

# Agenda

---

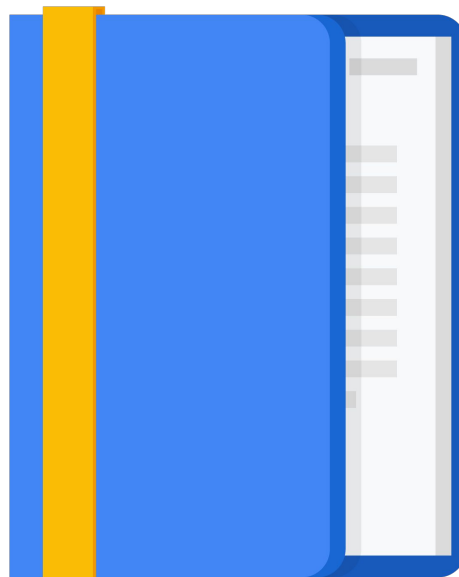
Processing Streaming Data

Cloud Pub/Sub

Cloud Dataflow Streaming  
Features

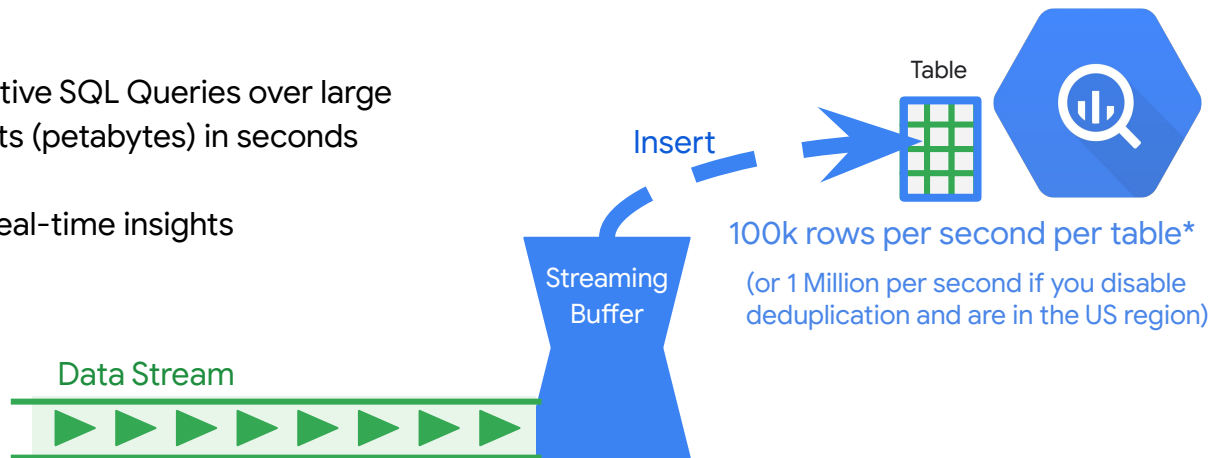
BigQuery and Bigtable Streaming  
Features

Advanced BigQuery Functionality



# BigQuery allows you to stream records into a table; query results incorporate latest data

- Interactive SQL Queries over large datasets (petabytes) in seconds
- Near-real-time insights



Note: Unlike load jobs, there is a cost for streaming inserts (see [quota and limits](#))

# Insert streaming data into a BigQuery table

```
export GOOGLE_APPLICATION_CREDENTIALS="/home/user/Downloads/[FILE_NAME].json"
```

## Credentials

The service must have  
Cloud IAM permissions  
set in the Web UI

```
pip install google-cloud-bigquery
```

## Install API

```
from google.cloud import bigquery
client = bigquery.Client(project='PROJECT_ID')

dataset_ref = bigquery_client.dataset('my_dataset_id')
table_ref = dataset_ref.table('my_table_id')
table = bigquery_client.get_table(table_ref)

# read data from Cloud Pub/Sub and place into row format
# static example customer orders in units:
rows_to_insert = [
    (u'customer 1', 5),
    (u'customer 2', 17),
]
errors = bigquery_client.insert_rows(table, rows_to_insert)
```

## Python

## Create a client

## Access dataset and table

## Perform insert


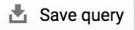
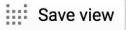
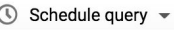

Get table access  
from API



Insert rows into table

Want to visualize insights? Explore Data Studio insights right from within BigQuery

Query editor

```
1 # which days did it rain in SF?
2 WITH rainy_sf AS (
3   SELECT
4     wban,
5     stn,
6     rain_drizzle,
7     fog,
8     PARSE_DATE("%F", CONCAT(year, '-', mo, '-', da)) AS date
9   FROM `bigquery-public-data.noaa_gsod.gsod2018`
10  WHERE wban = '93816'
11  ORDER BY rain_drizzle DESC, date
12 )
13
```

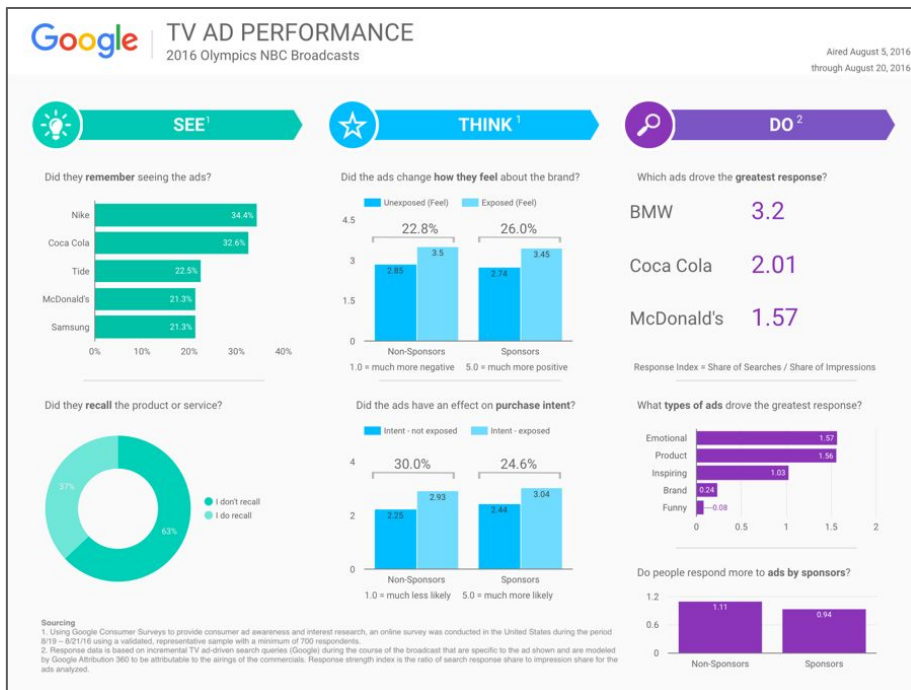
Query results  

Query complete (2.5 sec elapsed, 118.1 MB processed)

Job information Results JSON Execution details

Row	date	total_trips	rain_drizzle	fog
1	2018-01-07	1382	1	0
2	2018-01-08	805	1	0
3	2018-01-10	3459	1	1

# Build, collaborate, and share your dashboards



Tell a clear story with your data

Share and collaborate on reports with others

# Create new reports in the Data Studio UI

Data Studio beta Home

Start a new report

Blank

Acme Marketing  
Google Analytics

Ecommerce PPC  
Google Analytics + Adwords

AdWords Overview  
Google Adwords

ALL TEMPLATES

ALL OWNED BY ME SHARED WITH ME TRASH

Q welcome

REPORTS

DATA SOURCES

New Features!

