Data Processing Architecture

Google Cloud Platform Fundamentals: Big Data and Machine Learning

Version #1.1

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Agenda

1. Introduction
2. Fundamentals of GCP
3. Data Analysis on the Cloud
4. Scaling Data Analysis
5. Machine Learning
6. Data Processing Architecture

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Agenda

Message-oriented architectures

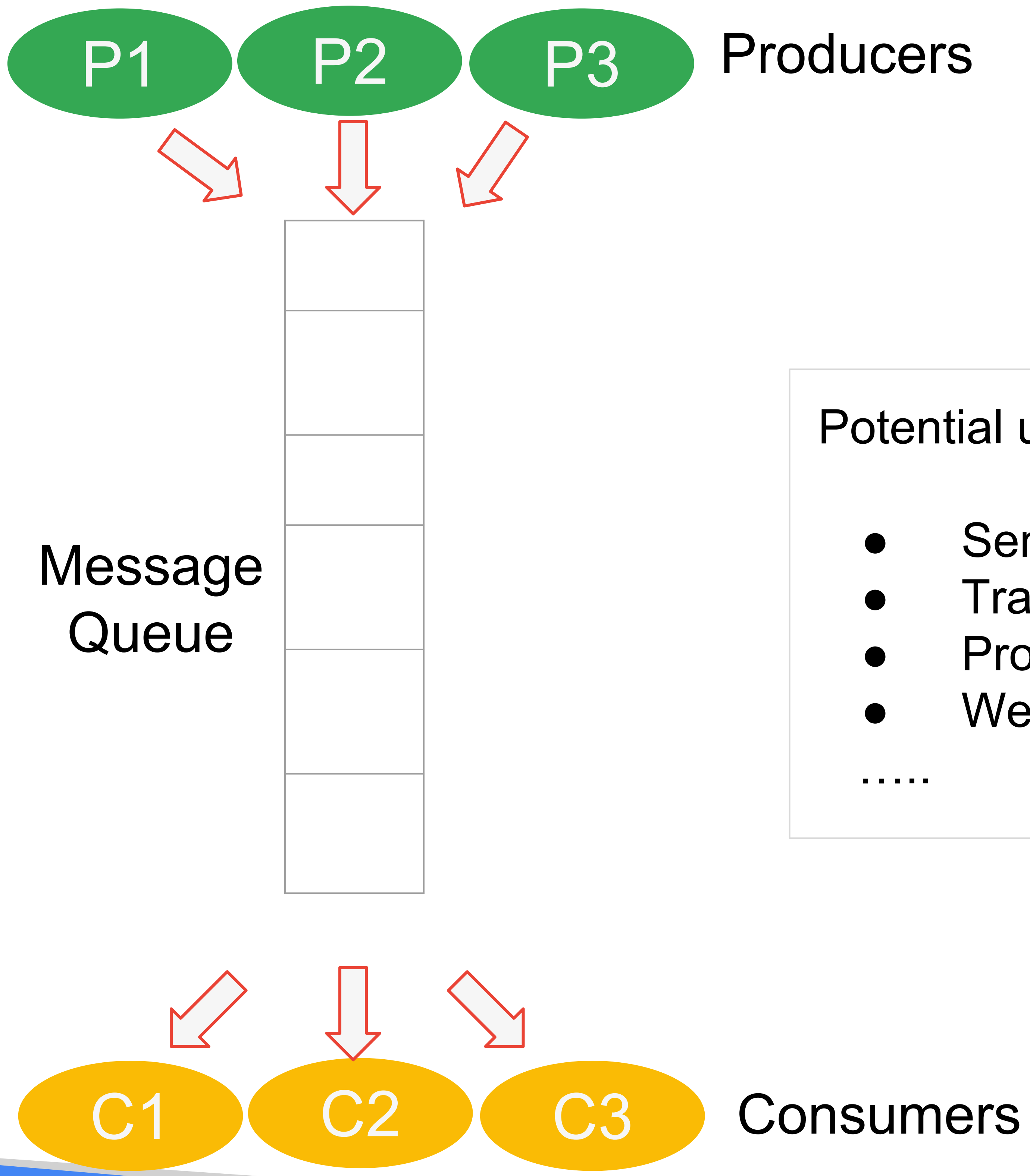
Serverless data pipelines

GCP Reference Architecture

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Asynchronous processing is useful for long-lived tasks or to have loose coupling between two systems



Potential use cases:

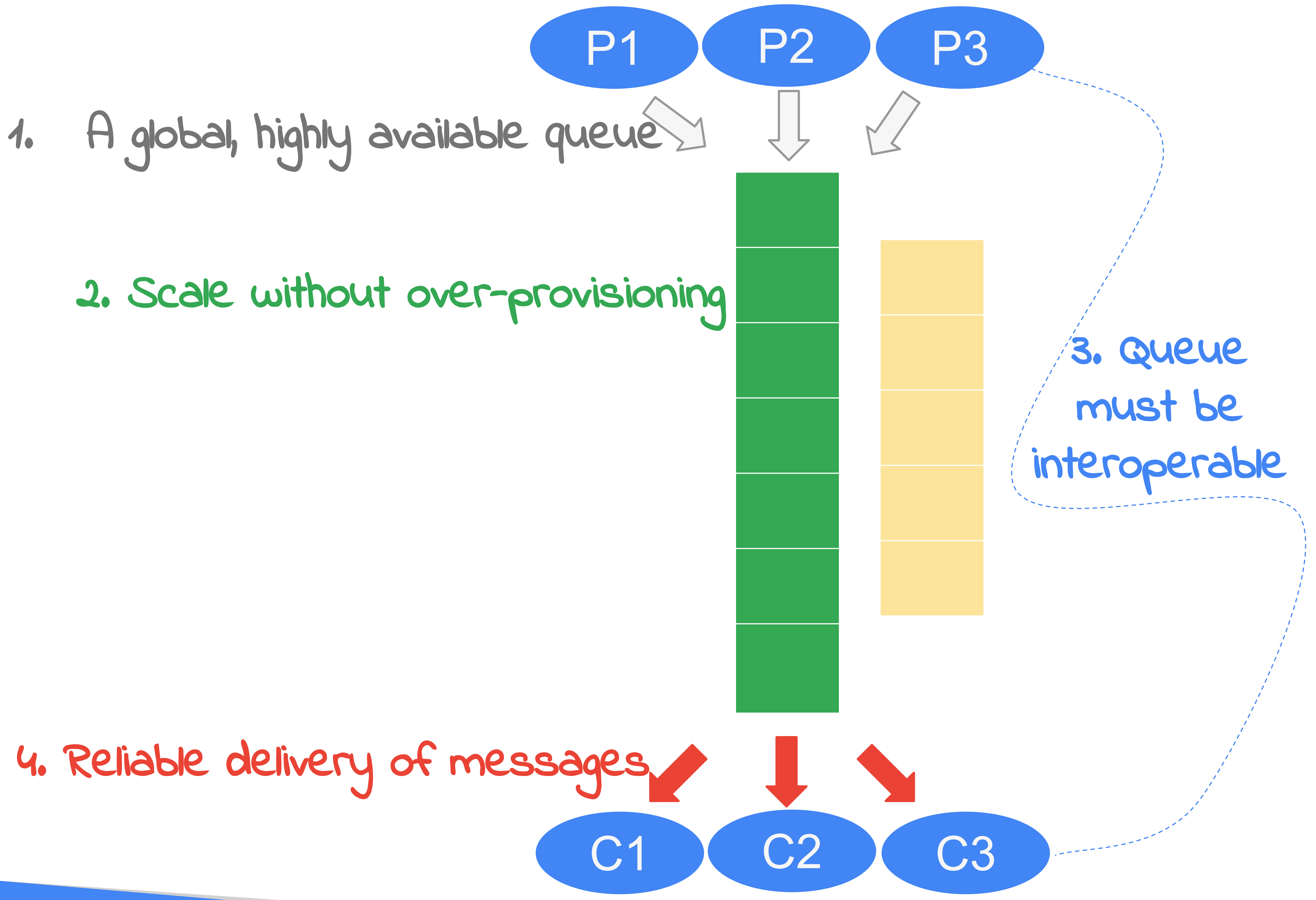
- Send an SMS
- Train ML model
- Process data from multiple sources
- Weekly reports

…..

