Google Dataflow 小試

Simon Su @ LinkerNetworks
{Google Developer Expert}

I’m Simon Su...

```
var simon = {/** I am at GCPUG.TW **/};
simon.aboutme = 'http://about.me/peihsinsu';
simon.nodejs = ‘http://opennodes.arecord.us';
simon.googleshare = 'http://gappsnews.blogspot.tw'
simon.nodejsblog = ‘http://nodejs-in-example.blogspot.tw';
simon.blog = ‘http://peihsinsu.blogspot.com';
simon.slideshare = ‘http://slideshare.net/peihsinsu/';
simon.email = ‘simonsu.mail@gmail.com’;
simon.say(‘Good luck to everybody!');
```

https://www.facebook.com/groups/GCPUG.TW/

https://plus.google.com/u/0/communities/116100913832589966421

- Data scientist
- Data engineer
- Frontend engineer

Google Cloud in Big Data Solution

Google Focused Cloud

Now

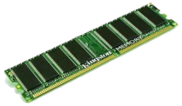
Assembly required

1st Wave
Colocation
Your kit, someone else's building. Yours to manage.

2nd Wave
Virtualized Data Centers
Standard virtual kit for Rent. Still yours to manage.

Storage

Processing


Memory

Network

Next

True On Demand Cloud

3rd Wave
An actual, global elastic cloud
Invest your energy in great apps.

Clusters

Containers

Distributed Storage, Processing & Machine Learning

Google Cloud Family

Application
Runtime Services

Enabling No-Touch Operations

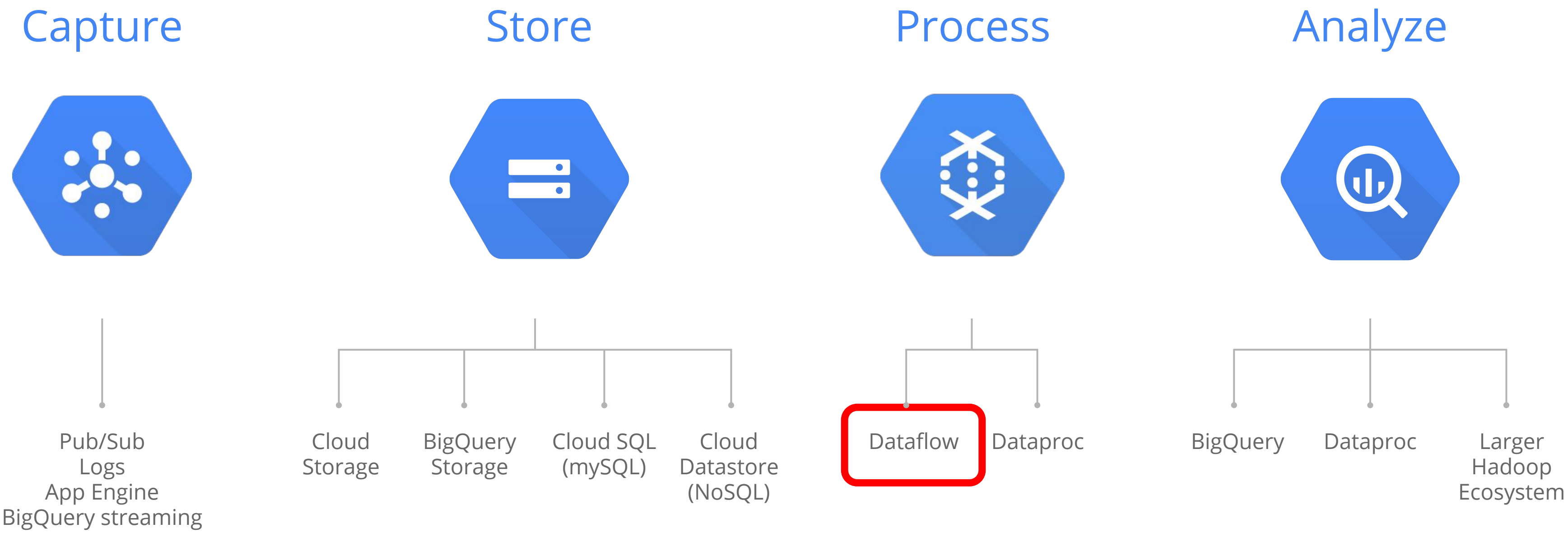
Data Services

**Breakthrough Insights,
Breakthrough Applications**

Foundation
Infrastructure & Operations

The Gear that Powers Google

GCP tools for data processing and analysis



Capture

Pub/Sub
Logs
App Engine
BigQuery streaming

Store

Cloud Storage

BigQuery Storage

Cloud SQL (mySQL)

Cloud Datastore (NoSQL)

