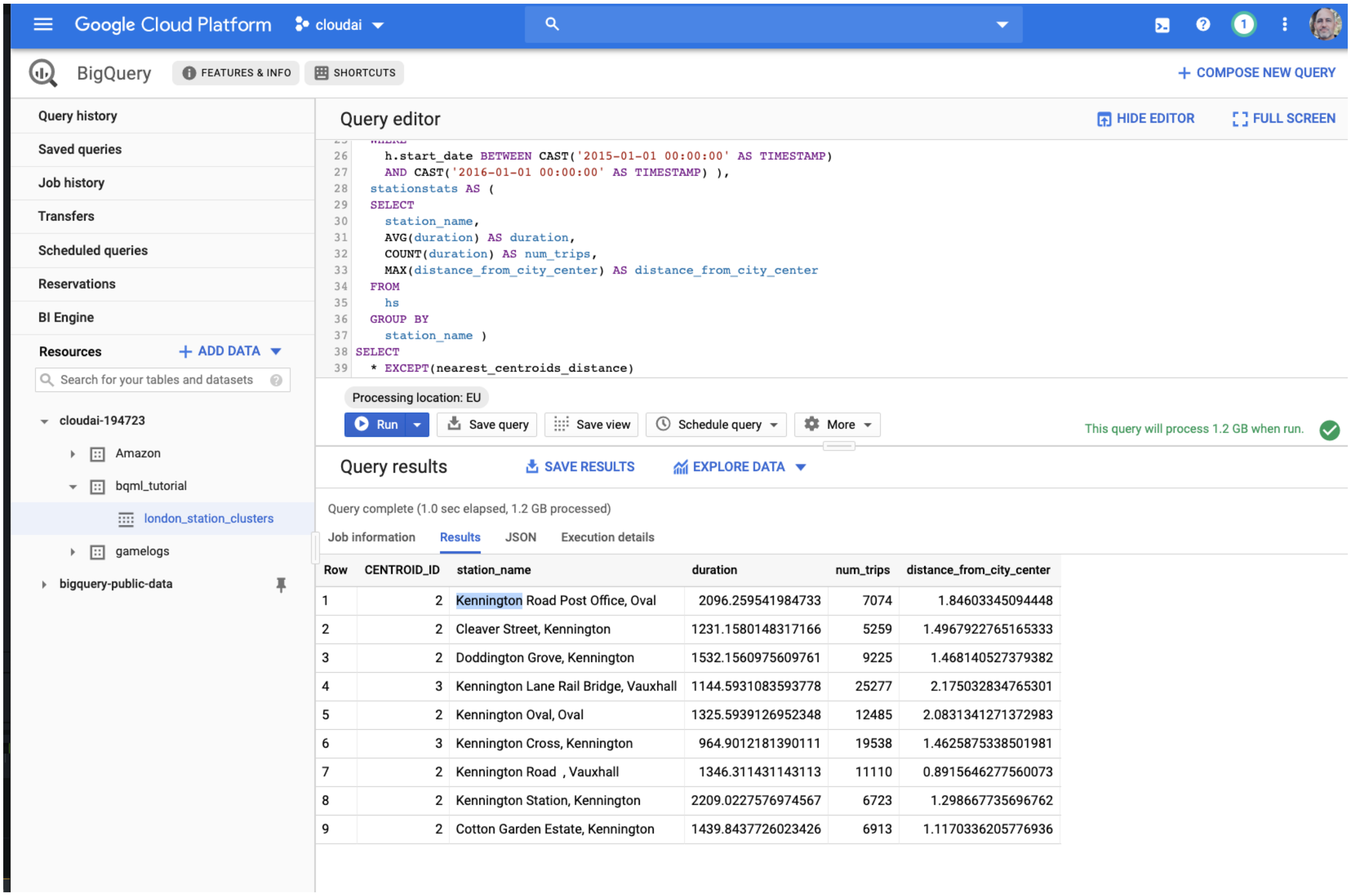
Cloud Data Warehouse

The advantage of the cloud is infinite compute and infinite storage. Cloud-native data warehouse systems also allow for serverless workflows that can directly integrate Machine Learning on the data lake. They are also ideal for developing Business Intelligence solutions.

GCP BigQuery

There is a lot to like about GCP BigQuery. It is serverless, it has integrated Machine Learning, and it is easy to use. This next section has a walkthrough of a k-means clustering tutorial.

