Analyzing Weather Data in BigQuery

**Overview**

In this lab you will analyze historical weather observations using BigQuery and use weather data in conjunction with other datasets.

**What you'll learn**

In this lab, you will:

* Carry out interactive queries on the BigQuery console.
* Combine and run analytics on multiple datasets.

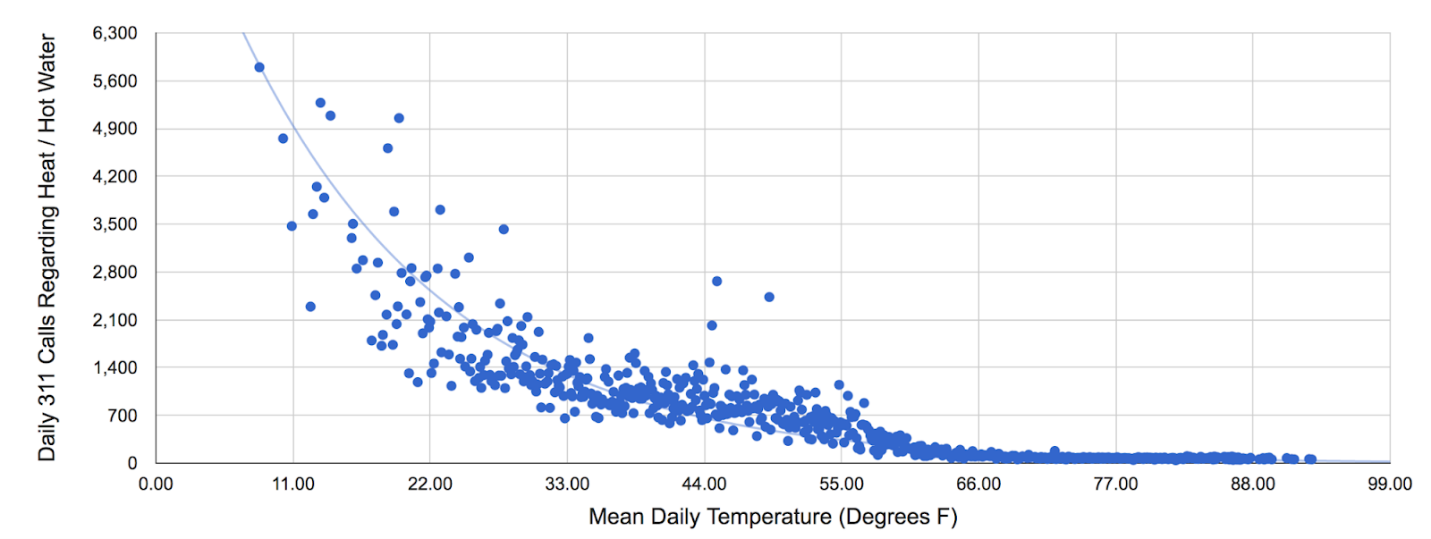
**Introduction**

This lab uses two public datasets in BigQuery: weather data from NOAA and citizen complaints data from New York City.

You will encounter, for the first time, several aspects of Google Cloud that are of great benefit to scientists:

1. **Serverless.** No need to download data to your machine in order to work with it - the dataset will remain on the cloud.
2. **Ease of use.** Run ad-hoc SQL queries on your dataset without having to prepare the data, like indexes, beforehand. This is invaluable for data exploration.
3. **Scale.** Carry out data exploration on extremely large datasets interactively. You don't need to sample the data in order to work with it in a timely manner.
4. **Shareability**. You will be able to run queries on data from different datasets without any issues. BigQuery is a convenient way to share datasets. Of course, you can also keep your data private, or share them only with specific persons -- not all data needs to be public.

