Msc. Big Data Analytics. **Date – 22/1/2021**

BigQuery (Google Cloud Platform) .

Abstract: BigQuery is Google Cloud's totally managed, petabyte-scale, and efficient analytics data warehouse that allows you to run analytics over immense amounts of data in close to real time.BigQuery is an enterprise information warehouse that solves this drawback by sanctioning super-fast SQL queries victimisation the process power of Google's infrastructure.

Introduction.

Google recently free BigQuery as a publized on the market service for any business or developer to use. This unleash created it potential for those outside of Google to utilize the facility of Dremel for his or her massive processing needs.BigQuery provides the core set of options on the market in Dremel to 3rd party developers. It will thus via a REST API, statement interface, Web UI, access management, data schema management and therefore the integration with Google Cloud Storage.

BigQuery design.

BigQuery's serverless design decouples capability and register and permits them to scale autonomously on interest. This construction offers each tremendous ability and price controls for shoppers since they do not ought to keep their expensive register assets prepared for action perpetually. This is often completely completely different from standard hub primarily based cloud info storeroom arrangements or on-premise massively multiprocessing (MPP) frameworks. this system likewise permits shoppers of any size to hold their info into {the info|the knowledge|the information} distribution center and start breaking down their information utilizing normal SQL while not stressing over data set tasks and framework planning.

Storage is Colossus, Google’s global storage system.

BigQuery leverages the columnar storage format and compression algorithmic rule to store data in Colossus, optimized for reading massive amounts of structured data.

Colossus additionally handles replication, recovery (when disks crash) and distributed management (so there's no single purpose of failure). Colossus permits BigQuery users to scale to dozens of petabytes of data keep seamlessly, while not paying the penalty of attaching rather more expensive cipher resources as in traditional data warehouses.

Loading data.

There ar many ways in which to ingest data into BigQuery:

- Batch load a collection of data records.
- Stream individual records or batches of records.
- Use queries to get new data and append or write the results to a table.
- Use a third-party application or service.

Batch loading: With batch loading, you load the supply data into a BigQuery table in an exceedingly single batch operation. For instance, the info supply might be a CSV file, AN outer data set, or a bunch of log documents.

Streaming Data: With streaming, you send the information every record successively or in batches. you'll either compose code that calls the streaming API squarely, otherwise you will utilize Dataflow with the Apache Beam SDK to line up a streaming pipeline.

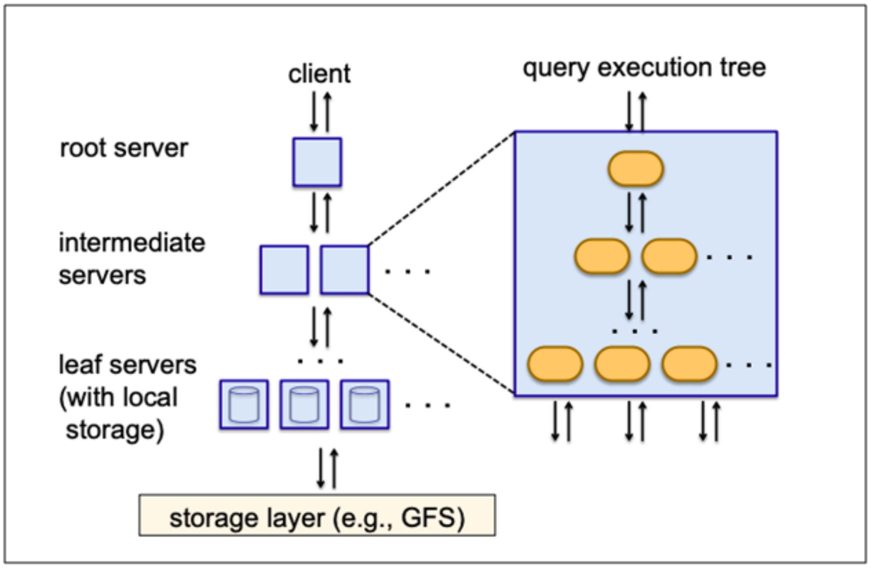
Generated data: you'll use data manipulation language (DML) statements to perform bulk inserts into AN existing table or store question ends up in a brand new table.

Exporting Data:

The subsequent limits apply to jobs created mechanically by exporting data using the bq command-line tool or the Cloud Console. the bounds conjointly apply to export jobs submitted programmatically by using the load-type jobs.insert API methodology.

How the Query Gets Executed?

BigQuery depends on Borg for data processing. Borg simultaneously instantiates hundreds of Dremel jobs across required clusters made up of thousands of machines. In addition to assigning compute capacity for Dremel jobs, Borg handles fault-tolerance as well.