The interface, when queried, intuitively gives back results. A key reason for this is the use of SQL and the direct integration with both Google Cloud and Google Data Studio



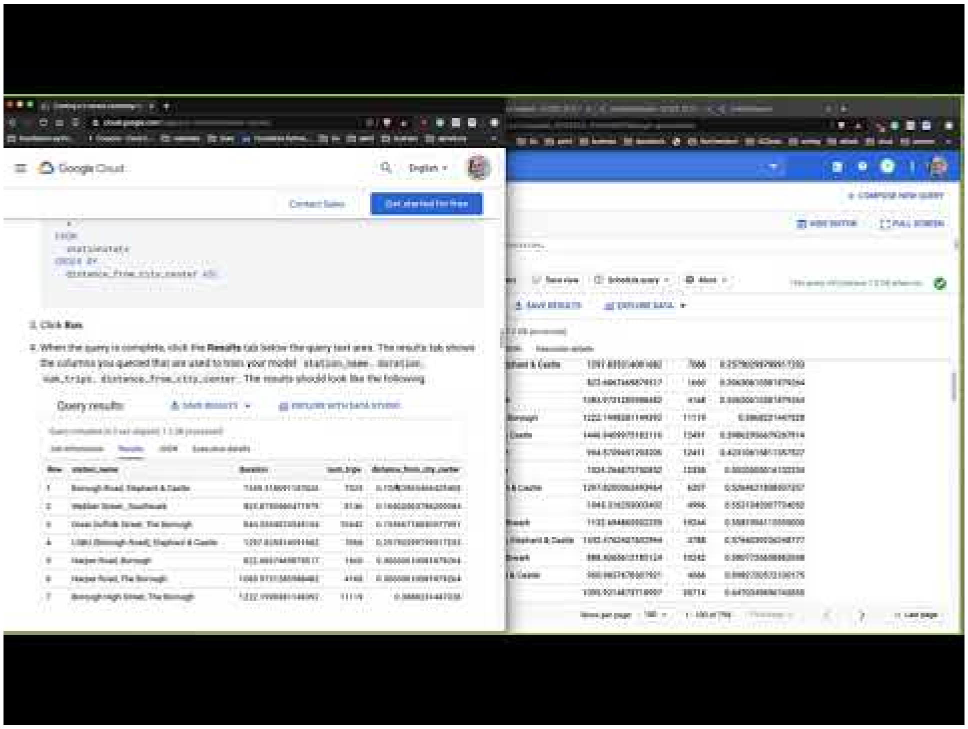
```
    h.start_date BETWEEN CAST('2015-01-01 00:00:00' AS TIMESTAMP)
    AND CAST('2016-01-01 00:00:00' AS TIMESTAMP) ),
  stationstats AS (
  SELECT
    station_name,
    AVG(duration) AS duration,
    COUNT(duration) AS num_trips,
    MAX(distance_from_city_center) AS distance_from_city_center
  FROM
    hs
  GROUP BY
    station_name )
SELECT
  * EXCEPT(nearest_centroids_distance)
```

This query will process 1.2 GB when run.

Query complete (1.0 sec elapsed, 1.2 GB processed)

Row	CENTROID_ID	station_name	duration	num_trips	distance_from_city_center
1	2	Kennington Road Post Office, Oval	2096.259541984733	7074	1.84603345094448
2	2	Cleaver Street, Kennington	1231.1580148317166	5259	1.4967922765165333
3	2	Doddington Grove, Kennington	1532.1560975609761	9225	1.468140527379382
4	3	Kennington Lane Rail Bridge, Vauxhall	1144.5931083593778	25277	2.175032834765301
5	2	Kennington Oval, Oval	1325.5939126952348	12485	2.0831341271372983
6	3	Kennington Cross, Kennington	964.9012181390111	19538	1.4625875338501981
7	2	Kennington Road , Vauxhall	1346.311431143113	11110	0.8915646277560073
8	2	Kennington Station, Kennington	2209.0227576974567	6723	1.298667735696762
9	2	Cotton Garden Estate, Kennington	1439.8437726023426	6913	1.1170336205776936

Learn to use Google BigQuery in the following screencast.



```
  FROM
    stationstats
  ORDER BY
    distance_from_city_center ASC
```