Process

Dataflow

Dataproc

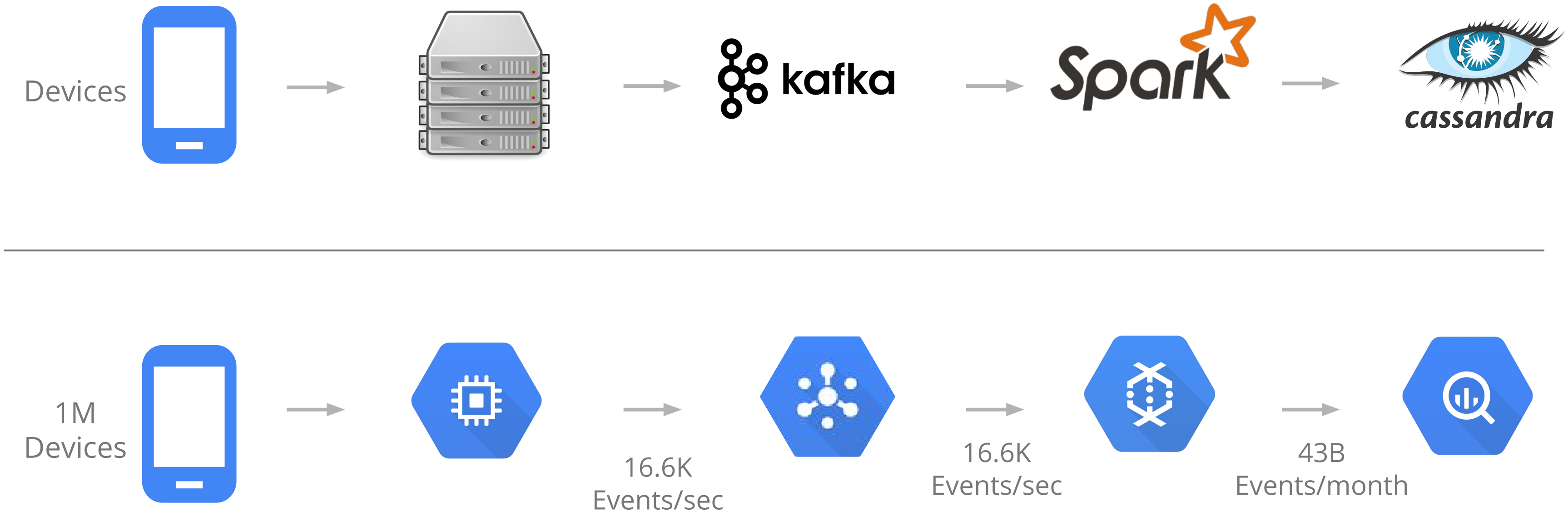
Analyze

BigQuery

Dataproc

Larger Hadoop Ecosystem

Common big data processing flow



Devices

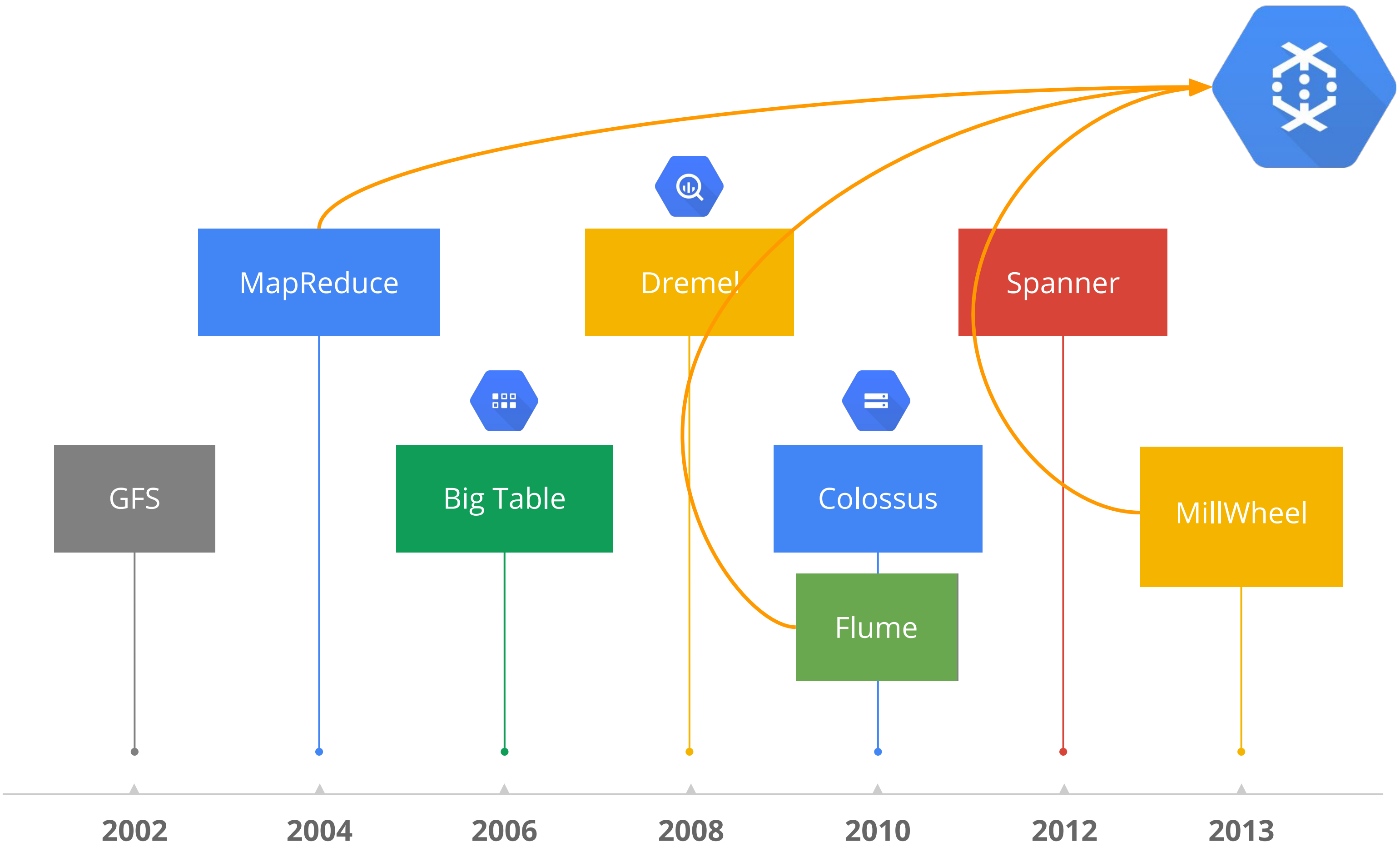
1M Devices

16.6K Events/sec

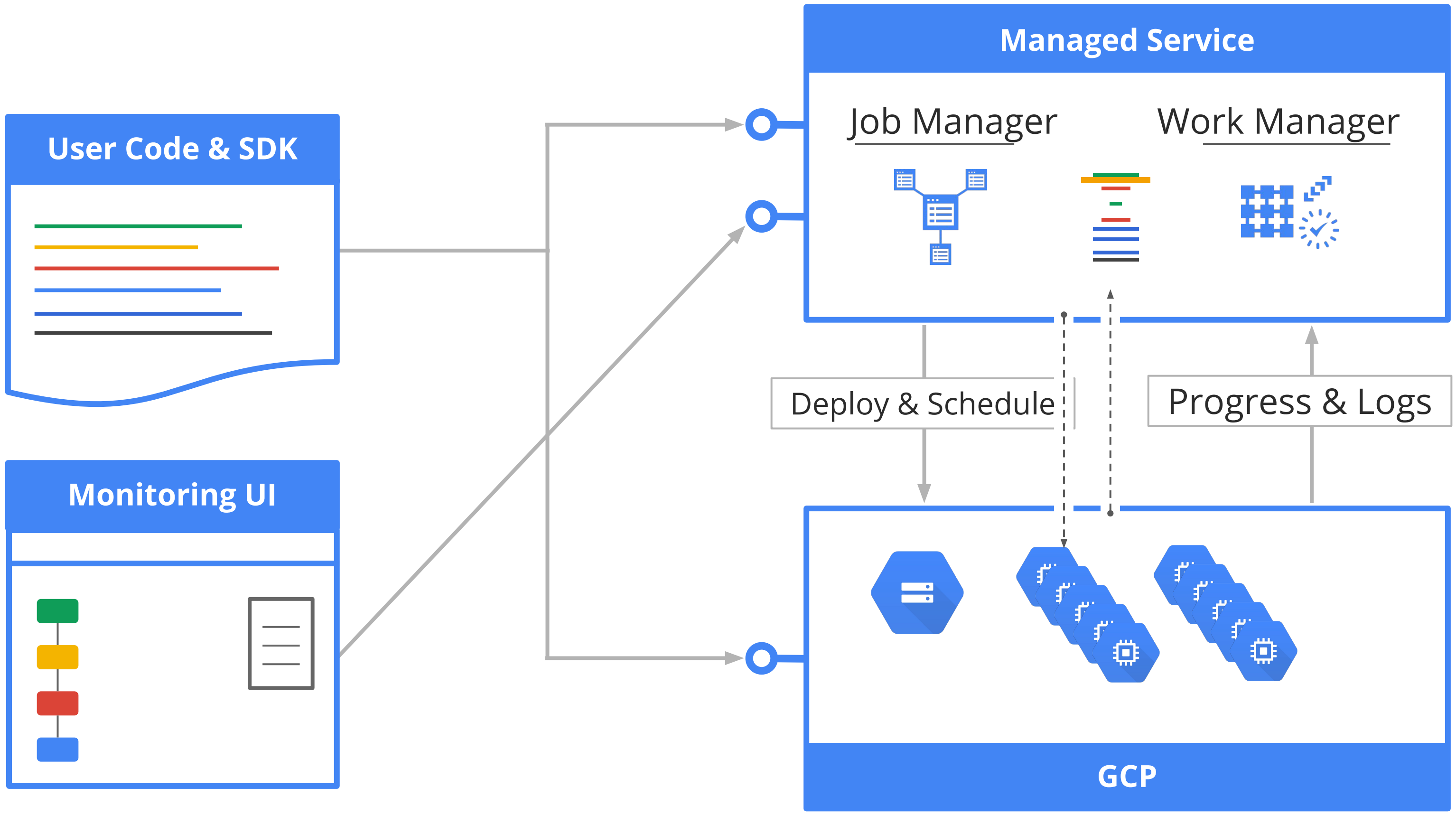
16.6K Events/sec

43B Events/month

Google Big Data Evolution History



Look into Dataflow



Dataflow use case

ETL

- Movement
- Filtering
- Enrichment
- Shaping

Analysis

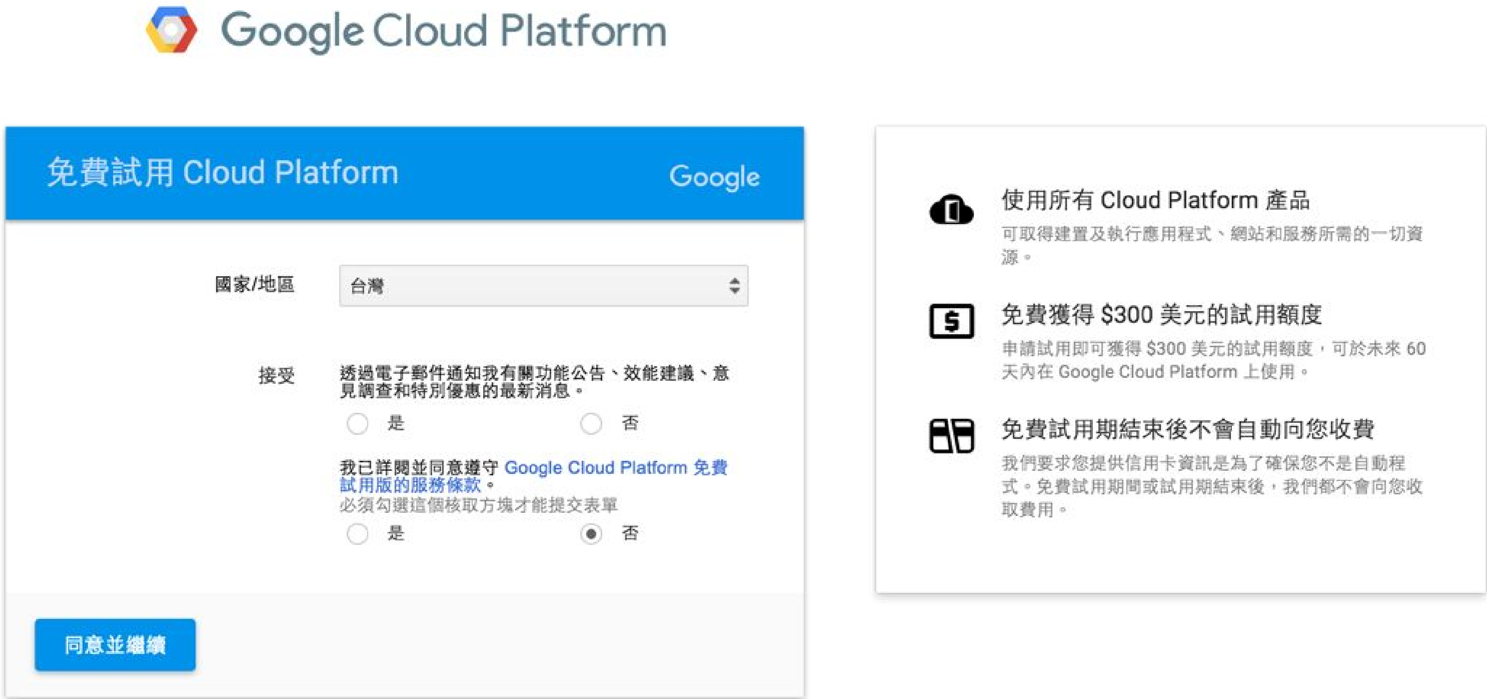
- Reduction
- Batch computation