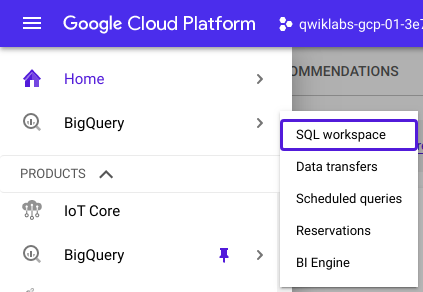
In this lab, you will find what types of municipal complaints are correlated with weather. For example, you will find (not surprisingly) that complaints about residential furnaces are most common when it is cold outside:



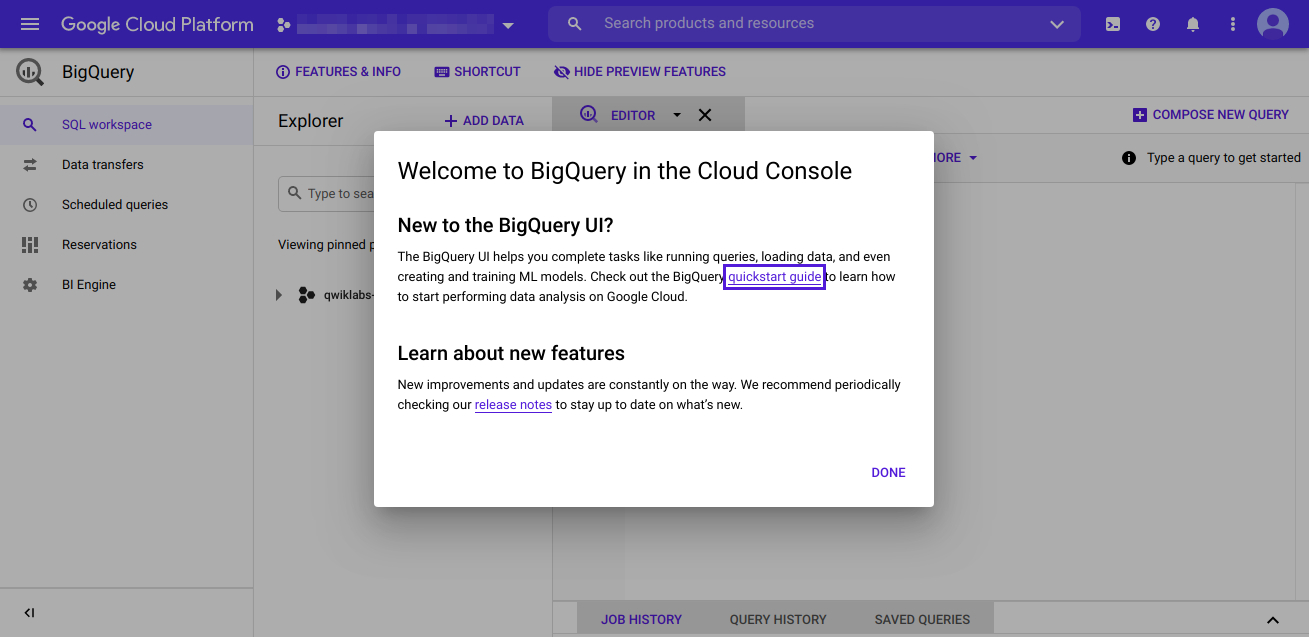
## Explore weather data

### Open BigQuery Console

In the Google Cloud Console, select **Navigation menu** > **BigQuery**:

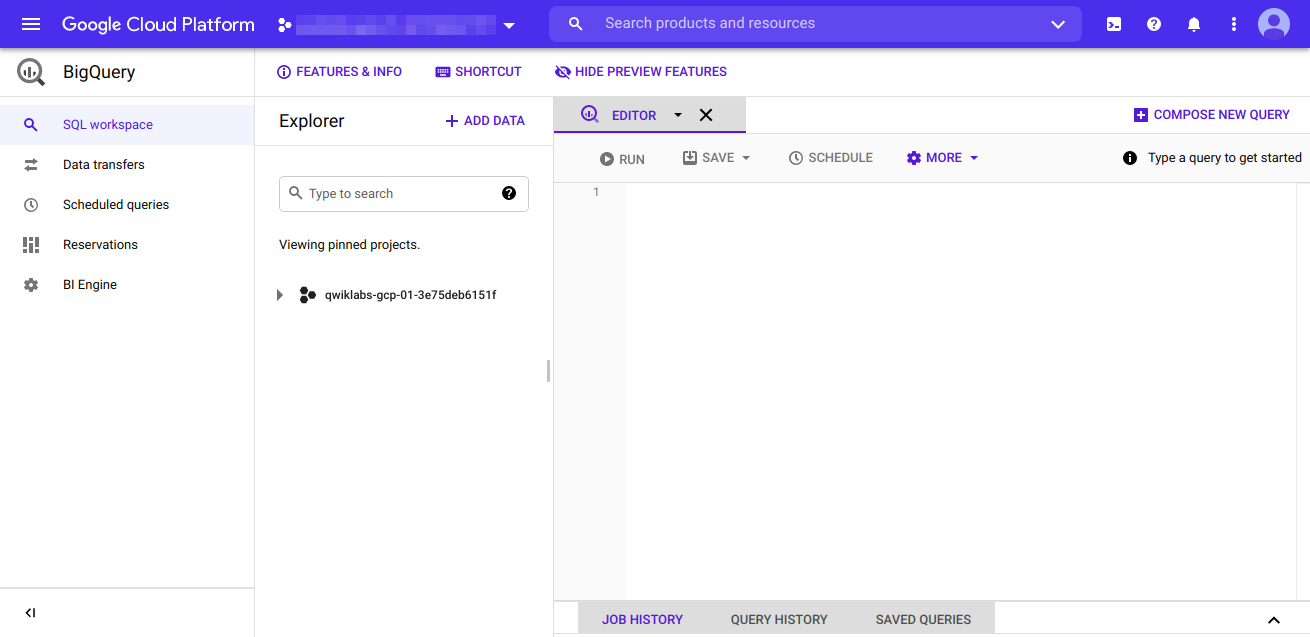


The **Welcome to BigQuery in the Cloud Console** message box opens. This message box provides a link to the quickstart guide and the release notes.

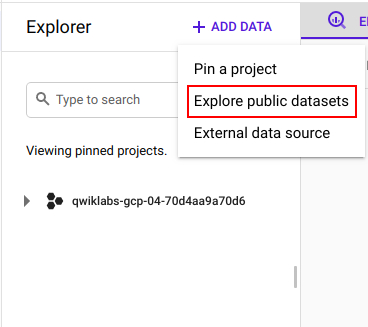


Click **Done**.

The BigQuery console opens.



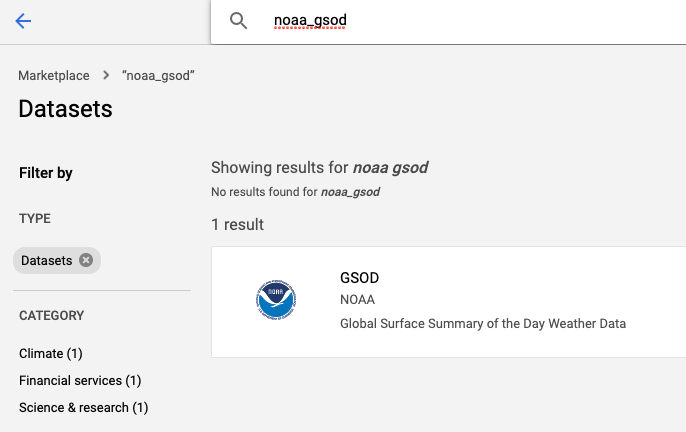
In the middle pane, click **ADD DATA** > **Explore public datasets**.



The Datasets window opens.

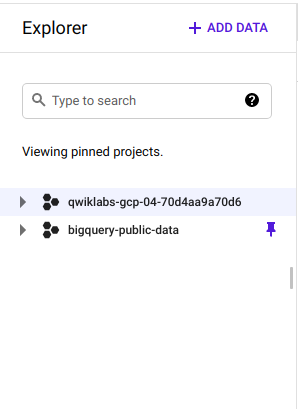
In the **Search** bar, type "noaa\_gsod" then press **Enter**.

