



Real-time Dashboards with Pub/Sub,
Dataflow, and Data Studio

Agenda

Modern data pipeline challenges

Message-oriented architectures

Serverless data pipelines

- Designing streaming pipelines with Apache Beam
- Implementing streaming pipelines on Cloud Dataflow

Data Visualization with Data Studio

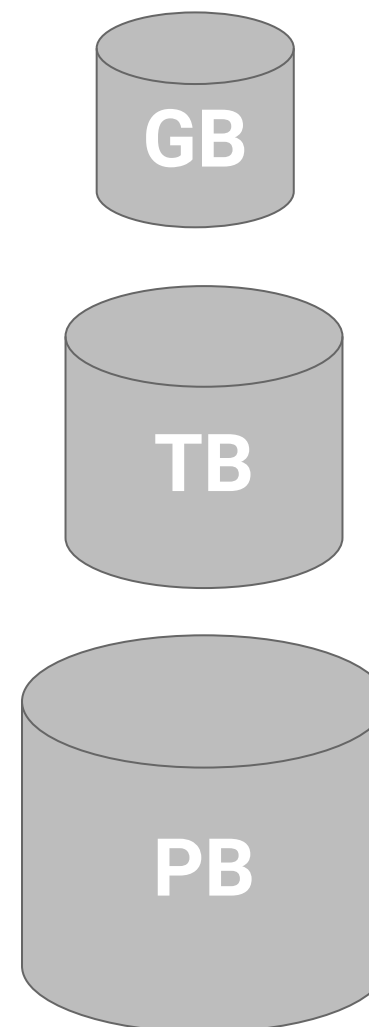
- Building collaborative dashboards
- Tips and tricks to create charts with the Data Studio UI

Modern big data pipelines face many challenges

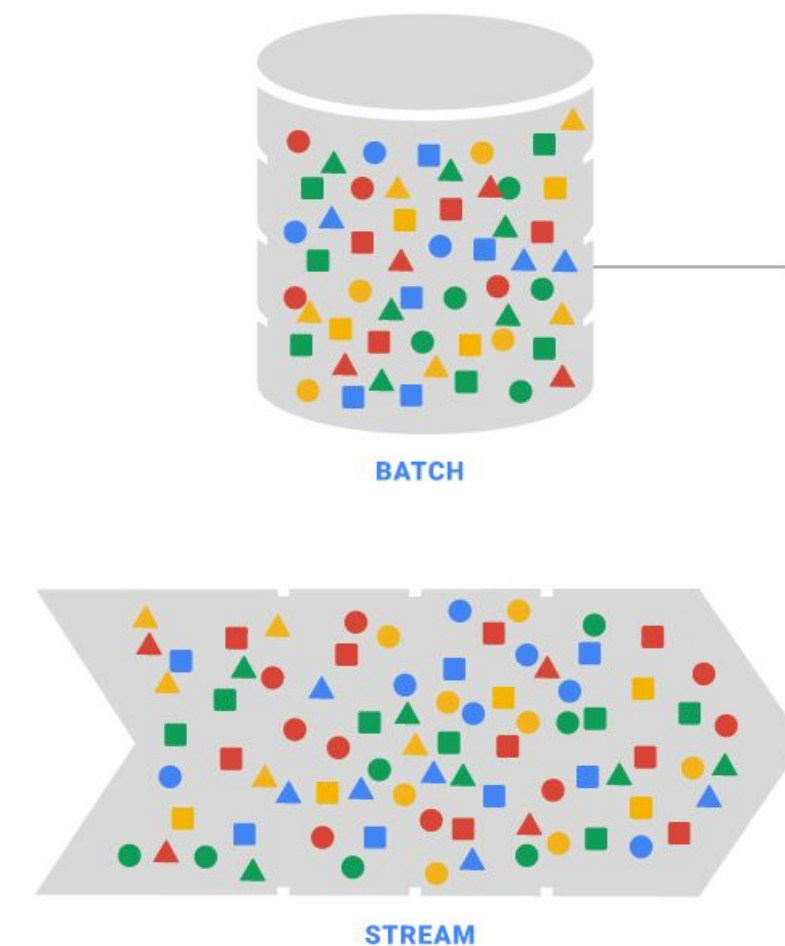
Variety



Volume



Velocity



Agenda

Modern data pipeline challenges

Message-oriented architectures

Serverless data pipelines

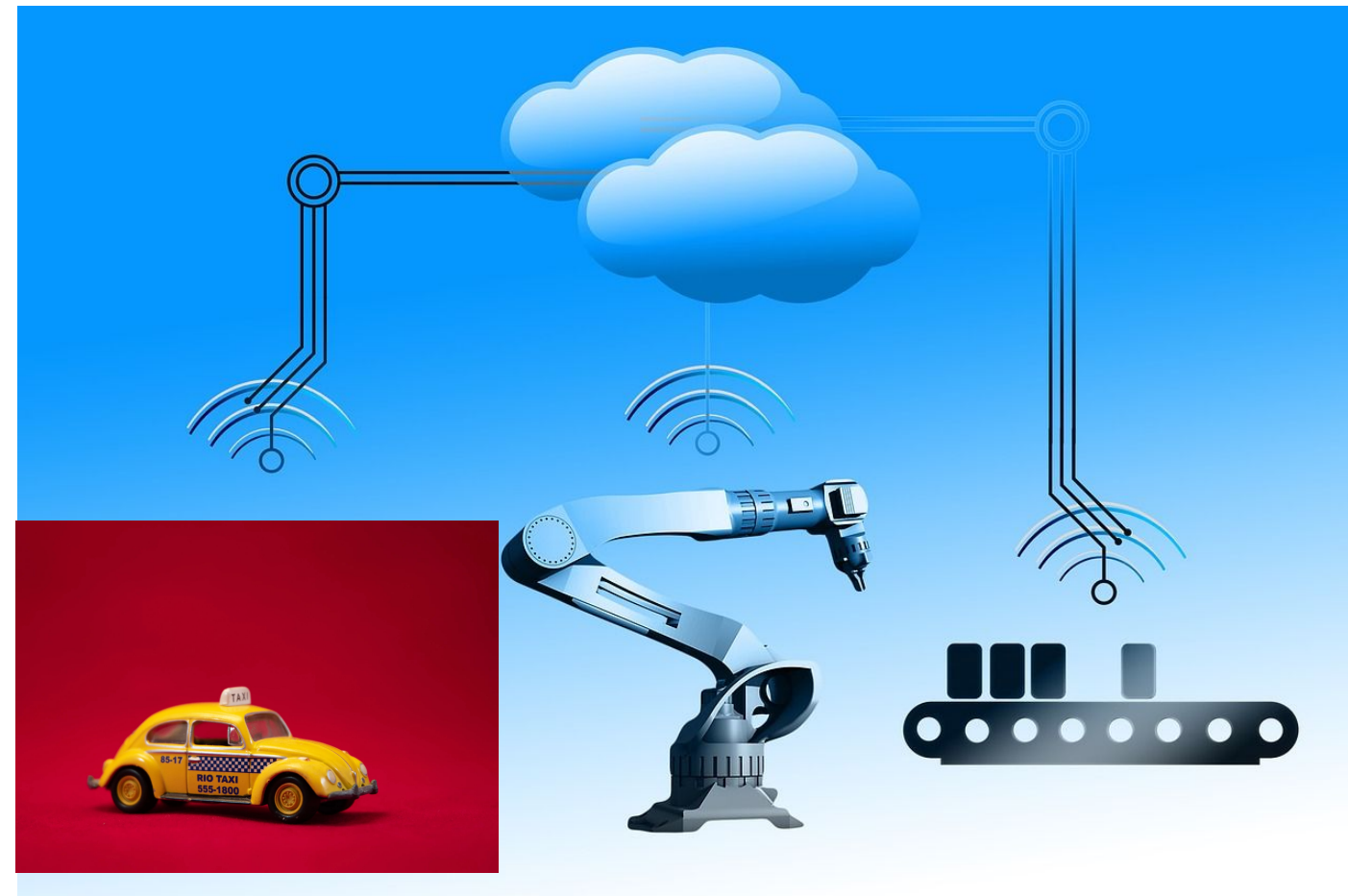
- Designing streaming pipelines with Apache Beam
- Implementing streaming pipelines on Cloud Dataflow

Data Visualization with Data Studio

- Building collaborative dashboards
- Tips and tricks to create charts with the Data Studio UI

IoT devices present new challenges to data ingestion

Distributed Messages



- Data streaming from various processes or devices
- Distributing event notifications (ex: new user sign up)
- Scale to handle volume
- Reliable (no duplicates)

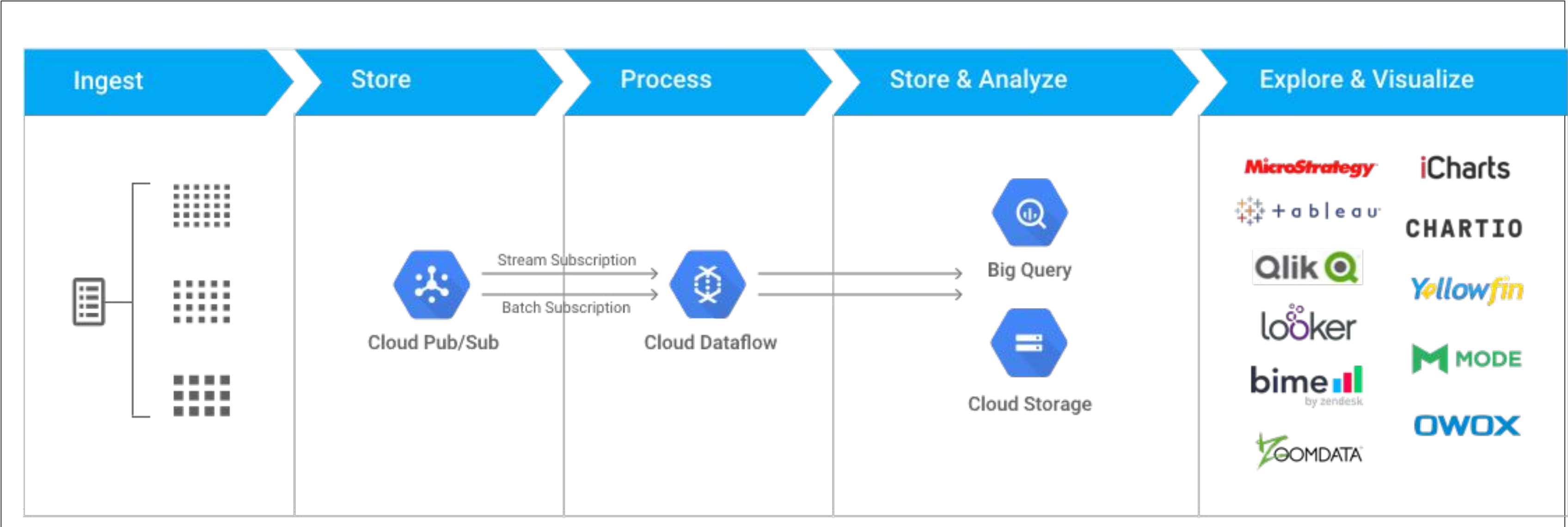
Cloud Pub/Sub offers reliable, real-time messaging

