Isaac Araya Solano | Carnet: 2018151703

Resumen #7

An Inside Look at the Google BigQuery

Google Big Query is a fully-managed and cloudbased interactive query service for massive datasets. BigQuery is the external implementation of one of the company's core technologies whose code name is Dremel.

Dremel is a query service that allows you to run SQL-like queries against very, very large data sets and get accurate results in mere seconds.

BigQuery: Externalization of Dremel

The difference between Dremel and Google BigQuery. BigQuery is the public implementation of Dremel that was recently launched to general availability. BigQuery provides the core set of features available in Dremel to third party developers. It does so via a REST API, a command line interface, a Web UI, access control and more, while maintaining the unprecedented query performance of Dremel.

Columnar Storage and Tree Architecture of Dremel

Columnar Storage

Dremel stores data in its columnar storage, which means it separates a record into column values and stores each value on different storage volume. Columnar storage has the following advantages:

- Traffic minimization: only required columns are scanned
- Higher compression ratio: Columnar storage can achieve a compression ratio of 1:10 Columnar storage has the disadvantage of not working efficiently when updating existing records. In the case of Dremel, it simply doesn't support any update operations.

Tree Architecture

One of the challenges Google had in designing Dremel was how to dispatch queries and collect results across tens of thousands of machines in a matter of seconds. The challenge was resolved by using the Tree architecture. The architecture forms a massively parallel distributed tree for pushing down a query to the tree and then aggregating the results from the leaves at a blazingly fast speed.

Dremel: Key to Run Business at "Google Speed"

Examples of applications include1:

- Analysis of crawled web documents
- Tracking install data for applications in the Android Market
- Crash reporting for Google products
- OCR results from Google Books
- Spam analysis
- Debugging of map tiles on Google Maps

- Tablet migrations in managed Bigtable instances
- Results of tests run on Google’s distributed build system
- Disk I/O statistics for hundreds of thousands of disks
- Resource monitoring for jobs run in Google’s data centers
- Symbols and dependencies in Google’s codebase As you can see from the list, Dremel has been an important

BigQuery versus MapReduce

This is the difference:

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

BigQuery and MapReduce Comparison

	BigQuery	**MapReduce**
What is it?	It is a query service for large datasets.	Programming model for processing large datasets
Common uses:	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
OLAP/BI:	Yes	No
Data Mining:	Partially	Yes
Very fast Response:	Yes	No
Easy to use for non-programmers:	Yes	No
Programming complex data processing logic:	No	Yes
Processing unstructured data:	Partially	Yes
Handling large results/Join large table:	No	Yes
Updating existing data:	No	Yes

For example, users may want to apply these criteria to decide what technology to use:

Use BigQuery

- Finding particular records with specified conditions.
- Quick aggregation of statistics with dynamically-changing conditions.
- Trial-and-error data analysis.

Use MapReduce

- Executing a complex data mining on Big Data which requires multiple iterations and paths of data processing with programmed algorithms.
- Executing large join operations across huge datasets.
- Exporting large amount of data after processing.

Of course, you can make the best use of both technologies by combining them to build a total solution. For example,

- Use MapReduce for large join operations and data conversions, then use BigQuery for quick aggregation and ad-hoc data analysis on the result dataset.
- Use BigQuery for a preflight check by quick data analysis, then write and execute MapReduce code to execute a production data processing or data mining.

Data Warehouse Solutions and Appliances for OLAP/BI

In OLAP/BI, you roughly have the following three alternatives for increasing the performance of Big Data handling.

- Relational OLAP (ROLAP): Based on relational databases. Hardware dependant for speed.
- Multidimensional OLAP (MOLAP): Cube or data marts based on dimensions. Demands extensive time of BI engineers before it can be used.
- Full scan: optimal solution for ad hoc queries or trial-and-error data analysis. Disk I/O Throughput is the key for full scan performance.

Traditional data warehouse solutions and appliances have tried to achieve better disk I/O throughput with the following technologies:

- In-memory database: Fill the DB appliance with memory modules.Best solution if you do not have cost restrictions.
- Columnar Storage
- Parallel disk I/O

BigQuery solves the parallel disk I/O problem by utilizing the cloud platform’s economy of scale.

BigQuery’s Unique Abilities

- Cloud-Powered Massively Parallel Query Service: lower costs
- How to Import Big Data: done by 2 steps, upload data to GCP and import files by command-line, web UI or API.