1 result, GSOD dataset, displays.



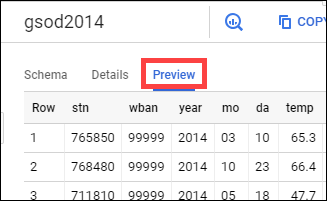
Click the GSOD dataset and then click **VIEW DATASET**.

In the BigQuery console you see two projects in the middle pane, one named your Qwiklabs project ID, and one named **bigquery-public-data**.



In the middle pane of the BigQuery console, select **bigquery-public-data** > **noaa\_gsod** > **gsod2014** table.

In the Table (gsod2014) window, click the *Preview* tab.



Examine the columns and some of the data values.

Click **COMPOSE NEW QUERY** in the upper right.

Paste the following in the Query editor textbox:

SELECT

-- Create a timestamp from the date components.

stn,

TIMESTAMP(CONCAT(year,"-",mo,"-",da)) AS timestamp,

-- Replace numerical null values with actual null

AVG(

IF

(temp=9999.9,

NULL,

temp)) AS temperature,

AVG(

IF

(wdsp="999.9",

NULL,

CAST(wdsp AS Float64))) AS wind\_speed,

AVG(

IF

(prcp=99.99,

0,

prcp)) AS precipitation

FROM

`bigquery-public-data.noaa\_gsod.gsod20\*`

WHERE

CAST(YEAR AS INT64) > 2010

AND CAST(MO AS INT64) = 6

AND CAST(DA AS INT64) = 12

AND (stn="725030" OR -- La Guardia

stn="744860") -- JFK

GROUP BY

stn,

timestamp

ORDER BY

timestamp DESC,

stn ASCcontent\_copy

Click **Run**. Look at the result and try to determine what this query does.

This query gives you the average temperature, wind\_speed and precipitation for stations La Guardia (725030) and JFK (744860) for every year on 12th June post 2010, in descending order of date and ascending order of station ID (stn).

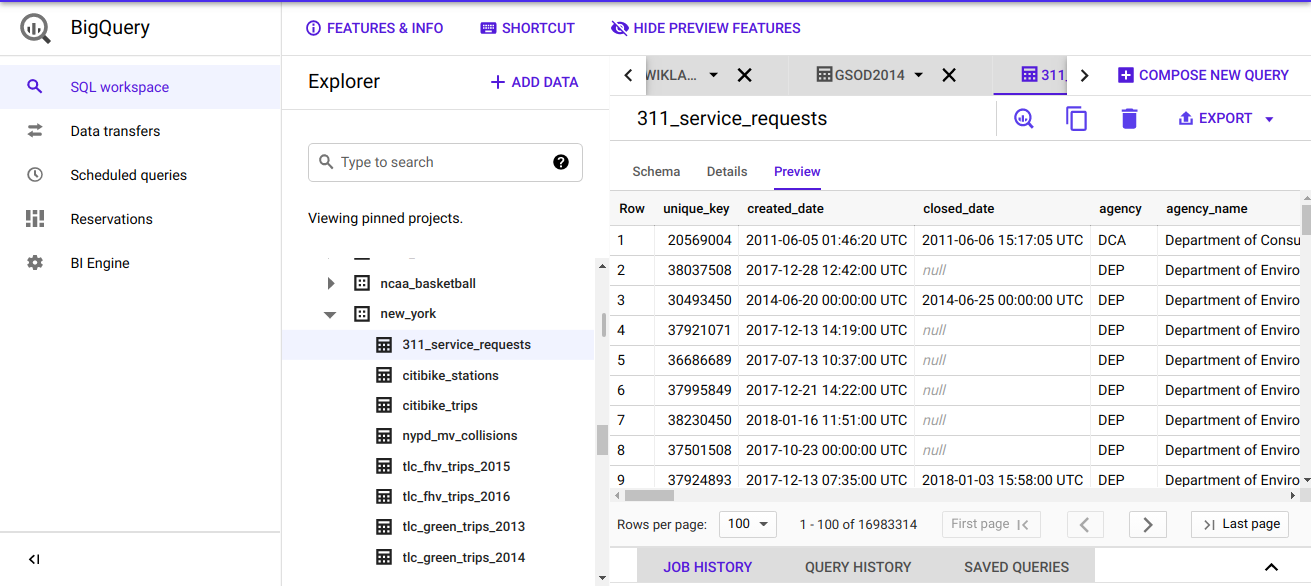
Click **Check my progress** below to verify you're on track in this lab.