- Continuous computation

Orchestration

- Composition
- External orchestration
- Simulation

Getting Start - Installation

Get your GCP project



Google Cloud Platform

免費試用 Cloud Platform

Google

國家/地區：台灣

接受

透過電子郵件通知我有關功能公告、效能建議、意見調查和特別優惠的最新消息。

○ 是　○ 否

我已詳閱並同意遵守 Google Cloud Platform 免費試用版的服務條款。

必須勾選這個核取方塊才能提交表單

○ 是　● 否

使用所有 Cloud Platform 產品

可取得建置及執行應用程式、網站和服務所需的一切資源。

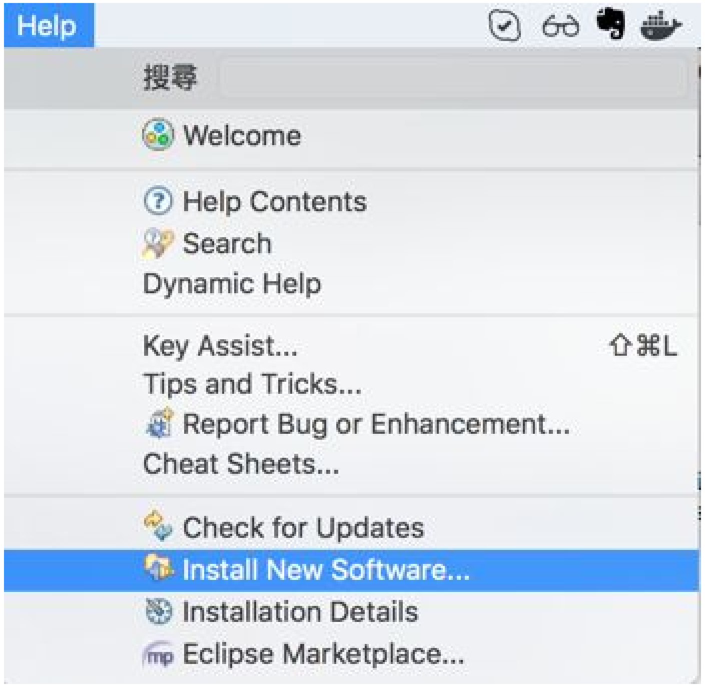
免費獲得 $300 美元的試用額度

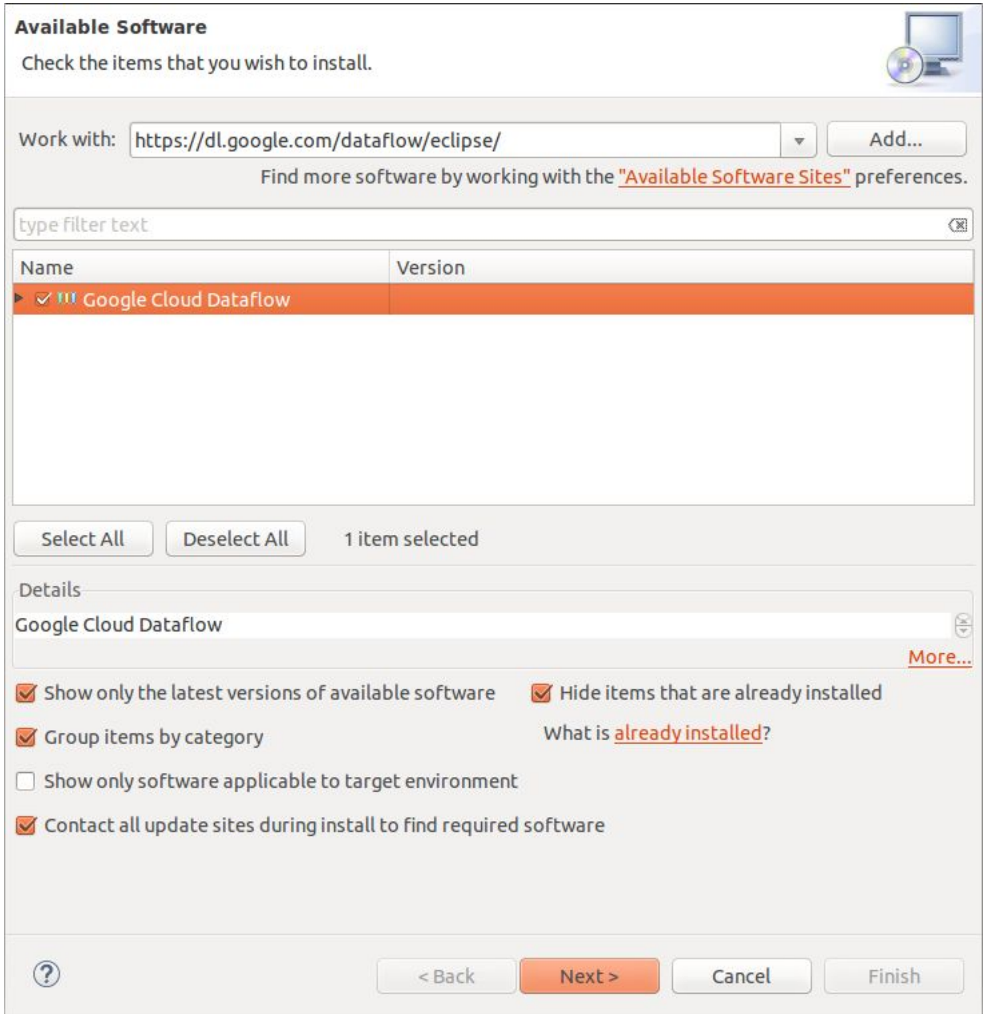
申請試用即可獲得 $300 美元的試用額度，可於未來 60 天內在 Google Cloud Platform 上使用。

