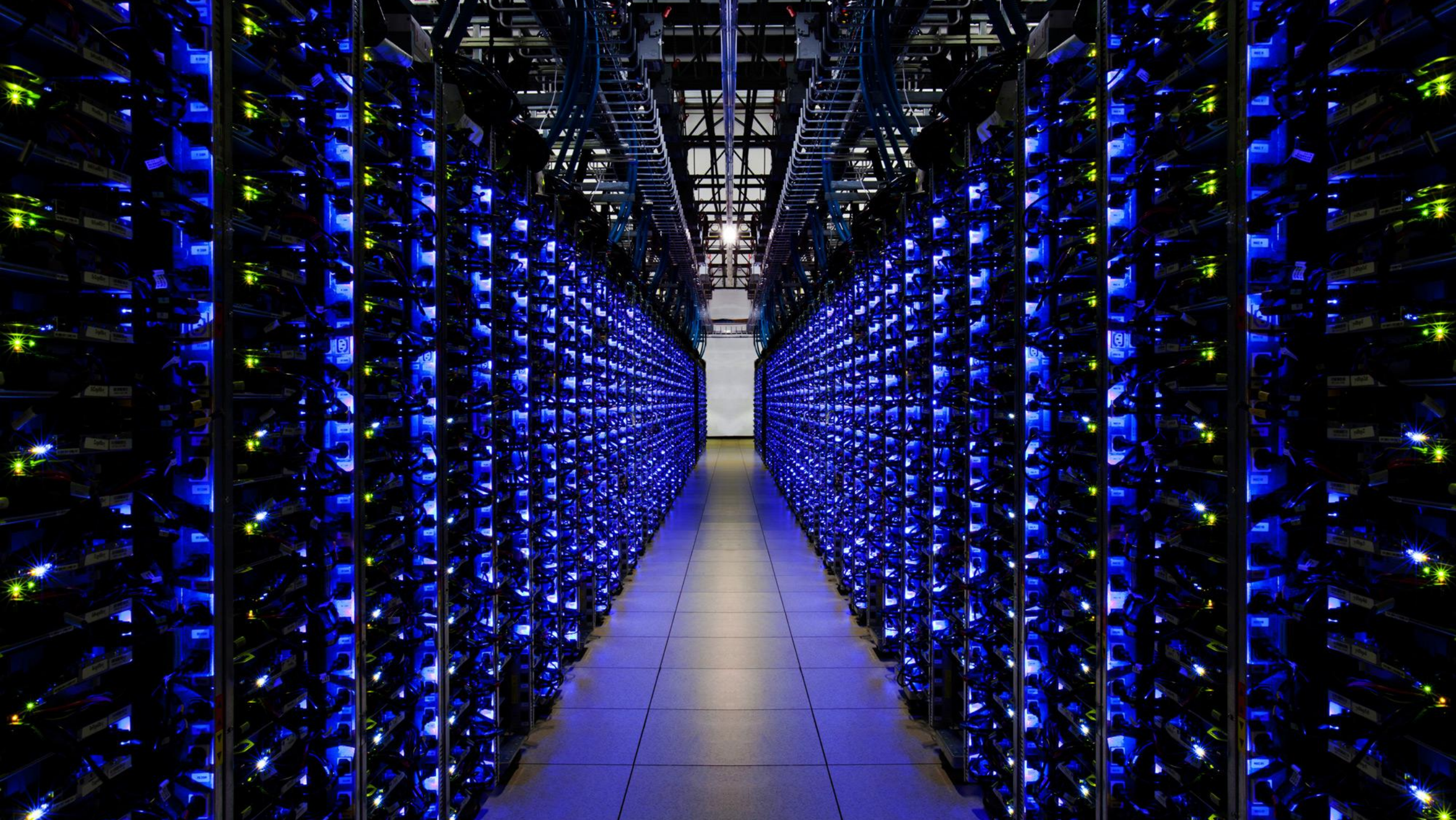

Google
Developers



System Log Analysis Using Google BigQuery

Gustavo Franco
Site Reliability Engineer @ Google



System Log Analysis using Google BigQuery



Mapreduce



Sawzall



System Log Analysis using Google BigQuery



Dremel



BigQuery



System Log Analysis using Google BigQuery

 Petabytes of data

 < 1 minute responses



System Log beyond syslog



Web server logs



Application logs



```
SELECT
   taskname,
   datacenter
FROM
   systemlog.today
WHERE
   severity='warning'
```



taskname	datacenter
"Crawler"	"SD"

1 row in result set (10.2 s)
Scanned 100GB



Agenda

- Do I need BigQuery?
- Mapreduce is flexible but heavy
- BigQuery Architecture
- Demo
- Pricing
- Homework
- Q&A

