaa_gsod.gsod*` AS weather
 WHERE prcp < 99.9  -- Filter unknown values
   AND prcp > 0      -- Filter stations/days with no precipitation
   AND _TABLE_SUFFIX >= '2018'
 ORDER BY date DESC -- Where has it rained/snowed recently
 LIMIT 10
```

Query results

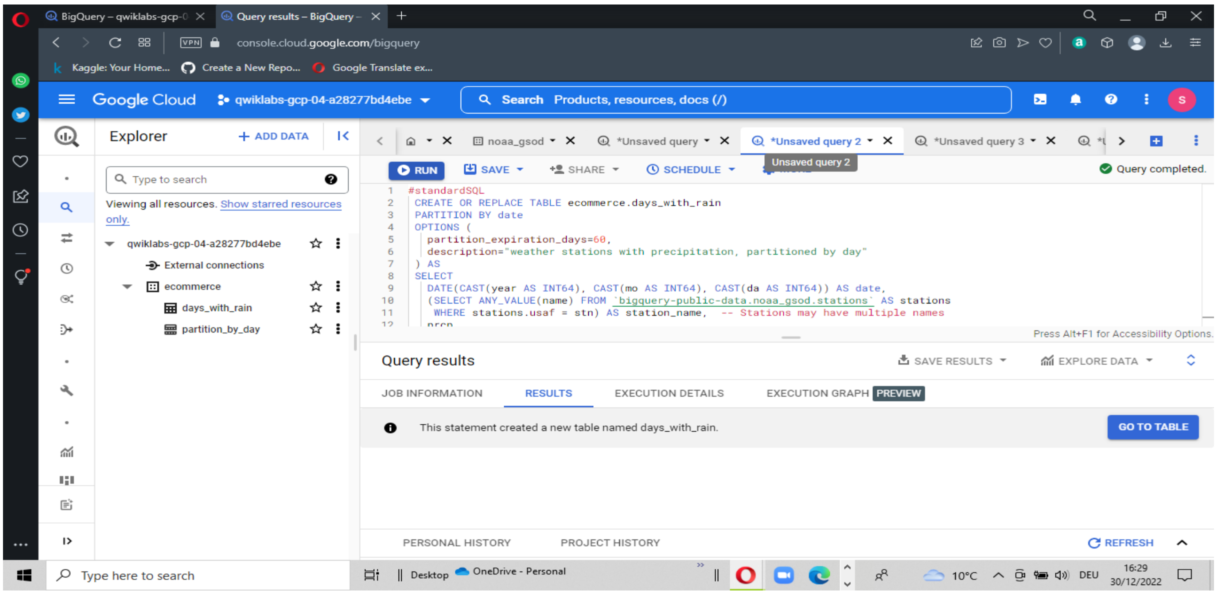
Row	date	station_name	prcp
1	2022-12-25	PANGBORN MEM	0.05
2	2022-12-25	PORTLAND/HILLSBORO	0.6
3	2022-12-25	GRAND MARIAS	0.01
4	2022-12-25	BOISE AIR TERMINAL/GOWEN ...	0.08
5	2022-12-25	OMAK	0.05

NB: The table wildcard * used in the FROM clause to limit the amount of tables referred to in the *TABLE_SUFFIX* filter.

Task 5. create a partitioned table

- Modify the previous query to create a table with the below specifications:
    - Table name: ecommerce.days_with_rain
    - Use the date field as your PARTITION BY
    - For OPTIONS, specify partition_expiration_days = 60
    - Add the table description = "weather stations with precipitation, partitioned by day".