3. Click **Run**.

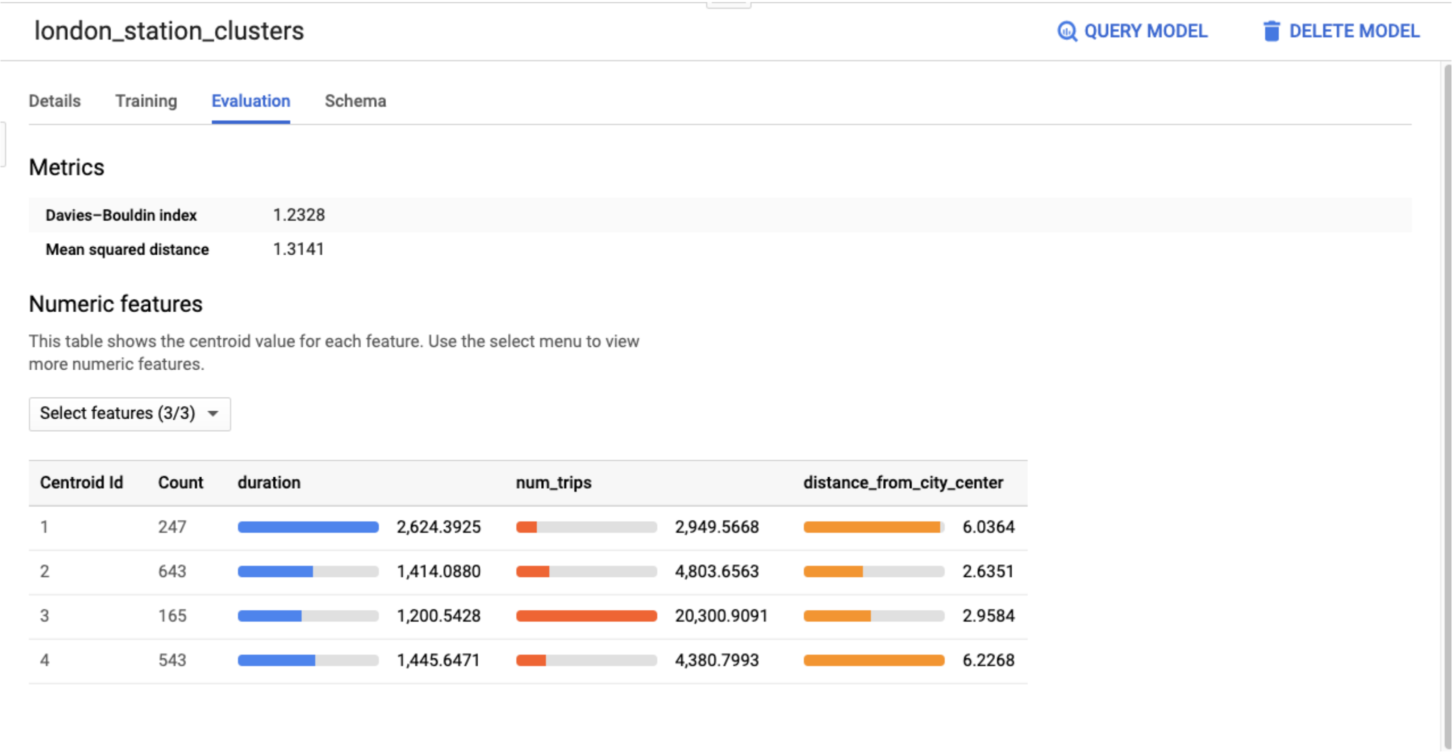
4. When the query is complete, click the **Results** tab below the query text area. The results tab shows the columns you queried that are used to train your model: station_name, duration, num_trips, distance_from_city_center. The results should look like the following.