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

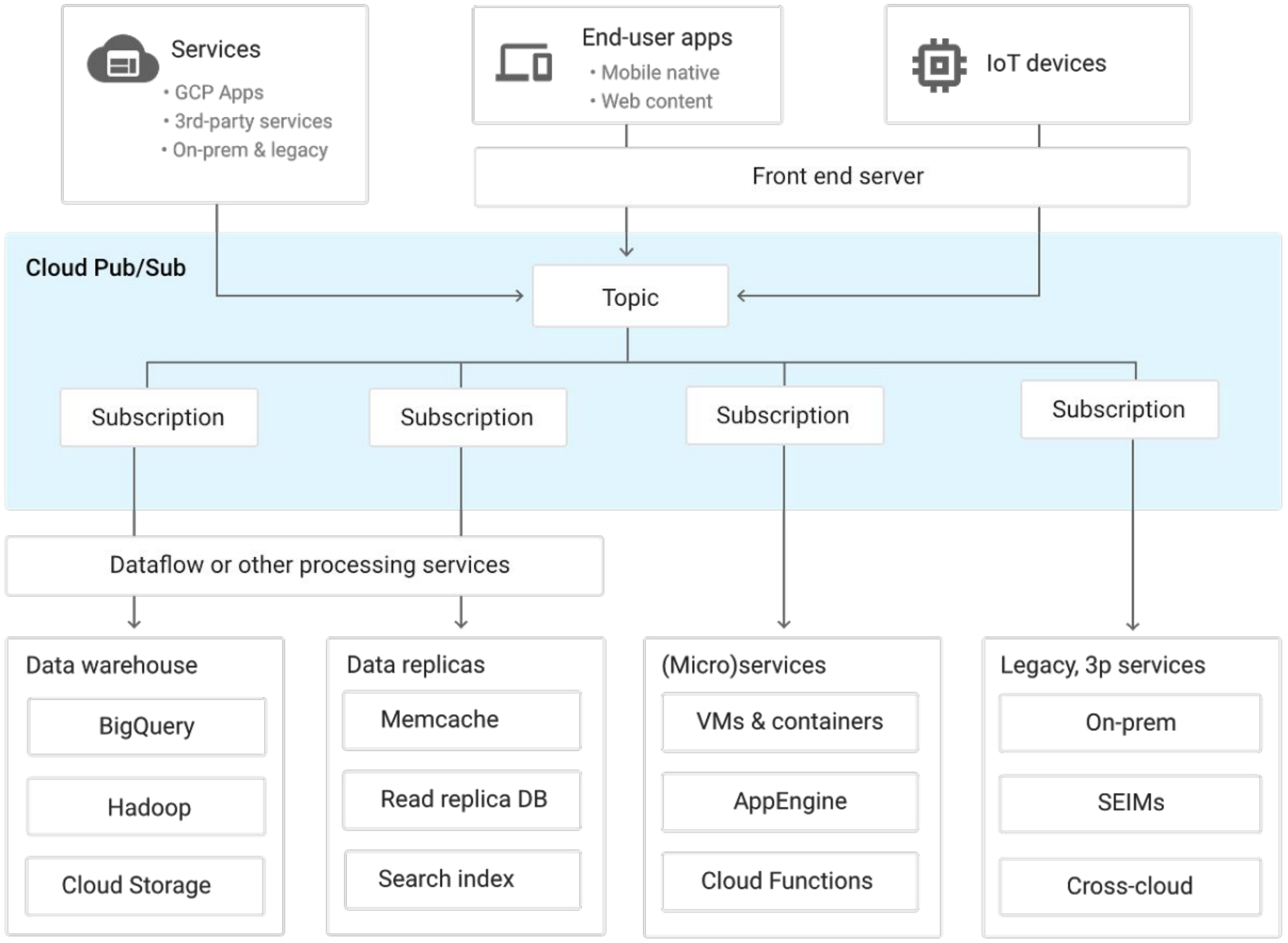
Google Cloud

For robust asynchronous processing, you need:


P1
P2
P3
1. A global, highly available queue
2. Scale without over-provisioning
3. Queue must be interoperable
4. Reliable delivery of messages
C1
C2
C3

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Pub/Sub provides a no-ops, serverless global message queue



© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Agenda

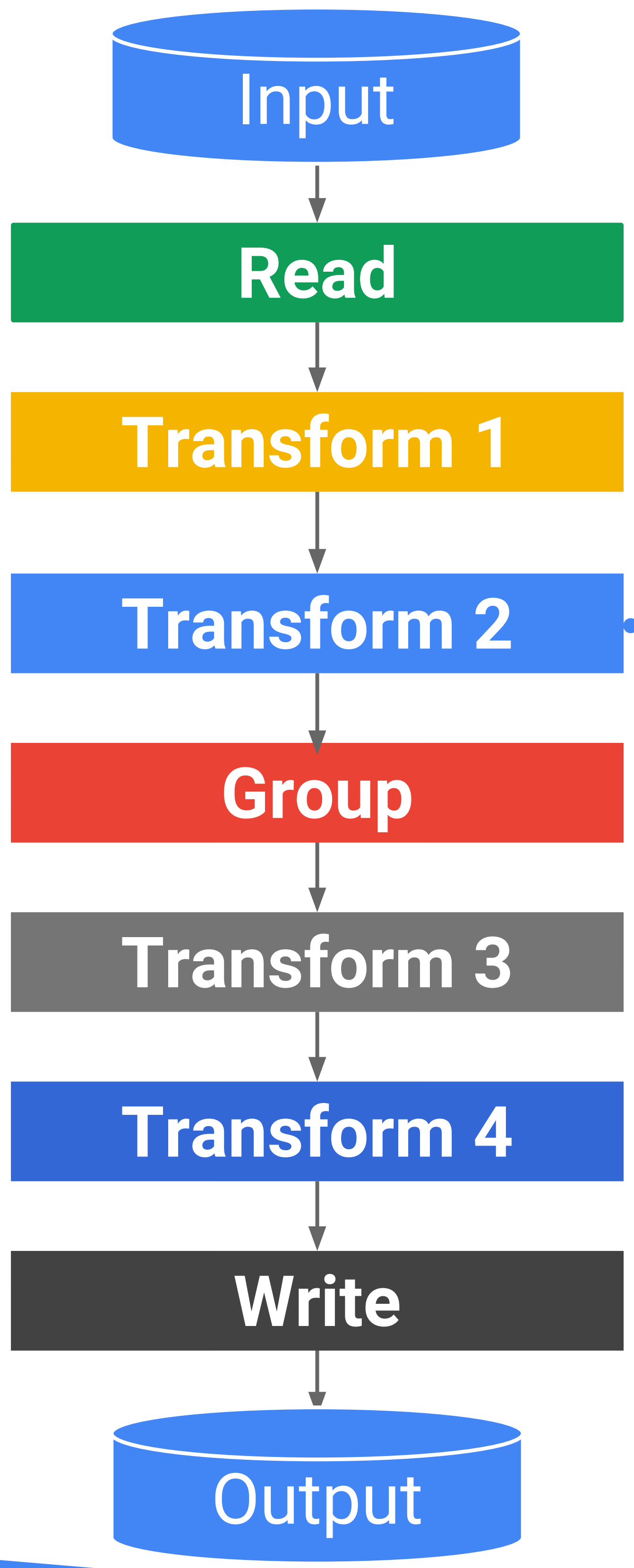
Message-oriented architectures

Serverless data pipelines

GCP Reference Architecture

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Dataflow offers NoOps data pipelines in Java and Python



```
p = beam.Pipeline(options=options)

  lines = p | beam.io.ReadFromText('gs://...')

  traffic = lines | beam.Map(parse_data).with_output_types(unicode)

                  | beam.Map(get_speedbysensor) # (sensor, speed)

                  | beam.GroupByKey() # (sensor, [speed])

                  | beam.Map(avg_speed) # (sensor, avgspeed)

                  | beam.Map(lambda tup: '%s: %d' % tup))

   output = traffic | beam.io.WriteToText('gs://...]')

p.run();
```