Video tutorials  
Learn by watching!

User settings

Previous 30 days

Owner Last opened by me

Welcome to Data Studio! (Start here) Google Data Studio Mar 31, 2017

Earlier

Owner Last opened by me

[Sample] Acme Marketing Website Google Data Studio Dec 12, 2016

Copy of Welcome to Data Studio! (Start here) Rick Elliott Jul 26, 2016

Copy of Welcome to Data Studio! (Start here) Rick Elliott Jul 26, 2016

Copy of Welcome to Data Studio! (Start here) Rick Elliott Jul 25, 2016

Copy of Welcome to Data Studio! (Start here) Rick Elliott Jul 1, 2016

+

# Select data sources to build your visualizations

Untitled Report

File View Page Help

VIEW

+

## Data source picker

Select:  
Google Merchandise Store: Master View

Add a data source

A data source provides data for charts. Select an existing data source or click CREATE NEW DATA SOURCE.

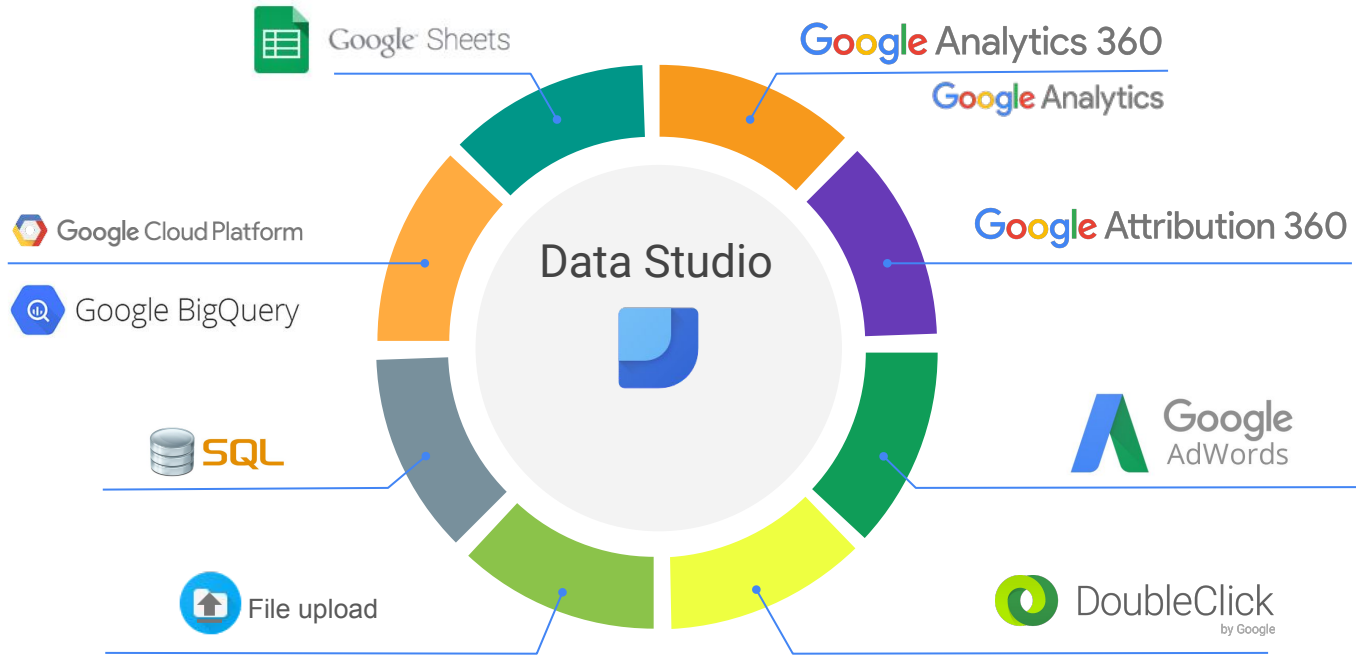
OKAY, GOT IT

Select Data Source

- YouTube Analytics
- Benchmark Calculations Retail
- Data Studio test - Dates
- Finestkind's Online Academy - Clas...
- AdWords
- HACH Europe ICS
- 1) Akiri campaigns\_stats // CHAN...
- Analytics Academy
- [Datamart] GA Data Sharing
- Test GAIQ from plx
- sduncan.dda\_segment
- datastudio\_metrics.usage.all
- Google Merchandise Store: Master...
- Data Studio test - Sheet1 2017-04...

CREATE NEW DATA SOURCE

# Connect to multiple different types of data sources





# Add the data source to your report

The screenshot shows the Google Cloud Reporting interface. At the top, a dark header bar contains the text "Untitled Report" and a menu with "File", "View", "Page", and "Help". On the right side of the header are icons for refresh, print, view, and user profile. The main workspace is a large gray area. In the center, a white dialog box is displayed with the title "You are about to add a data source to this report". Inside the dialog, it shows "Google Merchandise Store: Master View" with a document icon, an arrow pointing to a bar chart icon, and the text "Note that **Report Editors** can create charts using this data source, and can add dimensions and metrics not currently included in the report." At the bottom of the dialog are two buttons: "CANCEL" and "ADD TO REPORT". To the right of the main workspace is a sidebar. The top section of the sidebar is titled "Add a data source" and contains the text "A data source provides data for charts. Select an existing data source or click CREATE NEW DATA SOURCE." Below this is a search bar labeled "Select Data Source". A list of data sources follows, including "Google Merchandise Store: Master...", "Data Studio test - Sheet1", "Finestkind's Online Academy - Clas...", "YouTube Analytics", "Benchmark Calculations Retail", "Data Studio test - Dates", "Finestkind's Online Academy - Clas...", "AdWords", "HACH Europe ICS", "1) Akiri campaigns\_stats // CHAN...", "Analytics Academy", "[Dataman] GA Data Sharing", "Test GAIQ from plx", and "sduncan.dda\_segment". At the bottom of the sidebar is a button labeled "CREATE NEW DATA SOURCE".



---

# Streaming Analytics and Dashboards

## Objectives

- Connect to a BigQuery data source from Google Data Studio
- Create reports and charts to visualize BigQuery data

