



Google
Developers



Crunching Big Data with Big Query

Ryan Boyd, Developer Advocate
Jordan Tigani, Software Engineer

How BIG is big?

1 million rows?

1 million
1 million
1 million
1 million
1 million
1 million
1 million
1 million
1 million
1 million
1 million

10 million rows?

100 million rows?

500 million rows?

Big Data at Google

 **60** hours

 **100 million** gigabytes

 **425 million** users









Google's internal technology: Dremel

Big Data at Google - Finding top installed market apps

```
SELECT
  top(appId, 20) AS app,
  count(*) AS count
FROM installlog.2012;
ORDER BY
  count DESC
```

Result in ~20 seconds!



Big Data at Google - Finding slow servers

```
SELECT
  count(*) AS count, source_machine AS machine
FROM product.product_log.live
WHERE
  elapsed_time > 4000
GROUP BY
  source_machine
ORDER BY
  count DESC
```

Result in ~20 seconds!



BigQuery gives you this power



Store data with reliability, redundancy and consistency



Go from data to meaning



At scale ...



Quickly!



How are developers using it?



Game and social media analytics



Infrastructure monitoring



Advertising campaign optimization



Sensor data analysis



Agenda

- Show the power
- Loading your data
- Running your queries
- Underlying architecture design
- Advanced queries





Let's dive in!

BigQuery UI

The screenshot displays the Google BigQuery interface. On the left, there is a sidebar with a 'COMPOSE QUERY' button and a navigation menu. The main area is titled 'Table Details: revisions' and includes a 'Query Table' button. Below the title, there is a 'Preview' section with a dropdown arrow and the text 'Click to preview table data'. This section contains a table with 13 columns: Row, title, id, language, wp_namespace, is_redirect, contributor_ip, contributor_id, contributor_username, timestamp, is_minor, is_bot, and revision_id. The table lists five rows of data. Below the table is a 'Schema' section with a table listing the column names and their data types.

COMPOSE QUERY

Query History
Job History

RB BigQuery Jazz

- GeoLiteCity
- campaign_finance
- wikipedia_geo
 - 2008_election_results
 - revisions
- publicdata:samples
- googledata:borgmachines
- googledata:buganizer
- googledata:forbin
- googledata:grants
- googledata:grid
- googledata:sponge
- googledata:spore
- googledata:svs

Table Details: revisions

Query Table

Preview ▼ Click to preview table data

Row	title	id	language	wp_namespace	is_redirect	contributor_ip	contributor_id	contributor_username	timestamp	is_minor	is_bot	revision_id
1	AccessibleComputing	10	en	0	false		99	RoseParks	980043141	0	null	233192
2	AccessibleComputing	10	en	0	false		0	Conversion script	1014651791	1	null	862220
3	AccessibleComputing	10	en	0	false		7543	Ams80	1051305518	1	null	15898945
4	AccessibleComputing	10	en	0	false		516514	Nzd	1149350141	1	null	56681914
5	AccessibleComputing	10	en	0	false		750223	Rory096	1157685364	0	null	74466685

Schema

title	STRING
id	INTEGER
language	STRING
wp_namespace	INTEGER
is_redirect	BOOLEAN
contributor_ip	STRING
contributor_id	INTEGER
contributor_username	STRING
timestamp	INTEGER
is_minor	INTEGER
is_bot	STRING
revision_id	INTEGER