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

Map

Group-By

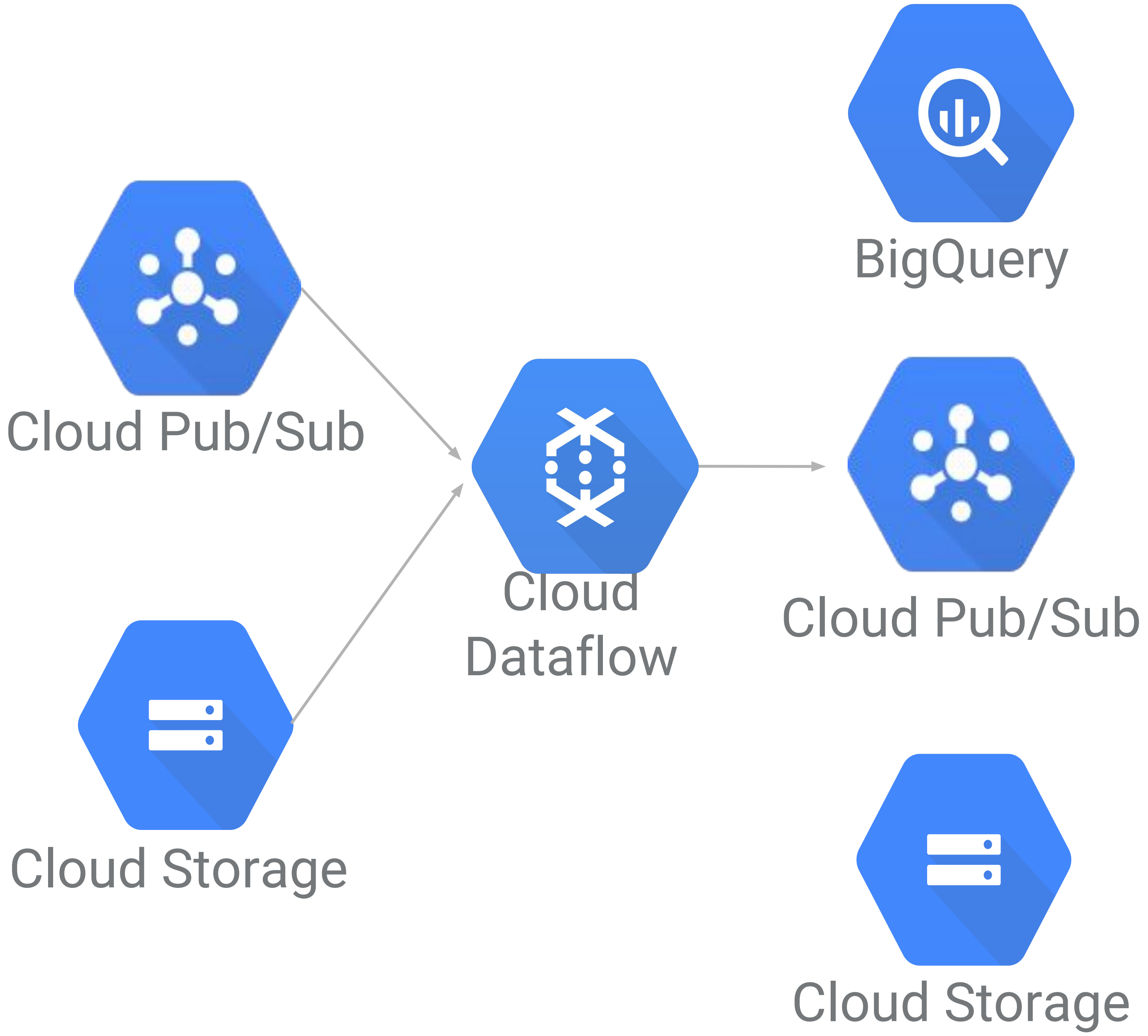
Reduce

Each of these steps is run in parallel and autoscaled by execution framework

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Same code does real-time and batch



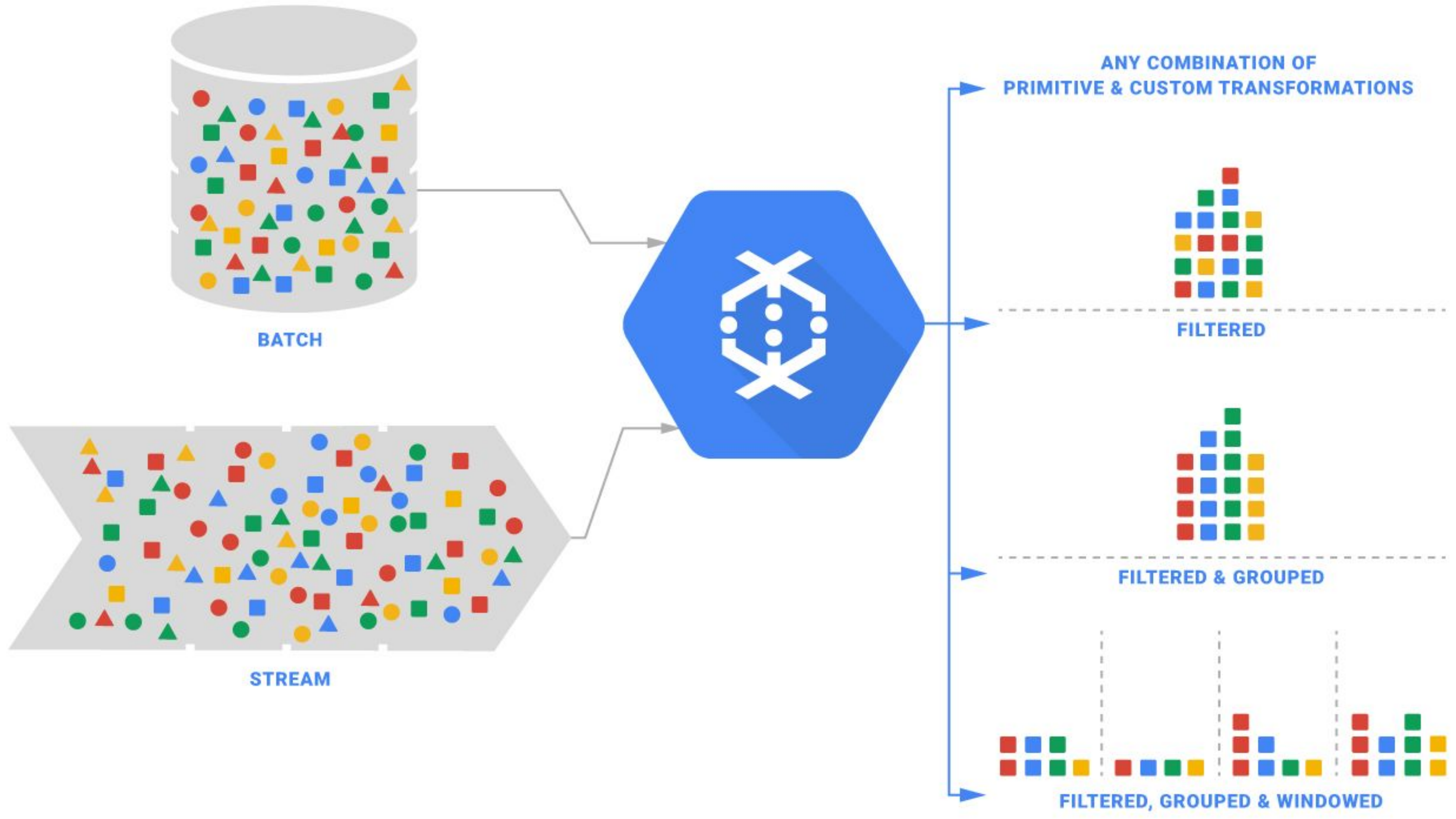
```
options = PipelineOptions(pipeline_args)
options.view_as(StandardOptions).streaming = True
p = beam.Pipeline(options=options)
   lines = p | beam.io.ReadStringsFromPubSub(input_topic)
   traffic = (lines
      | beam.Map(parse_data).with_output_types(unicode)
      | beam.Map(get_speedbysensor) # (sensor, speed)
      | beam.WindowInto(window.FixedWindows(15, 0))
      | beam.GroupByKey() # (sensor, [speed])
      | beam.Map(avg_speed) # (sensor, avgspeed)
      | beam.Map(lambda tup: '%s: %d' % tup))
   traffic | beam.io.WriteStringsToPubSub(output_topic)

p.run()
```