# Cloud Bigtable

Cloud  
Bigtable

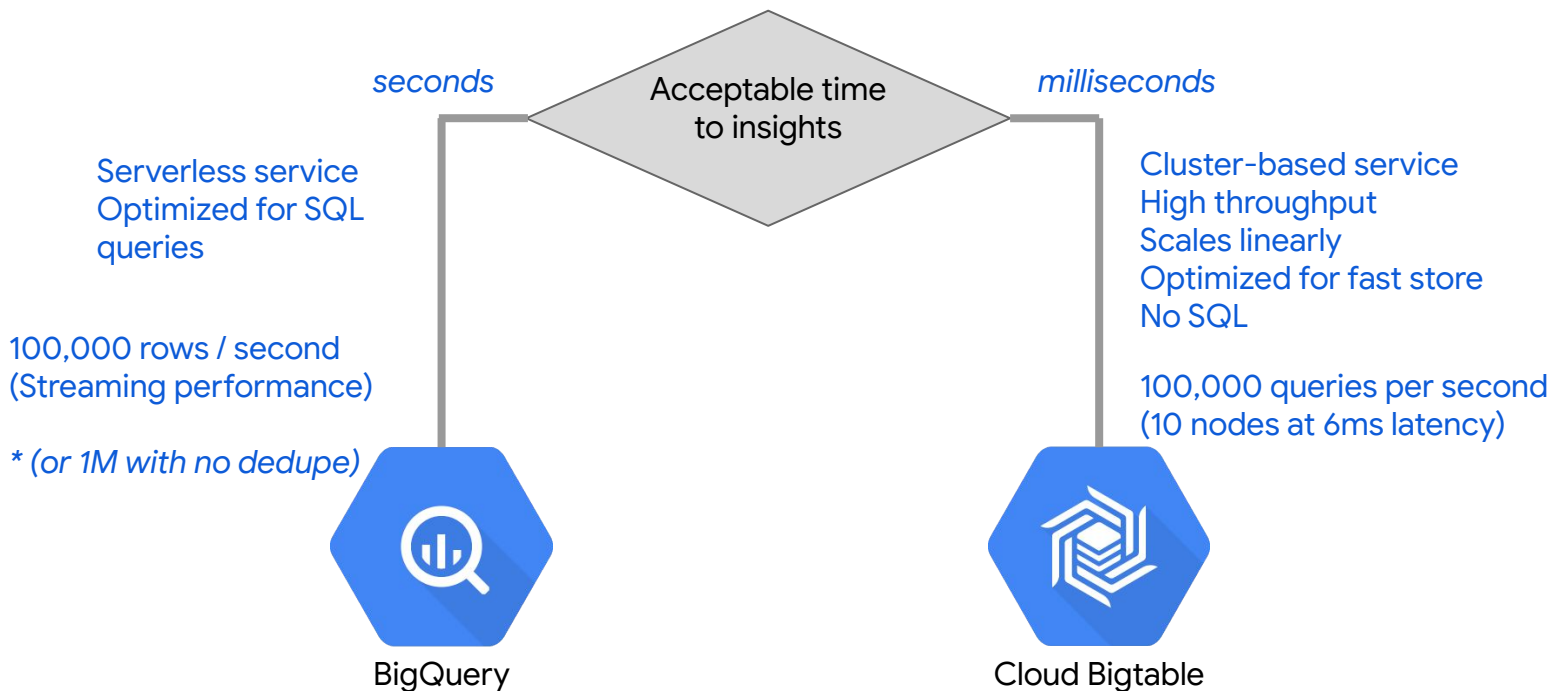


Qualities that Cloud Bigtable contributes to Data Engineering solutions:

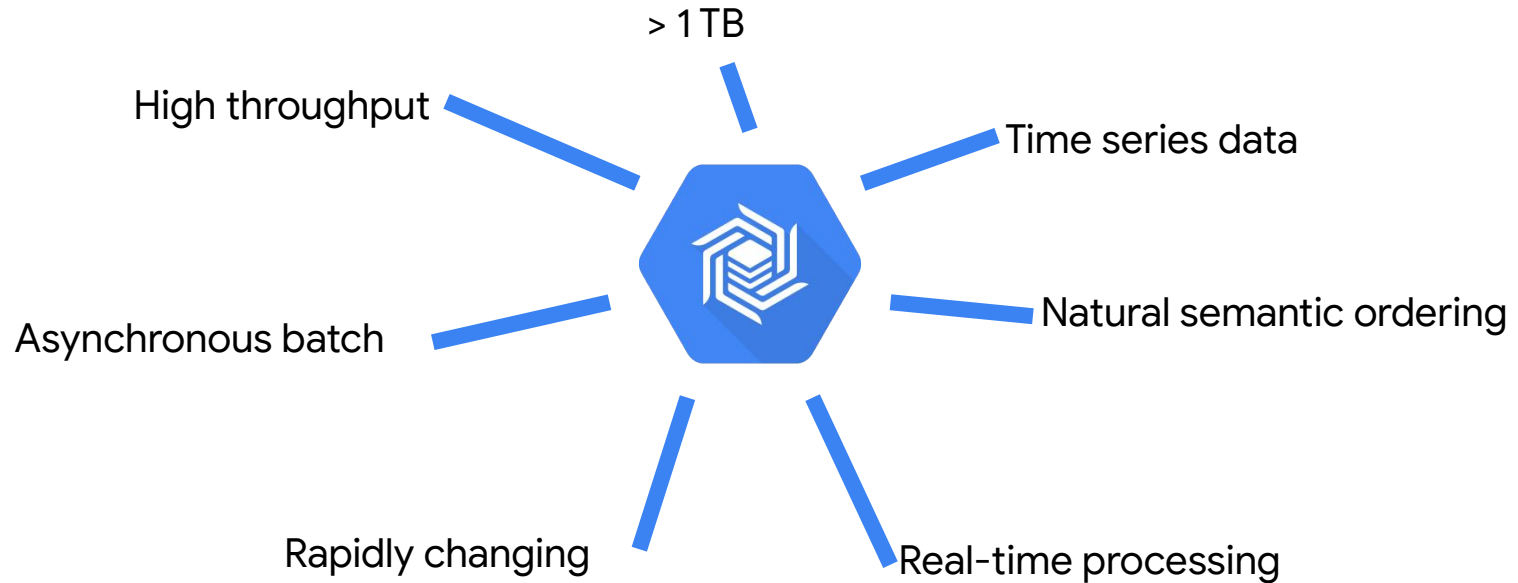
NoSQL Queries over large datasets (petabytes) in milliseconds

Very fast for specific cases

# How to choose between Cloud Bigtable and BigQuery



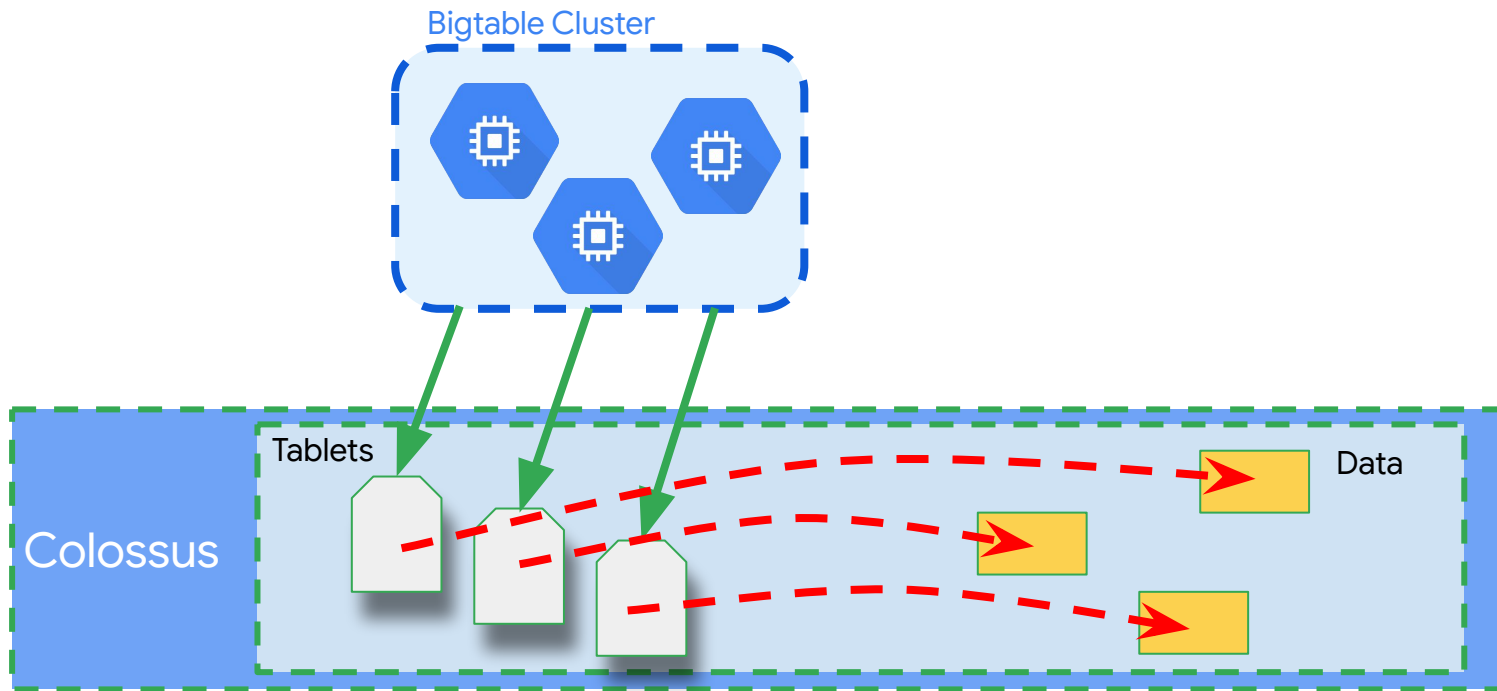
# Consider Cloud Bigtable for these requirements



