**Data in BigQuery**

* **Outer import methods for the dataset √**
* **Latency: Import data latency and query latency √**

**Loading**

There are many different ways to load data into BigQuery. We can batch load a set of data records, stream individual data, or use queries to generate new data and then append or overwrite the data results into the table. We may also use a third-party application or service.[6]

In addition, Google also has related services to help with ingesting the data, and the BigQuery platform provides public datasets for use as well. A public dataset is any dataset stored in BigQuery and is made available to the public through the Google Cloud Public Dataset Program where people can access and integrate them into the projects and applications. Apart from that, BigQuery is able to run queries on certain forms of external data without loading the data into BigQuery storage.[6][8]

**Latency**

On top of these, the associated data latency is also an important metric while considering how to import data into BigQuery. However, not all use cases require low-latency data. Higher latency is acceptable for use cases that do not require real-time data, such as generating quarterly sales reports. Normally, more complex use cases like fraud detection or recommendation engines require analysis on real-time or near real-time data, where low latency of data will be a priority measurement in terms of the data metrics.

Streaming offers the lowest latency of data being available for analysis while periodic load jobs have a higher latency because new data is only available after each load job finishes. Moreover, query performance for external data sources might not be as high as query performance for data stored in BigQuery.[6]

**Project Dataset**

The dataset we used in the project is the COVID-19 epidemiological database which we accessed through the BigQuery Marketplace. The dataset is about 11GB, and it covers many areas such as demographics, economy, epidemiology, geography, health and so on. Moreover, the data merge daily time series, from over 20000+ global sources, using a consistent set of region keys where All regions are assigned a unique location key.[1]

**Discussion**

* **Alternatives for BigQuery √**
* **BigQuery Use cases √**
* **Pros and cons of BigQuery √**

From the study above, we can see that Google BigQuery is a serverless, highly scalable data warehouse that comes with a built-in query engine and it is easy to start analytics of Bigdata in the field of cloud with the help of it. Datasets may be collected from various sources and can be imported in numerous ways.  In our project, the data samples we used are taken from the publicly available datasets on the BigQuery marketplace. During the upload process of the CSV file, a corresponding table will be automatically created in the Datawarehouse of BigQuery. Then the results will be displayed based on the given SQL queries. After querying the data, the queried results are saved to the Google Data Studio for data visualization from which we can easily observe and analyze the query results by adjusting data parameters.

Additionally, the data result also showed that BigQuery has the ability of smooth data management as well as the data handling capability where data from several sources can be processed cost-effectively without any infrastructure developments and database administrators.

**BigQuery Use Cases**

In general, one common use case for BigQuery would be in the business intelligence area where businesses can import the data in XML, JSON, CSV, or maybe a wide range of other formats by using a number of data collection methods for BigQuery as they need.

BigQuery can filter unnecessary noise from the data and then organize that data into a semi-structured or structured data warehouse, which is very helpful for flexible analysis of data and generation of data insights. As the system gets more data, BigQuery becomes more capable of creating predictive models that can produce accurate forecasts of key business factors (i.e., prices or costs) with high accuracy.

For example, when Bounteous partnered with Domino's Pizza of Canada, they imported customer loyalty and purchase data from the offline databases, merged the data with Google Analytics in BigQuery, then developed purchase propensity models to score audiences for marketing campaigns that drive revenue increases; When with Hotels, Bounteous used BigQuery to develop statistical models to understand which travel amenities customers are likely to be interested in, to provide them with a more personalized booking experience; For PBS, Bounteous need a solution to analyze a very large scale dataset with over 330 million sessions, 800 million pageviews, and 17.5 million video episode plays per year, and BigQuery was able to develop machine learning models to Analyze user trends and create audience clusters to inform new content decisions.[7]

Clearly, BigQuery's ability to perform standard data management routines makes it a desired data analysis tool for automating data processes.

**BigQuery Pros and Cons**

Using BigQuery, A full data warehouse is up and running in minutes with virtually zero ongoing operational overhead. Programmatically bulk load or stream data where queried data can be visualized on multiple data visualization platforms for in-depth analysis and insights. All without any index storage, hardware, or any typical local data warehouse management issues. Moreover, BigQuery integrates CDN and high availability, where data security and automatic backups are already handled within the system.[9][10]

Despite the advantages, BigQuery is not without its faults. One is that BigQuery has too much control over infrastructure management which means users cannot allocate resources themselves, such as storing data in different clusters or needing to separate processes across servers. Another is the cost of maintaining a data warehouse. While Google BigQuery is reasonably priced, and storage-compute-separation pricing gives businesses more control over pricing, not paying attention to the scale of usage still can lead to high costs in the long run. For example, queries that have not been performance-tuned and query returning a lot of redundant data can become costly very quickly.[9][10]

**BigQuery Alternatives**

Overall, Google BigQuery is a very powerful cloud data analytics tool, it is not the only option for the Cloud Data Warehouse Software. When deciding, Users have to evaluate and make considerations from various aspects, such as the costs, performance, the ability to handle real-time workloads, and maybe other parameters to decide which cloud data warehouse software is the best to fit their needs.