Video Link: https://www.youtube.com/watch?v=elec2DXqw3Q

Even better, you can directly train Machine Learning models using a SQL statement. This workflow shows an emerging trend with Cloud Database services in that they let you both query the data and train the model. In this example, the `kmeans` section is where the magic happens.

```
CREATE OR REPLACE MODEL
  bqml_tutorial.london_station_clusters OPTIONS(model_type='kmeans',
    num_clusters=4) AS
WITH
  hs AS (
  SELECT
    h.start_station_name AS station_name,
  IF
    (EXTRACT(DAYOFWEEK
      FROM
        h.start_date) = 1
      OR EXTRACT(DAYOFWEEK
      FROM
        h.start_date) = 7,
      "weekend",
      "weekday") AS isweekday,
    h.duration,
    ST_DISTANCE(ST_GEOGPOINT(s.longitude,
        s.latitude),
      ST_GEOGPOINT(-0.1,
        51.5))/1000 AS distance_from_city_center
  FROM
    `bigquery-public-data.london_bicycles.cycle_hire` AS h
  JOIN
    `bigquery-public-data.london_bicycles.cycle_stations` AS s
  ON
    h.start_station_id = s.id
  WHERE
    h.start_date BETWEEN CAST('2015-01-01 00:00:00' AS TIMESTAMP)
    AND CAST('2016-01-01 00:00:00' AS TIMESTAMP) ),
  stationstats AS (
  SELECT
    station_name,
    isweekday,
    AVG(duration) AS duration,
    COUNT(duration) AS num_trips,
    MAX(distance_from_city_center) AS distance_from_city_center
  FROM
    hs
  GROUP BY
    station_name, isweekday)
SELECT
  * EXCEPT(station_name, isweekday)
FROM
  stationstats
```

Finally, when the k-means clustering model trains, the evaluation metrics appear as well in the console.



Metrics

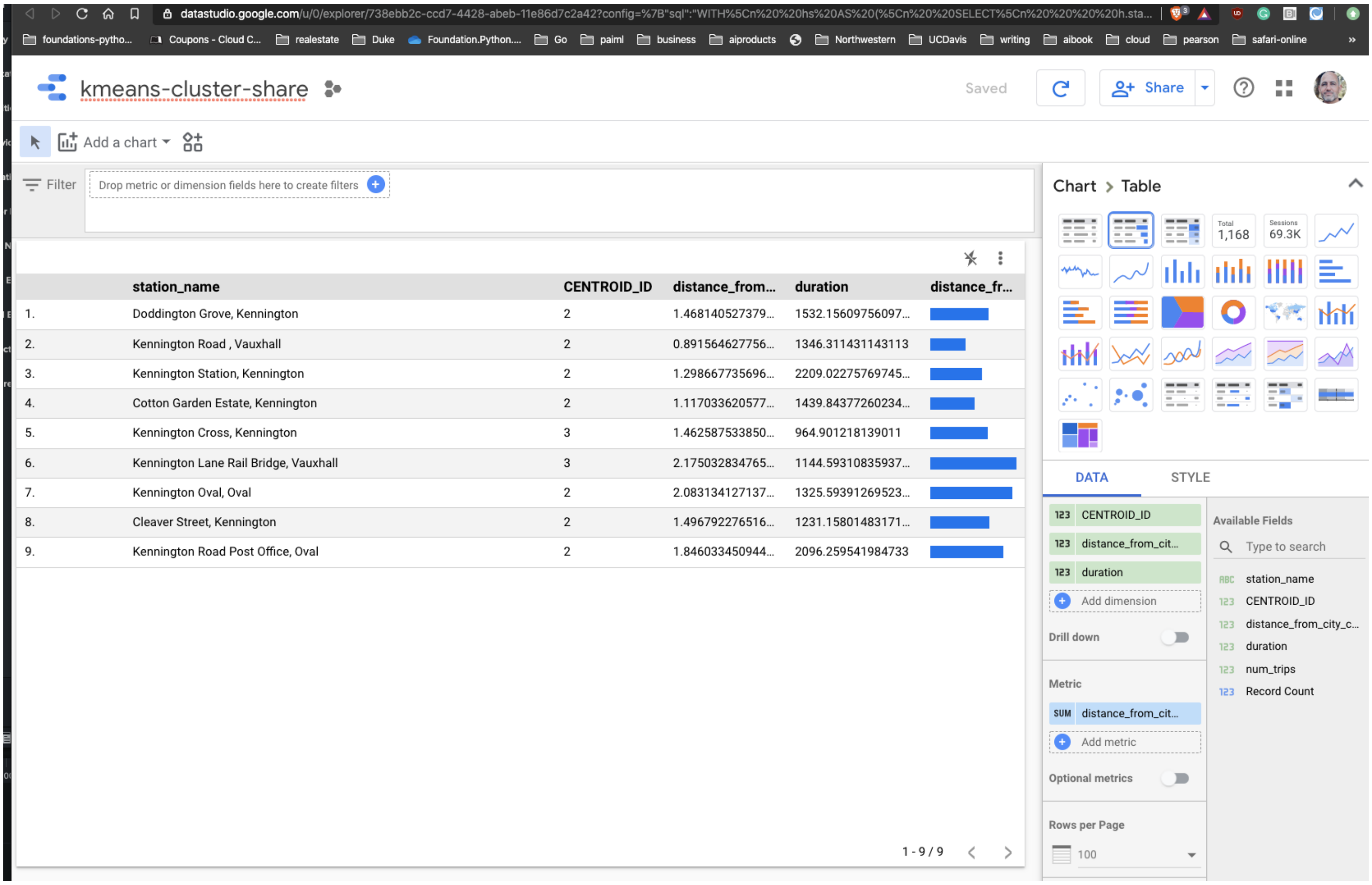
Davies–Bouldin index	1.2328
Mean squared distance	1.3141