Distributed Messaging with Cloud Pub/Sub



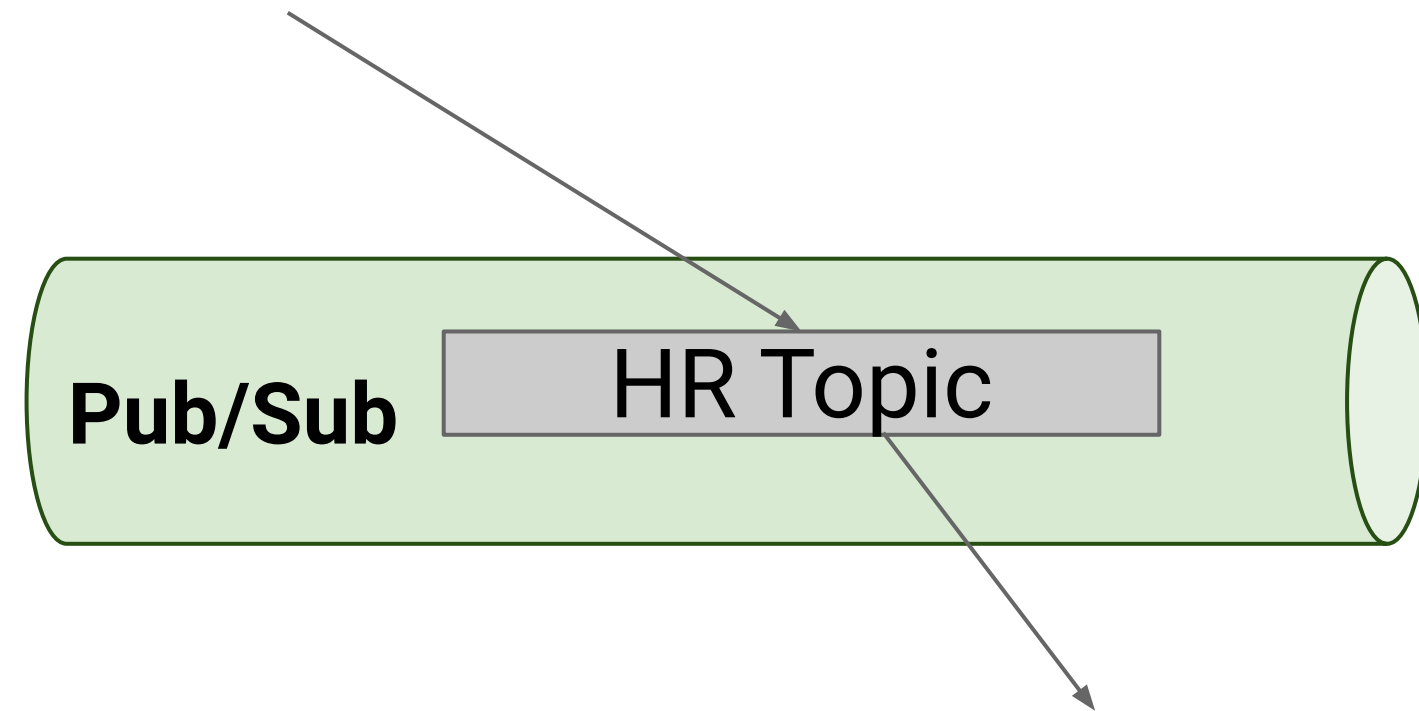
- At-least-once delivery
- Exactly-once processing
- No provisioning, auto-everything
- Open APIs
- Global by default
- End-to-end encryption

Google Cloud Serverless Big Data Pipeline



Pub/Sub topics are like radio antennas

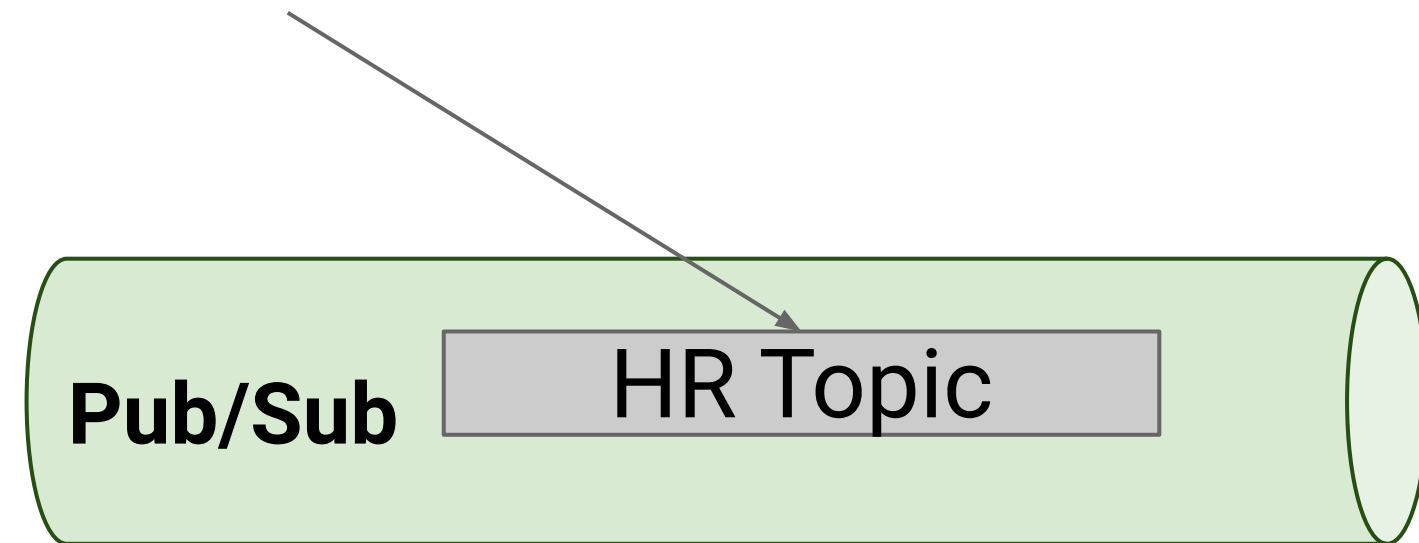
Publishers



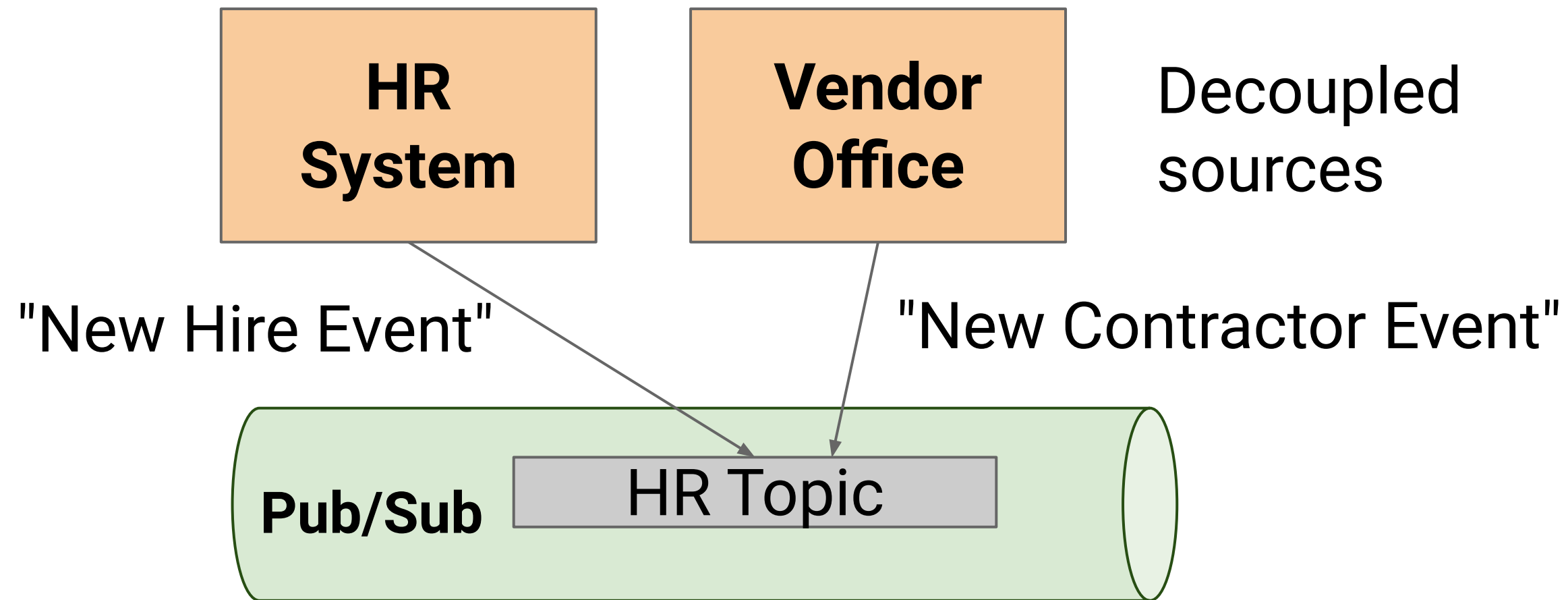
Subscribers

Scenario: HR messaging system

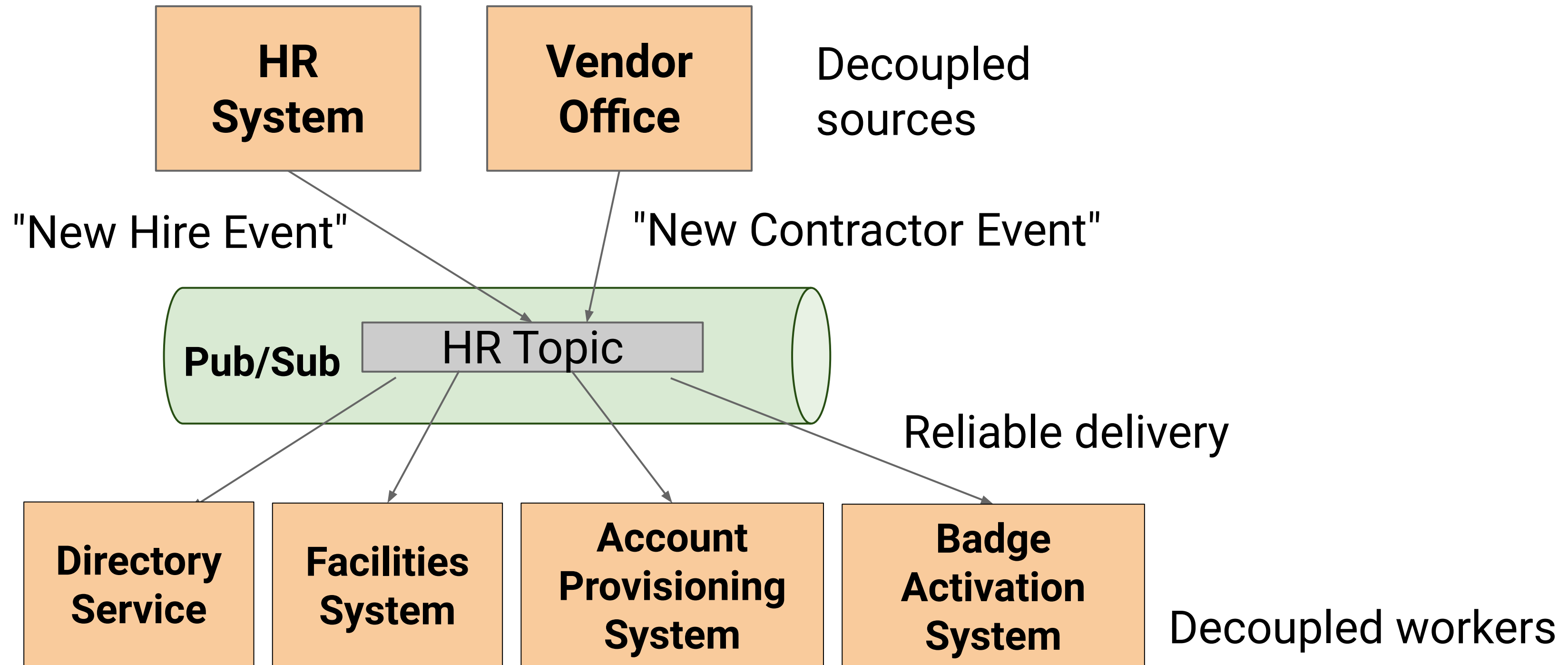
"New Hire Event"