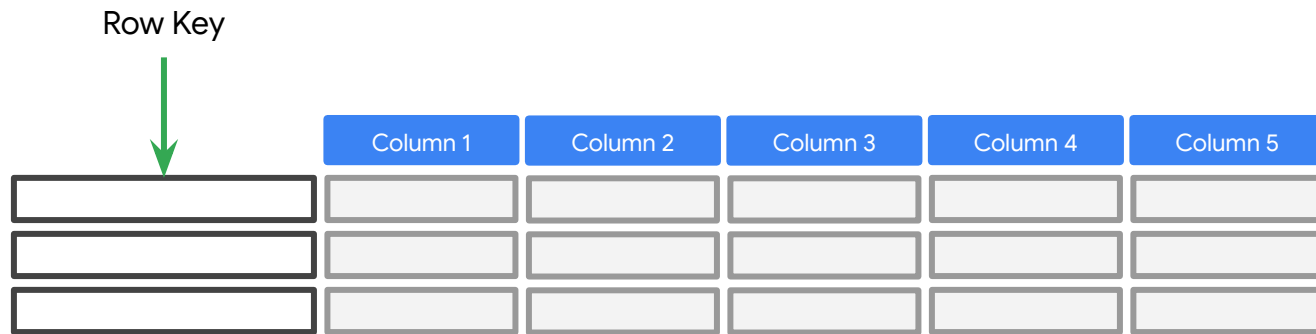
# The most common use of Cloud Bigtable is...

Productionize a real-time lookup as part of an application, where speed and efficiency are desired beyond that of other databases.

# How does Cloud Bigtable work?



# Cloud Bigtable design idea is "simplify for speed"

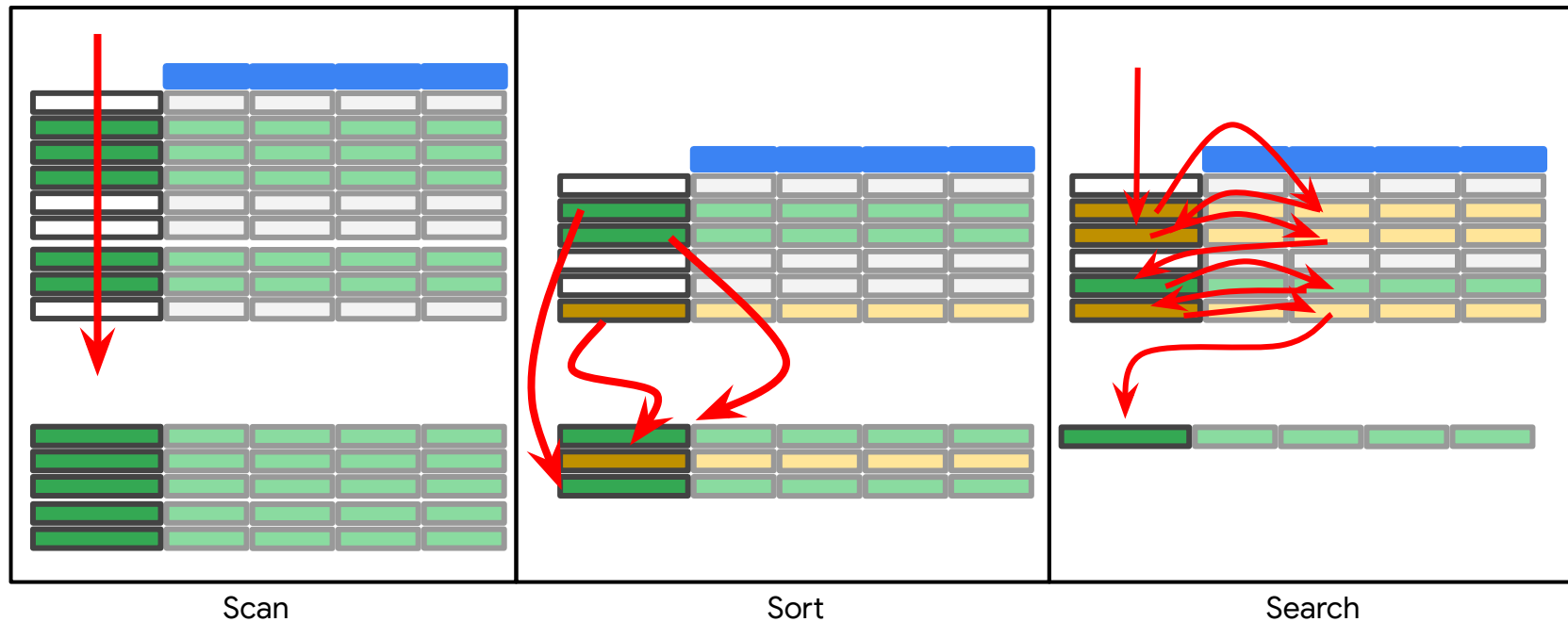


The Row Key is the index.

And you get only one.



But speed depends on your data and Row Key

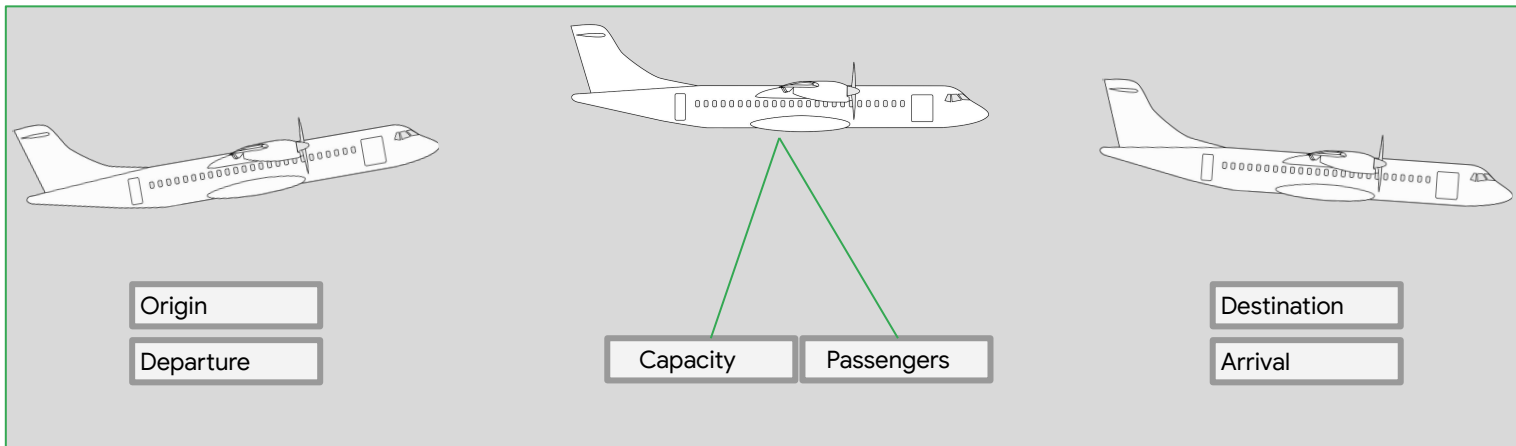


# Flights of the world: Reviewing the data

Make

Model

Age



# What is the best Row Key?

Query: All flights originating in Atlanta and arriving between March 21st and 29th

