on point

An Analysis of Google BigQuery

Mauricio Valdez, *IT Manager*

BigQuery is a highly scalable, robust big data processing warehouse that is available for all organizations within Google's big data catalog of services. Organizations looking to query large data sets now have the option of using this service to accomplish tasks that, until recently, required expensive infrastructure and complex applications. BigQuery is fully managed so there is no need to host any servers or applications to use its services. Google makes this service available and can be accessed by logging into the Google Cloud Platform (GCP) and selecting the service. In this whitepaper, we will cover BigQuery key features and access a publicly available data set for demonstration.

BigQuery comes with many features that increase productivity, security, and collaboration. Because BigQuery is hosted in the GCP, the BigQuery service allows users to analyze large data sets in a relatively short time. The data can be stored within the Google cloud, or accessed from alternate locations. It is even possible to stream live data into BigQuery for real-time analysis. BigQuery takes security into account by ensuring all data is encrypted when it is being processed. The use of account permissions also protects the data and allows users to select who can see the data. All data sets used in BigQuery can be shared with other users and can be replicated in the GCP to ensure it's available as needed, where needed. Specific share permissions can be set up to limit who has access to view your data sets and who can query the data.

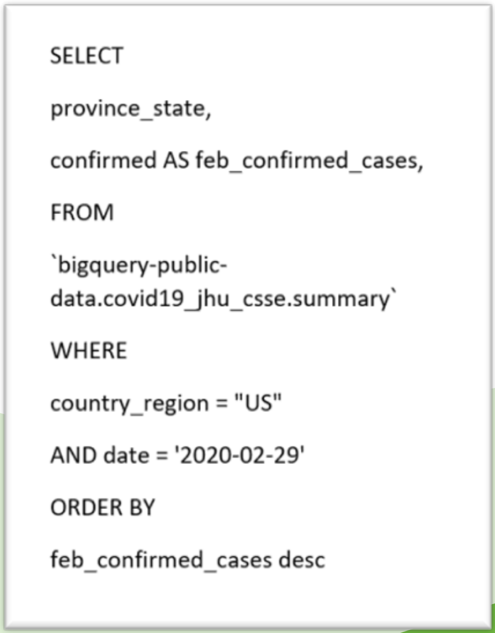
Along with great features, BigQuery also offers many benefits for its users. The low cost is a great incentive for organizations to use this service. As a no-host solution, organizations do not need to invest in on-premise infrastructure, hardware, or physical servers. BigQuery users can pay as-needed with flexible options available, such as on-demand or flat rate options. There is no cost per hour for BigQuery, only a cost for queries through the data warehouse service. An additional cost may arise for storage, but that is only when a user exceeds 10 GB during their first month, and that cost is minimal at $0.02 per GB. While additional services may incur costs as well, getting started is very affordable.

The first step in accessing GCP is to ensure you have a Google account. If you do not have one, the process to register for an account is very simple. Go to https://accounts.google.com and click on the "create an account" button. Once an account has been set up you will need to sign into the GCP at https://cloud.google.com. The new account will also provide you a $300 credit for your initial GCP trial. If you do not add a credit card when setting up the account, you can create projects in the Sandbox, but that environment has use and functionality limitations and is for trial-use only. Once you have your Google account and you have signed into the GCP, you will need to make sure that you click on the blue button labeled "Console" to ensure you see all the available options. Once in the GCP console, you can create a new project, search for BigQuery, and select the link that will take you to the BigQuery console.

Once inside the BigQuery console, you will see a column on the left-hand side of your screen and two windows split in the middle. The top pane is the Query Editor for queries to be run against the data sets, and the bottom pane is for query results. The left column contains links to tools used for queries and further analysis. It is important to mention that BigQuery can access many publicly available datasets and Google offers a catalog of available data sets for use by the general public. For this demonstration, we accessed a publicly available dataset for COVID-19 data. The dataset is the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at John Hopkins University (JHU). To access this data set, please copy the following URL link into your browser: **https://console.cloud.google.com/bigquery?project=august-cascade-215413&p=bigquery-public-data&d=covid19_jhu_csse&page=dataset.** When you paste this link into your browser it will automatically load the JHU CSSE dataset into your BigQuery session.