There are also many other alternatives and competitors to Google BigQuery. Of these, the mainstream and reliable, as well as the most similar functionally to BigQuery, are Snowflake, Amazon Redshift, and Microsoft Azure Synapse Analytics. Although they differ in cost or implementation, they still have most of the common characteristics such as high scalability. They all use massively parallel processing (MPP) to process the storage structures for multiple operations at the same time. This not only accelerates the expansion and contraction of storage and computing resources but also implements the storage of data into compressed, columnar format, providing minimizing storage footprint and query characteristics for data. They guarantee reliable data replication, backup, recovery, and fast retrieval even in the event of a system outage or failure.[11][12]

**Conclusion**

Data is one of the most valuable assets of firms, organizations, or other businesses. The processing of data has a tremendous impact on the long-term success of an organization or industry. However, processing and storing a large scale of different data types (a form of text, audio, video, etc.) using traditional techniques and methods are very complicated, time-consuming, and costly. Moreover, the traditional data warehouses have some limitations in handling such kinds of data efficiently, which is why considering using cloud data warehouses.

In the project, we use the Bigquery of google Datawarehouse for analysing the COVID-19 large-scale dataset. This kind of Data (structured, semi-structured, and non-structured) from real-time sources can be loaded to the google cloud platform in a number of methods. By using BigQuery, we can flexibly and efficiently extract the necessary information from the data.

As demonstrated in the paper, BigQuery is a very powerful, enabling data exploration and analysis tool for querying, analyzing, and managing data with build-in features like machine learning, geospatial analysis, and business intelligence. [16] It is a serverless and fully managed platform, it manages the underlying software for clients as well as the infrastructure that prioritizes the high scalability as well as the high availability. [13]

BigQuery maximizes flexibility by separating the compute engine that analyzes data from the storage choices. Either storing and analyzing data within BigQuery is all convenient or using BigQuery to assess the data where it lives. BigQuery's federated queries makes it capable to read data from external sources while its streaming supports the continuous data updates as well. [16] BigQuery also provides extremely high cost-effectiveness and full-scan performance for data queries as it uniquely incorporates a massively parallel query engine that enables it to efficiently analyze large-scale datasets. Besides, it solves the parallel disk I/O problem by utilizing the cloud platform’s economy of scale. [14]

Moreover, BigQuery has native integrations with many data visualization tools such as Data Studio, Excel sheet Looker, Power BI, Tableau, etc. This means that BigQuery can visualize and perform real-time analysis directly using the streaming data without having to manually dealing with the queried results.

In a world where data acquisition is growing at an incredible rate, tools like BigQuery help to generate value from data. However, despite its unique advantages and powerful features, BigQuery is not perfect. For example, it’s not recommended for frequently changing data as its storage location and resource allocation is tied to Google's own services and by BigQuery’s processing limitations, it's best not to use it as a primary data warehouse; BigQuery mainly designed to handle massive amounts of data at super-fast speeds. So, for businesses that only have small data sets, BigQuery may not be the right solution.

In addition to BigQuery, there are many other accessible cloud data analytics tools similar to it. For example, Snowflake, Amazon Redshift and Microsoft Azure Synapse Analytics are also good choices for cloud data warehouse software.

**Reference**

[6] Introduction to loading data

<https://cloud.google.com/bigquery/docs/loading-data>

[7] The Benefits of Using BigQuery with Google Analytics Data.

<https://www.bounteous.com/insights/2021/06/29/benefits-using-bigquery-google-analytics-data/>

[8] [Google BigQuery: The Definitive Guide by Valliappa Lakshmanan, Jordan Tigani](https://learning.oreilly.com/library/view/google-bigquery-the/9781492044451/).

<https://www.oreilly.com/library/view/google-bigquery-the/9781492044451/ch01.html>

[9] Introduction to Google BigQuery

<https://dzone.com/articles/introduction-to-google-bigquery>

[10] BigQuery, a serverless data warehouse

<https://medium.com/sajari/bigquery-a-serverless-data-warehouse-9a4d623f3f43>

[11] Top 5 BigQuery Alternatives: A Challenge of Complexity

<https://blog.panoply.io/top-5-bigquery-alternatives>

[12] Top 10 BigQuery Alternatives & Competitors

<https://www.g2.com/products/google-bigquery/competitors/alternatives>

[13]  A Deep Dive Into Google BigQuery Architecture

<https://panoply.io/data-warehouse-guide/bigquery-architecture>

[14] Google BigQuery – Architecture Deep Dive and Key Features

[https://sarasanalytics.com/blog/what-is-google-bigquery](https://sarasanalytics.com/blog/what-is-google-bigquery#:~:text=Scalability%3A%20BigQuery%20relies%20on%20massively,thousands%20of%20machines%20per%20region)

[15] Google BigQuery Architecture: The Comprehensive Guide

<https://hevodata.com/blog/google-bigquery-data-warehouse>

[16] What is BigQuery?

<https://cloud.google.com/bigquery/docs/introduction>

<https://www.sisense.com/blog/the-benefits-of-combining-google-bigquery-and-bi/>

F1:

1. Stacked Bar chart of cumulative confirmed cases in each country’s population

(2) Pie chart of total confirmed cases as a percentage of the country's population

F2: Stacked bar chart of cumulative confirmed cases by age groups of each country

F3: Time-Series graph for the number of people confirmed until July 27, 2022图表, 折线图

描述已自动生成

F4:

Time-Series graph of