Origin	Arrival	Remaining columns...
ATL	20190321-1005	

Sort or  
Search

Arrival	Origin	Remaining columns...
20190321-1005	ATL	

Sort or  
Search

ATL#arrival#20190321-1005

Constructed Row Key	Remaining columns...
---------------------	----------------------

Scan




# Cloud Bigtable schema organization



## Column Families

Row Key	Flight_Information					Aircraft_Information			
	Origin	Destination	Departure	Arrival	Passengers	Capacity	Make	Model	Age
ATL#arrival#20190321-1121	ATL	LON	20190321-0311	20190321-1121	158	162	B	737	18
ATL#arrival#20190321-1201	ATL	MEX	20190321-0821	20190321-1201	187	189	B	737	8
ATL#arrival#20190321-1716	ATL	YVR	20190321-1014	20190321-1716	201	259	B	757	23

Queries that use the row key, a row prefix, or a row range are the most efficient

Query: Current arrival delay for flights from Atlanta

1

*ROW KEY BASED ON ATLANTA ARRIVALS*

*E.G.* ORIGIN#arrival

( ATL#arrival#20190321-1005 )

Puts latest flights at bottom of table

2

*REVERSE TIMESTAMP TO THE ROWKEY*

*E.G.* ORIGIN#arrival#RTS

( ATL#arrival#12345678 )

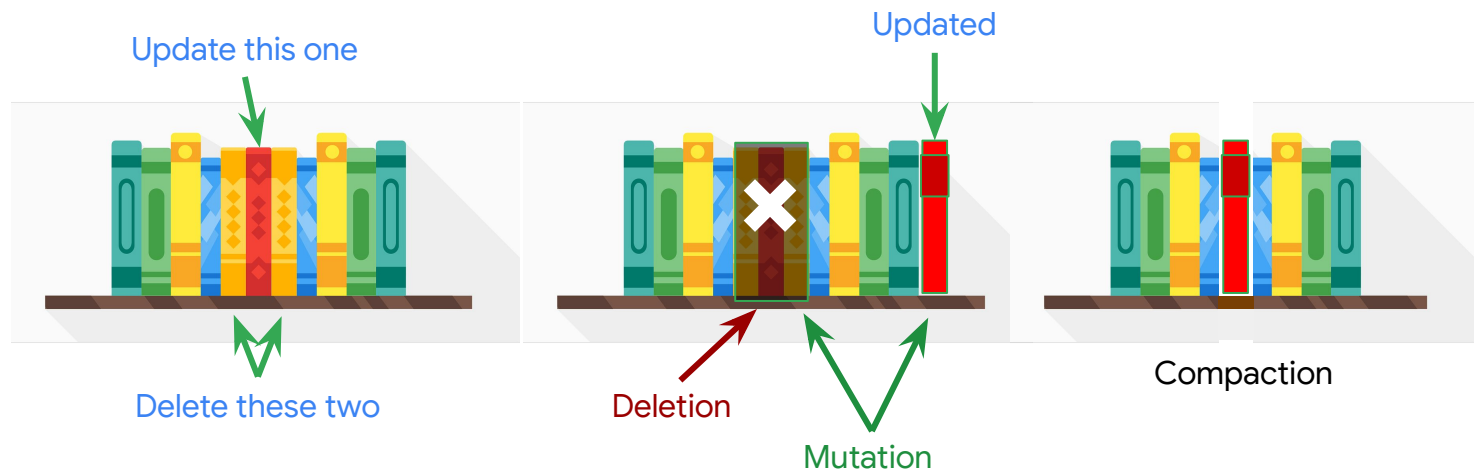
Puts latest flights at top of table

Use reverse timestamps when your most common query is for the latest values.

Query: Current arrival delay for flights from Atlanta

```
// key is ORIGIN#arrival#REVTS
String key = info.getORIGIN() //
    + "#arrival" //
    + "#" + (Long.MAX_VALUE - ts.getMillis()); // reverse timestamp
```

# What happens when data in Cloud Bigtable is changed?



# Optimizing data organization for performance



Group related data for  
more efficient reads

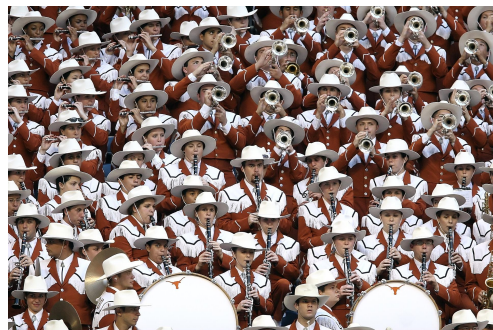
Example row key:

```
DehliIndia#2019031411841
```

Use column families



Distribute data evenly for  
more efficient writes

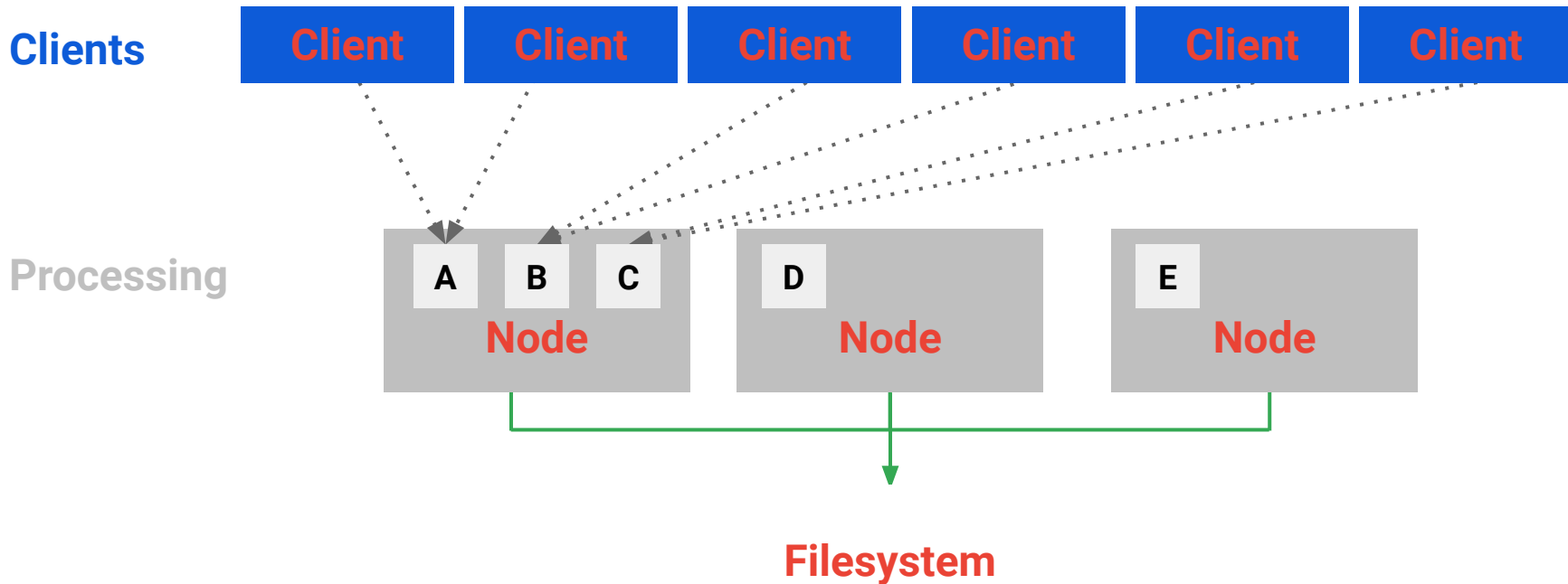


Place identical values in  
the same row or adjoining  
rows for more efficient  
compression