BigQuery utilizes queries that parse the data and deliver the information requested. Once it finds the data it organizes the data as instructed by the query. Queries are highly customizable and can offer many representations of the data, but they do require a working knowledge of their structure and how to write them. BigQuery uses standard SQL syntax for its queries, and knowledge of standard SQL will allow for greater analysis of the data set. As BigQuery users become familiar with queries, the data results generated can be full of valuable information. The more the user knows about queries and how to execute them successfully, the more value the data set will present to the user. The JHU CSSE sample query we used is listed in the box below. This query answered the following question: How many confirmed COVID-19 cases were there in the US, by state, in February 2020?

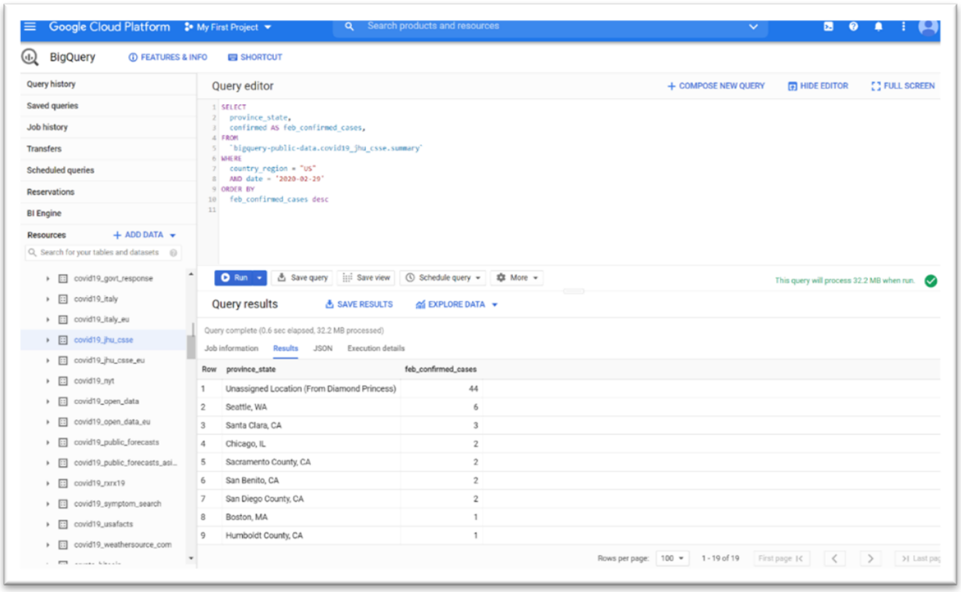
This sample query identifies the total number of cases in February 2020.



```
SELECT

province_state,

confirmed AS feb_confirmed_cases,

FROM

`bigquery-public-
data.covid19_jhu_csse.summary`

WHERE

country_region = "US"

AND date = '2020-02-29'

ORDER BY

feb_confirmed_cases desc
```

Exhibit 1: Sample Query

Once the query is loaded into the Query Editor it needs to be initiated or run against the data set. This is done by clicking on the blue “Run” button. This initiates the query, and the results are provided in the lower pane of the BigQuery console. The data is presented in a table format and depending on the size of the results, the table may have many columns and rows that extend over many pages. A better way to view this data is in visual format presentation via Google Data Studio. To view the data in Google Data Studio, click on the Explore Data tab in the center of the screen in the Query Results menu and select the Explore Data option. The data can be imported in Data Studio and presented in one of many ways: pie charts, bar graphs, line charts, etc. The data can then be arranged by selected tables and exported to a dashboard, website, or imported into presentations. See Exhibit 2 below for the results of the query above in table format and in a pie chart in Exhibit 3, generated using Google Data Studio. Exhibit 4 below isolates the COVID data per location as represented in the data set table used for this example.



