Google BigQuery

Johannes Ahlmann
johannes@sensatus.io

GDG

Agenda

- Why BigQuery
- Use Cases
- Architecture
- Alternatives
- Demo Time

(Material available on GH codinguncut/talks)

Use Cases

- bigquery-public-data:
  - crypto_bitcoin (1.1TB)
  - crypto_ethereum (1.4TB)
  - github_repos (3.7TB source code)
- fh-bigquery:
  - wikipedia_v3 (11TB pageviews)
  - reddit_comments (1.8TB)
  - reddit_extracts (5.6TB)
  - wikidata (3.5TB)
  - pypi (20TB package installs)
  - stackoverflow (5.2TB)
  - stackoverflow_archive (2TB)
  - wikidata (3.5TB)

```
SELECT sum(size_bytes)/pow(10,12) as size
FROM `<dataset>.__TABLES__`;
```

- gdelt-bq:gdeltv2 (79TB)
- githubarchive:day (6TB)
- httparchive:latest (27.5TB)

Input Data Formats

Data Sources

Google
Cloud Storage

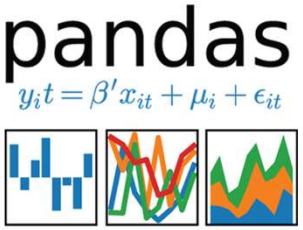
Google Drive

Google
Analytics

Google Cloud Dataflow

Google
BigQuery

Google Data Studio



Google
Sheets

DEVFEST 2019

GDG

Dremel

- Dremel: Interactive Analysis of Web-Scale Datasets
  - Interactive ad-hoc query system for in-situ nested data, near-realtime
  - Inspiration for Apache Drill, Apache Impala
  - Columnar
  - 1s Query Planning
  - Query Execution
  - Strongly-typed nested records
  - MR ~ minutes to hours, Dremel ~ seconds
- Cost
  - $5 / TB processed
  - active storage - $0.02/ GB/ month
  - long-term storage (90 days) $0.01 /GB/ month

Query Execution

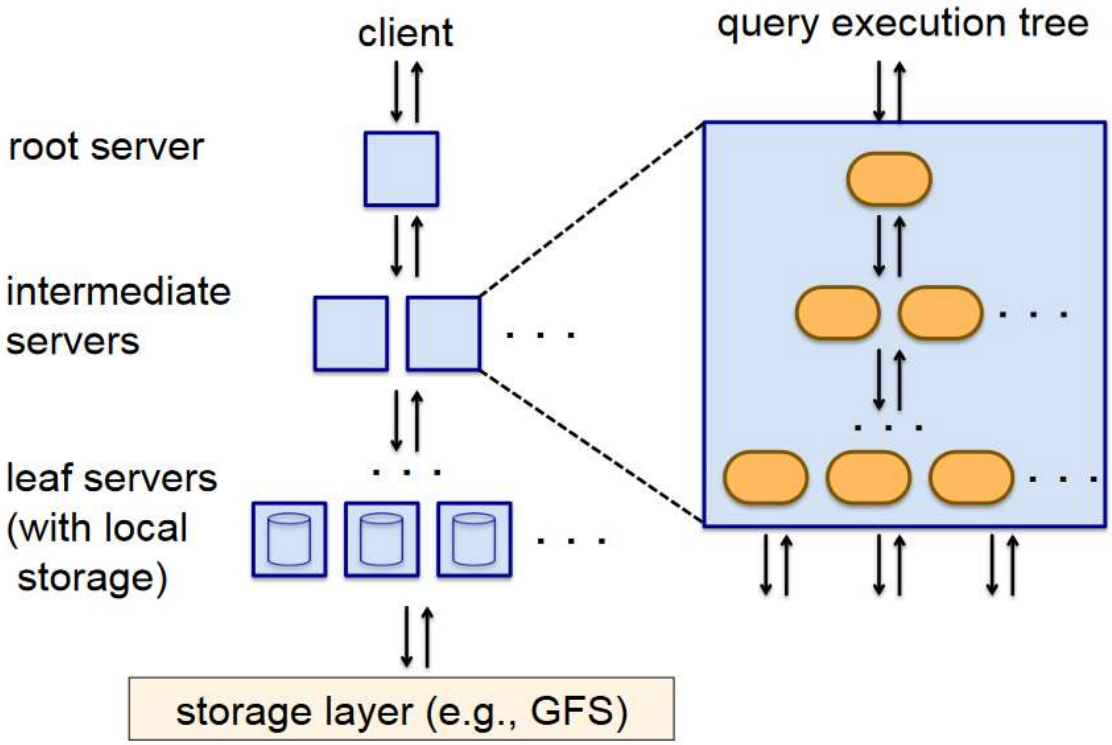
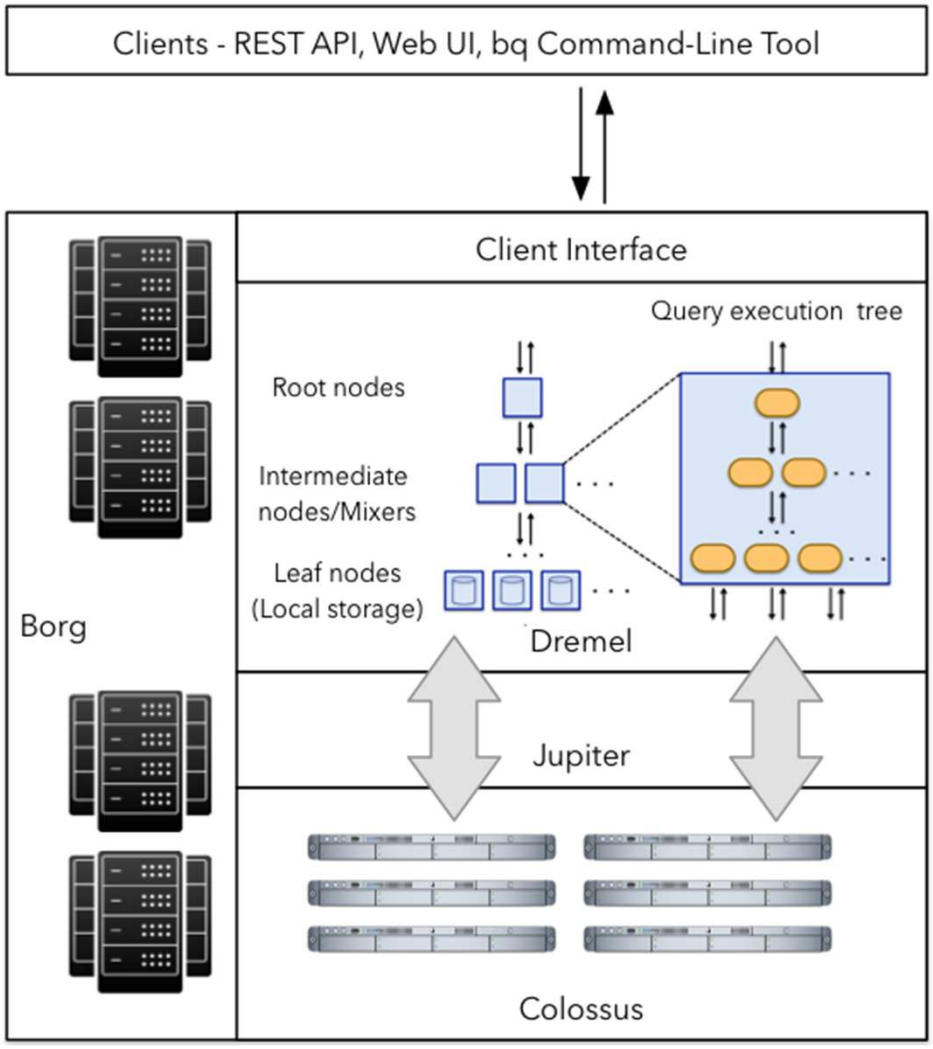