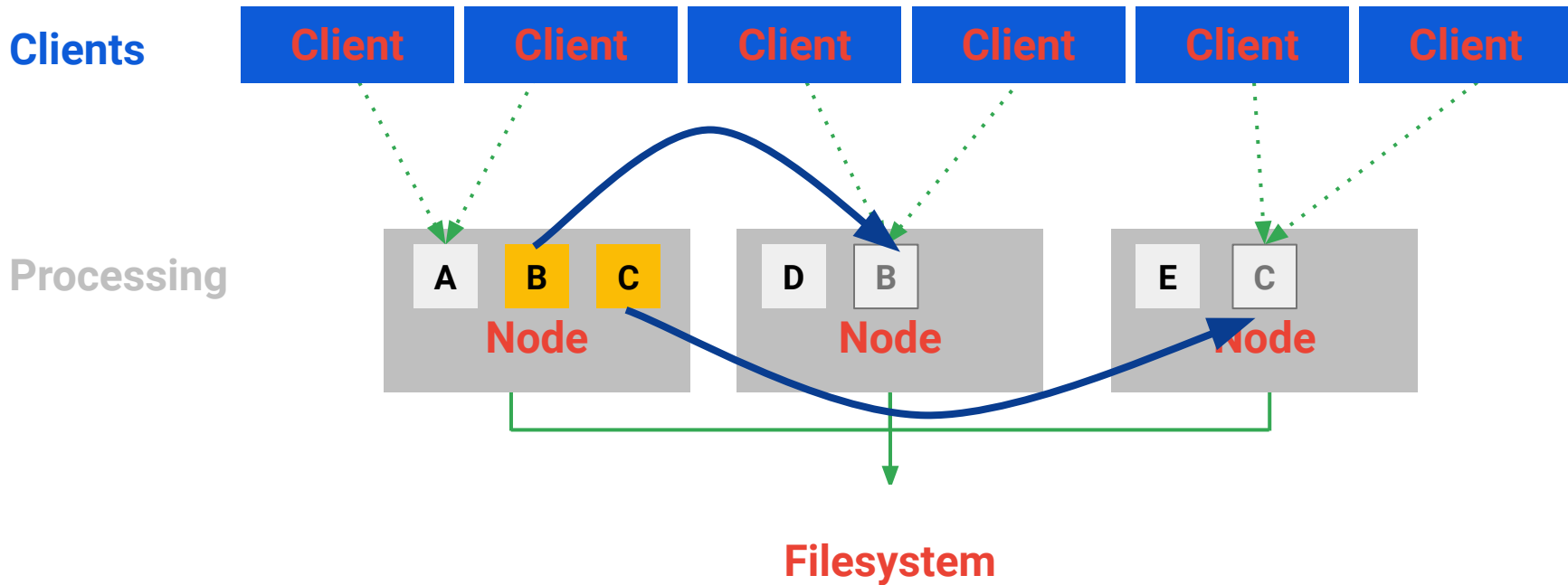
Use row keys to organize identical data



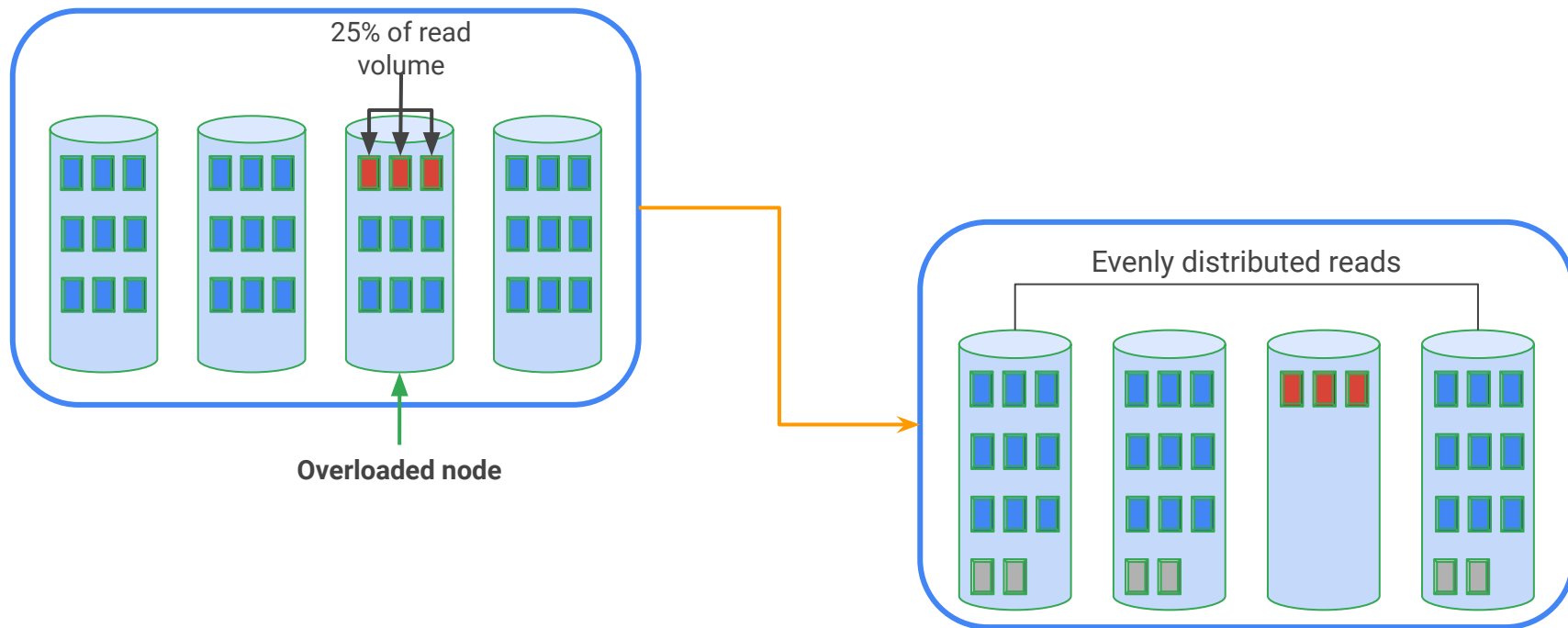
# Cloud Bigtable self-improves by learning access patterns...