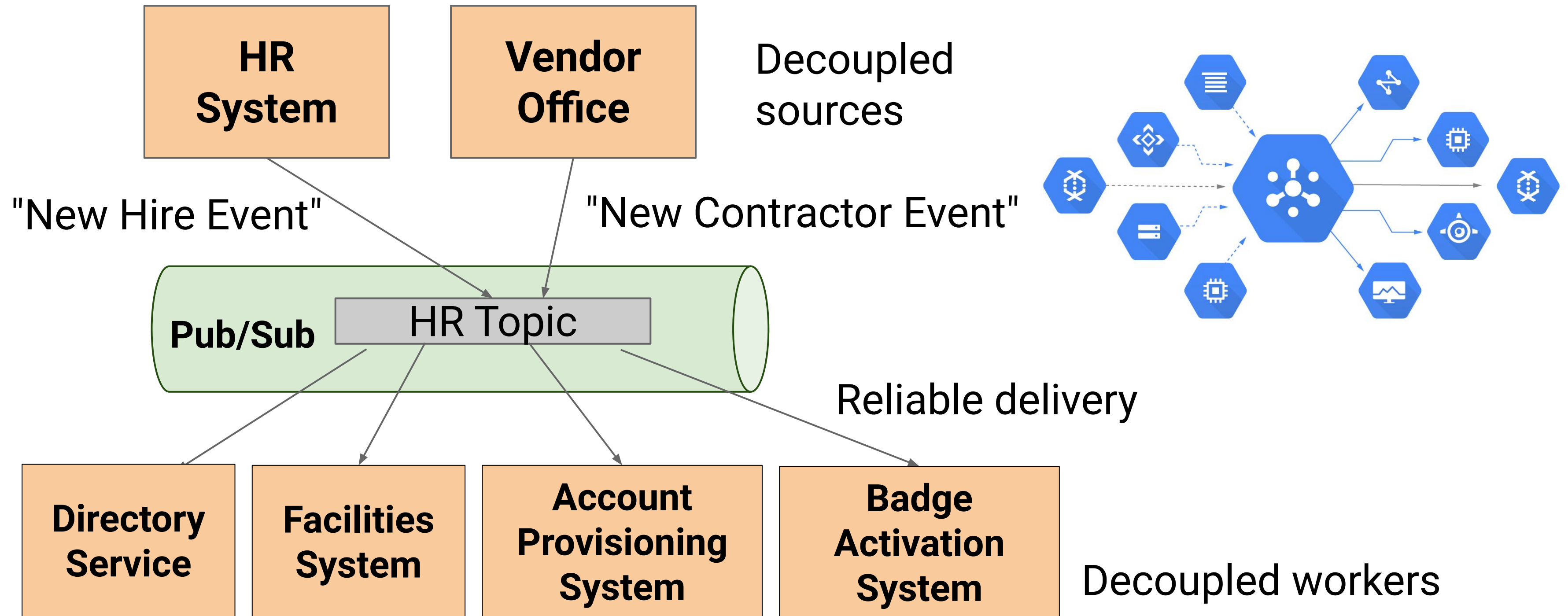
Scenario: HR messaging system



Scenario: HR messaging system



Scenario: HR messaging system



Agenda

Modern data pipeline challenges

Message-oriented architectures

Serverless data pipelines

- Designing streaming pipelines with Apache Beam
- Implementing streaming pipelines on Cloud Dataflow

Data Visualization with Data Studio

- Building collaborative dashboards
- Tips and tricks to create charts with the Data Studio UI

Data Engineers need to solve two distinct problems

Pipeline design



Implementation



Data Engineers need to solve two distinct problems

Pipeline design



- Will my code work with both batch and streaming data?
- Does the SDK support the transformations I need to do?
- Are there existing solutions?

Implementation



Data Engineers need to solve two distinct problems

Pipeline design with Apache Beam



- Will my code work with both batch and streaming data? Yes
- Does the SDK support the transformations I need to do? Likely
- Are there existing solutions? Choose from templates

Start with provided templates and build from there:

github.com/GoogleCloudPlatform/DataflowTemplates

- BigQuery to Datastore
- Bigtable to GCS Avro
- Bulk Compressor
- Bulk Decompressor
- Datastore Bulk Delete *
- Datastore to BigQuery
- Datastore to GCS Text *
- Datastore to Pub/Sub *
- Datastore Unique Schema Count

- GCS Avro to Bigtable
- GCS Avro to Spanner
- GCS Text to BigQuery *
- GCS Text to Datastore
- GCS Text to Pub/Sub (Batch)
- GCS Text to Pub/Sub (Streaming)
- Jdbc to BigQuery

- Pub/Sub to BigQuery *
- Pub/Sub to Datastore *
- Pub/Sub to GCS Avro
- Pub/Sub to GCS Text
- Pub/Sub to Pub/Sub
- Spanner to GCS Avro
- Spanner to GCS Text
- Word Count

Agenda

Modern data pipeline challenges

Message-oriented architectures

Serverless data pipelines

- Designing streaming pipelines with Apache Beam
- Implementing streaming pipelines on Cloud Dataflow

Data Visualization with Data Studio

- Building collaborative dashboards
- Tips and tricks to create charts with the Data Studio UI

What is Apache Beam?

Beam is an advanced unified & portable data processing programming model

- Programming model
- SDKs for writing data pipelines
- Runners to run distributed processing



Why Apache Beam?

- **Unified** - Use a single programming model for both batch and streaming use cases
- **Portable** - Execute pipelines on multiple execution environments
- **Extensible** - Write and share new SDKs, IO connectors, and transformation libraries

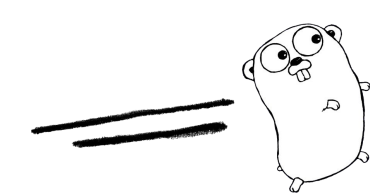
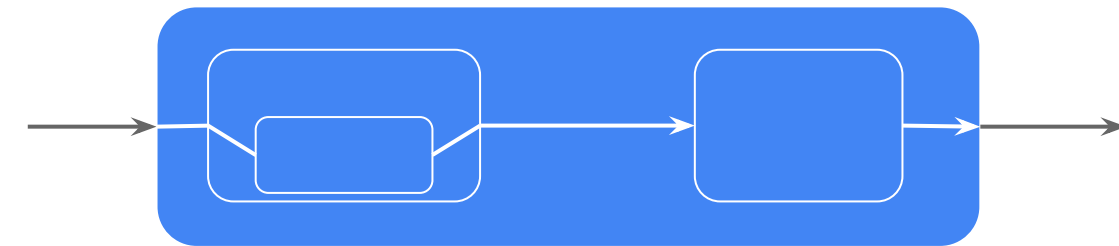
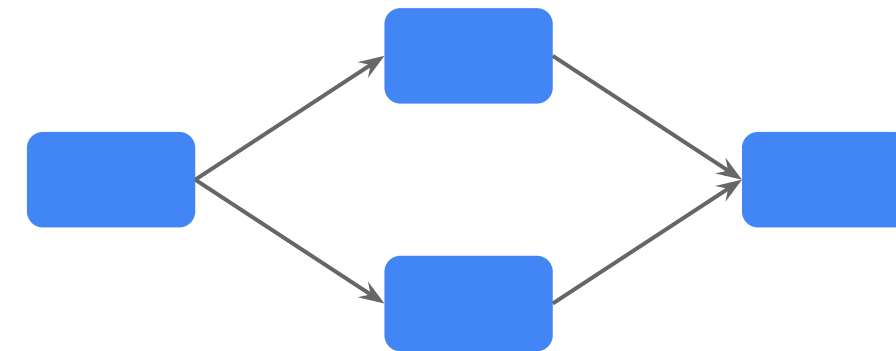
User code

Libraries of PTransforms, IO

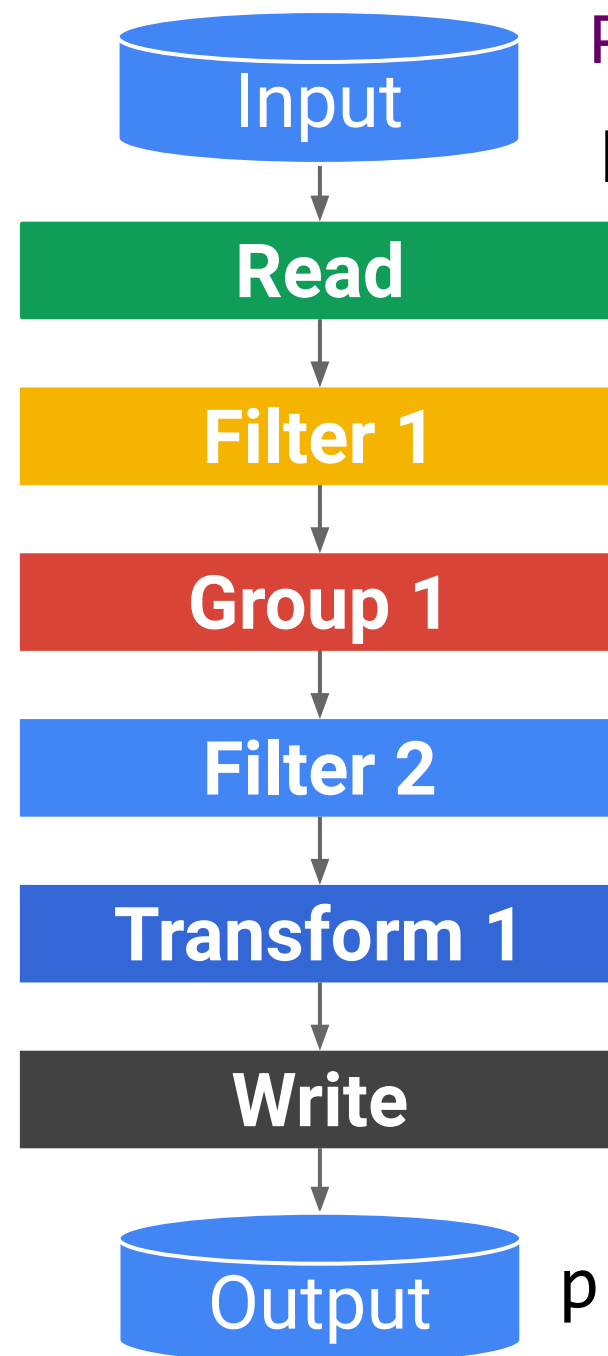
Language SDK

Beam Model representation

Runner



Dataflow offers NoOps data pipelines



```
Pipeline p = Pipeline.create();
```

```
p  
  .apply(TextIO.Read.from("gs://...  
  "))
```

```
  .apply(ParDo.of(new Filter1()))
```

```
  .apply(new Group1())
```

```
  .apply(ParDo.of(new Filter2()))
```

```
  .apply(new Transform1())
```

```
  .apply(TextIO.Write.to("gs://..."  
  );
```

