Google BigQuery

Investigation; Jeff Kao; October 13, 2017

What is BigQuery?

Overview

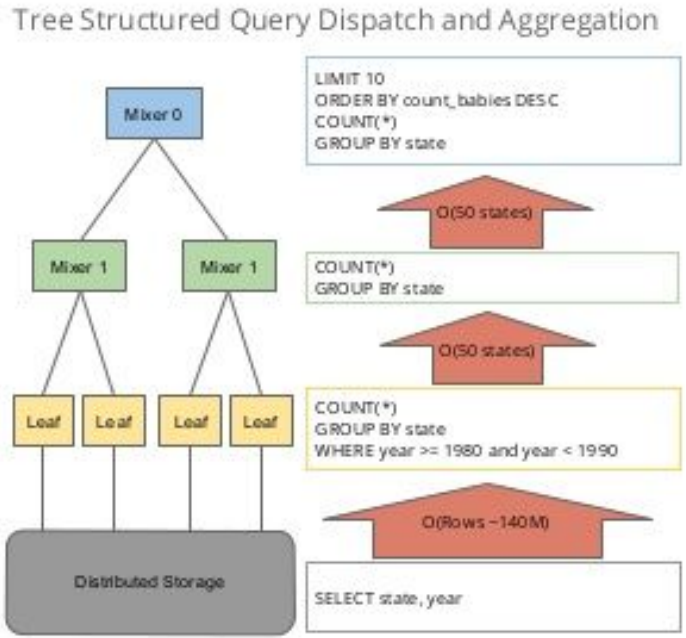
Publicly available since November 2011

Infrastructure as a Service (*IaaS*)

Super-fast SQL queries on large datasets (TBs).

Distributed, takes advantage of Google's infrastructure (Dremel, breaks job up into pieces and re-assembles the results).

How BigQuery works




```
SELECT
    state, COUNT(*) count_babies
FROM [publicdata:samples.natality]
WHERE
    year >= 1980 AND year < 1990
GROUP BY state
ORDER BY count_babies DESC
LIMIT 10
```

Valid: This query will process 4.06 TB when run.

RUN QUERY | Save Query | Save View | Format Query | Show Options

Query complete (35.3s elapsed, 4.06 TB processed)

Benefits

Solves challenges associated with warehousing and querying big data sets.

Everything happens under the hood.

Scalable.

No need to buy servers/upgrade software/monitor uptime/maintain infrastructure.

When to use BigQuery?

Use Cases

What it's used for:

After ingesting and processing the data, the resulting data set can be stored in BigQuery for analysis.

Fast ad-hoc queries.

Storage at the end of the big data pipeline.

What it's NOT used for:

Frequent/real-time read-write operations.

Very small datasets (minimum unit of charge by MB queried).

Access points

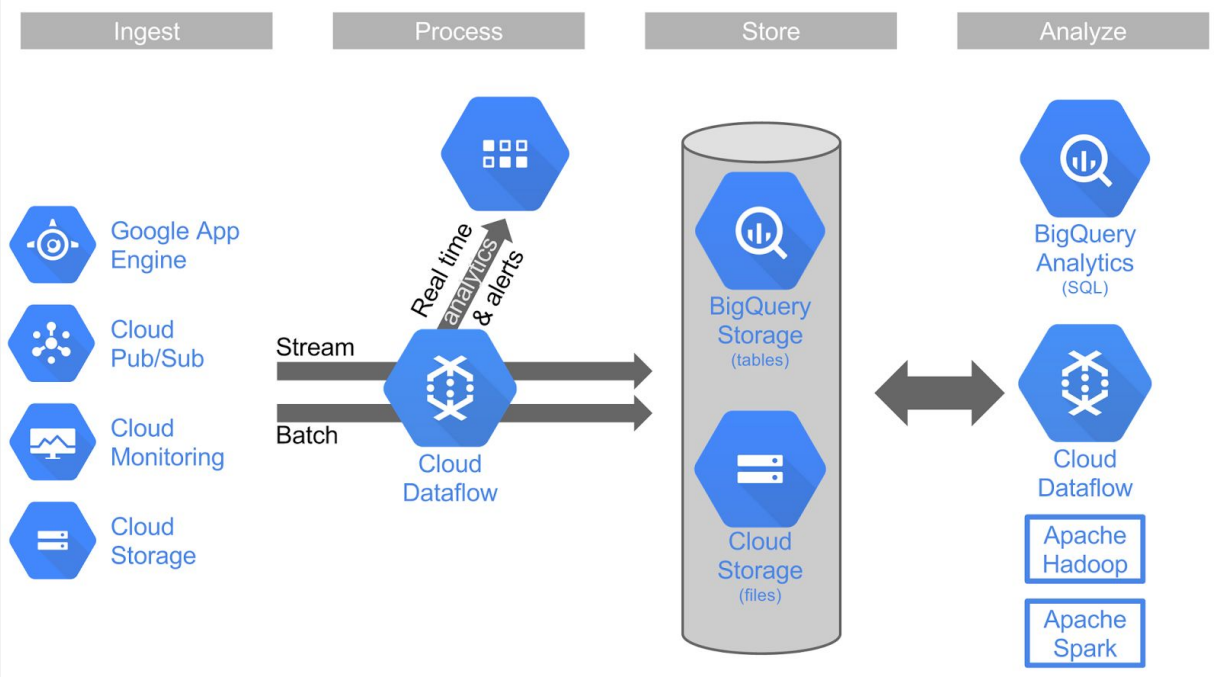
Web UI

Command-line tool

REST API

Client libraries such as Java, .NET, or Python

3rd Party Integrations (e.g., Tableau, MapReduce, Google Analytics, SAP Analytics Cloud)



BigQuery Web UI Demo

Other Fun Resources

Google Developer Advocate (https://medium.com/@hoffa)

Public BigQuery Datasets (https://www.reddit.com/r/bigquery/wiki/datasets)

Discussion/Sample Queries on BigQuery Datasets (https://www.reddit.com/r/bigquery/top/?sort=top&t=all)

Questions?