Explore weather data

Check my progress

**Explore New York citizen complaints data**

In the middle pane of the BigQuery Console, select the newly added **bigquery-public-data** project and select **new\_york** > **311\_service\_requests**. Then click on the **Preview** tab. Your console should resemble the following:



Examine the columns and some of the data values.

Click **COMPOSE NEW QUERY** in the upper right.

Paste the following into the Query editor:

SELECT

EXTRACT(YEAR

FROM

created\_date) AS year,

complaint\_type,

COUNT(1) AS num\_complaints

FROM

`bigquery-public-data.new\_york.311\_service\_requests`

GROUP BY

year,

complaint\_type

ORDER BY

num\_complaints DESCcontent\_copy

Click **Run**.

This query gives you the number of complaints num\_complaints for all individual complaint types complaint\_type created every year and ordered in descending order of count of complaints num\_complaints.

Look at the results to determine what the most common complaints are. You will try to determine if these complaints correlate to weather in a later part of this lab.

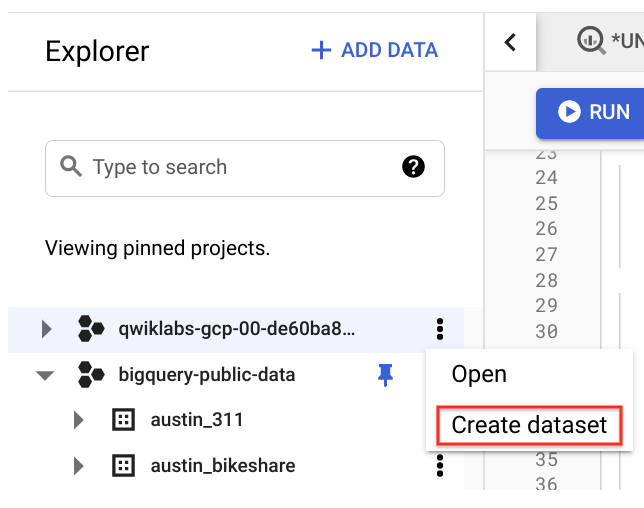
Click **Check my progress** below to verify you're on track in this lab.

Explore New York citizen complaints data

Check my progress

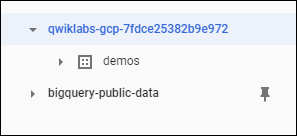
**Saving a new table of weather data**

In the middle pane of the BigQuery Console, select your qwiklabs project and click on the View action icon (three dots) then select **Create Dataset**.



In the Create dataset dialog, set the **Dataset ID** to "demos" and leave the other options at their default values.

Click **Create dataset**. Your project now has a dataset named "demos"



Click **COMPOSE NEW QUERY** and then run the following query:

SELECT

-- Create a timestamp from the date components.