Numeric features

This table shows the centroid value for each feature. Use the select menu to view more numeric features.

Centroid Id	Count	duration	num_trips	distance_from_city_center
1	247	2,624.3925	2,949.5668	6.0364
2	643	1,414.0880	4,803.6563	2.6351
3	165	1,200.5428	20,300.9091	2.9584
4	543	1,445.6471	4,380.7993	6.2268

Often a meaningful final step is to take the result and then export it to their Business Intelligence (BI) tool, data studio.



	station_name	CENTROID_ID	distance_from...	duration	distance_fr...
1.	Doddington Grove, Kennington	2	1.468140527379...	1532.15609756097...	
2.	Kennington Road , Vauxhall	2	0.891564627756...	1346.311431143113	
3.	Kennington Station, Kennington	2	1.298667735696...	2209.02275769745...	
4.	Cotton Garden Estate, Kennington	2	1.117033620577...	1439.84377260234...	
5.	Kennington Cross, Kennington	3	1.462587533850...	964.901218139011	
6.	Kennington Lane Rail Bridge, Vauxhall	3	2.175032834765...	1144.59310835937...	
7.	Kennington Oval, Oval	2	2.083134127137...	1325.59391269523...	
8.	Cleaver Street, Kennington	2	1.496792276516...	1231.15801483171...	
9.	Kennington Road Post Office, Oval	2	1.846033450944...	2096.259541984733	