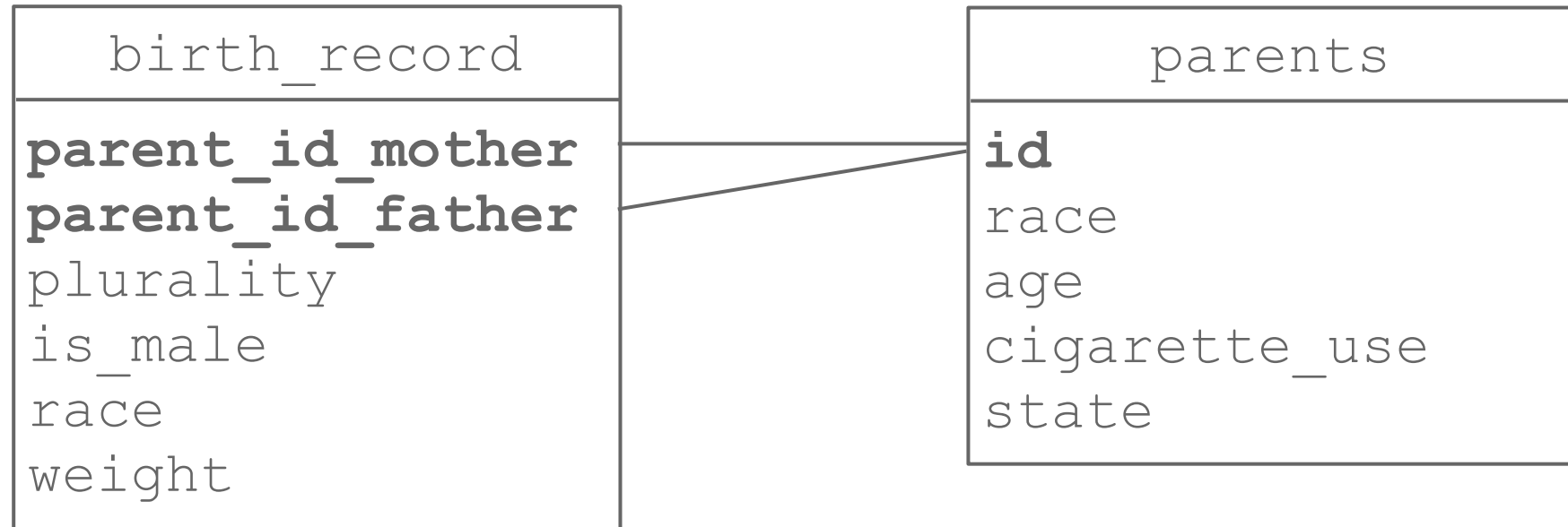


bigquery.cloud.google.com



Loading your Data

Ingestion: Data format



Ingestion: Data format

birth_record
mother_race
mother_age
mother_cigarette_use
mother_state
father_race
father_age
father_cigarette_use
father_state
plurality
is_male
race
weight



Ingestion: Data format

```
1969,1969,1,20,,AL,TRUE,1,7.813,AL,1,20,true  
1971,1971,5,7,,NY,FALSE,1,7.213,MA,5,7,true  
2001,2001,12,5,,CA,TRUE,2,6.427,CA,12,5,true
```

CSV





Running your Queries

Libraries

- Java
- Python
- .NET
- PHP
- JavaScript
- Apps Script
- ... more ...