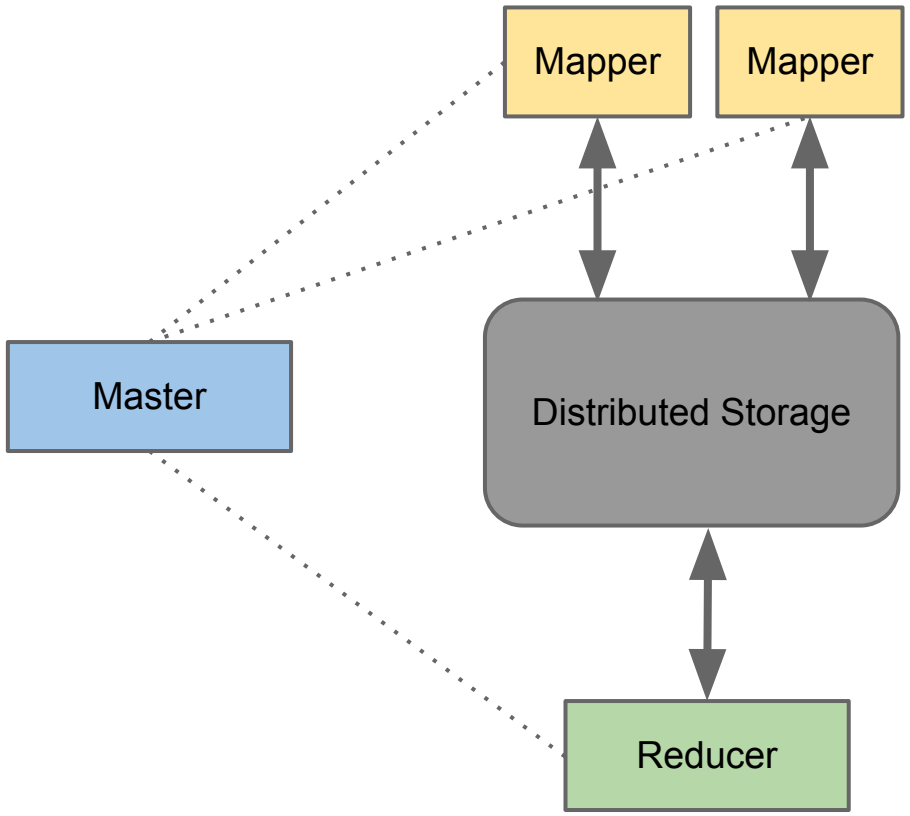


Do I need BigQuery?

- Using grep alone won't scale
- Let's consider {My,Postgre}SQL - OLAP x OLTP
- Hard to beat CAPEX. Bigquery is SaaS.
- Log analysis to enhance the product not just devops



MapReduce is Flexible but Heavy



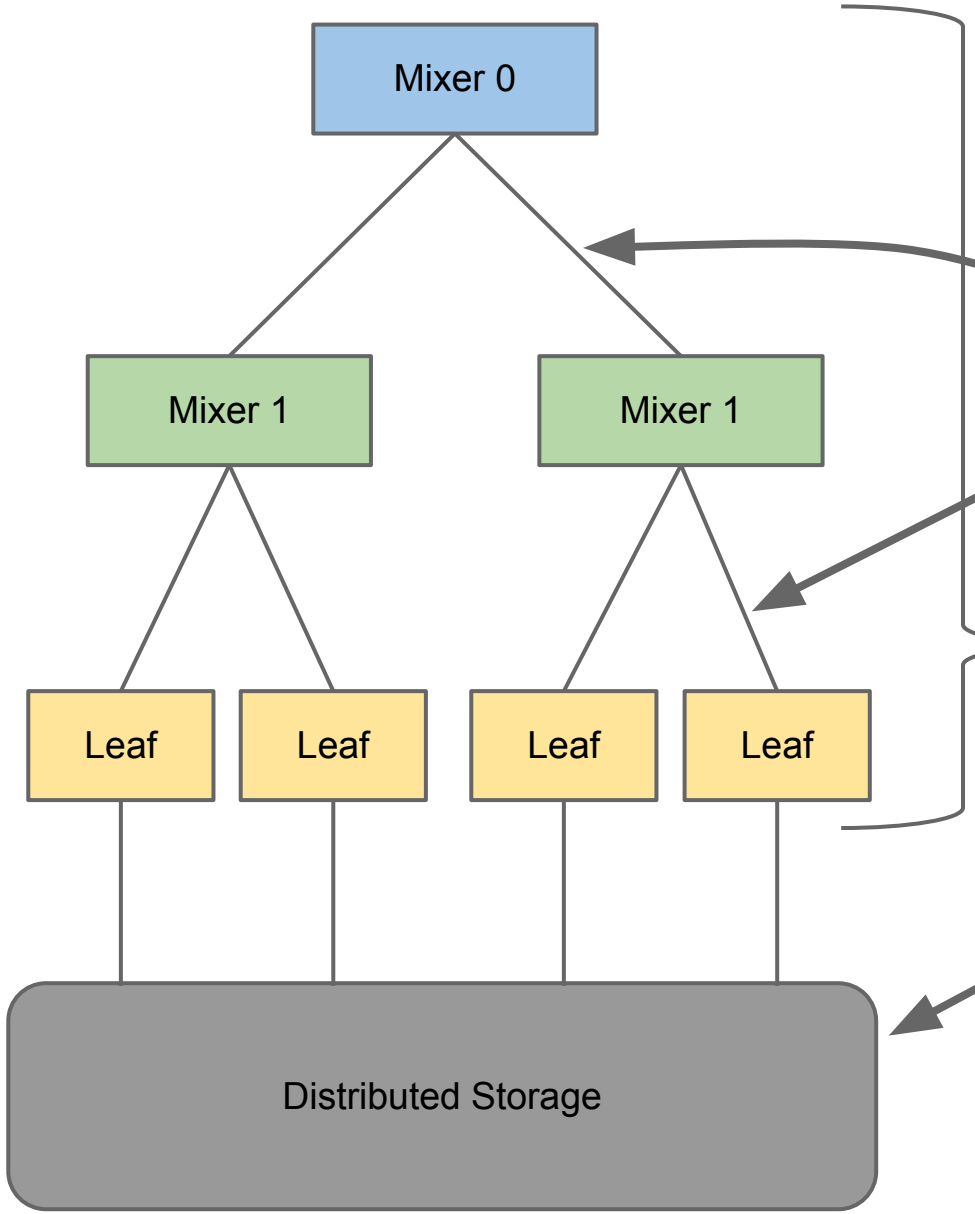
- Master constructs the plan and begins spinning up workers
- Mappers read and write to distributed storage
- Map -> Shuffle -> Reduce
- Reducers read and write to distributed storage