1 - 9 / 9

Dimension:
123 CENTROID_ID
123 distance_from_cit...
123 duration

Drill down

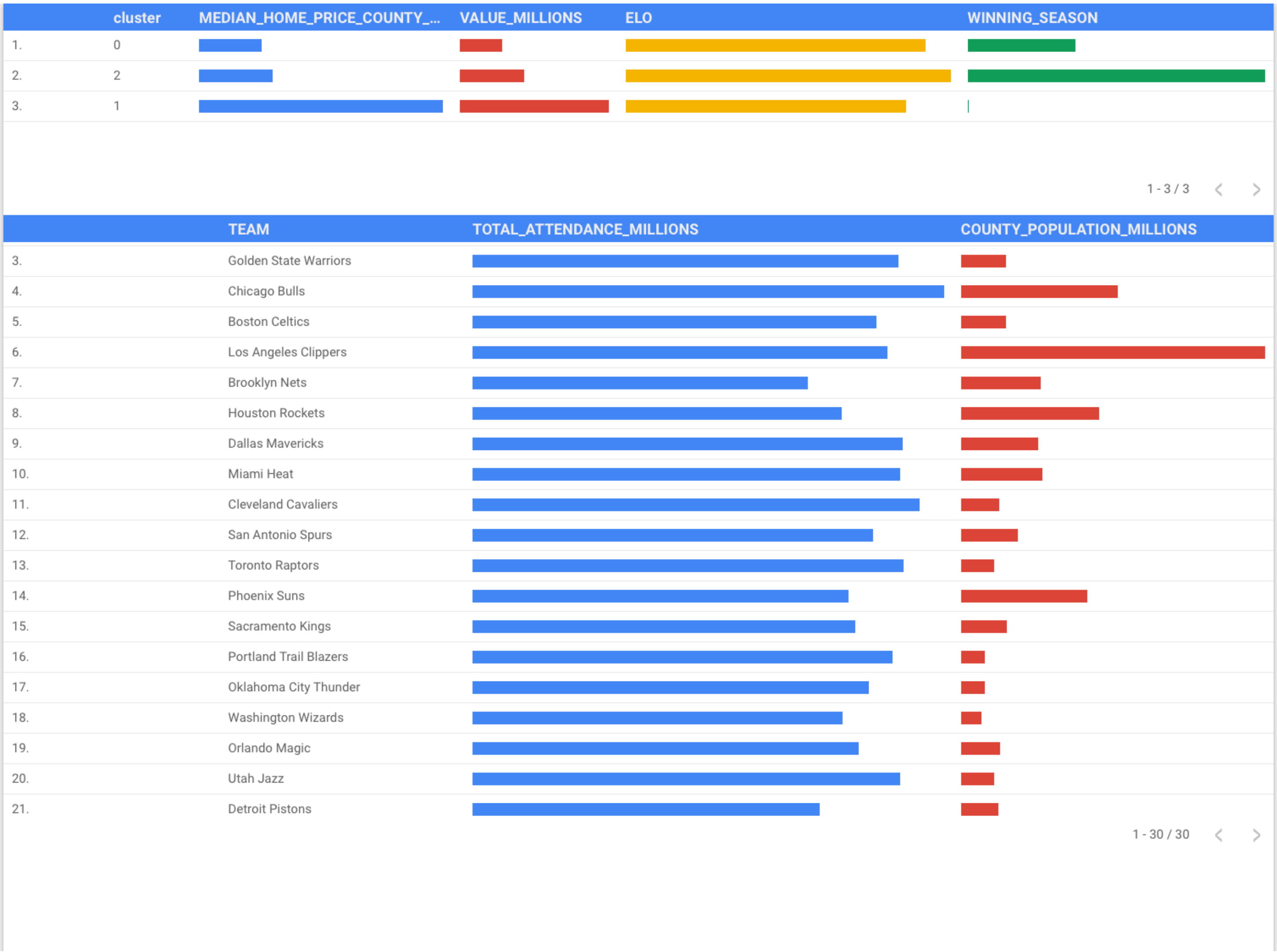
Metric
SUM distance_from_cit...

Optional metrics

Rows per Page
100

Available Fields
RBC station_name
123 CENTROID_ID
123 distance_from_city_c...
123 duration
123 num_trips
123 Record Count

The following is an excellent example of what a cluster visualization could look like in Google Big Query exported to Google Data Studio.



	cluster	MEDIAN_HOME_PRICE_COUNTY_...	VALUE_MILLIONS	ELO	WINNING_SEASON
1.	0				
2.	2				
3.	1				

	TEAM	TOTAL_ATTENDANCE_MILLIONS	COUNTY_POPULATION_MILLIONS
3.	Golden State Warriors		
4.	Chicago Bulls		
5.	Boston Celtics		
6.	Los Angeles Clippers		
7.	Brooklyn Nets		
8.	Houston Rockets		
9.	Dallas Mavericks		
10.	Miami Heat		
11.	Cleveland Cavaliers		
12.	San Antonio Spurs		
13.	Toronto Raptors		
14.	Phoenix Suns		
15.	Sacramento Kings		
16.	Portland Trail Blazers		
17.	Oklahoma City Thunder		
18.	Washington Wizards		
19.	Orlando Magic		
20.	Utah Jazz		
21.	Detroit Pistons		

You can view the report using this direct URL.