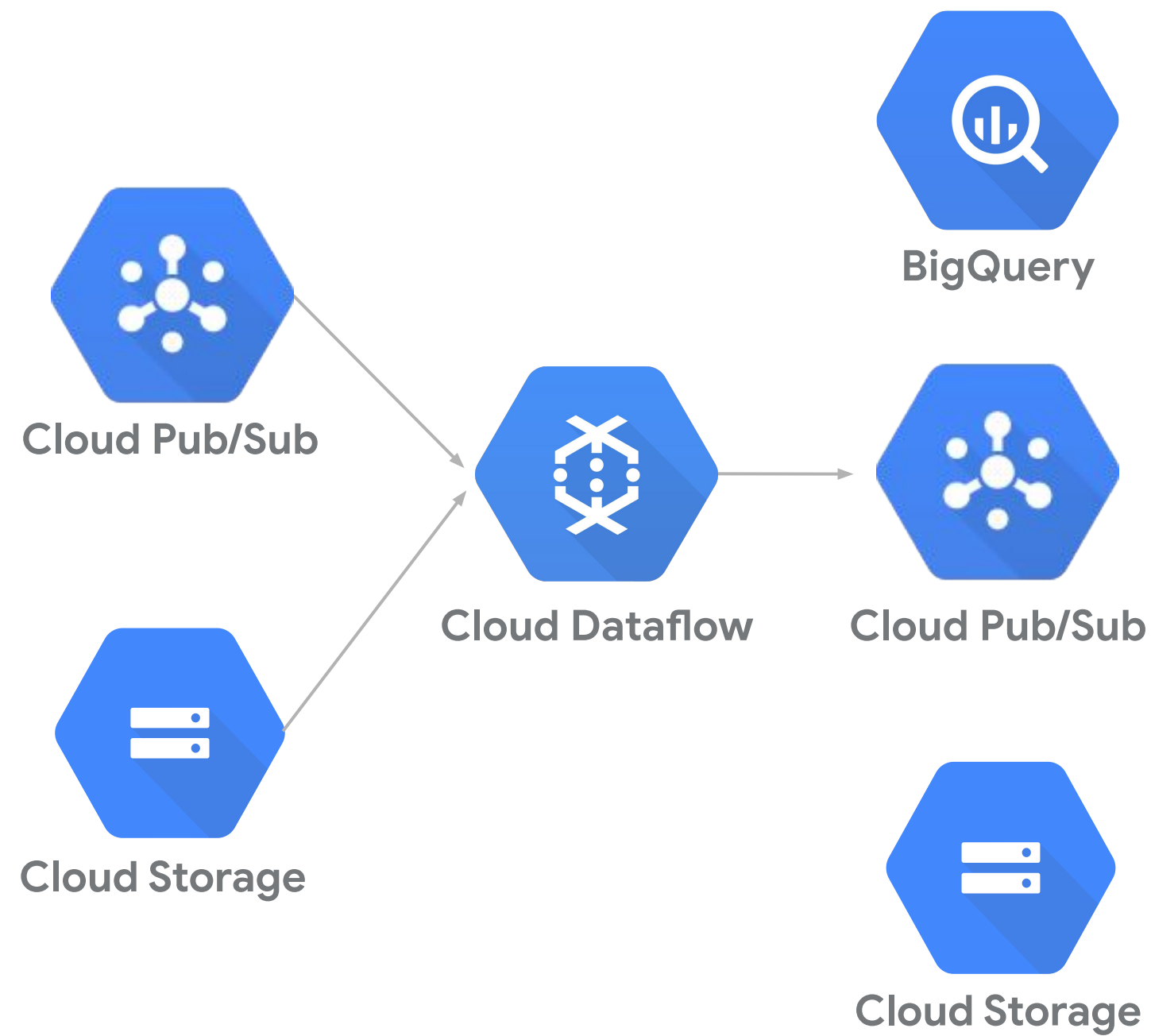
```
p.run();
```

Open-source API (Apache Beam) can be executed on Flink, Spark, etc. also

Parallel task
(autoscaled by execution framework)

```
class Filter1 extends DoFn<...> {  
  public void  
    processElement(ProcessContext c) {  
    ... = c.element();  
    ...  
    c.output(...);  
  }  
}
```

Same code does real-time and batch



```
Pipeline p = Pipeline.create();
p.begin()
    .apply(PubsubIO.Read.from("input_topic"))
    .apply(SlidingWindows.of(60, MINUTES))
    .apply(ParDo.of(new Filter1()))
    .apply(new Group1())
    .apply(ParDo.of(new Filter2()))
    .apply(new Transform1())
    .apply(PubsubIO.Write.to("output_topic"));
p.run();
```

Agenda

Modern data pipeline challenges

Message-oriented architectures

Serverless data pipelines

- Designing streaming pipelines with Apache Beam
- Implementing streaming pipelines on Cloud Dataflow

Data Visualization with Data Studio

- Building collaborative dashboards
- Tips and tricks to create charts with the Data Studio UI

Data Engineers need to solve two distinct problems

Pipeline design



- Will my code work with both batch and streaming data?
- Does the SDK support the transformations I need to do?
- Are there existing solutions?

Implementation



- How much maintenance overhead is involved?
- Is the infrastructure reliable?
- How is scaling handled?
- How can I monitor and alert?
- Am I locked in to a vendor?

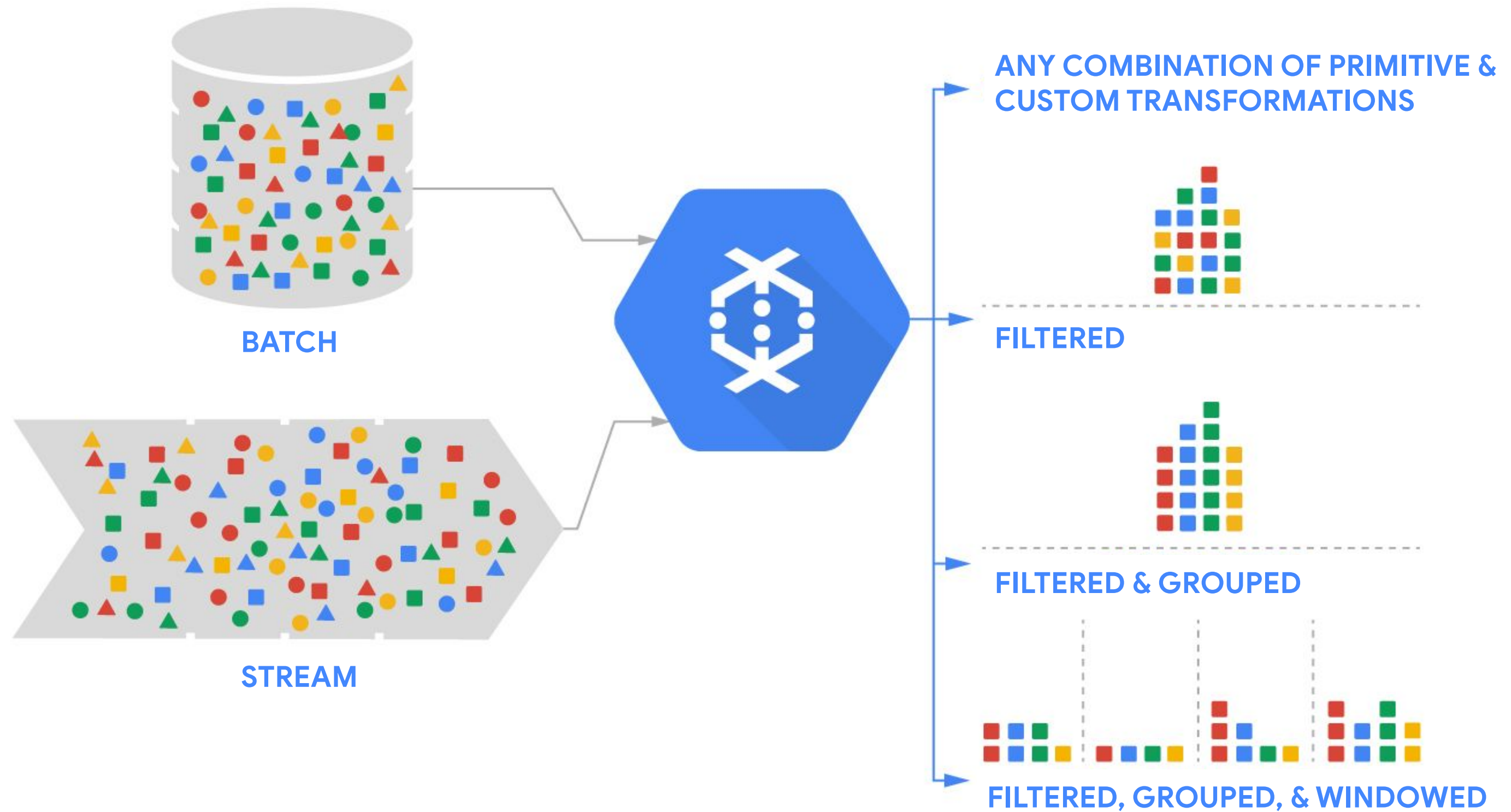
Data Engineers need to solve two distinct problems