It's REST

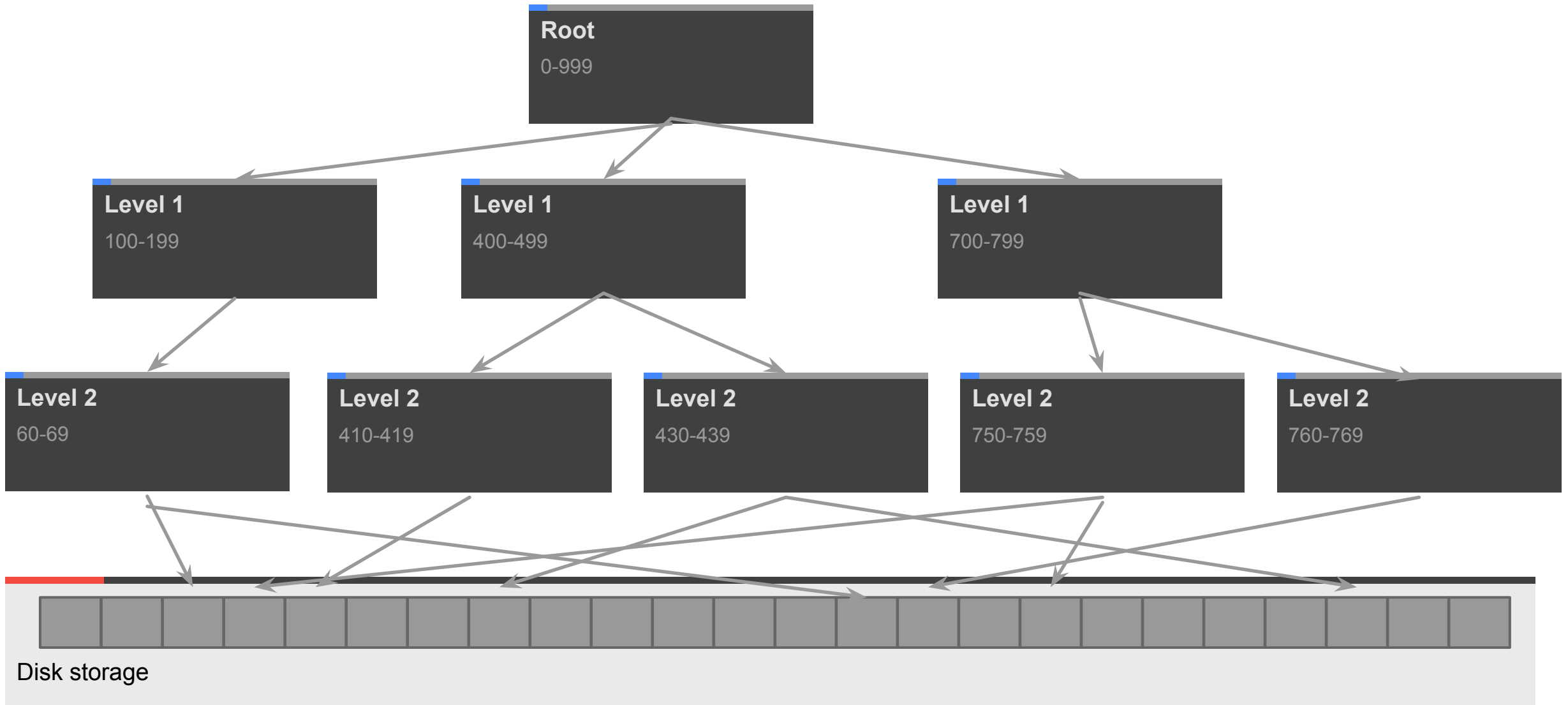




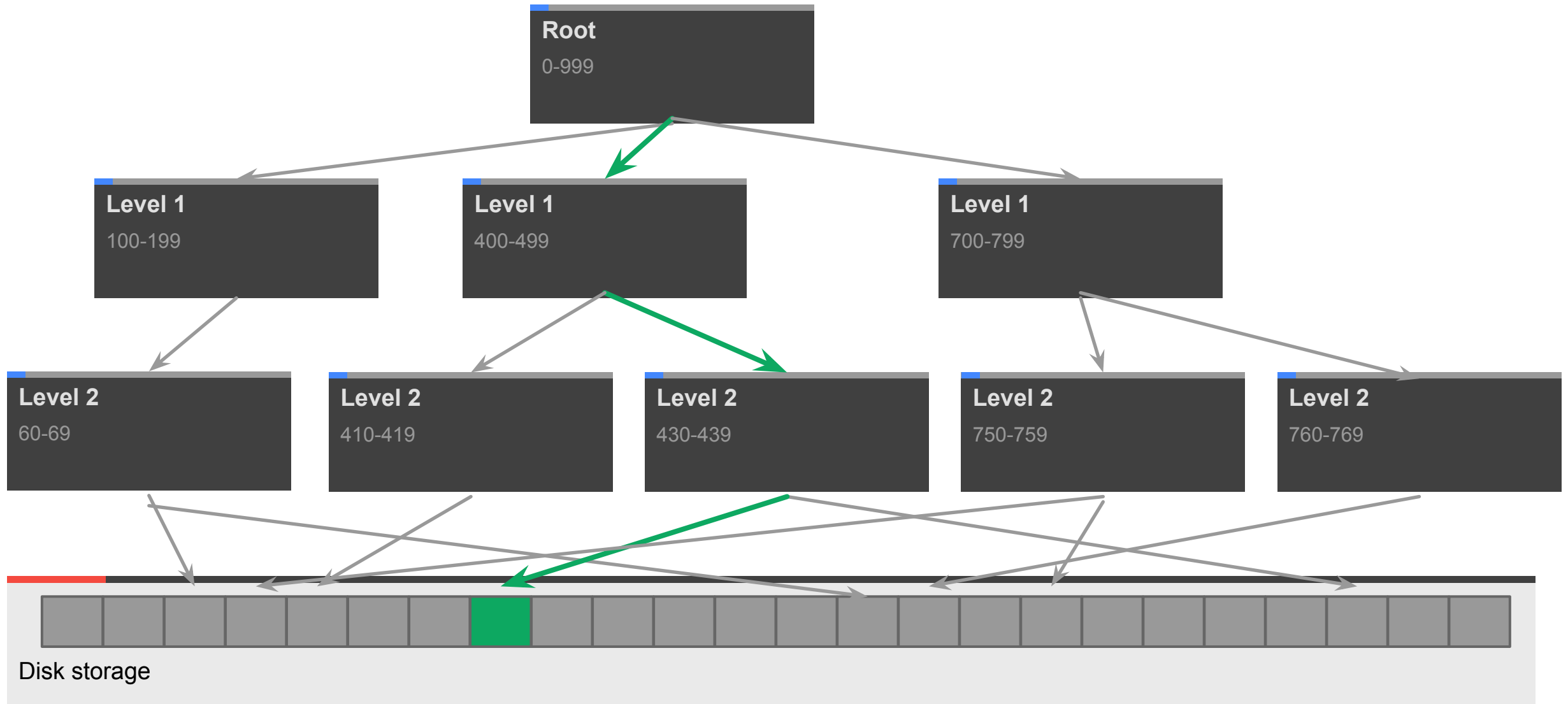
BigQuery architecture

Developing intuition about BigQuery

Relational Database Architecture: B-Tree



Relational Database Architecture: Finding a Value



“ If you do a table scan over a 1TB table,
you're going to have a bad time. ”

Anonymous

16th century Italian Philosopher-Monk



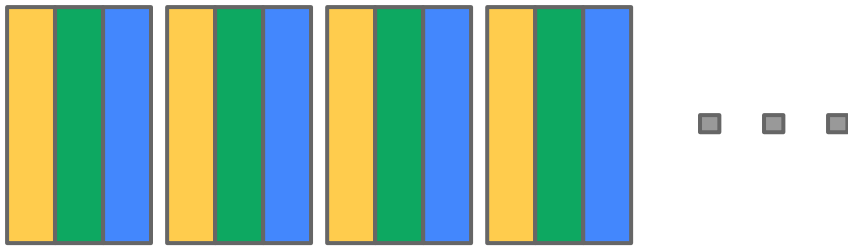
Goal: Perform a 1 TB table scan in 1 second

Parallelize Parallelize Parallelize!

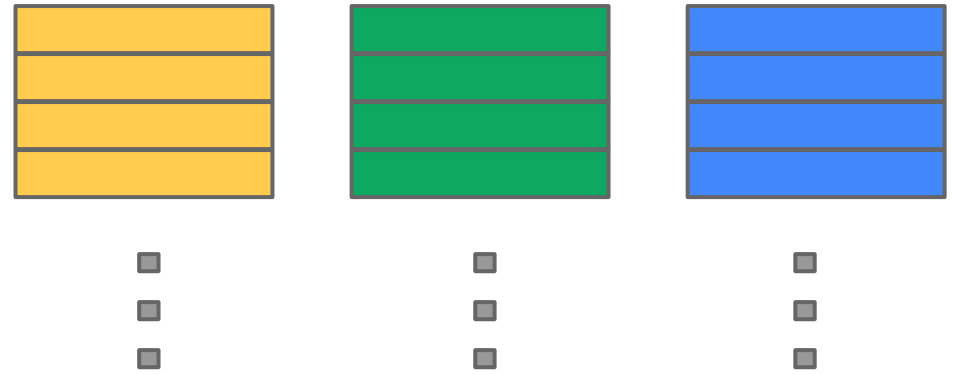
- Reading 1 TB/ second from disk:
 - 10k+ disks
- Processing 1 TB / sec:
 - 5k processors



