> Bases de Datos 2, Resumen 7
> Miguel Ku Liang - 2019061913

An Inside Look at Google BigQuery

How Google Handles Big Data Daily Operations

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

BigQuery: Externalization of Dremel

BigQuery is the implementation of Dremel that was launched to general availability. It provides core features of Dremel.

Dremel Can Scan 35 Billion Rows Without an Index in Tens of Seconds

Dremel shares Google's infrastrucutre to parallelize queries and run them on thousands of servers simultaneously. It has super high scalability.

Columnar Storage and Tree Architecture of Dremel

Dremel uses two core technologies:

- Columnar Storage: data is stored in a columnar storage, achieving very high compression ratio and scan throughput. Some advantages are: traffic minimization and higher compression ratio. One disadvantage is that it doesn't work efficiently when updating existing records.
- Tree Architecture: for dispatching queries and aggregating results.

Dremel: Key to Run Business at “Google Speed”

Some applications that use Dremel are: 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 and symbols and dependencies in Google’s codebase.

And what is BigQuery?

BigQuery is a service for any business or developer to use. It provides core features available in Dremel. They share the same architecture and performance characteristics.

BigQuery versus MapReduce

- 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 Limitations

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

BigQuery and MapReduce Comparison

- BigQuery: finding particular records with specified conditions, quick aggregation of statistics with dynamically-changin conditions and trial-and-error data analysis.
- MapReduce: executing complex data mining on Big Data, executing large join operations, exporting large amount of data.

Data Warehouse Solutions and Appliances for OLAP/BI

In OLAP/BI, you have three alternatives for increasing performance of Big Data handling:

- Relational OLAP (ROLAP): OLAP based on relational databases. To make relational databases faster, build indices before running OLAP queries.
- Multidimensional OLAP (MOLAP): OLAP designed to build data cubes or data marts based on dimensions predefined during the design phase.
- Full scan: accessing all records on disk drives without indexing or pre-aggregated values. Disk I/O throughput is the key to full scan performance. To achieve better throughput, use: in-memory database or flash storage, columnar storage and parallel disk I/O.

BigQuery's Unique Abilities

BigQuery provides high cost-effectiveness and full scan performance for ad hoc queries because of the combination of a massively parallel query engine.

How to Import Big Data

There are two steps:

- Upload the data to Google Cloud Storage
- Import the files to BigQuery

Why Use the Google Cloud Platform?

BigQuery requires no capacity planning, provisioning, 24x7 monitoring, nor manual security patch updates, reducing the cost of ownership. Growing datasets is not a problem. BigQuery's REST API enables you to build App Engine-based dashboards and mobile frontends.