BigQuery architecture

Dremel Architecture



- Partial Reduction
- Diskless data flow
- Long lived shared serving tree
- Columnar Storage



Demo



COMPOSE QUERY

Query History

Job History

lisa

- weblogs
  - lisa100M
  - lisa10M
- publicdata:samples

New Query

```
1 SELECT ip, request, count(*) as hits FROM weblogs.lisa10M where REGEXP_MATCH(ip,'^10.') group by request, ip
```

RUN QUERY

Query complete (5.3s elapsed, 295 MB processed)

Query Results

2:33pm, 6 Dec 2012

Download as CSV | Save as Table | Chart View

Row	ip	request	hits
1	10.0.0.4	GET /admin HTTP/1.0	512197
2	10.0.0.4	GET / HTTP/1.0	512089
3	10.0.0.4	GET /foo HTTP/1.0	511770
4	10.0.0.2	GET /foo HTTP/1.0	512653
5	10.0.0.1	GET /admin HTTP/1.0	513121
6	10.0.0.3	GET / HTTP/1.0	513149
7	10.0.0.3	GET /bigquery HTTP/1.0	510678
8	10.0.0.3	GET /admin HTTP/1.0	512260
9	10.0.0.2	GET / HTTP/1.0	513030
10	10.0.0.1	GET /foo HTTP/1.0	510990
11	10.0.0.4	GET /bigquery HTTP/1.0	512515
12	10.0.0.2	GET /bigquery HTTP/1.0	510577
13	10.0.0.2	GET /admin HTTP/1.0	513255
14	10.0.0.3	GET /foo HTTP/1.0	510535

COMPOSE QUERY

Query History

Job History

lisa

- weblogs
  - lisa100M
  - lisa10M
- publicdata:samples

New Query

```
1 SELECT ip, request, count(*) as hits FROM weblogs.lisa100M where REGEXP_MATCH(ip,'^10.') group by request, ip
```

RUN QUERY

Query complete (11.1s elapsed, 4.60 GB processed)

Query Results

2:36pm, 6 Dec 2012

Download as CSV | Save as Table | Chart View

Row	ip	request	hits
1	10.0.0.2	GET /foo HTTP/1.0	8181680
2	10.0.0.2	GET / HTTP/1.0	8185779
3	10.0.0.1	GET /admin HTTP/1.0	8182677
4	10.0.0.1	GET /bigquery HTTP/1.0	8182814
5	10.0.0.1	GET / HTTP/1.0	8184214
6	10.0.0.4	GET /admin HTTP/1.0	8178455
7	10.0.0.1	GET /foo HTTP/1.0	8183848
8	10.0.0.3	GET /foo HTTP/1.0	8184429
9	10.0.0.3	GET /bigquery HTTP/1.0	8179769
10	10.0.0.4	GET /bigquery HTTP/1.0	8181731
11	10.0.0.4	GET /foo HTTP/1.0	8182757
12	10.0.0.2	GET /bigquery HTTP/1.0	8178372
13	10.0.0.3	GET /admin HTTP/1.0	8179085
14	10.0.0.3	GET / HTTP/1.0	8177457

How does it work?

- Install gsutil (optional) and bq
- Log to CSV then upload to Cloud Storage
- bq load
- Query



How does it work?

- Logrotate then have daily tables
- Append data is supported
- Copy tables too
- Same for gzipped logs



Pricing

- https://cloud.google.com/pricing/big-query
- First 100GB of data processed per month is at no charge (Dec '12)
- Query Processing + Storage in BigQuery
- Cloud Storage pricing is different; also has a free tier



Homework

- http://goo.gl/JkhFC
- Dashboards using Google Visualization API



Q&A -
Homework: http://goo.gl/JkhFC

System Log Analysis using Google BigQuery



http://cloud.google.com/

contact stratus+lisa@google.com


Google
Developers