Data access: Column Store



Record Oriented Storage



Column Oriented Storage

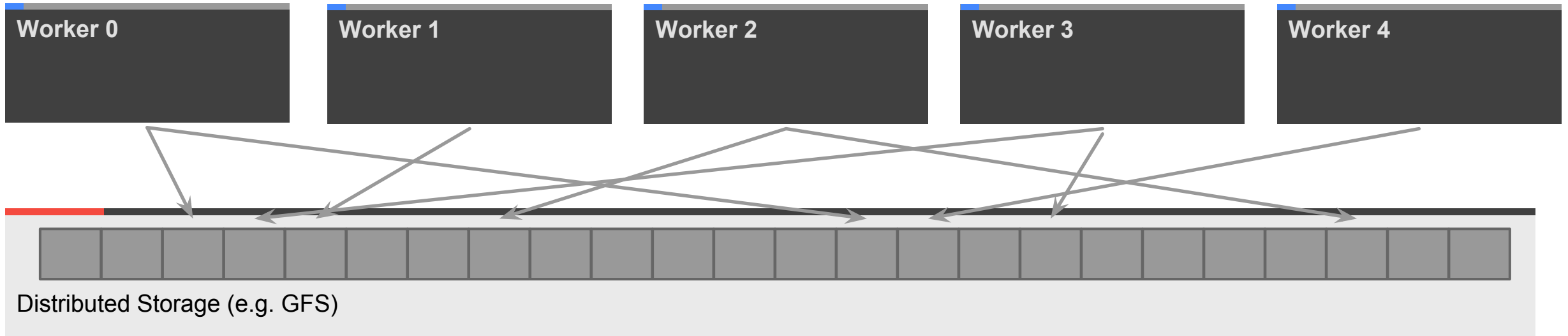
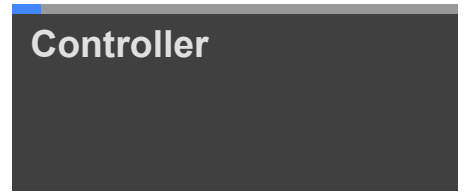


"Why not MapReduce?"

Anonymous
Reddit User

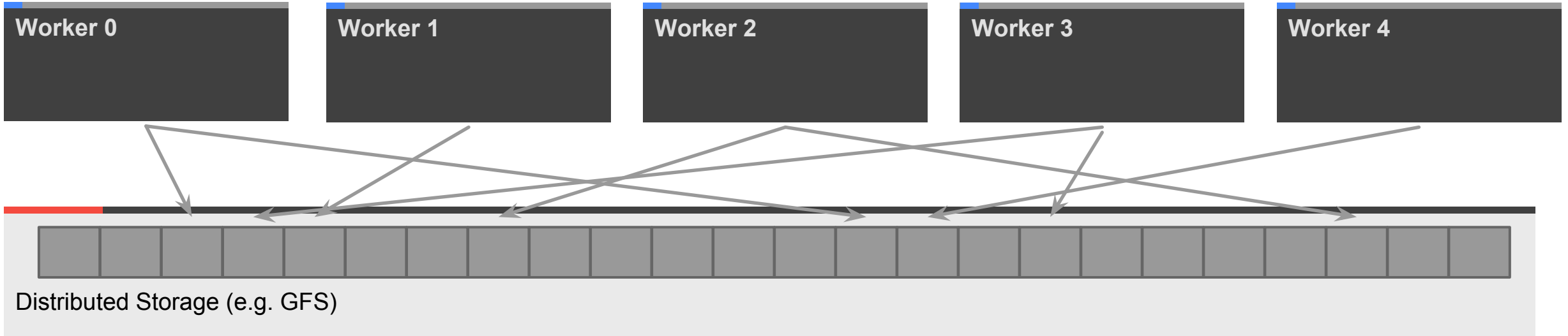
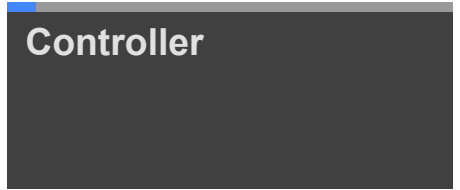


MapReduce... how does it work?