# ...and rebalances data accordingly

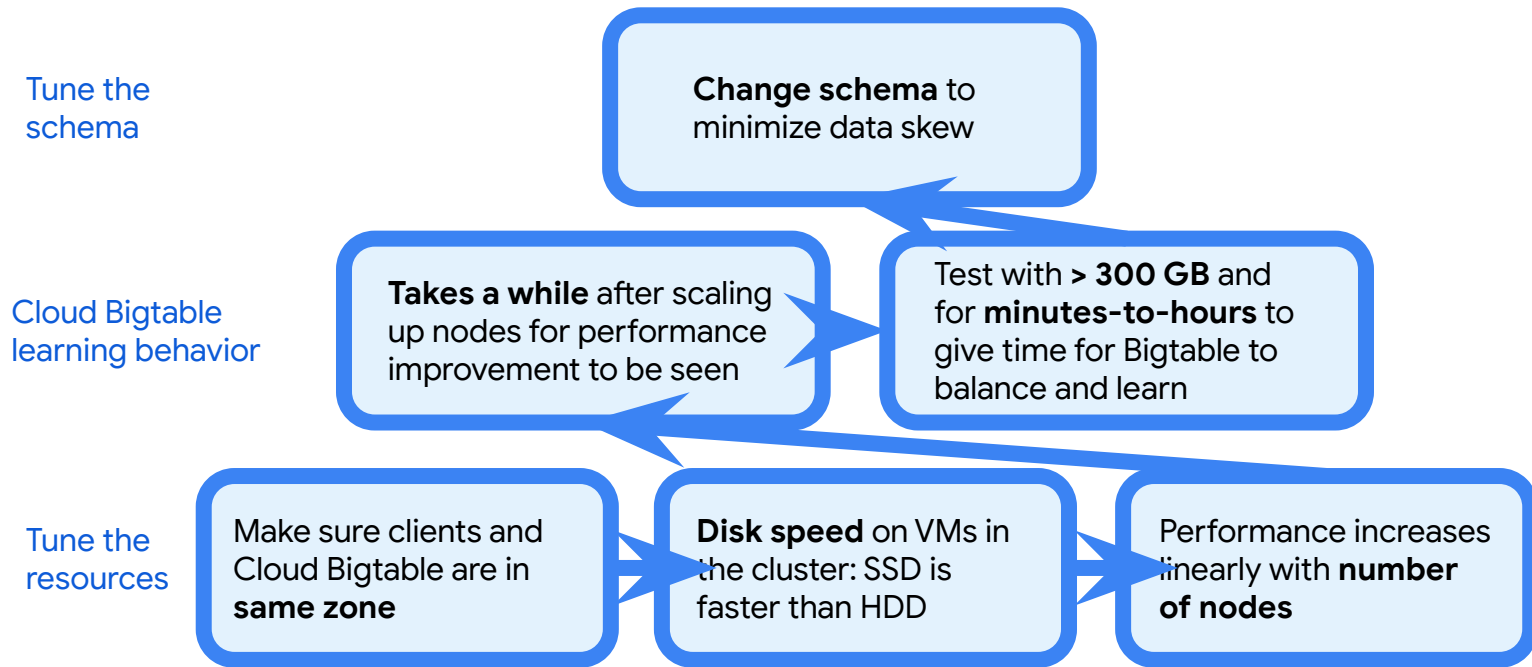


## Rebalance strategy: distribute reads

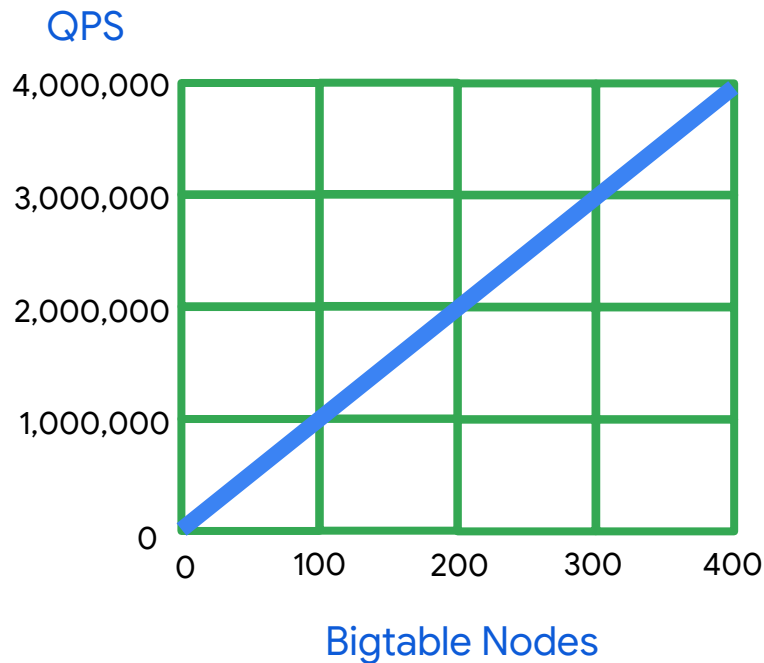
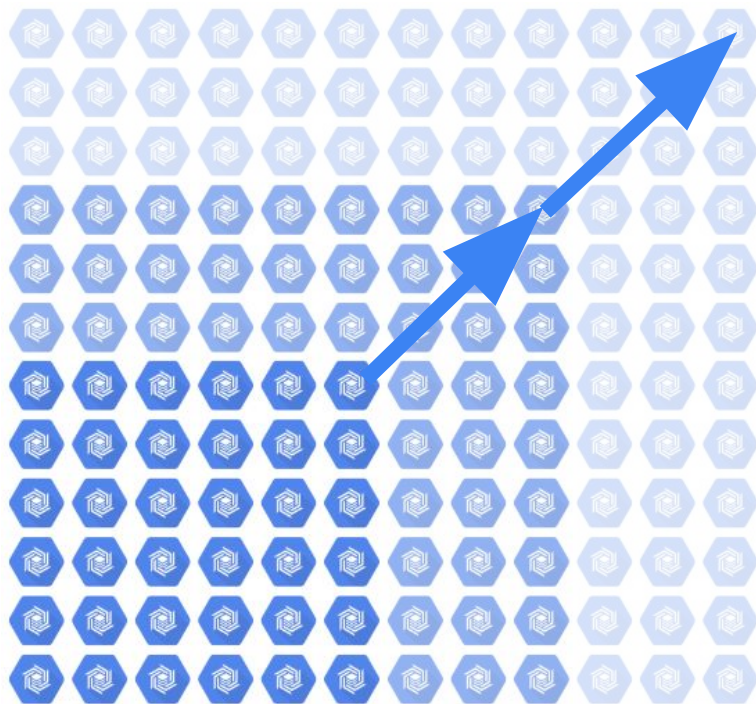