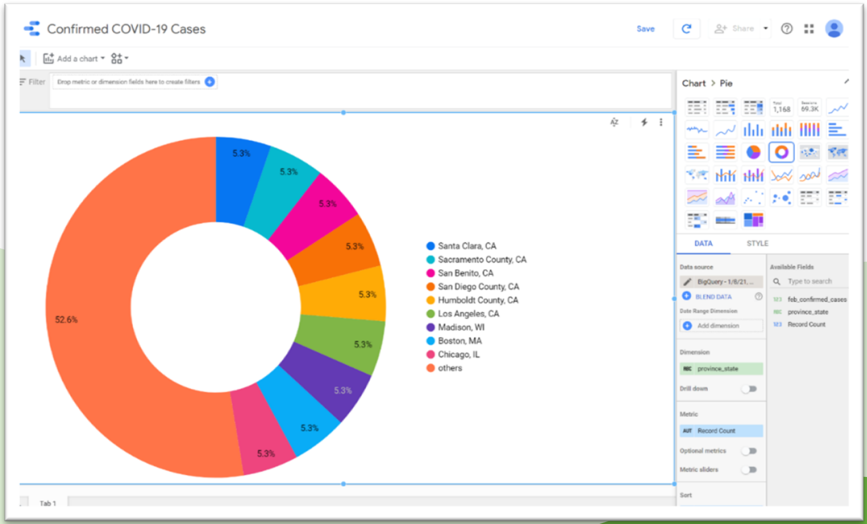
```
SELECT
  province_state,
  confirmed AS feb_confirmed_cases,
FROM
  `bigquery-public-data.covid19_jhu_csse.summary`
WHERE
  country_region = "US"
  AND date = '2020-02-29'
ORDER BY
  feb_confirmed_cases desc
```

This query will process 32.2 MB when run.

Query complete (0.6 sec elapsed, 32.2 MB processed)

Row	province_state	feb_confirmed_cases
1	Unassigned Location (From Diamond Princess)	44
2	Seattle, WA	6
3	Santa Clara, CA	3
4	Chicago, IL	2
5	Sacramento County, CA	2
6	San Benito, CA	2
7	San Diego County, CA	2
8	Boston, MA	1
9	Humboldt County, CA	1

Exhibit 2: Data Represented in a Table Format



Data source: BigQuery - 1/8/21, ...

Dimension: province_state

Metric: Record Count

Exhibit 3: Data Represented in a Pie Chart

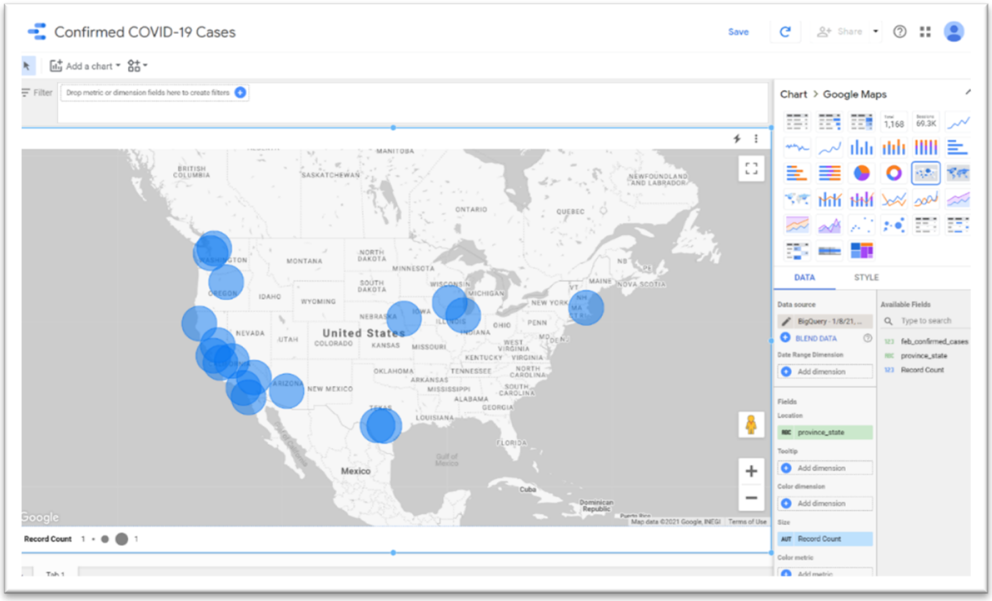


Exhibit 4: An Example of Isolating the Data

In the demonstration above, we used the publicly-available data set offered by JHU CSSE and hosted on the Google Cloud Platform, but BigQuery also allows you to import your own data in various formats such as CSV, JASON, Parquet, and Avro. This allows you to use your own data and BigQuery to analyze the data, and Data Studio to represent the data in a visual format as shown above. To use your own data, you need to create a specific project, create a new data set, and import the data. If the data is in the correct format without errors, BigQuery will create the schema automatically. This paper only presents the tip of the iceberg with what can be done with BigQuery—there are many other more advanced options within the program that can be discovered by doing a BigQuery deep dive.