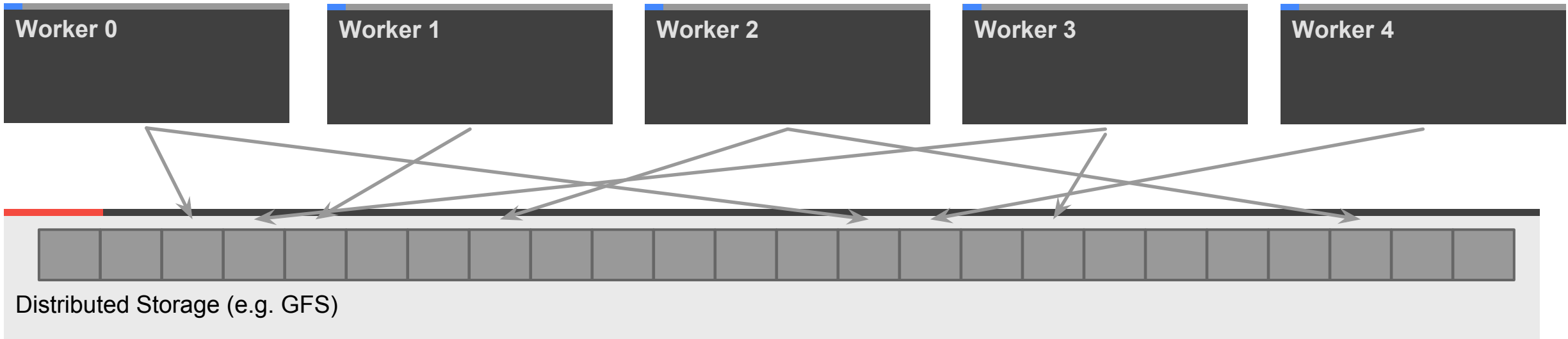
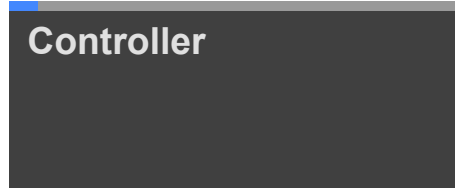
MapReduce

1. Map!



MapReduce

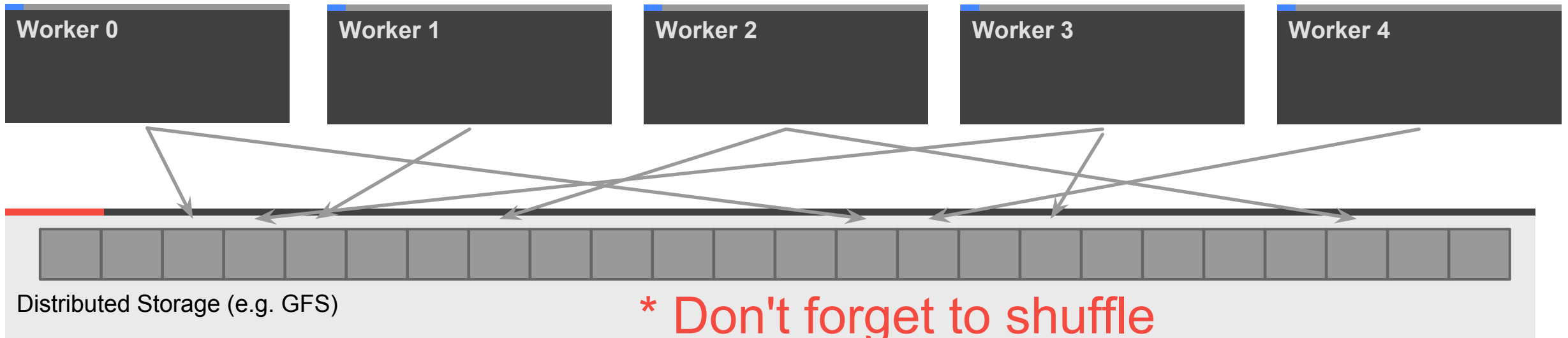
2. Reduce!