TIMESTAMP(CONCAT(year,"-",mo,"-",da)) AS timestamp,

-- Replace numerical null values with actual nulls

AVG(

IF

(temp=9999.9,

NULL,

temp)) AS temperature,

AVG(

IF

(visib=999.9,

NULL,

visib)) AS visibility,

AVG(

IF

(wdsp="999.9",

NULL,

CAST(wdsp AS Float64))) AS wind\_speed,

AVG(

IF

(gust=999.9,

NULL,

gust)) AS wind\_gust,

AVG(

IF

(prcp=99.99,

NULL,

prcp)) AS precipitation,

AVG(

IF

(sndp=999.9,

NULL,

sndp)) AS snow\_depth

FROM

`bigquery-public-data.noaa\_gsod.gsod20\*`

WHERE

CAST(YEAR AS INT64) > 2008

AND (stn="725030" OR -- La Guardia

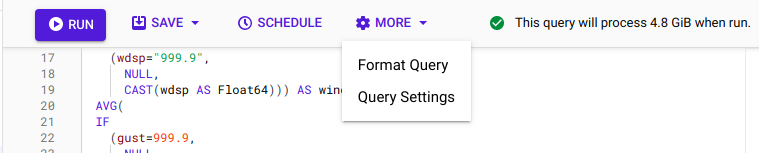
stn="744860") -- JFK

GROUP BY

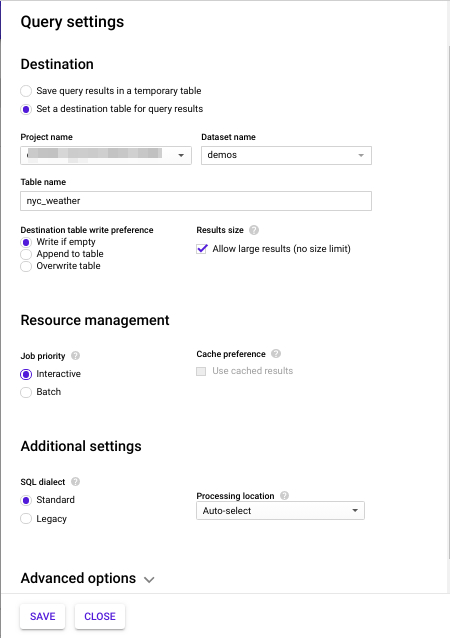
timestampcontent\_copy

This query gives you the averages of temperature, wind\_speed, visibility, wind\_gust, precipitation and snow\_depth for stations La Guardia (725030) and JFK (744860) for every year post 2008 grouped by date.

Along the top of the Query editor, click **More** > **Query settings**.



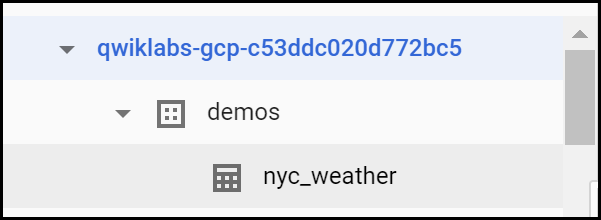
In the Query settings dialog, set the following fields. Leave all others at their default value.