Implementation with Google Cloud Dataflow

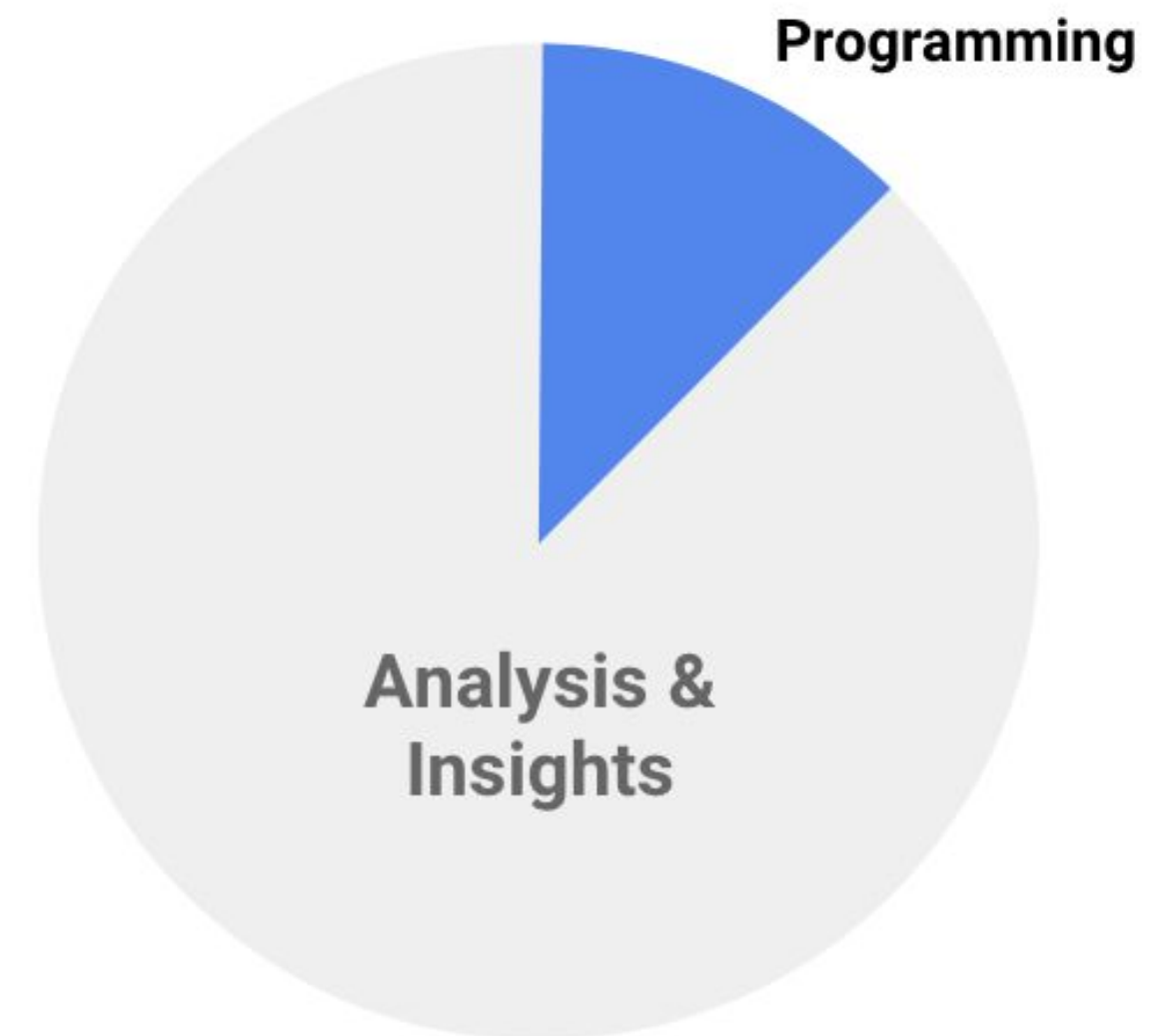
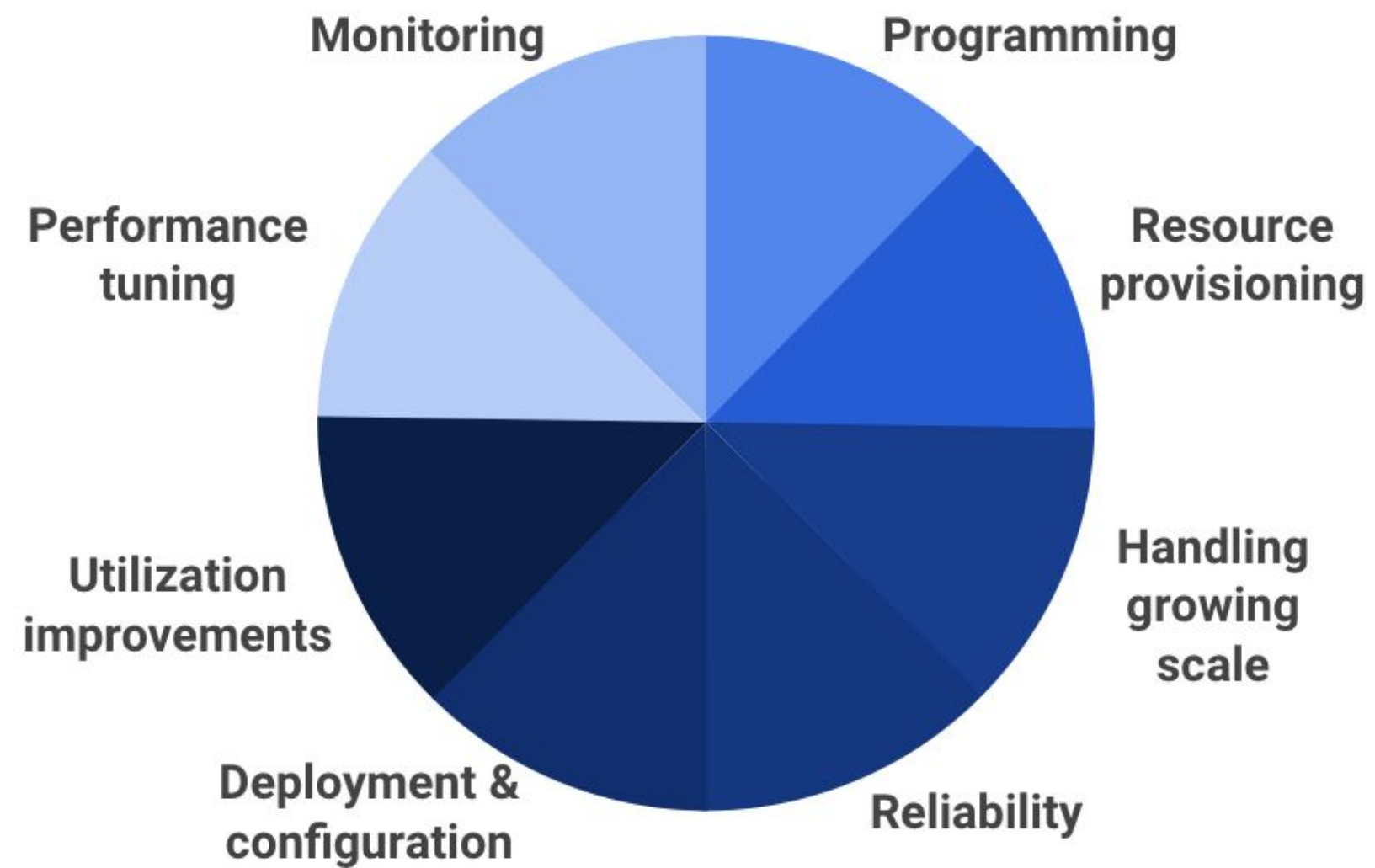


- How much maintenance overhead is involved? Little
- Is the infrastructure reliable? Built on Google infrastructure
- How is scaling handled? Autoscale workers
- How can I monitor and alert? Integrated with Stackdriver
- Am I locked in to a vendor? Run Apache Beam elsewhere

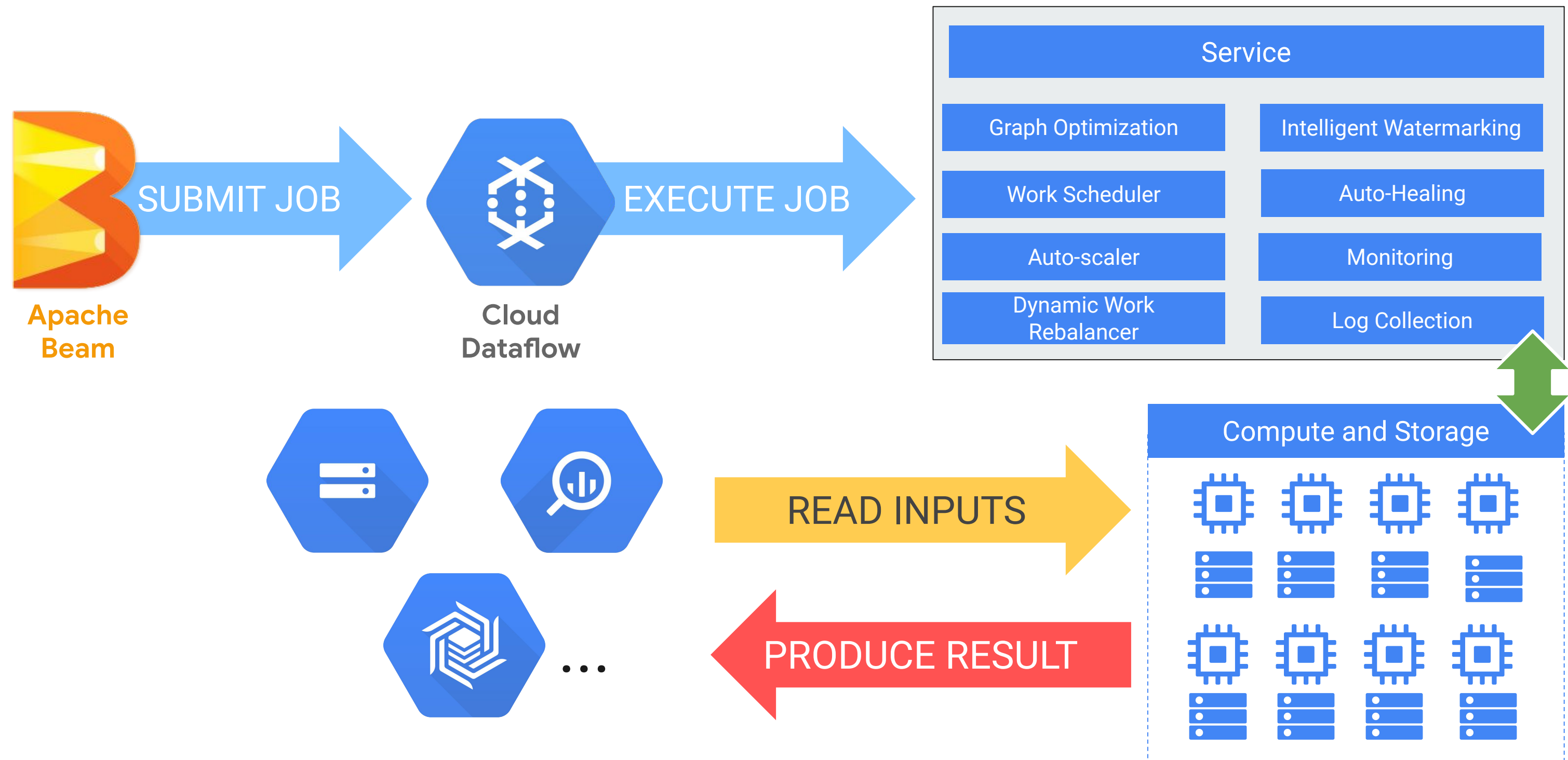
Dataflow does ingest, transform, and load



Why Serverless?




Workflow with Dataflow



Cloud Dataflow



- Serverless, fully managed data processing
- Unified batch and streaming processing + autoscale
- Open source programming model using  beam
- Intelligently scales to millions of QPS

Agenda

Modern data pipeline challenges

Message-oriented architectures

Serverless data pipelines

- Designing streaming pipelines with Apache Beam
- Implementing streaming pipelines on Cloud Dataflow

Data Visualization with Data Studio

- Building collaborative dashboards
- Tips and tricks to create charts with the Data Studio UI