MapReduce

Controller

3. Profit!*†‡



* Don't forget to shuffle

† Multiple passes may be required

‡ Void where prohibited

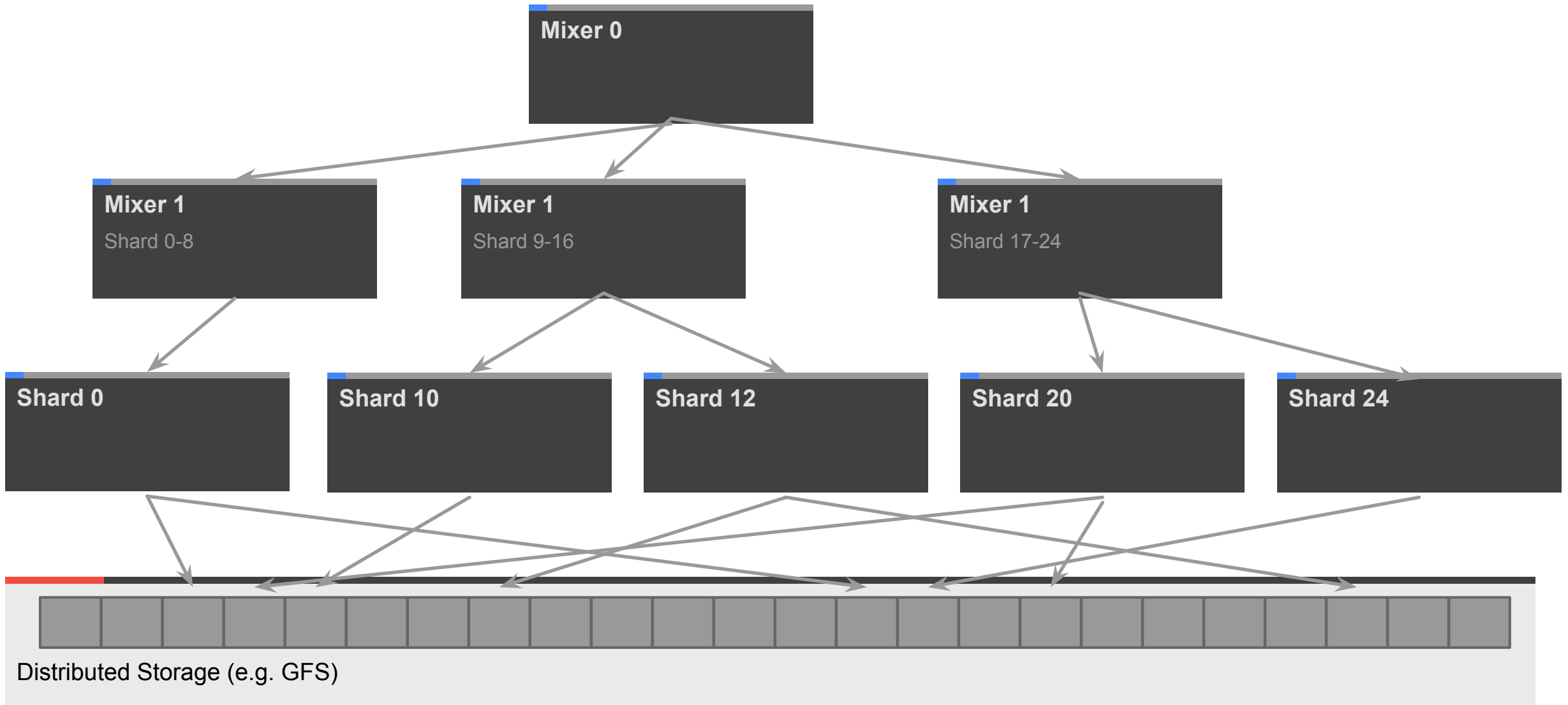


Gray's third law for big data:
“Bring computations to the data, rather
than data to the computations.”

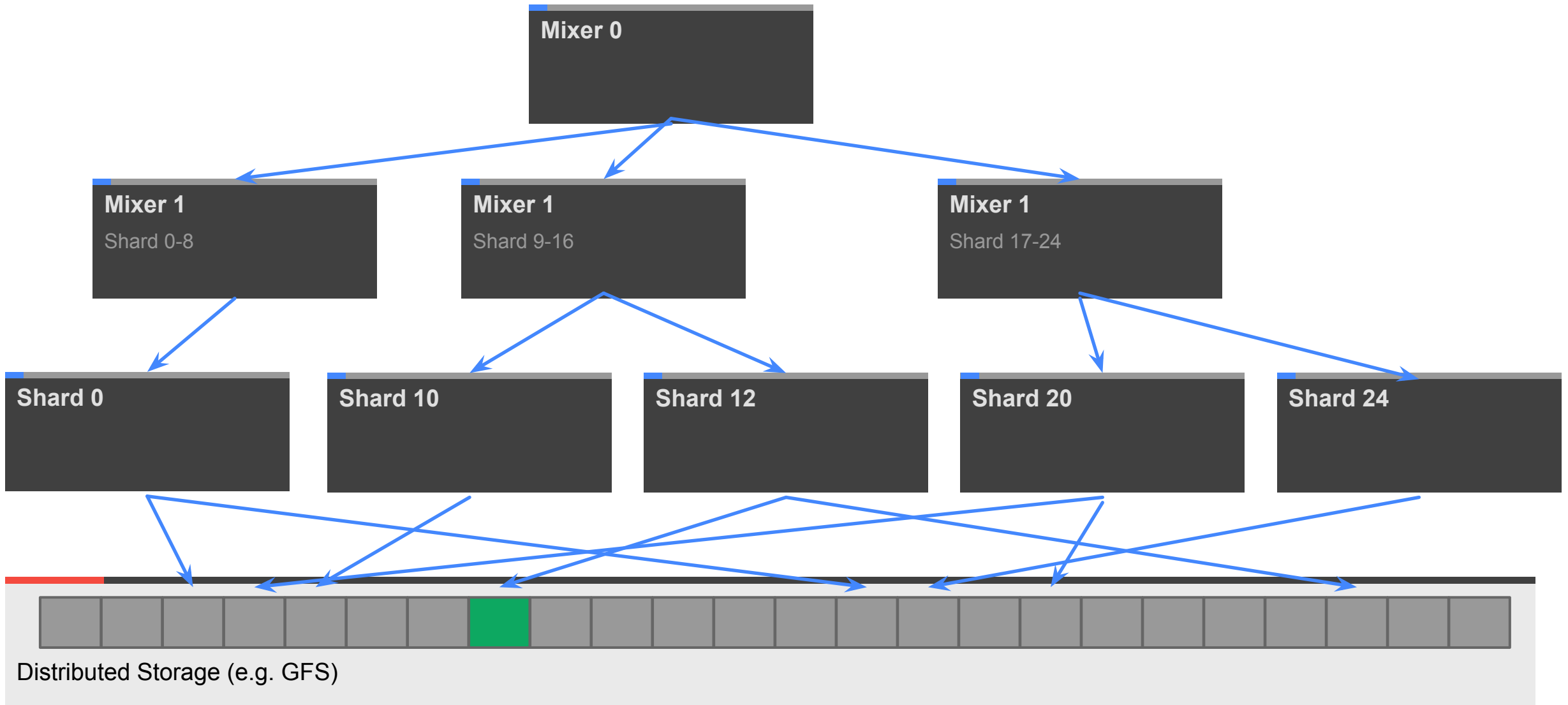
Jim Gray
Database Pioneer



BigQuery Architecture: Computation tree



BigQuery Architecture: Finding a value



BigQuery SQL Example: Simple aggregates

```
SELECT COUNT(foo), MAX(foo), STDDEV(foo)  
FROM ...
```