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Dataflow does ingest, transform, and load; consider using it instead of Spark


BATCH
STREAM
ANY COMBINATION OF PRIMITIVE & CUSTOM TRANSFORMATIONS
FILTERED
FILTERED & GROUPED
FILTERED, GROUPED & WINDOWED

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Agenda

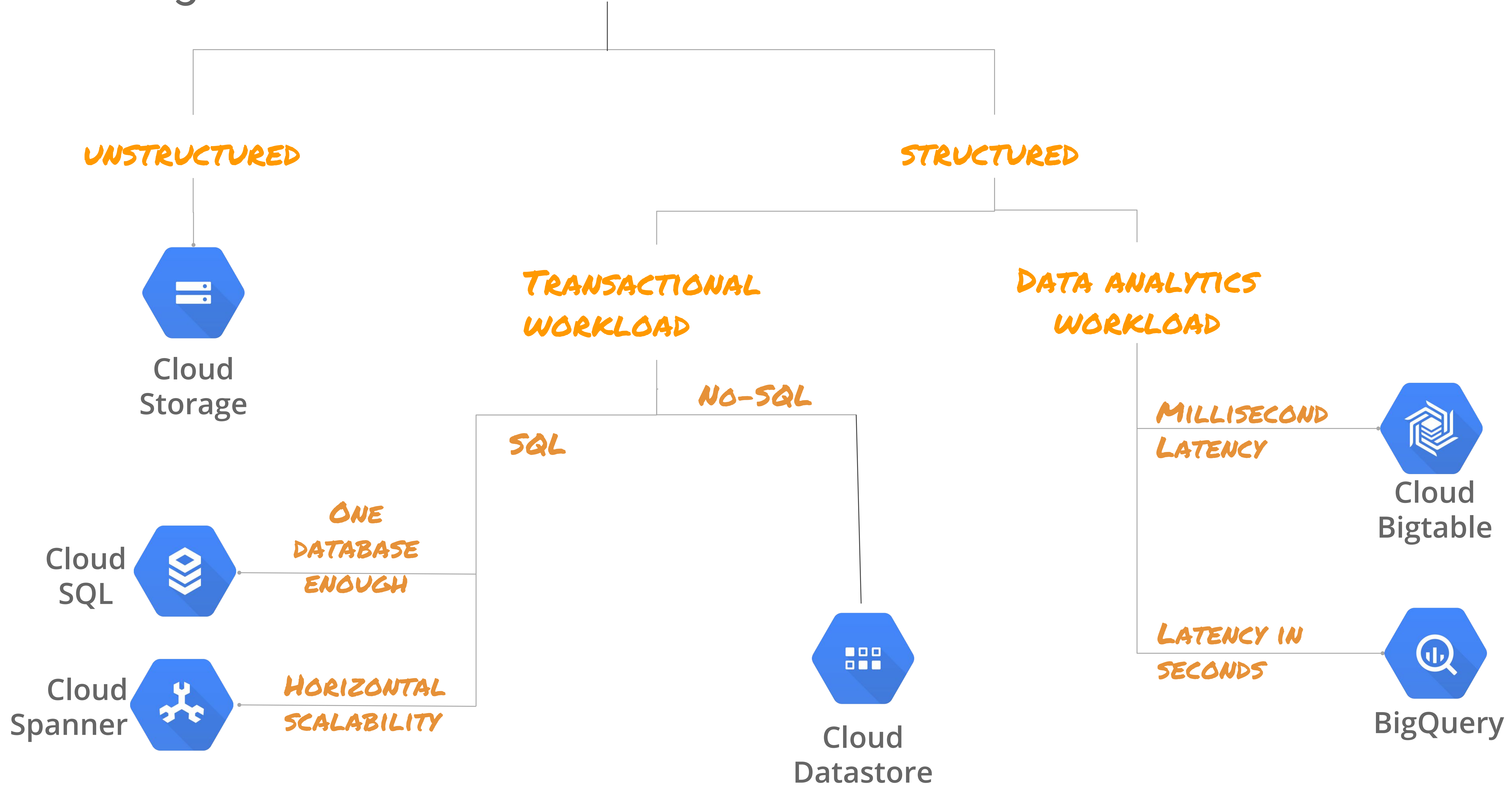
Message-oriented architectures

Serverless data pipelines

GCP Reference Architecture

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

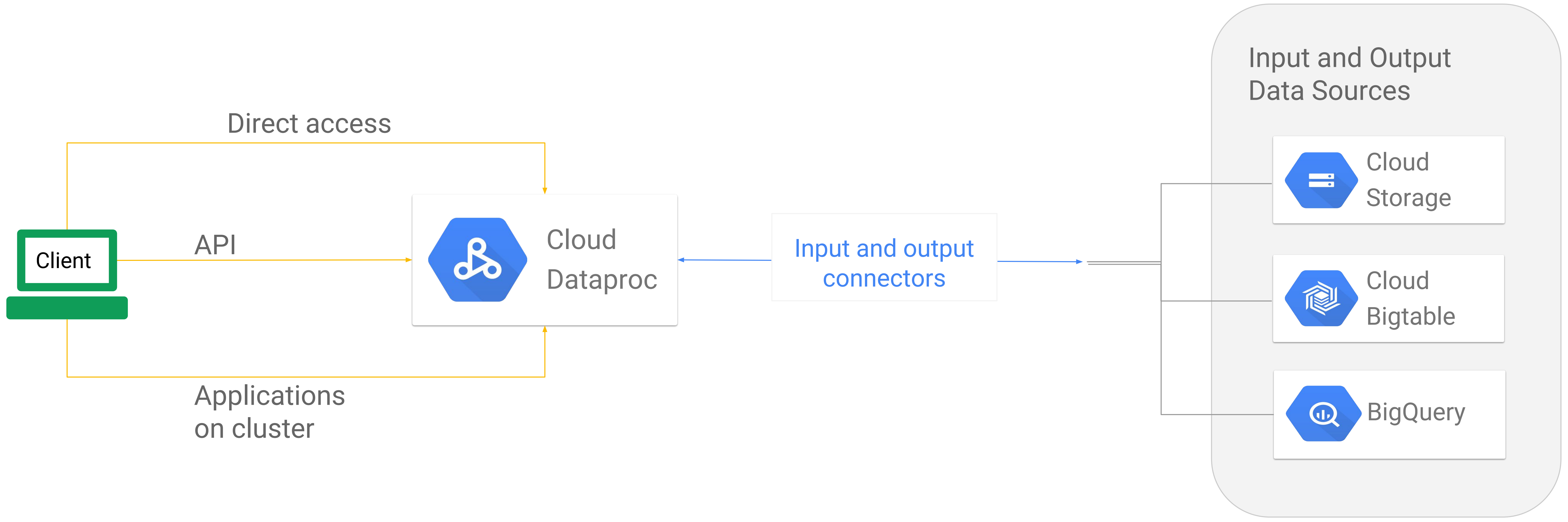
Choosing where to store data on GCP



© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

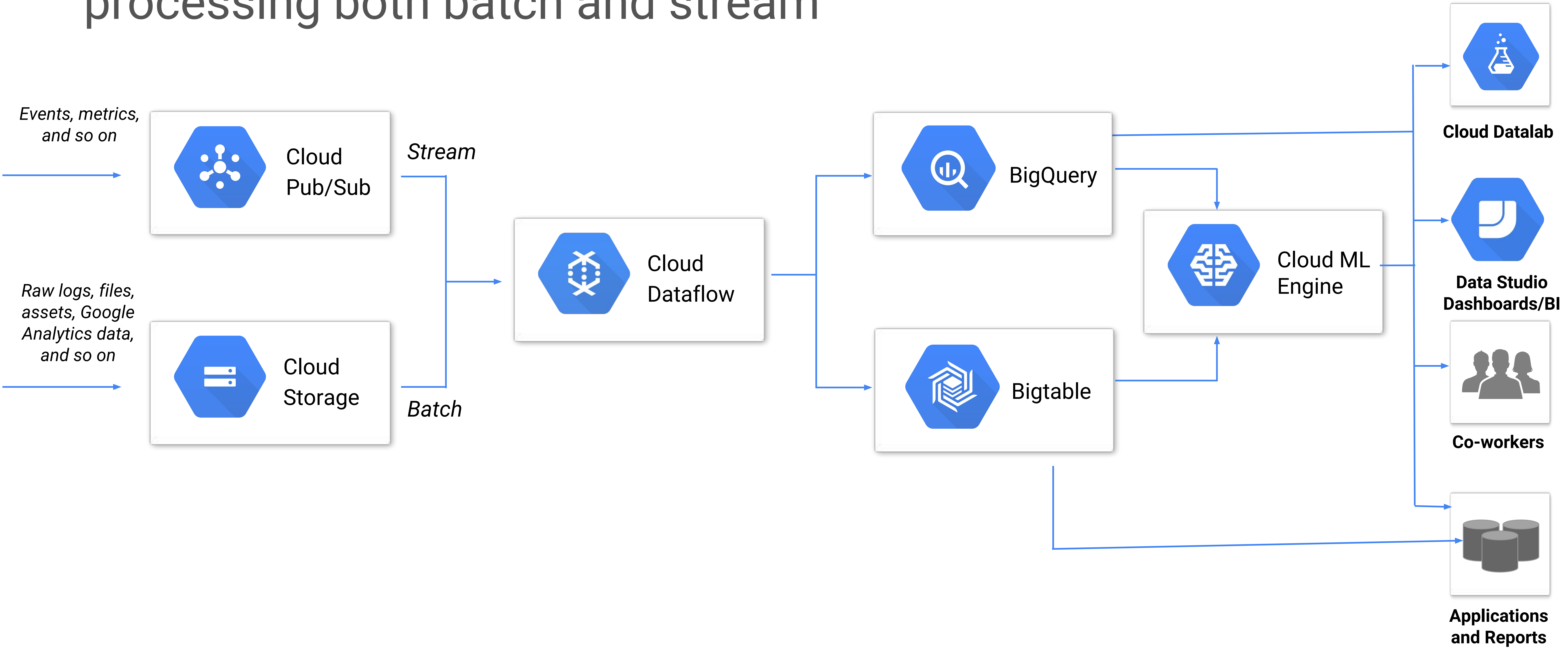
Google Cloud

Run Spark/Hadoop jobs on Cloud Dataproc



© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

On GCP, you can have the same data processing pipeline for processing both batch and stream



© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Module Review

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Google Cloud

Module review

Match the use case on the left with the product on the right

A. Decoupling producers and consumers of data in large organizations and complex systems

B. Scalable, fault-tolerant multi-step processing of data

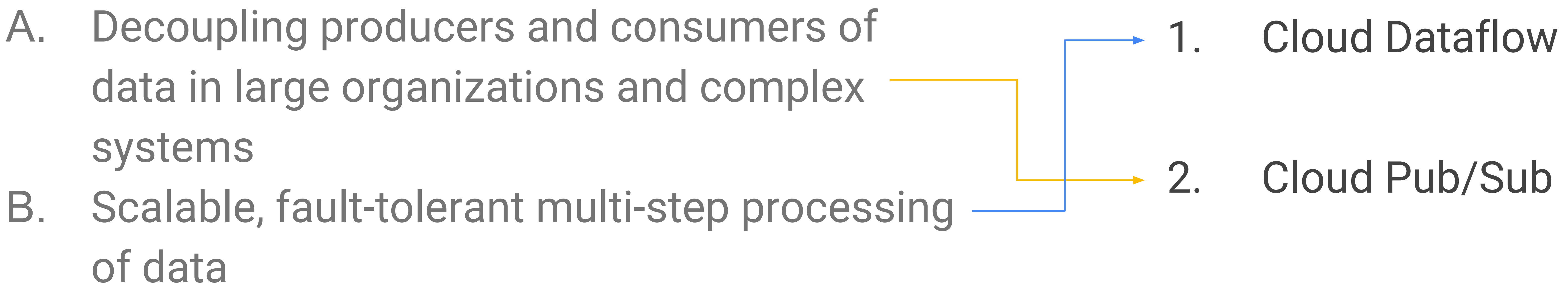
1. Cloud Dataflow

2. Cloud Pub/Sub

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Module review answers

Match the use case on the left with the product on the right



A. Decoupling producers and consumers of data in large organizations and complex systems

B. Scalable, fault-tolerant multi-step processing of data

1. Cloud Dataflow

2. Cloud Pub/Sub

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Proprietary + Confidential

Resources (1 of 2)

Cloud Pub/Sub	https://cloud.google.com/pubsub/
Cloud Dataflow	https://cloud.google.com/dataflow/
Processing media using Cloud Pub/Sub and Compute Engine	https://cloud.google.com/solutions/media-processing-pub-sub-compute-engine

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

Proprietary + Confidential

Resources (2 of 2)

Reverse Geocoding of Geolocation Telemetry in the Cloud Using the Maps API	https://cloud.google.com/solutions/reverse-geocoding-geolocation-telemetry-cloud-maps-ap
Using Cloud Pub/Sub for Long-running Tasks	https://cloud.google.com/solutions/using-cloud-pub-sub-long-running-tasks

© 2017 Google Inc. All rights reserved. Google and the Google logo are trademarks of Google Inc. All other company and product names may be trademarks of the respective companies with which they are associated.

cloud.google.com

Images by Connie Zhou