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
```

Run

Save query

Save view

Schedule query

More

Query results

SAVE RESULTS

EXPLORE IN DATA STUDIO

Query complete (2.5 sec elapsed, 118.1 MB processed)

Job information

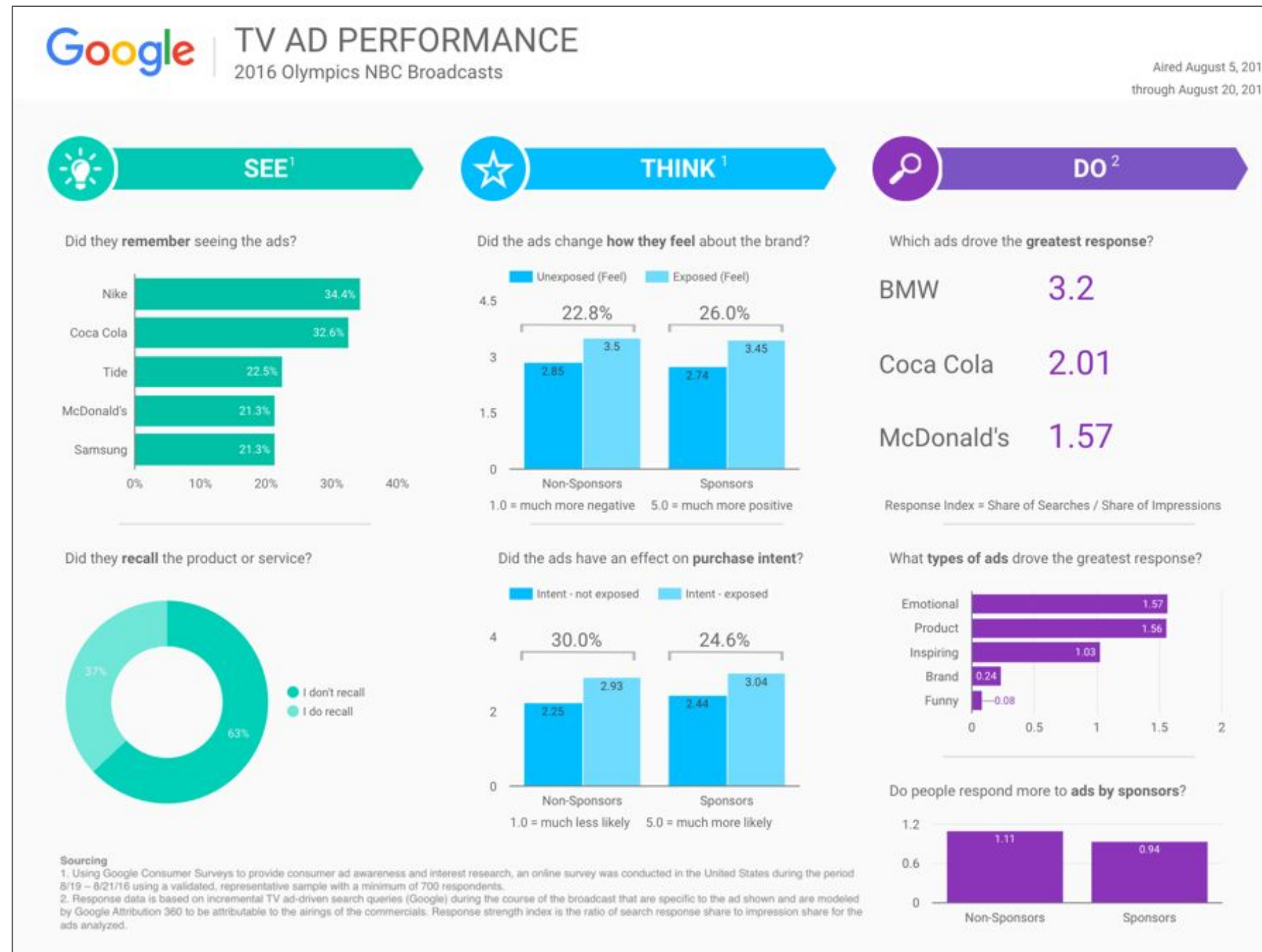
Results

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

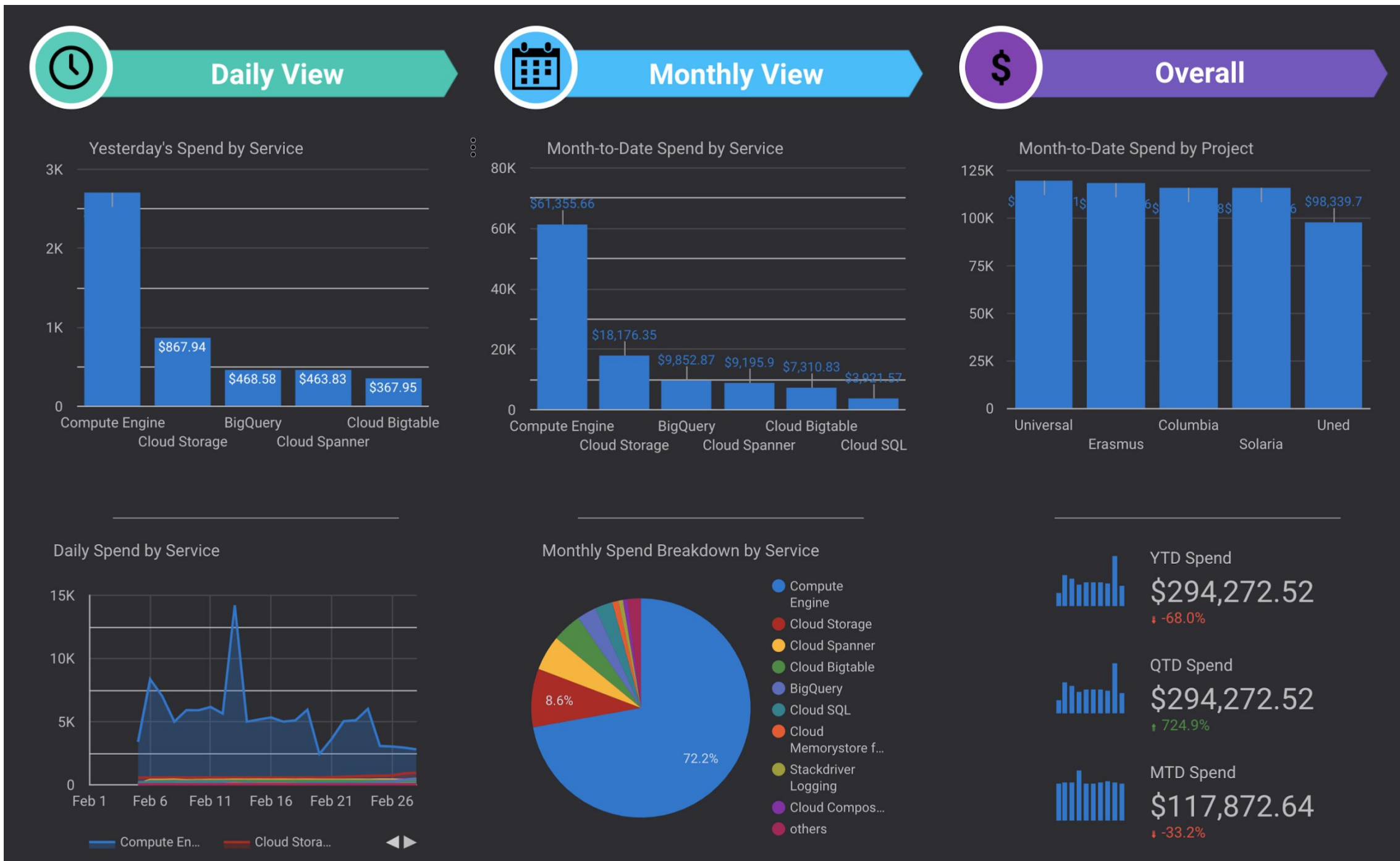
Build, collaborate, and share your dashboards



Tell a clear story with your data

Share and collaborate on reports
with others

Access templates like this GCP Billing Dashboard



Create new reports in the Data Studio UI

Data Studio beta

Home

Start a new report

+

Blank

ACME

53,206

66,104

327,396

47.29%

How are site sessions trending?

What are the top countries by sessions?

Which channels are driving engagement?

Acme Marketing
Google Analytics

Ecommerce PPC Dashboard

\$170.1K

1,024

\$297.8

25.5K

Top 5 Products - Ad Sources

Top 5 Campaigns - AdWords

Ecommerce PPC
Google Analytics + Adwords

Google Adwords

Click Through Rate & Cost

Conversion Rate & Cost

Cost Per Click

Top Campaigns

Device Breakdown

AdWords Overview
Google Adwords

ALL TEMPLATES

ALLOWNED BY MESHARED WITH METRASH

welcome

XAZ

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

Oct 18, 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

+

Data Studio Home

Select data sources to build your visualizations

Untitled Report
File View Page Help