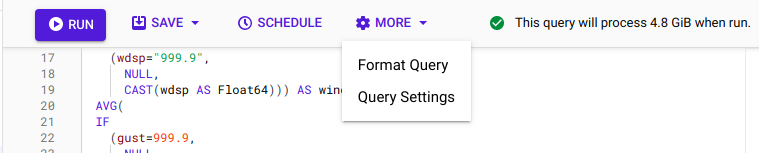
Click **Save**

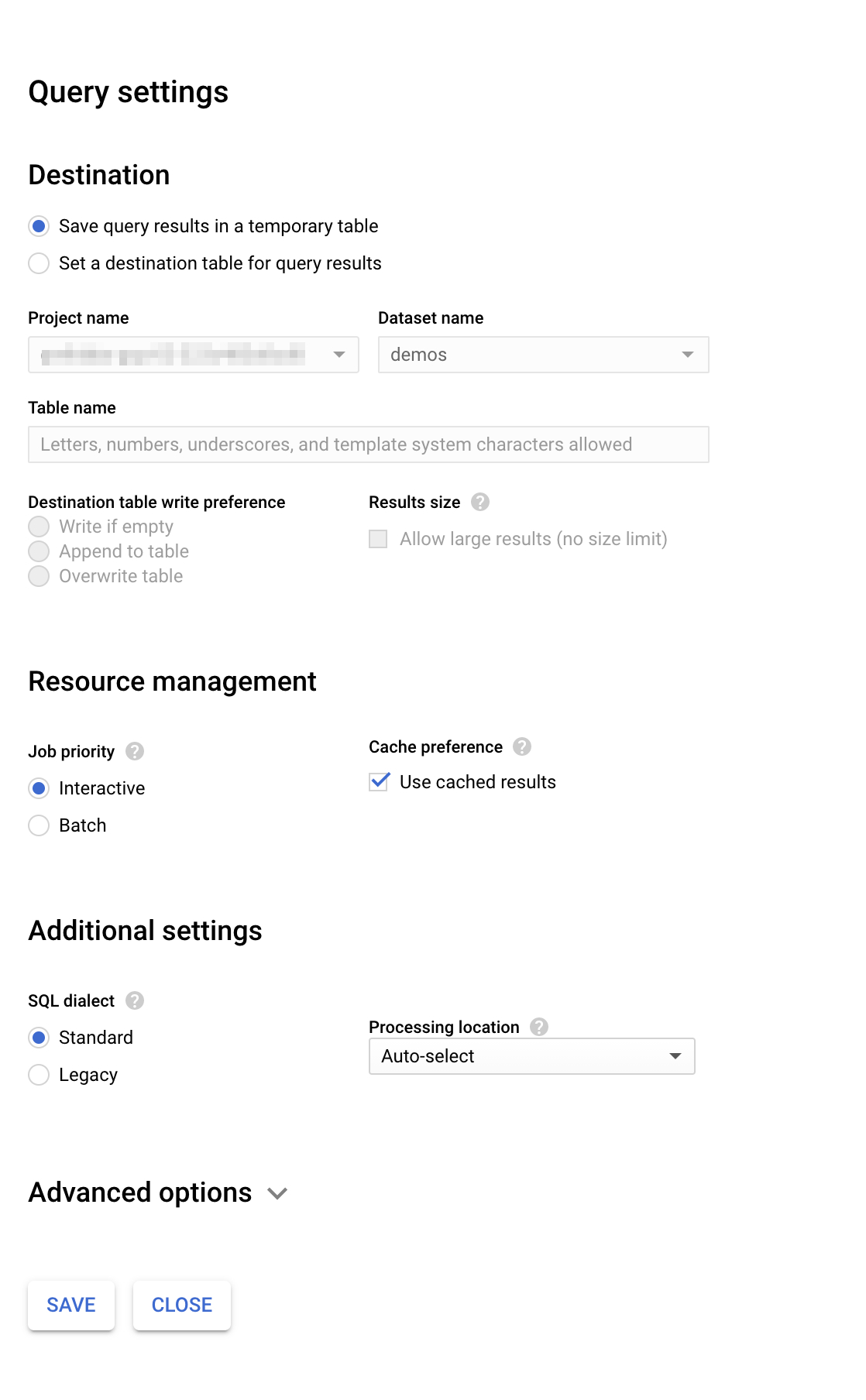
Click **Run**.

The results are now saved in the dataset you created (demos).



Navigate back to **More** > **Query settings** and, in the *Destination field* select **Save query results in a temporary table**. This removes the demos dataset as a destination for future queries.





Click **Save** to close the query.

Click **Check my progress** below to verify you're on track in this lab.

Saving a new table of weather data

Check my progress

**Summary**

In this lab you did ad-hoc queries on two datasets. You were able to query the data without setting up any clusters, creating any indexes, etc. All without ever leaving your browser!

**Congratulations!**

You learned how to run some very interesting queries on BigQuery!



Next steps / learn more

* For more fun analysis of the NYC data and how it is correlated with weather, see [Reto Meier's blog post](https://medium.com/google-cloud/four-seasons-and-5-boroughs-in-one-post-d8c90afc7071)
* [Learn more about BigQuery public data sets](https://cloud.google.com/bigquery/public-data/).

Copyright 2020 Google LLC All rights reserved. Google and the Google logo are trademarks of Google LLC. All other company and product names may be trademarks of the respective companies with which they are associated.