BigQuery SQL Example: Complex Processing

```
SELECT ... FROM ....  
WHERE REGEXP_MATCH(url, "\.com$")  
      AND user CONTAINS 'test'
```



BigQuery SQL Example: Nested SELECT

```
SELECT COUNT(*) FROM  
  (SELECT foo ..... )  
GROUP BY foo
```

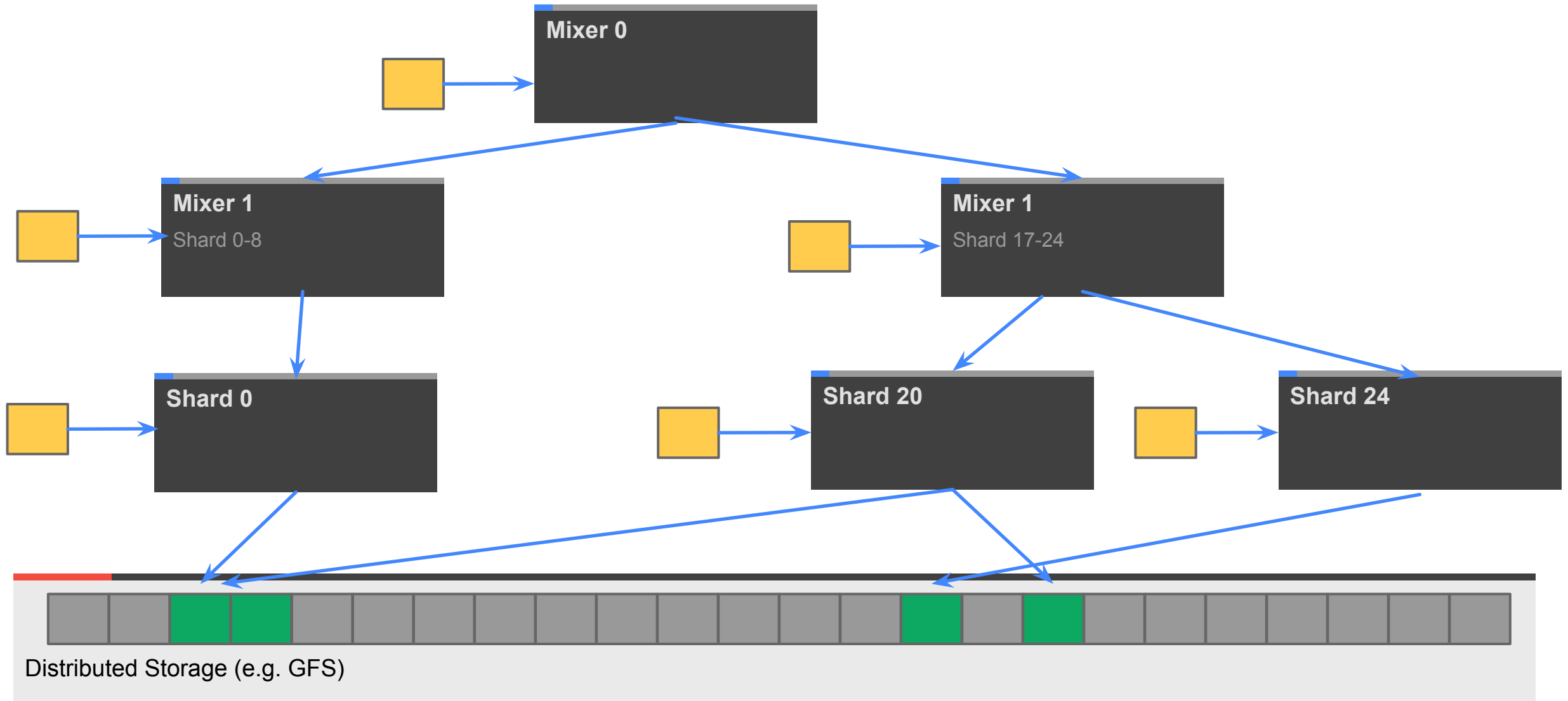


BigQuery SQL Example: Small JOIN

```
SELECT huge_table.foo  
FROM huge_table  
JOIN small_table  
ON small_table.foo = small_table.foo
```



BigQuery Architecture: Small Join



BigQuery SQL Example: Response too large

```
SELECT foo, bar FROM huge_table
```

Where huge_table is very large and no filter is applied.
Fix with:

```
... LIMIT 100
```



BigQuery SQL Example: Internal response too large

```
SELECT ... FROM ... GROUP BY user_id
```

Where number of unique users is very large.

Fix with:

```
... WHERE HASH(user_id) % 10 = 0
```



BigQuery SQL Example: Internal response too large II

```
SELECT user_id, COUNT(user_id) ...  
GROUP BY user_id  
ORDER BY user_id DESC
```

Where number of unique users is very large.