VIEW

+ person

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.

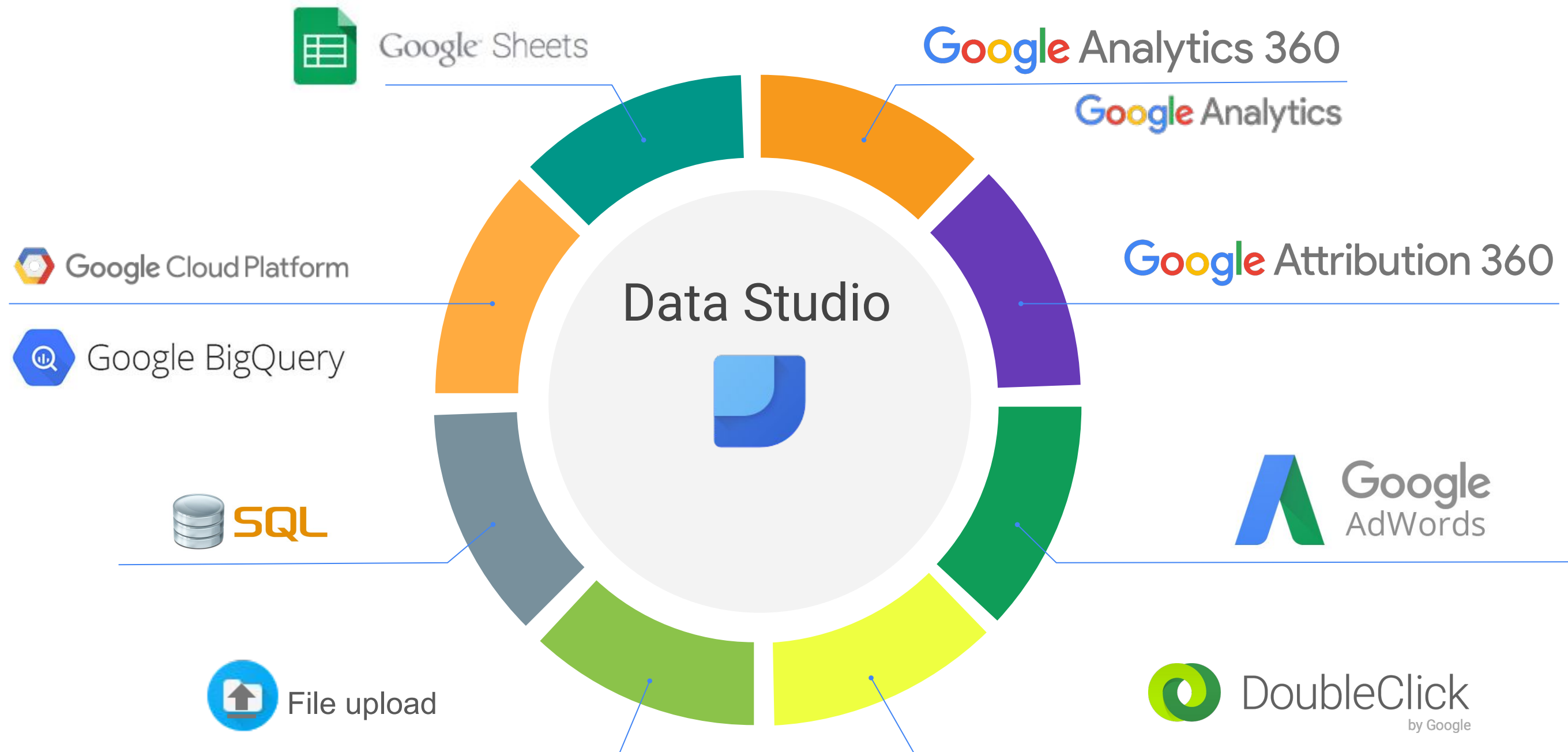
OKAY, GOT IT

Select Data Source

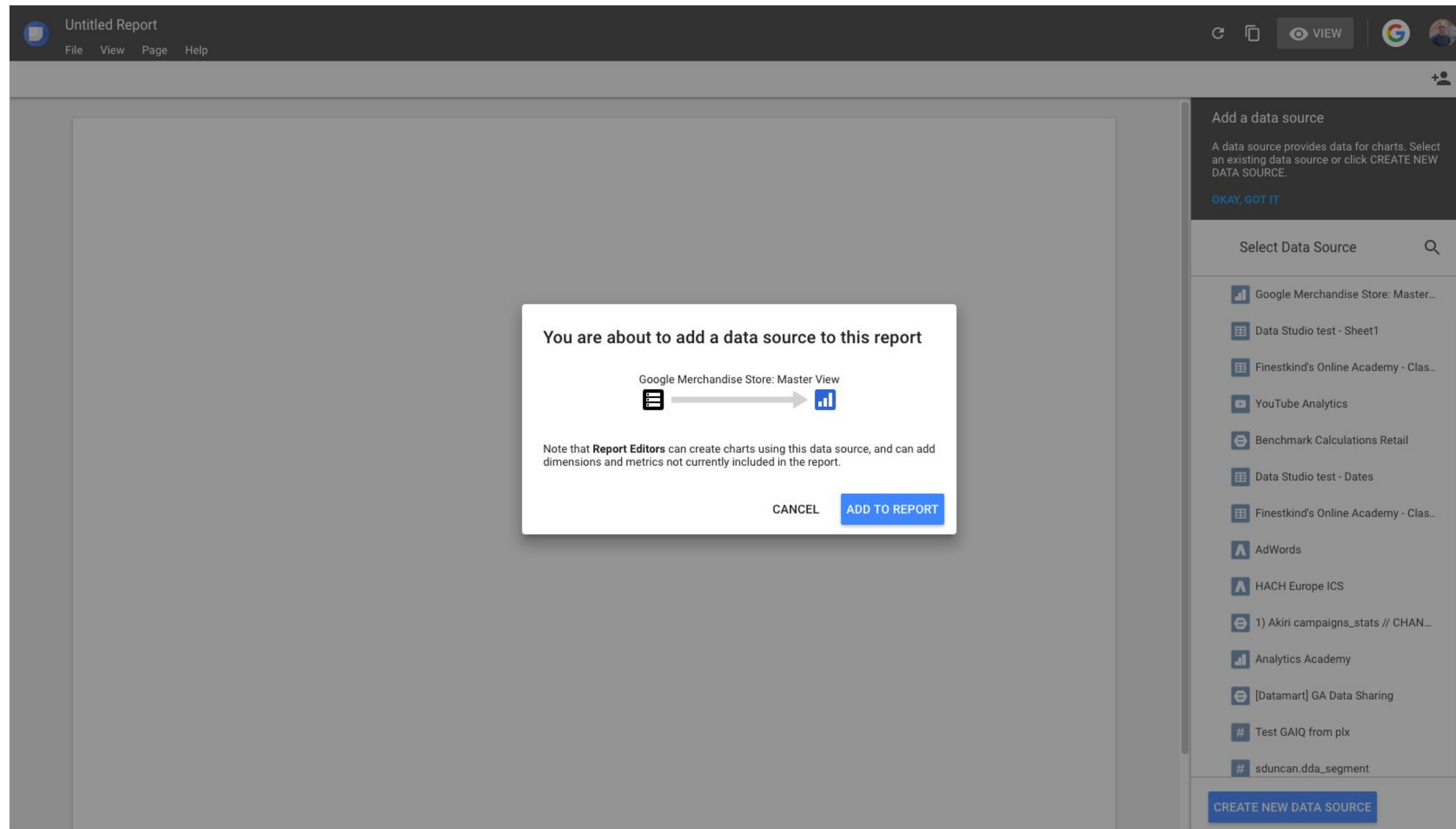
- YouTube Analytics
- Benchmark Calculations Retail
- Data Studio test - Dates
- Finestkind's Online Academy - Clas...
- AdWords
- HACH Europe ICS
- 1) Akiri campaigns_stats // CHAN...
- Analytics Academy
- [Datamart] GA Data Sharing
- Test GAIQ from plx
- sduncan.dda_segment
- datastudio_metrics.usage.all
- 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



Agenda

Modern data pipeline challenges

Message-oriented architectures

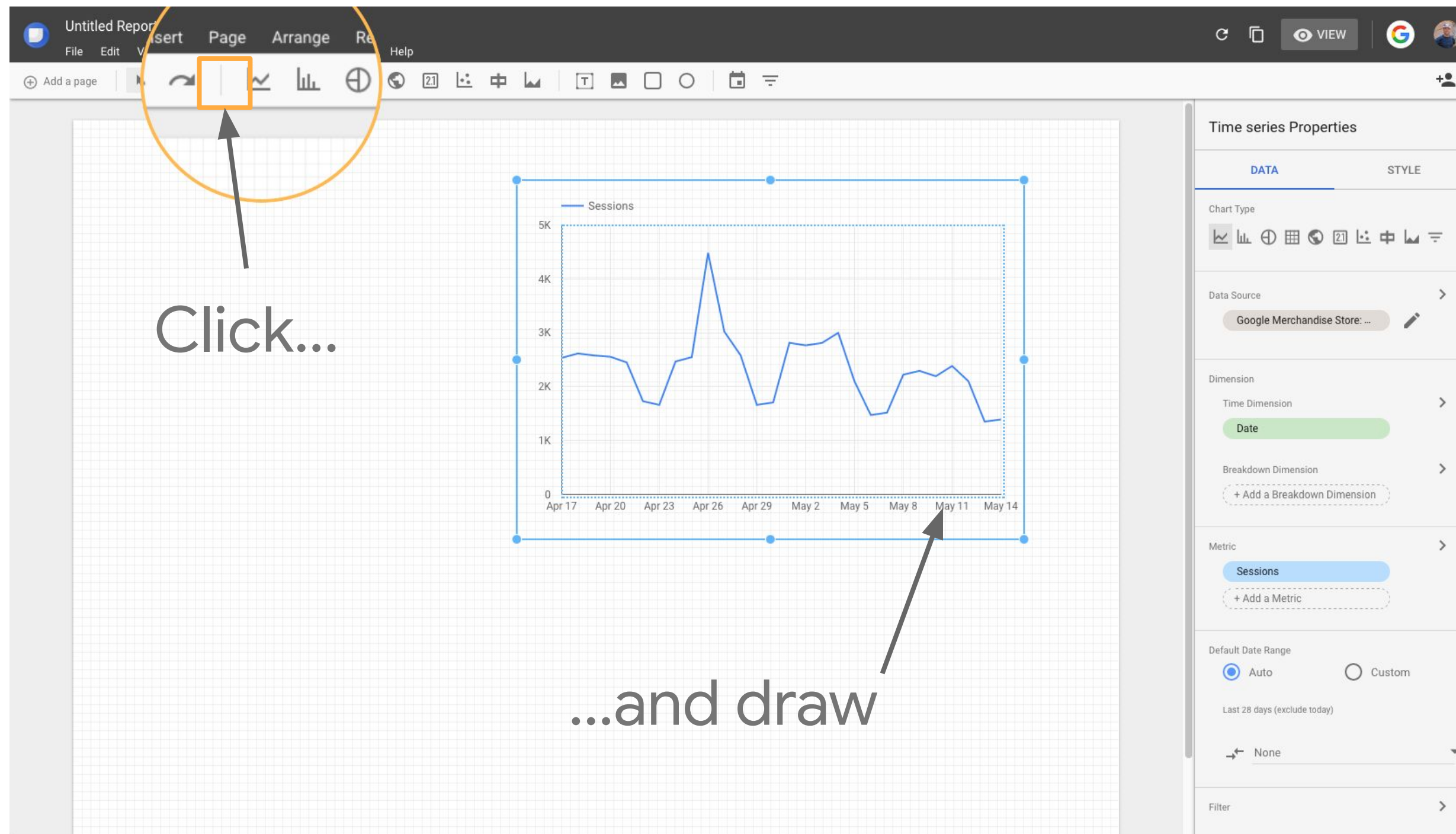
Serverless data pipelines

- Designing streaming pipelines with Apache Beam
- Implementing streaming pipelines on Cloud Dataflow

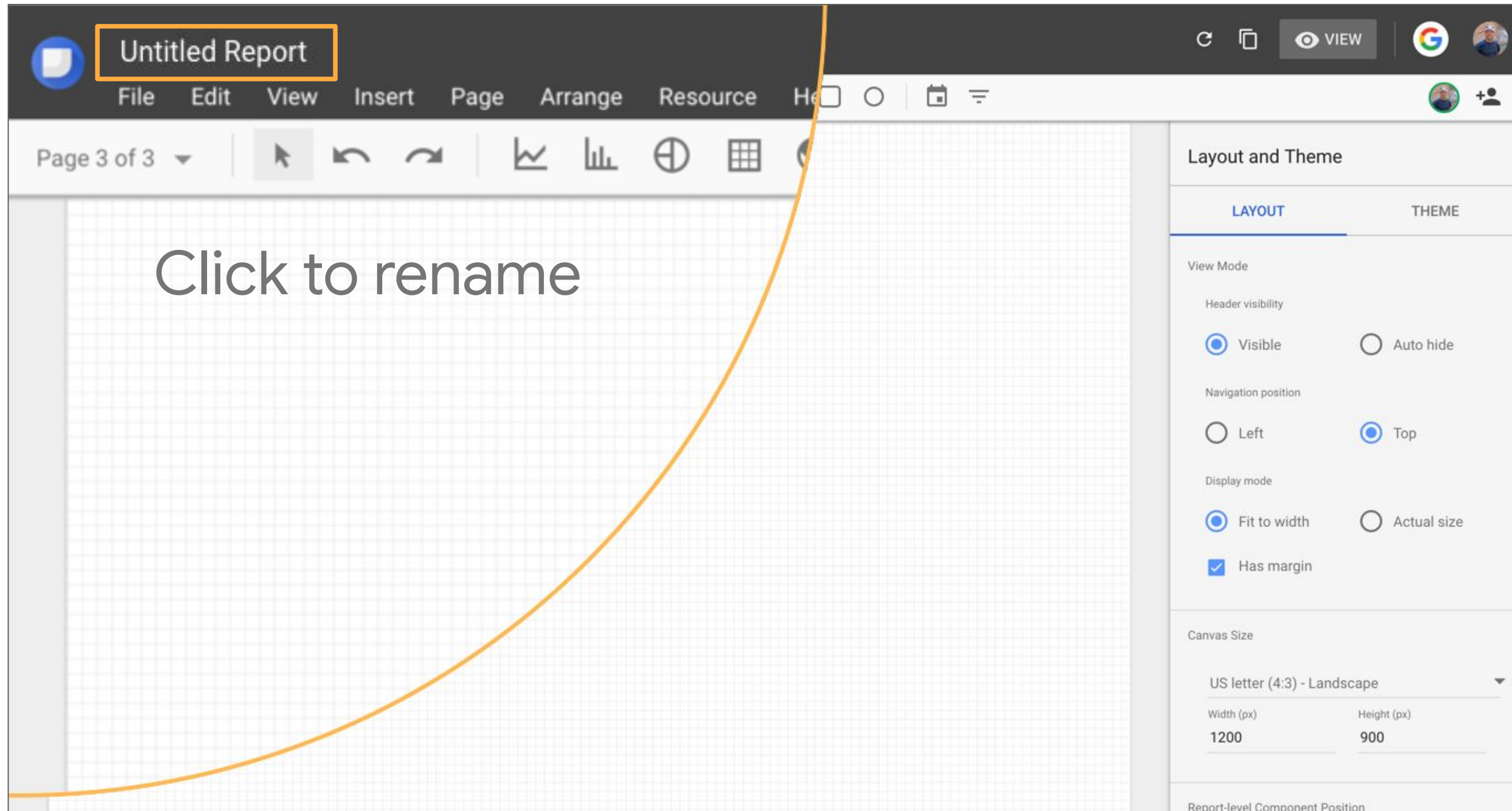
Data Visualization with Data Studio

- Building collaborative dashboards
- Tips and tricks to create charts with the Data Studio UI

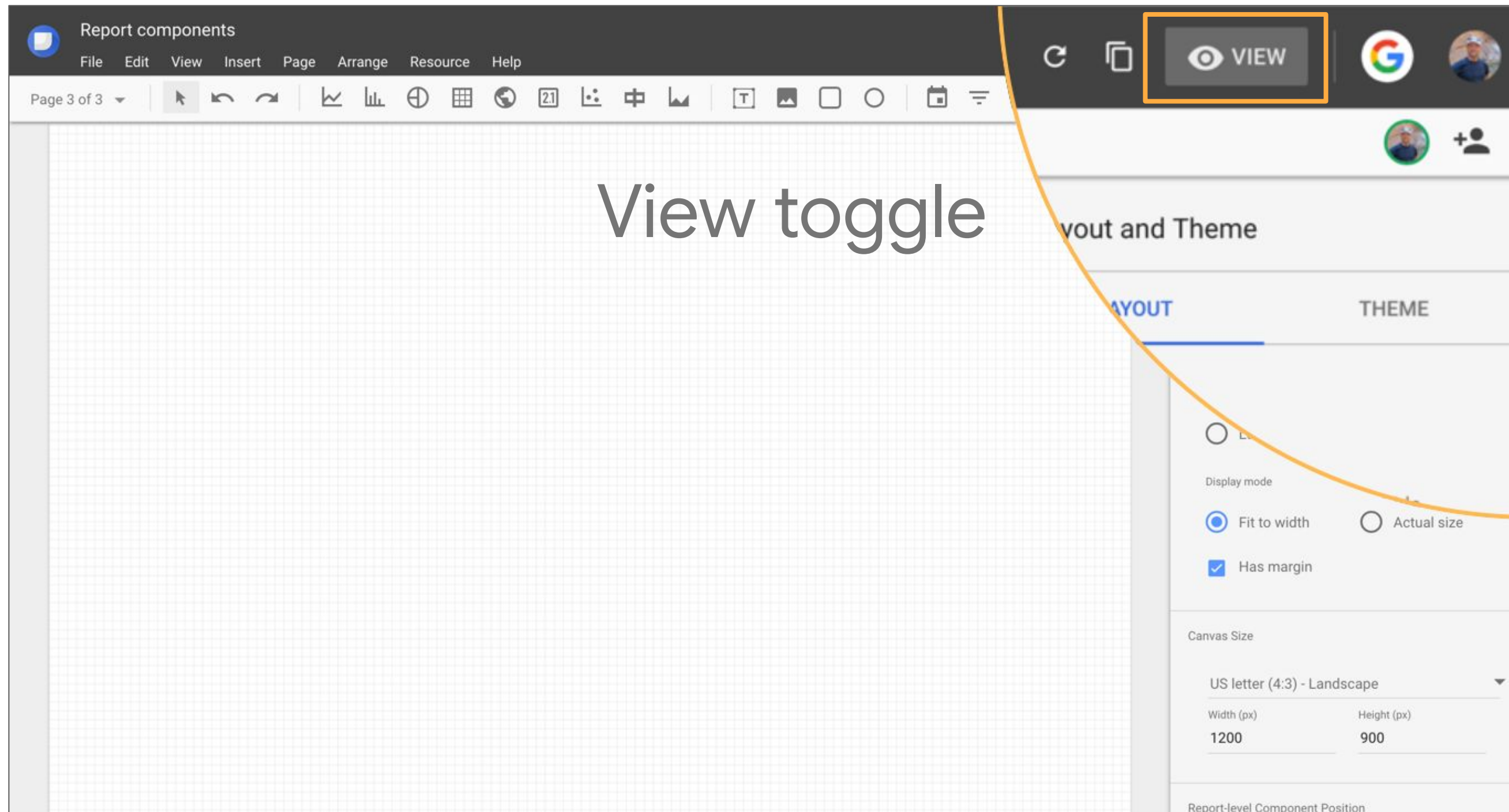
Create charts to visualize data relationships



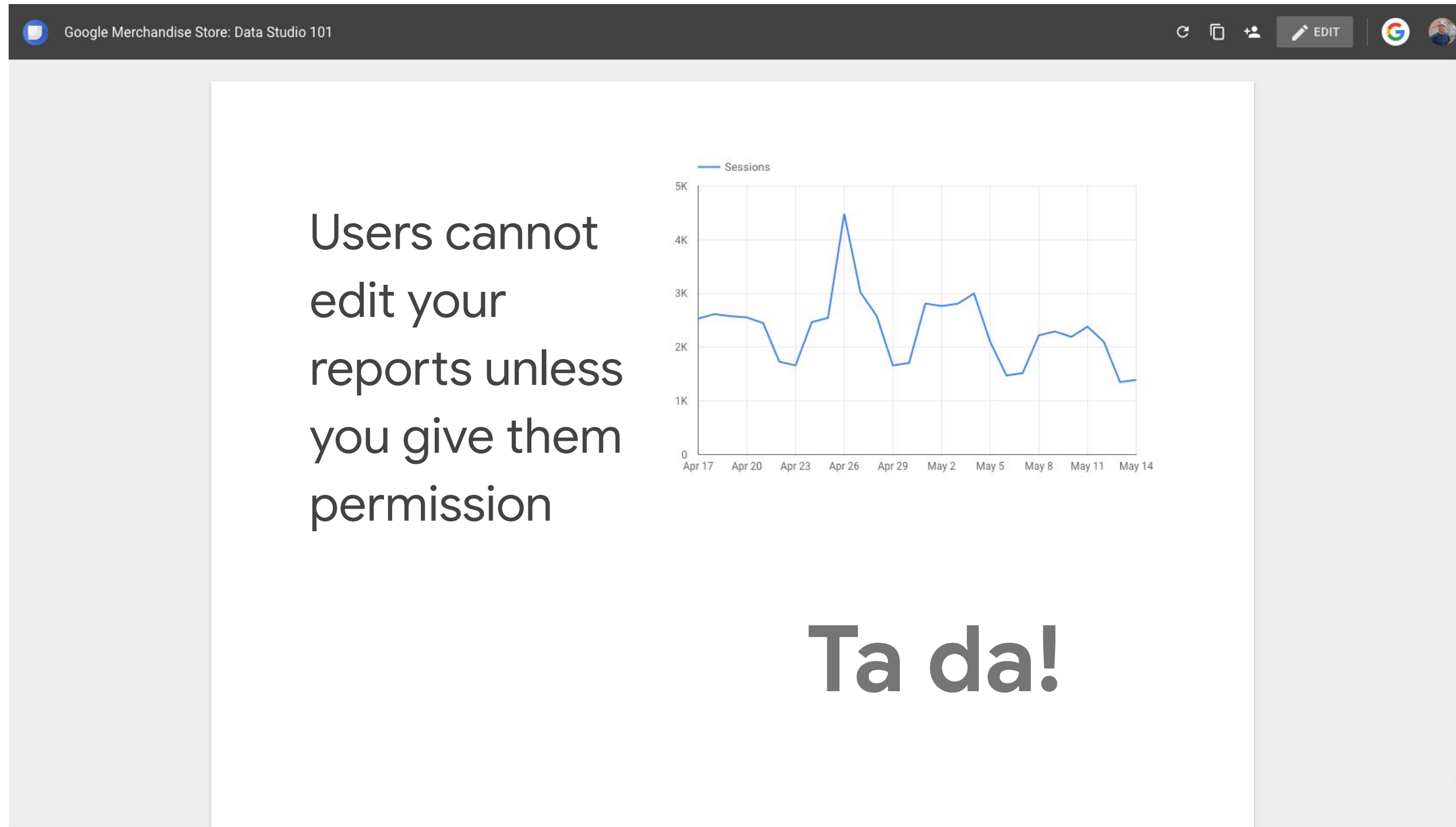
Add a descriptive name to your report



View the end-user version of the report



View your report as an end-user



Understand the date source shade

AdWords

USING OWNER'S CREDENTIALS

CREATE REPORT

← EDIT CONNECTION

Dimensions

Index	Field	Aggregation
1	Calculated field	Auto
2	Metro area code	None
3	Country code	None