Architecture forms a gigantically parallel distributed tree for pushing down a query to the tree and aggregating the results from the leaves at a blazingly fast speed

BigQuery versus MapReduce

The difference between the two:

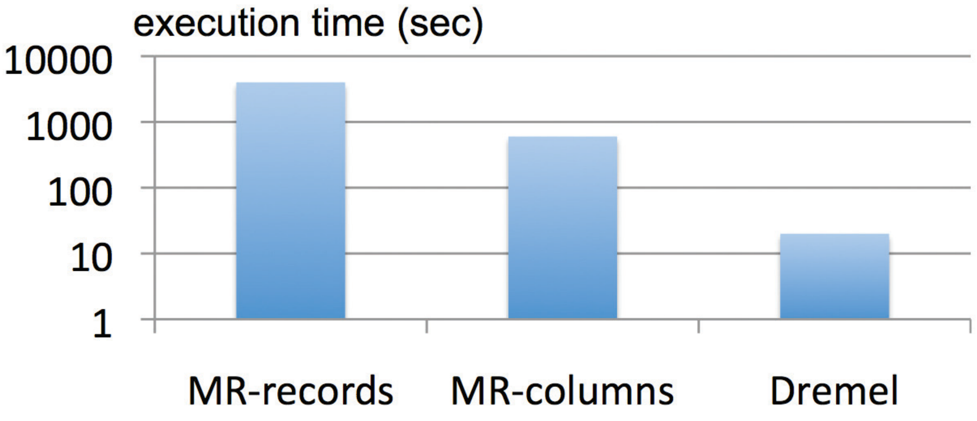
- Dremel is designed as an interactive data analysis tool for large datasets.
- MapReduce is designed as a programming framework to batch process large datasets.

MapReduce is a distributed computing technology that allows you to implement custom "mapper" and "reducer" functions programmatically and run batch processes with them on hundreds or thousands of servers concurrently. By using MapReduce, enterprises can cost-effectively apply parallel data processing on their Big Data in a highly scalable manner, without bearing the burden of designing a large distributed computing cluster from scratch.

MapReduce is designed as a batch processing framework, so it's not suitable for ad hoc and trial-and-error data analysis.

BigQuery is designed to handle structured data using SQL. For example, you must define a table in BigQuery with column definition, and then import data from a CSV (comma separated values) file into Google Cloud Storage and then into BigQuery.	MapReduce is a better choice when you want to process unstructured data programmatically. The mappers and reducers can take any kind of data and apply complex logic to it. MapReduce can be used for applications such as data mining.

The following figure shows a comparison of execution times between MapReduce and Dremel. As you can see, there is a difference in orders of magnitude.



Key Differences	**BigQuery**	**MapReduce**
What is it?	Query service for large datasets	Programming model for processing large datasets
Common use cases	Ad hoc and trial-and- error interactive query of large dataset for quick analysis and troubleshooting	Batch processing of large dataset for time-consuming data conversion or aggregation
Sample use cases		
OLAP/BI use case	Yes	No
Data Mining use case	Partially (e.g. preflight data analysis for data mining)	Yes
Very fast response	Yes	No (takes minutes - days)
Easy to use for non-programmers (analysts,tech support, etc)	Yes	No (requires Hive/Tenzing)
Programming complex data processing logic	No	Yes
Processing unstructured data	Partially (regular expression matching on text)	Yes
Data handling		

Handling large results / Join large table	No (as of Sept 2012)	Yes
Updating existing data	No	Yes

How to do Machine Learning on BigQuery?

BigQuery ML enables data scientists and data analysts to build and operationalize ML models on planet-scale structured or semi-structured data, directly inside BigQuery, using simple SQL—in a fraction of the time. Export BigQuery ML models for online prediction into Cloud AI Platform or your own serving layer

Real-time analytics

BigQuery's high-speed streaming insertion API provides a strong foundation for real-time analytics, creating your latest business data immediately offered for analysis. you'll be able to conjointly leverage Pub/Sub and Dataflow to stream data into BigQuery.

Conclusion

BigQuery could be a query service that permits us to run SQL-like queries against multiple terabytes of data during a matter of seconds. If you've got structured data, BigQuery is the most suitable choice to go for. whereas MapReduce is appropriate for long-running batch processes like data mining, BigQuery is that the most suitable option for impromptu OLAP/BI queries that need results as quick as possible

Reference

- ➔ An Inside Look at Google BigQuery by Kazunori Sato, Solutions Architect, Cloud Solutions team
- ➔ https://cloud.google.com/bigquery/docs