Fix with:

```
SELECT TOP(user_id, 20), count(user_id) ...
```



Advanced Query Demo

Using GitHub timeline dataset

Wikipedia:

"**GitHub** is a [web-based hosting service](#) for software development projects ... GitHub is ... the most popular open source hosting site"



Summary

- What is big data, anyway?
- BigQuery's Not MapReduce
- What's BigQuery good for?
- How to think about query execution



SELECT questions FROM audience

**SELECT 'Thank You!'
FROM ryan, jordan**



<http://developers.google.com/bigquery>

@ryguyrg

<http://profiles.google.com/ryan.boyd>

@tigani

<https://plus.google.com/115600841849663767233>



Google
Developers

Presentation Bullet Slide Layout

- Titles are formatted as Open Sans with bold applied and font size is set at 30pts
 - Vertical position for title is .3”
 - Vertical position for bullet text is 1.54”
- Title capitalization is title case
- Subtitle capitalization is title case
- Titles and subtitles should never have a period at the end



Bullet Slide With Subtitle Placeholder

Subtitle Placeholder

- Titles are formatted as Open Sans with bold applied and font size is set at 30pts
 - Vertical position for title is .3”
 - Vertical position for subtitle is 1.1”
 - Vertical position for bullet text is 2”
- Title capitalization is title case
- Subtitle capitalization is title case
- Titles and subtitles should never have a period at the end



Color Palette

Flat Color



Secondary



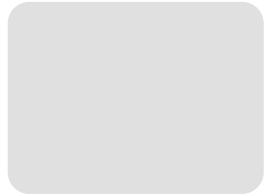
Gradient



Grays



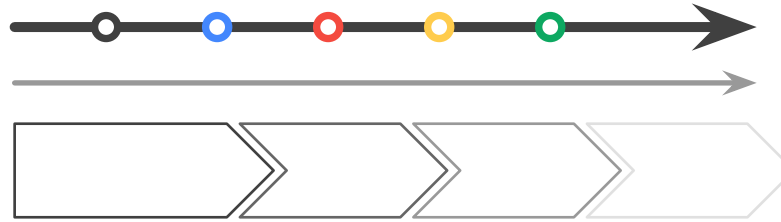
Graphic Element Styles and Arrows



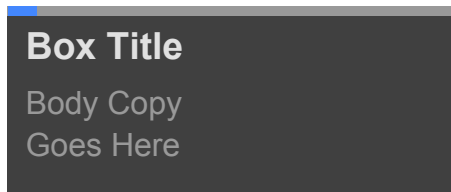
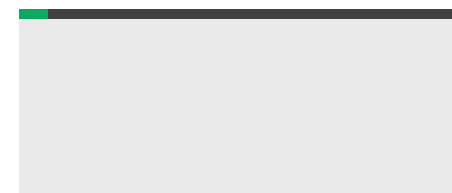
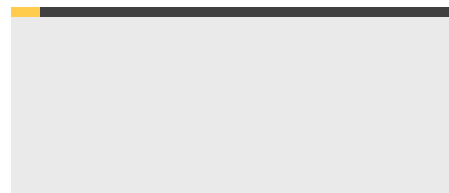
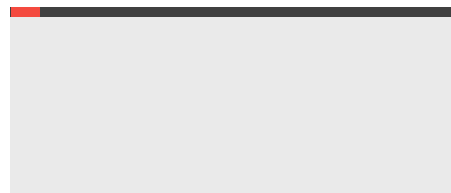
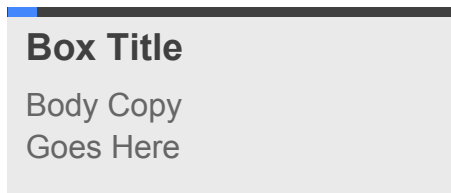
Rounded Boxes



Code Boxes



Arrows



Content Container Boxes



Pie Chart Example

Subtitle Placeholder

Chart Title

source: place source info here



Column Chart Example

Subtitle Placeholder

source: place source info here



Line Chart Example

Subtitle Placeholder

source: place source info here



Table Option A

Subtitle Placeholder

	Column 1	Column 2	Column 3	Column 4
Row 1	placeholder	placeholder	placeholder	placeholder
Row 2	placeholder	placeholder	placeholder	placeholder
Row 3	placeholder	placeholder	placeholder	placeholder
Row 4	placeholder	placeholder	placeholder	placeholder
Row 5	placeholder	placeholder	placeholder	placeholder
Row 6	placeholder	placeholder	placeholder	placeholder
Row 7	placeholder	placeholder	placeholder	placeholder



Table Option B

Subtitle Placeholder

Header 1	placeholder	placeholder	placeholder
Header 2	placeholder	placeholder	placeholder
Header 3	placeholder	placeholder	placeholder
Header 4	placeholder	placeholder	placeholder
Header 5	placeholder	placeholder	placeholder





Segue Slide

Subtitle Placeholder

“ This is an example of
quote text. ”

Name

Company



Code Slide With Subtitle Placeholder

Subtitle Placeholder

```
<script type='text/javascript'>  
  // Say hello world until the user starts questioning  
  // the meaningfulness of their existence.  
  function helloWorld(world) {  
    for (var i = 42;--i >= 0;) {  
      alert ('Hello' + String(world));  
    }  
  }  
</script>  
<style>  
p { color: pink }  
p { color: blue }  
u { color: 'umber' }  
</style>
```

HTML