4	Value / conv.	Auto
5	Avg. CPP	Auto

Understand the date source shade

AdWords

USING OWNER'S CREDENTIALS

CREATE REPORT

← EDIT CONNECTION

Metrics

Index	Field	Aggregation
1	Calculated field	Auto
2	Metro area code	None
3	Country code	None
4	Value / conv.	Auto
5	Avg. CPP	Auto

Understand the date source shade

AdWords

USING OWNER'S CREDENTIALS

CREATE REPORT

← EDIT CONNECTION

Index	Field	Type	Aggregation
1	Calculated field	123 Number	Auto
2	Metro area code	Metro Code	None
3	Country code	Country Code	None
4	Value / conv.	123 Number	Auto
5	Avg. CPP	123 Currency (USD - US Dollar (\$))	Auto

Calculated fields can be either metrics or dimensions

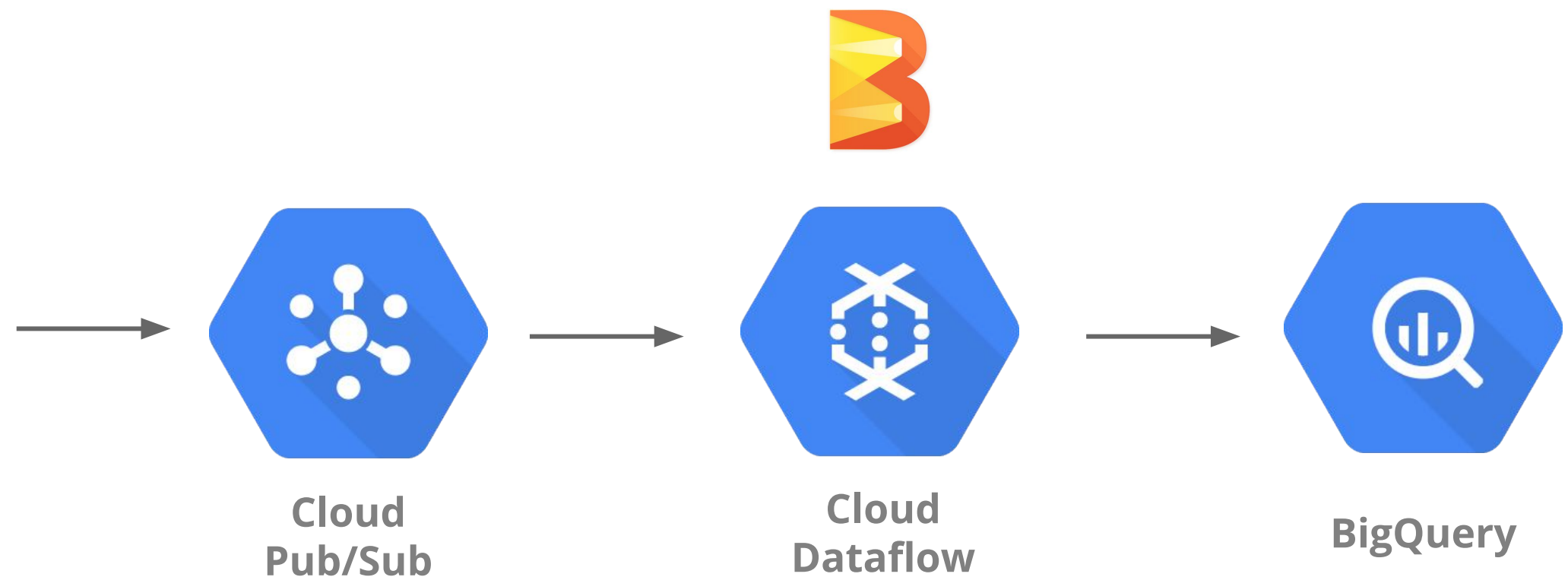
Sharing and Collaborating on Dashboards



- Data Studio uses Google Drive for sharing and storing files.
 - When you share a report with view permission, no login is required to view the report. A Google login is required to edit a report.
- Sharing a report does NOT share direct access to any added data sources.
- Data sources must be shared separately from reports.

Lab: Real-time dashboards with Pub/Sub, Dataflow, and Data Studio

How can I monitor streaming insights for my business?



Lab

Streaming Data into BigQuery with Pub/Sub and DataFlow

- Setup streaming taxi cab topic in Pub/Sub
- Create Dataflow job from template
- Stream and monitor pipeline in BigQuery
- Analyze results and create views
- Visualize key metrics in Data Studio