免費試用期結束後不會自動向您收費

我們要求您提供信用卡資訊是為了確保您不是自動程式。免費試用期間或試用期結束後，我們都不會向您收取費用。

Install Eclipse Plugin for Dataflow




Available Software

Check the items that you wish to install.

Work with: https://dl.google.com/dataflow/eclipse/

Find more software by working with the "Available Software Sites" preferences.

type filter text

Name	Version
☑ Google Cloud Dataflow	

1 item selected

Details

Google Cloud Dataflow

☑ Show only the latest versions of available software

☑ Hide items that are already installed

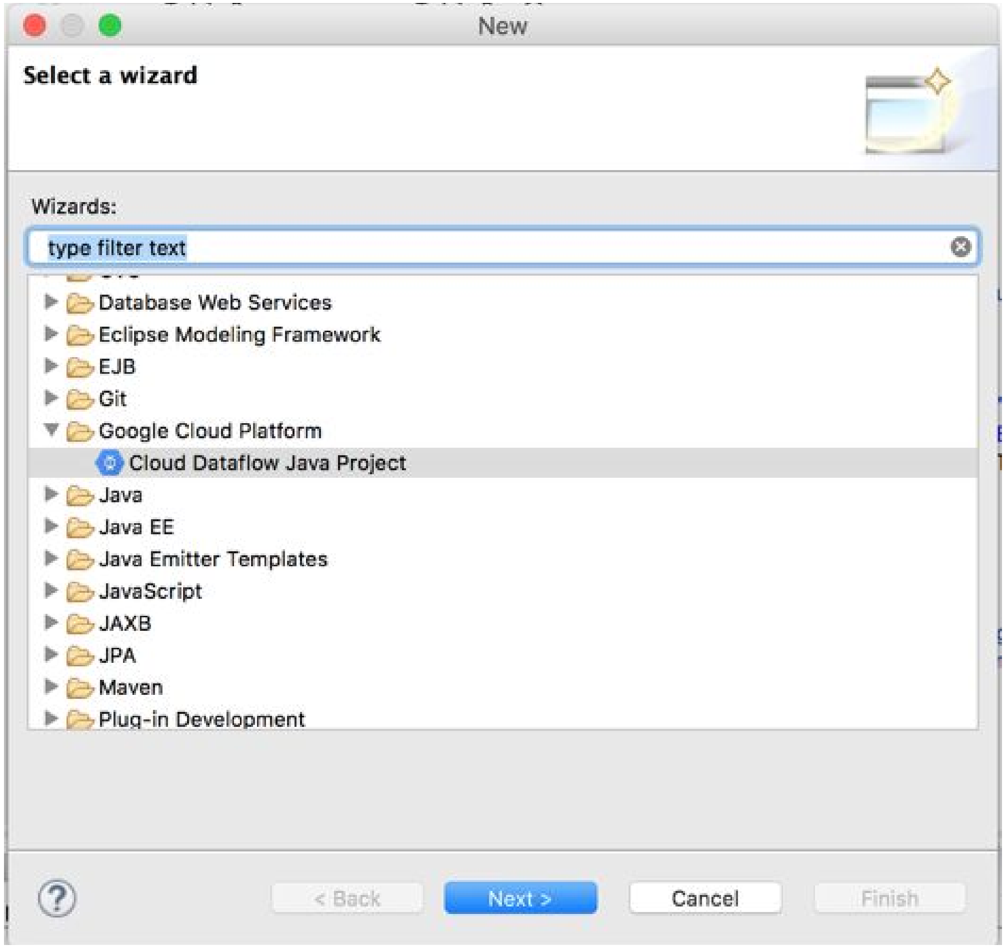
☑ Group items by category

What is already installed?

□ Show only software applicable to target environment

☑ Contact all update sites during install to find required software

Verify your installation



New

Select a wizard

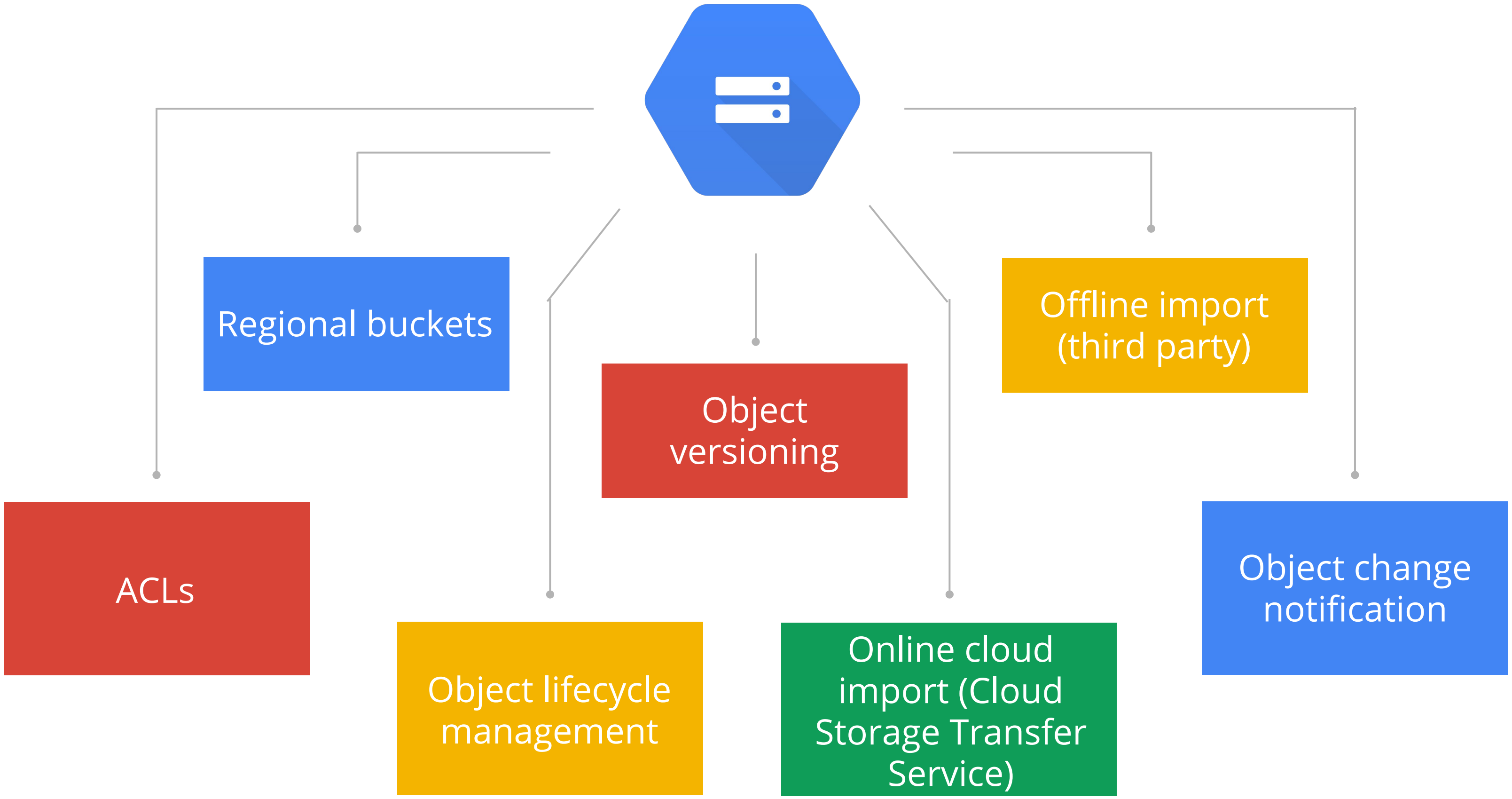
Wizards:

type filter text

- Database Web Services
- Eclipse Modeling Framework
- EJB
- Git
- Google Cloud Platform
  - Cloud Dataflow Java Project
- Java
- Java EE
- Java Emitter Templates
- JavaScript
- JAXB
- JPA
- Maven
- Plug-in Development

Getting start with GCS

Google Cloud Storage Features



Regional buckets

Object versioning

Offline import (third party)

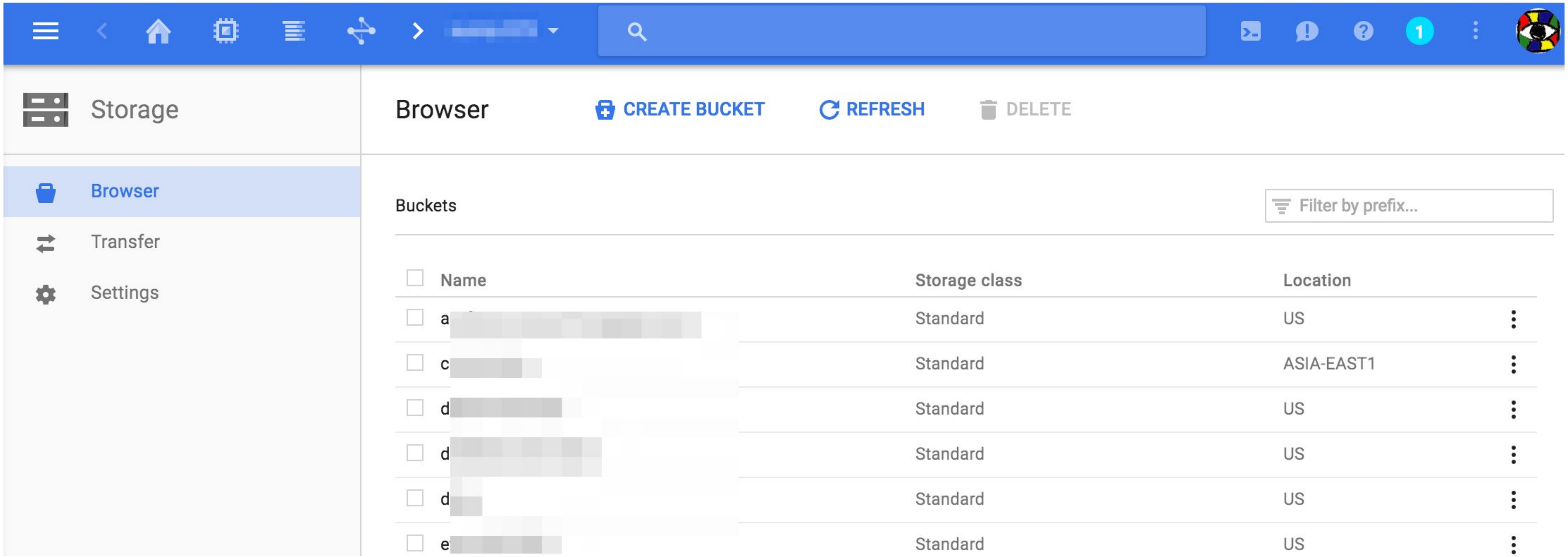
ACLs

Object lifecycle management

Online cloud import (Cloud Storage Transfer Service)

Object change notification

Create your bucket for Dataflow use

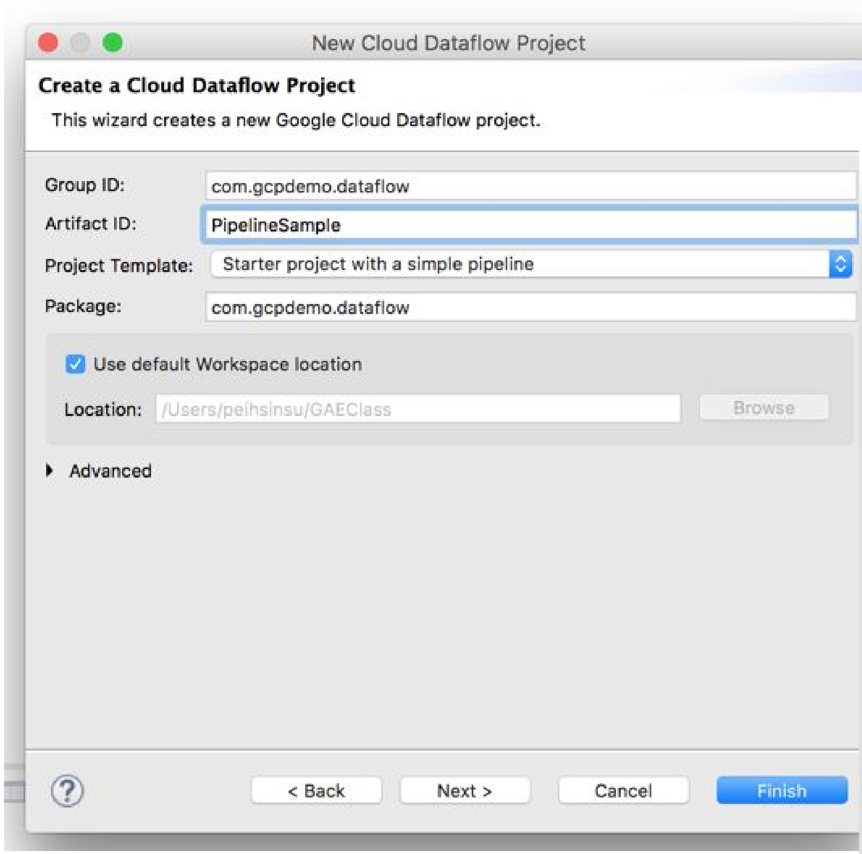


Buckets

□	Name	Storage class	Location
□	a	Standard	US
□	c	Standard	ASIA-EAST1
□	d	Standard	US
□	d	Standard	US
□	d	Standard	US
□	e	Standard	US

Run Dataflow in Local

Create dataflow project



New Cloud Dataflow Project

Create a Cloud Dataflow Project

This wizard creates a new Google Cloud Dataflow project.