Summary of GCP BigQuery

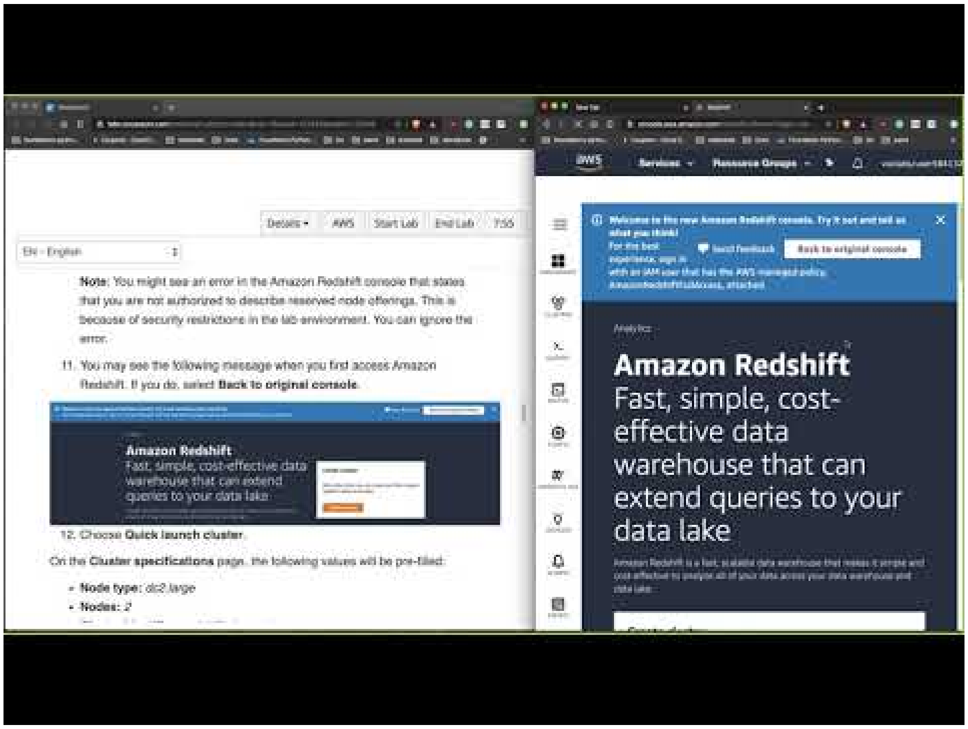
In a nutshell, GCP BigQuery is a useful tool for Data Science and Business Intelligence. Here are the key features.

- Serverless
- Large selection of Public Datasets
- Integrated Machine Learning
- Integration with Data Studio
- Intuitive
- SQL based

AWS Redshift

AWS Redshift is a Cloud data warehouse designed by AWS. The key features of Redshift include the ability to query exabyte data in seconds through the columnar design. In practice, this means excellent performance regardless of the size of the data.

Learn to use AWS Redshift in the following screencast.



Note: You might see an error in the Amazon Redshift console that states that you are not authorized to describe reserved node offerings. This is because of security restrictions in the lab environment. You can ignore the error.

11. You may see the following message when you first access Amazon Redshift. If you do, select **Back to original console**.

12. Choose **Quick launch cluster**.

On the **Cluster specifications** page, the following values will be pre-filled:

- **Node type:** dc2.large
- **Nodes:** 2

Welcome to the new Amazon Redshift console. Try it out and tell us what you think!
For the best experience, sign in with an IAM user that has the AWS managed policy AmazonRedshiftFullAccess attached

Analytics

Amazon Redshift
Fast, simple, cost-effective data warehouse that can extend queries to your data lake

Amazon Redshift is a fast, scalable data warehouse that makes it simple and cost-effective to analyze all of your data across your data warehouse and data lake.

Video Link: https://www.youtube.com/watch?v=vXSH24AJzrU

Key actions in a Redshift Workflow

In general, the key actions are as described in the Redshift getting started guide. These are the critical steps to setup a workflow.

- Cluster Setup
- IAM Role configuration (what can role do?)
- Setup Security Group (i.e. open port 5439)
- Setup Schema

```
create table users(
userid integer not null distkey sortkey,
username char(8),
```

- Copy data from S3

```
copy users from 's3://awssampledbuswest2/tickit/allusers_pipe.txt'
credentials 'aws_iam_role=<iam-role-arn>'
delimiter '|' region 'us-west-2';
```

- Query

```
SELECT firstname, lastname, total_quantity
FROM
(SELECT buyerid, sum(qtysold) total_quantity
FROM  sales
GROUP BY buyerid
ORDER BY total_quantity desc limit 10) Q, users
WHERE Q.buyerid = userid
ORDER BY Q.total_quantity desc;
```

Summary of AWS Redshift

The high-level takeaway for AWS Redshift is the following.

- Mostly managed
- Deep Integration with AWS
- Columnar
- Competitor to Oracle and GCP Big Query
- Predictable performance on massive datasets

Summary

This chapter covers storage, including object, block, filesystem, and Databases. A unique characteristic of Cloud Computing is the ability to use many tools at once to solve a problem. This advantageous trait is heavily at play with the topic of Cloud Storage and Cloud Databases.