# Optimizing Cloud Bigtable Performance

# Optimizing Cloud Bigtable Performance

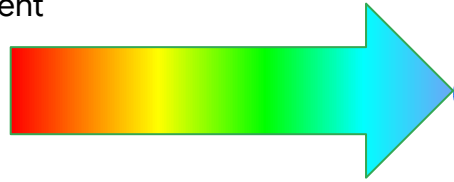


# Throughput can be controlled by node count



# Features for Cloud Bigtable streaming

Incoming  
streaming data  
is independent



Writing

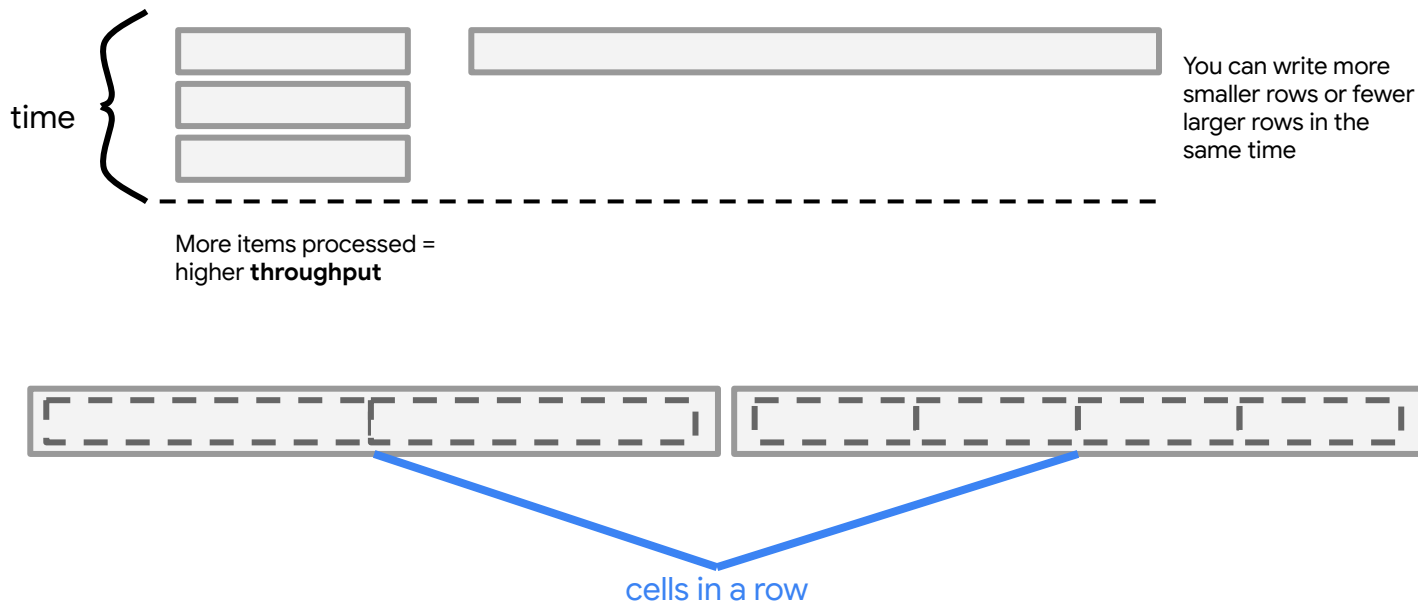


Reading



Application  
reading of data  
is controllable

# Schema design is the primary control for streaming





# Use Cloud Bigtable replications to improve availability

## Why perform replication?

- Isolate serving applications from batch reads
- Improve availability
- Provide near-real-time backup
- Ensure your data has a global presence

```
gcloud bigtable clusters create CLUSTER_ID \  
  --instance=INSTANCE_ID \  
  --zone=ZONE \  
  [--num-nodes=NUM_NODES] \  
  [--storage-type=STORAGE_TYPE]
```

Batch analytic  
read-only Cluster

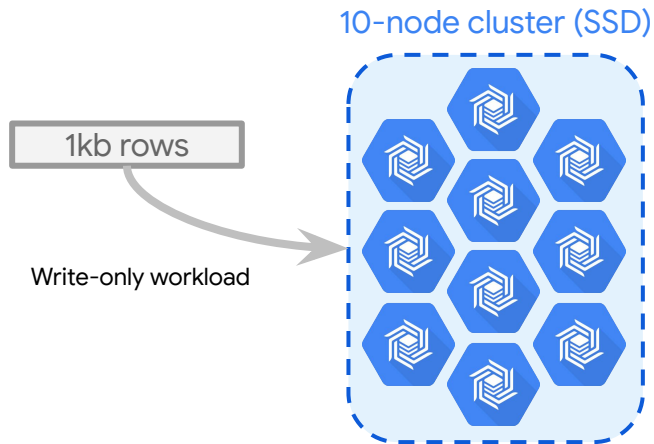


App Traffic Cluster

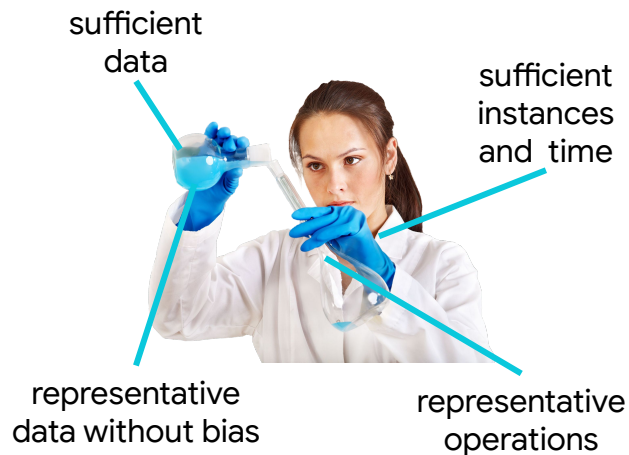


# Run performance tests carefully for Cloud Bigtable streaming

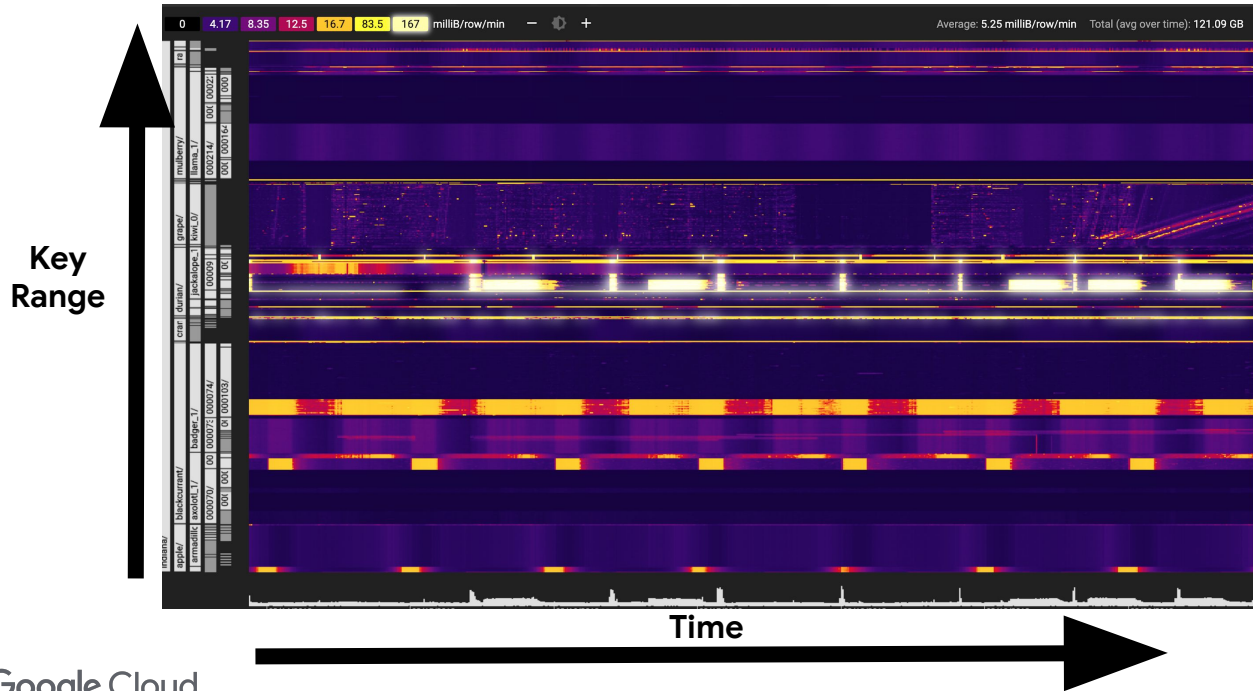
## Ideal



## Real



## Key Visualizer exposes read/write access patterns over time and key space



- find/prevent hotspots
- find rows with too much data
- see if your key schema is balanced



---

# Streaming Data Pipelines into Bigtable

## Objectives

- Launch a Dataflow pipeline to read from PubSub and write into Bigtable
- Open an HBase shell to query the Bigtable database