Group ID: com.gcpdemo.dataflow

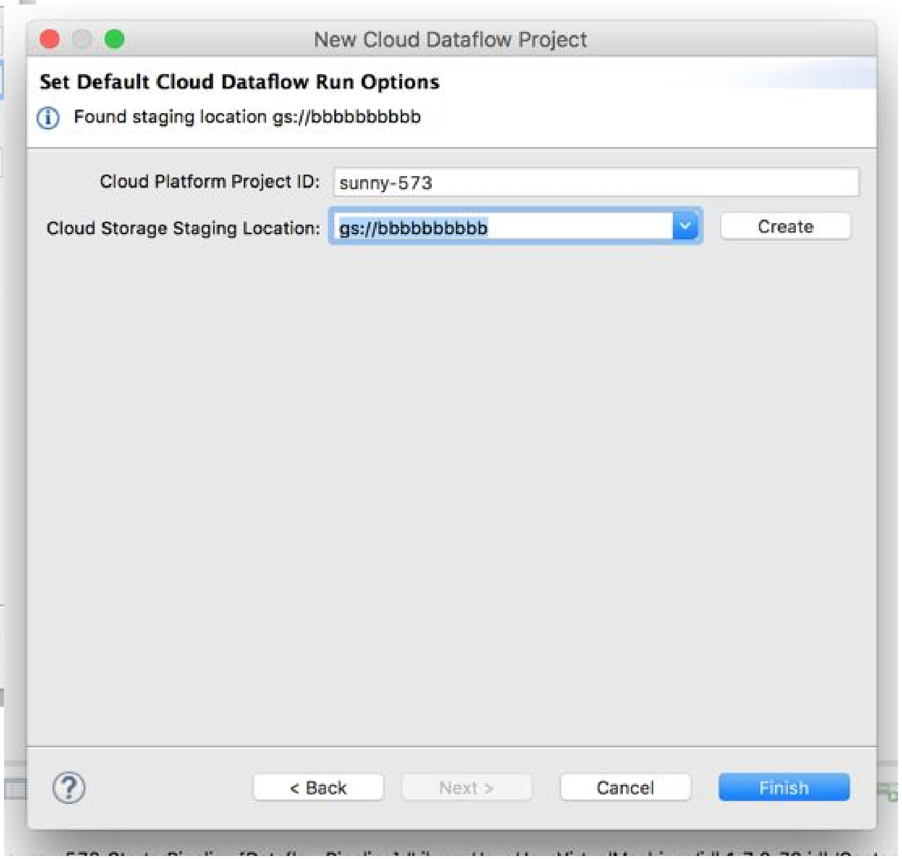
Artifact ID: PipelineSample

Project Template: Starter project with a simple pipeline

Package: com.gcpdemo.dataflow

☑ Use default Workspace location

Location: /Users/peihsinsu/GAEClass



New Cloud Dataflow Project

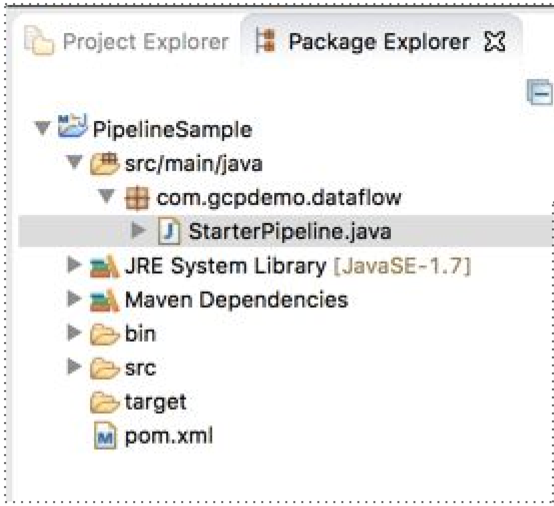
Set Default Cloud Dataflow Run Options

Found staging location gs://bbbbbbbbbb

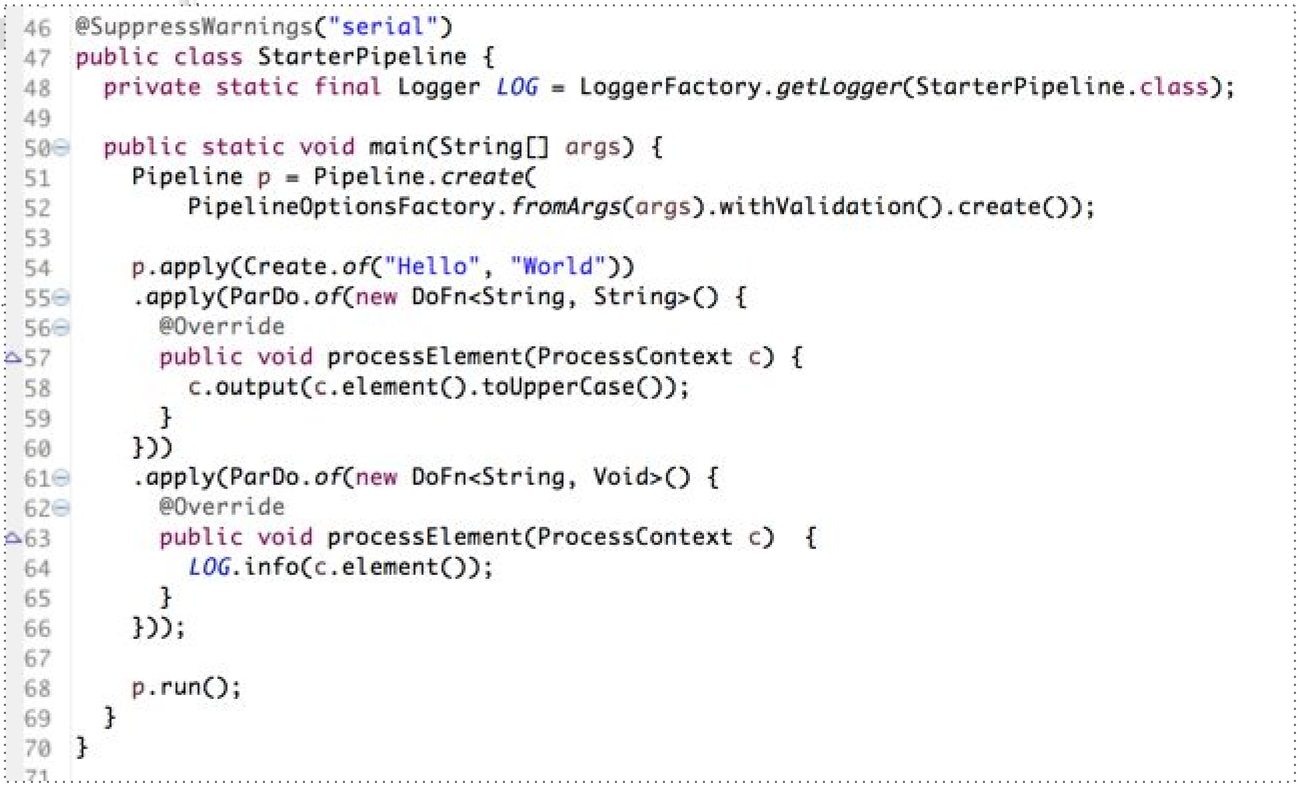
Cloud Platform Project ID: sunny-573

Cloud Storage Staging Location: gs://bbbbbbbbbb

Dataflow Sample



- PipelineSample
  - src/main/java
    - com.gcpdemo.dataflow
      - StarterPipeline.java
  - JRE System Library [JavaSE-1.7]
  - Maven Dependencies
  - bin
  - src
  - target
  - pom.xml