Figure 7: System architecture and execution inside a server node

Architecture

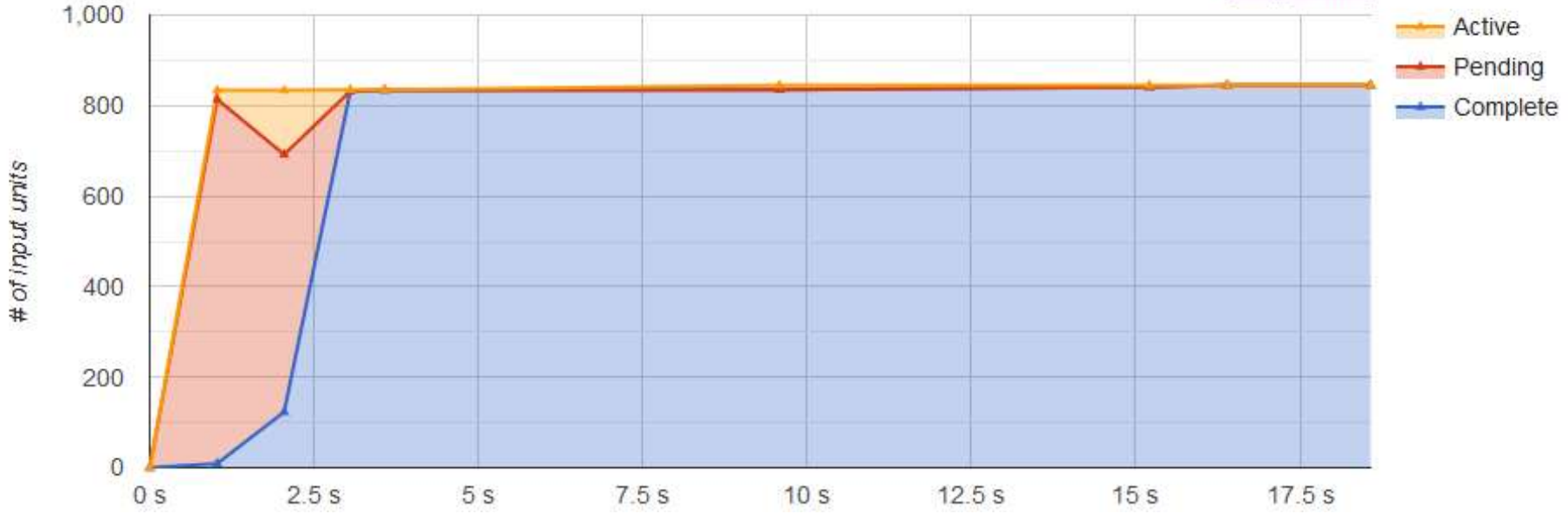


Results | Details

Job ID	asdf-167312:US.bquijob_56dbc3e_16eacbde2e0
Creation Time	Nov 27, 2019, 12:02:34 PM
Start Time	Nov 27, 2019, 12:02:34 PM
End Time	Nov 27, 2019, 12:02:53 PM
User	johannes@fluquid.com
Bytes Processed	11.8 GB
Bytes Billed	11.8 GB
Slot Time (ms)	341 K
Query Priority	Interactive
Destination Table	asdf-167312:_a10c5db64880e8cc031098b2d6a612efbe4613d0.anon35465fa0ce25fb17d754a62d9b8c3ddd46e7bdfb

Timeline

Collapse chart



Alternatives

Postgres

Presto

Demo Time