solution:



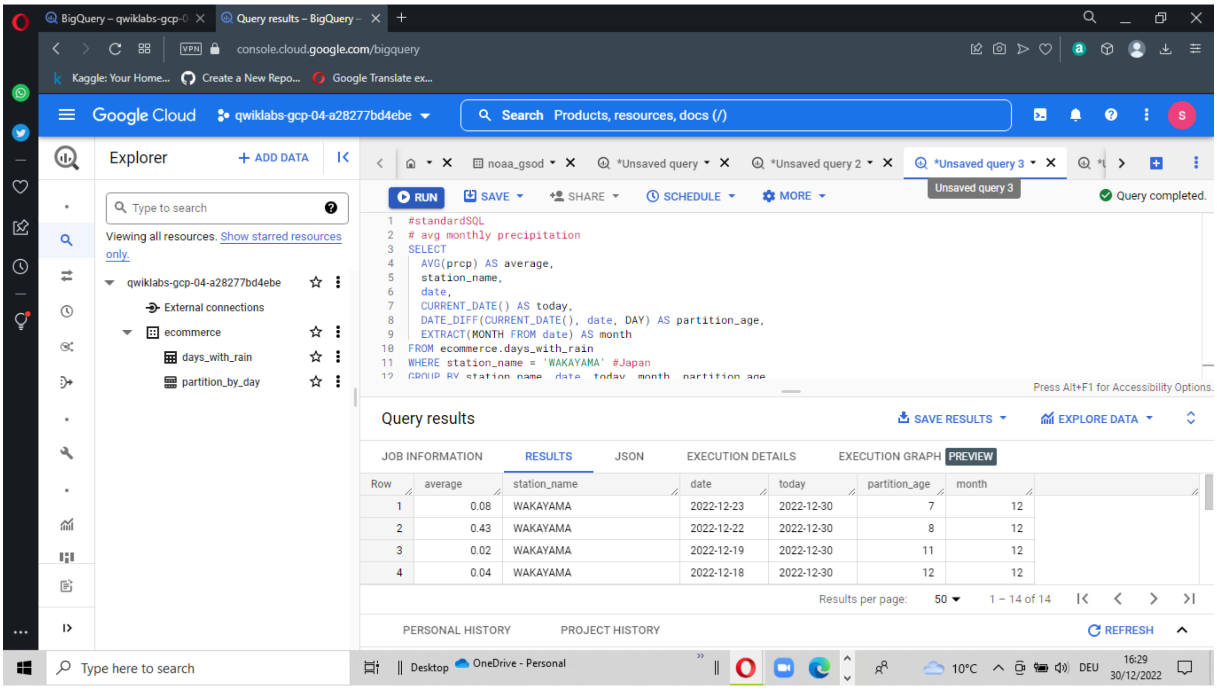
```sql
#standardSQL
CREATE OR REPLACE TABLE ecommerce.days_with_rain
PARTITION BY date
OPTIONS (
  partition_expiration_days=60,
  description="weather stations with precipitation, partitioned by day"
) AS
SELECT
  DATE(CAST(year AS INT64), CAST(mo AS INT64), CAST(da AS INT64)) AS date,
  (SELECT ANY_VALUE(name) FROM `bigquery-public-data.noaa_gsod.stations` AS stations
   WHERE stations.usaf = stn) AS station_name,  -- Stations may have multiple names
  prcp
```

Query completed.

This statement created a new table named days_with_rain.

Below is a query which tracks the average rainfall for the NOAA weather station in [Wakayama, Japan] which has significant precipitation.

solution:



```sql
#standardSQL
# avg monthly precipitation
SELECT
  AVG(prcp) AS average,
  station_name,
  date,
  CURRENT_DATE() AS today,
  DATE_DIFF(CURRENT_DATE(), date, DAY) AS partition_age,
  EXTRACT(MONTH FROM date) AS month
FROM ecommerce.days_with_rain
WHERE station_name = 'WAKAYAMA' #Japan
GROUP BY station_name, date, today, month, partition_age
```

Query completed.