The benefits of using BigQuery are many, as mentioned before, but these benefits increase when large data sets are used and the vast, scalable infrastructure and robust processing power of the Google Cloud Platform are utilized to make the process of analyzing large data sets possible and at much faster speeds. This efficiency allows for faster decision-making and more time to be spent on analysis within organizations. For example, BigQuery can process terabytes of data in seconds and petabytes of data in minutes. This makes it extremely beneficial when working with large data sets. Additionally, the fact that it is a Google-hosted solution, available via the Internet 24/7/365, and not dependent on on-premise hardware, allows for users to work from any location at any time. Once the data has been uploaded into BigQuery and processed correctly, it can then be transferred and shared with many other GCP services for further review and analysis.

There are thousands of organizations using BigQuery and GCP services across all industries. As shown in the demonstration above, the academic and healthcare industry is using BigQuery to analyze COVID-19 data sets as reported by the US States. Integrate this with Google Data Studio, a web-based data visualization service, and you get a visual interpretation of the impact COVID-19 is having in North America, as shown in Exhibit 5 below. BigQuery services are used to identify specific trends and patterns, and to better understand an organization's data, in this case, confirmed COVID cases, their concentration per state, and the speed at which the virus is spreading. In the end, all interpretive results from data sets can be beneficial to an organization to help it understand information and patterns that can be very beneficial for decision-makers.

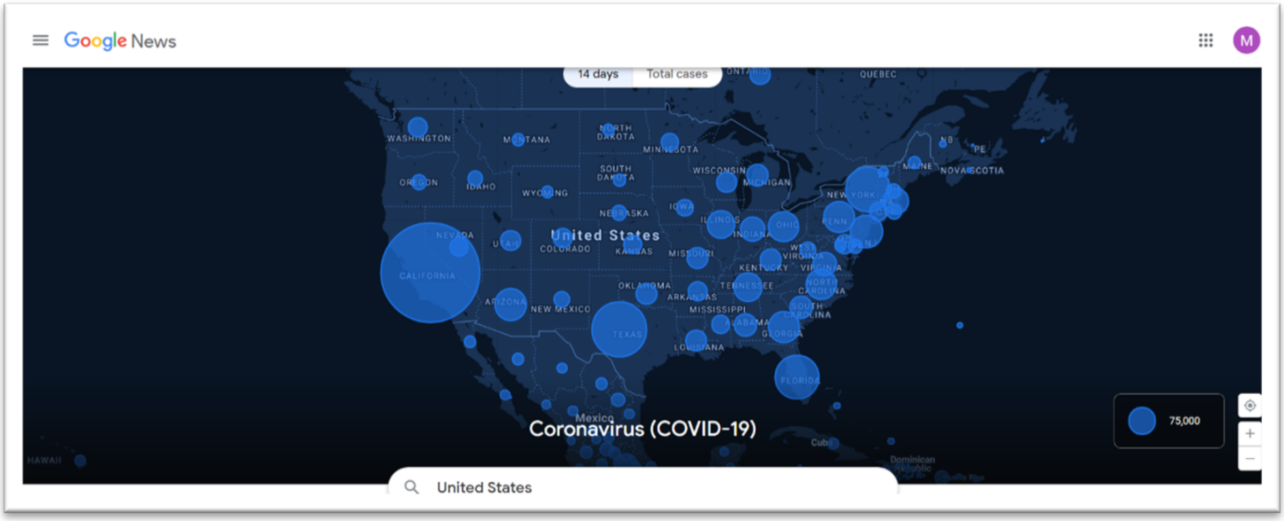


Exhibit 5: The Impact of COVID-19 in North America[1]

About OnPoint

OnPoint Consulting, Inc. (OnPoint) delivers secure IT infrastructure, enterprise systems, cybersecurity, and program management solutions for the U.S. federal government. Our specialized strategy, cyber, and technology capabilities are changing the way our clients improve performance, effectively deliver results, and manage risk. OnPoint holds ISO 9001:2015, ISO 20000-1:2011, and ISO 27001:2013 certifications, and a CMMI Maturity Level 3 rating in both services and development.

OnPoint is a part of the Publicis Sapient platform, with access to industry-leading AI tools and teams. Contact us at innovation@onpointcorp.com or visit onpointcorp.com to learn more about us and our services.

[1] "Coronavirus (COVID-19)." Google News, 1 February 2021. https://news.google.com/covid19/map?hl=en-US&gl=US&ceid=US%3Aen&mid=%2Fm%2F09c7w0