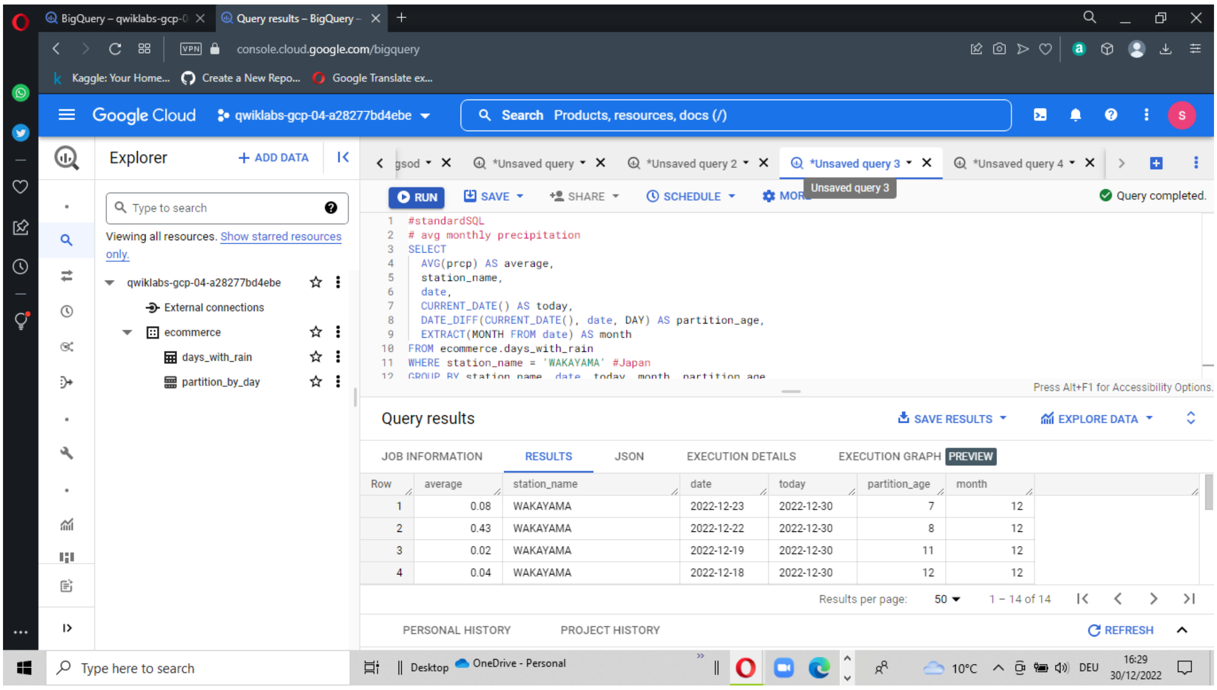
Row	average	station_name	date	today	partition_age	month
1	0.08	WAKAYAMA	2022-12-23	2022-12-30	7	12
2	0.43	WAKAYAMA	2022-12-22	2022-12-30	8	12
3	0.02	WAKAYAMA	2022-12-19	2022-12-30	11	12
4	0.04	WAKAYAMA	2022-12-18	2022-12-30	12	12

Results per page: 50 1 – 14 of 14

Task 6. Confirm the oldest partition_age is at or below 60 days

Update the ORDER BY clause to show the oldest partitions first

Solution:



```
#standardSQL
# avg monthly precipitation
SELECT
  AVG(prcp) AS average,
  station_name,
  date,
  CURRENT_DATE() AS today,
  DATE_DIFF(CURRENT_DATE(), date, DAY) AS partition_age,
  EXTRACT(MONTH FROM date) AS month
FROM ecommerce.days_with_rain
WHERE station_name = 'WAKAYAMA' #Japan
GROUP BY station_name, date, today, month, partition_age
```

Query completed.

Query results

Row	average	station_name	date	today	partition_age	month
1	0.08	WAKAYAMA	2022-12-23	2022-12-30	7	12
2	0.43	WAKAYAMA	2022-12-22	2022-12-30	8	12
3	0.02	WAKAYAMA	2022-12-19	2022-12-30	11	12
4	0.04	WAKAYAMA	2022-12-18	2022-12-30	12	12

Results per page: 50 1 – 14 of 14