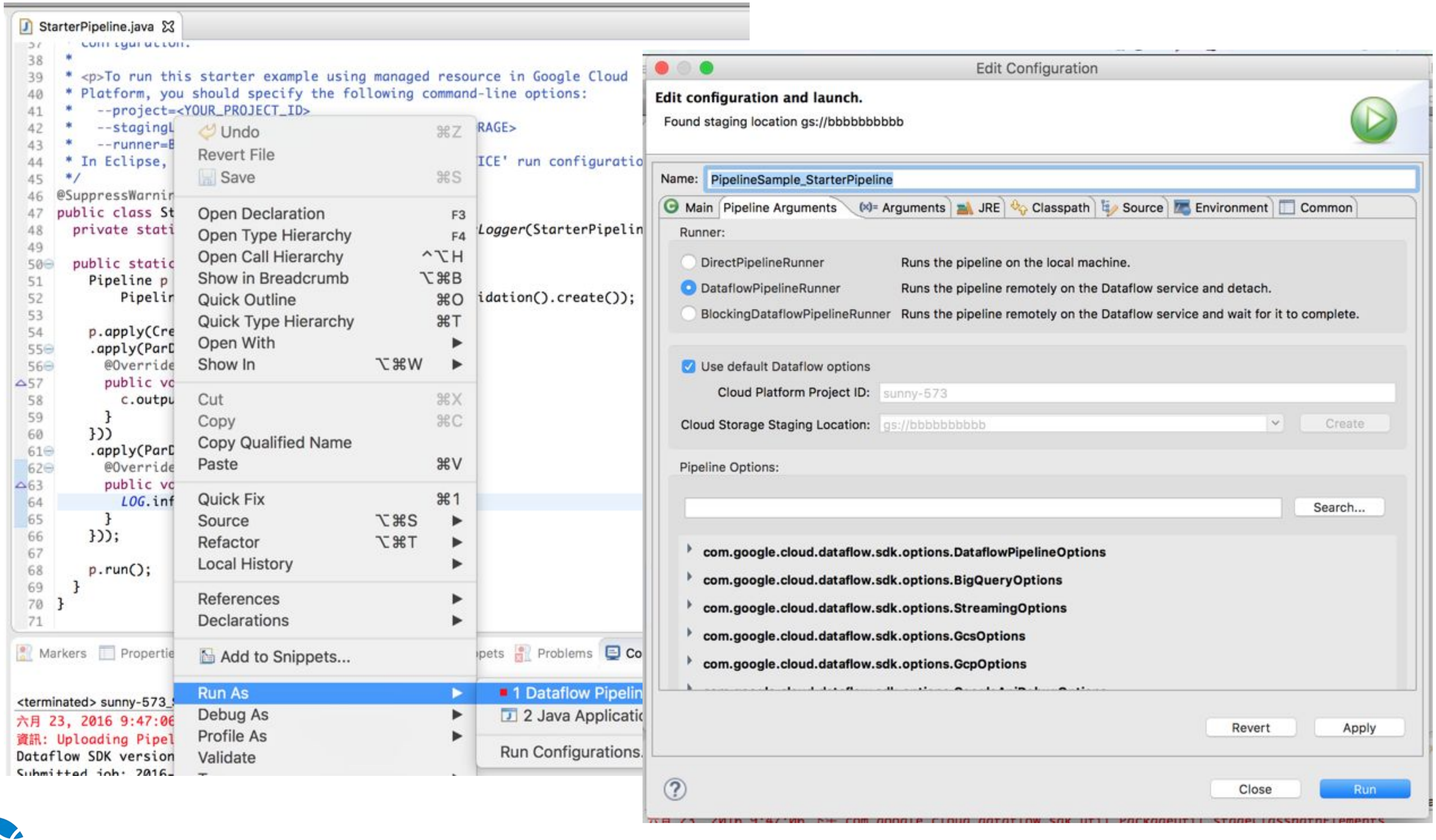
```java
@SuppressWarnings("serial")
public class StarterPipeline {
  private static final Logger LOG = LoggerFactory.getLogger(StarterPipeline.class);

  public static void main(String[] args) {
    Pipeline p = Pipeline.create(
        PipelineOptionsFactory.fromArgs(args).withValidation().create());

    p.apply(Create.of("Hello", "World"))
    .apply(ParDo.of(new DoFn<String, String>() {
      @Override
      public void processElement(ProcessContext c) {
        c.output(c.element().toUpperCase());
      }
    }))
    .apply(ParDo.of(new DoFn<String, Void>() {
      @Override
      public void processElement(ProcessContext c)  {
        LOG.info(c.element());
      }
    }));

    p.run();
  }
}
```

Execute Dataflow



```
 *
 * <p>To run this starter example using managed resource in Google Cloud
 * Platform, you should specify the following command-line options:
 *   --project=<YOUR_PROJECT_ID>
 *   --stagingL                                        RAGE>
 *   --runner=B
 * In Eclipse,                                         ICE' run configuratio
 */
@SuppressWarnin
public class St
  private stati                                        Logger(StarterPipelin

  public static
    Pipeline p
        Pipelir                                        idation().create());

    p.apply(Cre
    .apply(ParD
      @Override
      public vo
        c.outpu
      }
    }))
    .apply(ParD
      @Override
      public vo
        LOG.inf
      }
    }));

    p.run();
  }
}
```

```
<terminated> sunny-573_
六月 23, 2016 9:47:06
資訊: Uploading Pipel
Dataflow SDK version
Submitted job: 2016-
```

Edit configuration and launch.

Found staging location gs://bbbbbbbbbb

Name: PipelineSample_StarterPipeline

Runner:

- ○ DirectPipelineRunner — Runs the pipeline on the local machine.
- ● DataflowPipelineRunner — Runs the pipeline remotely on the Dataflow service and detach.
- ○ BlockingDataflowPipelineRunner — Runs the pipeline remotely on the Dataflow service and wait for it to complete.

☑ Use default Dataflow options

Cloud Platform Project ID: sunny-573

Cloud Storage Staging Location: gs://bbbbbbbbbb

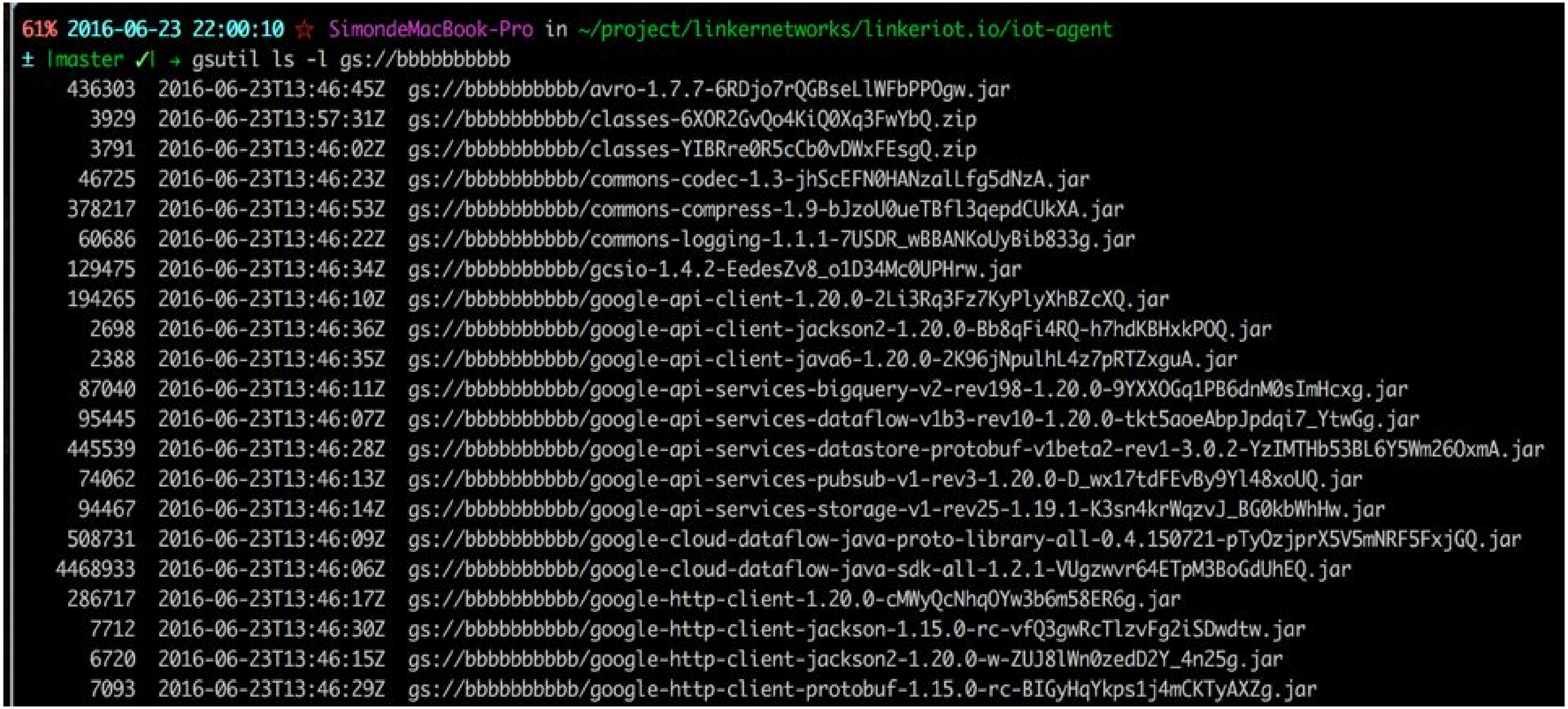
Pipeline Options:

- com.google.cloud.dataflow.sdk.options.DataflowPipelineOptions
- com.google.cloud.dataflow.sdk.options.BigQueryOptions
- com.google.cloud.dataflow.sdk.options.StreamingOptions
- com.google.cloud.dataflow.sdk.options.GcsOptions
- com.google.cloud.dataflow.sdk.options.GcpOptions

Lab 1: Ready your dataflow environment and create your first dataflow project

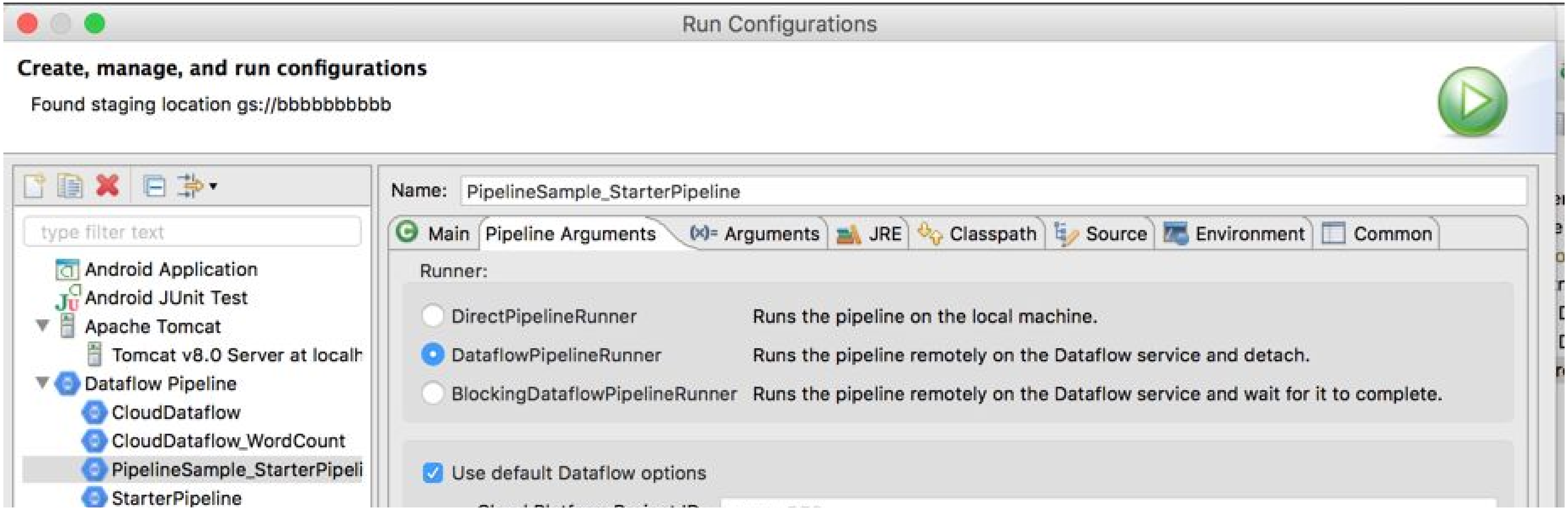
- Create GCP project
- Install Eclipse and Dataflow plugin
- Create first Dataflow project
- Run your project

After lab - What thing in GCS bucket



```
61% 2016-06-23 22:00:10 ☆ SimondeMacBook-Pro in ~/project/linkernetworks/linkeriot.io/iot-agent
± |master ✓| → gsutil ls -l gs://bbbbbbbbbb
    436303  2016-06-23T13:46:45Z  gs://bbbbbbbbbb/avro-1.7.7-6RDjo7rQGBseLlWFbPPOgw.jar
      3929  2016-06-23T13:57:31Z  gs://bbbbbbbbbb/classes-6XOR2GvQo4KiQ0Xq3FwYbQ.zip
      3791  2016-06-23T13:46:02Z  gs://bbbbbbbbbb/classes-YIBRre0R5cCb0vDWxFEsgQ.zip
     46725  2016-06-23T13:46:23Z  gs://bbbbbbbbbb/commons-codec-1.3-jhScEFN0HANzalLfg5dNzA.jar
    378217  2016-06-23T13:46:53Z  gs://bbbbbbbbbb/commons-compress-1.9-bJzoU0ueTBfl3qepdCUkXA.jar
     60686  2016-06-23T13:46:22Z  gs://bbbbbbbbbb/commons-logging-1.1.1-7USDR_wBBANKoUyBib833g.jar
    129475  2016-06-23T13:46:34Z  gs://bbbbbbbbbb/gcsio-1.4.2-EedesZv8_o1D34Mc0UPHrw.jar
    194265  2016-06-23T13:46:10Z  gs://bbbbbbbbbb/google-api-client-1.20.0-2Li3Rq3Fz7KyPlyXhBZcXQ.jar
      2698  2016-06-23T13:46:36Z  gs://bbbbbbbbbb/google-api-client-jackson2-1.20.0-Bb8qFi4RQ-h7hdKBHxkPOQ.jar
      2388  2016-06-23T13:46:35Z  gs://bbbbbbbbbb/google-api-client-java6-1.20.0-2K96jNpulhL4z7pRTZxguA.jar
     87040  2016-06-23T13:46:11Z  gs://bbbbbbbbbb/google-api-services-bigquery-v2-rev198-1.20.0-9YXXOGq1PB6dnM0sImHcxg.jar
     95445  2016-06-23T13:46:07Z  gs://bbbbbbbbbb/google-api-services-dataflow-v1b3-rev10-1.20.0-tkt5aoeAbpJpdqi7_YtwGg.jar
    445539  2016-06-23T13:46:28Z  gs://bbbbbbbbbb/google-api-services-datastore-protobuf-v1beta2-rev1-3.0.2-YzIMTHb53BL6Y5Wm26OxmA.jar
     74062  2016-06-23T13:46:13Z  gs://bbbbbbbbbb/google-api-services-pubsub-v1-rev3-1.20.0-D_wx17tdFEvBy9Yl48xoUQ.jar
     94467  2016-06-23T13:46:14Z  gs://bbbbbbbbbb/google-api-services-storage-v1-rev25-1.19.1-K3sn4krWqzvJ_BG0kbWhHw.jar
    508731  2016-06-23T13:46:09Z  gs://bbbbbbbbbb/google-cloud-dataflow-java-proto-library-all-0.4.150721-pTyOzjprX5V5mNRF5FxjGQ.jar
   4468933  2016-06-23T13:46:06Z  gs://bbbbbbbbbb/google-cloud-dataflow-java-sdk-all-1.2.1-VUgzwvr64ETpM3BoGdUhEQ.jar
    286717  2016-06-23T13:46:17Z  gs://bbbbbbbbbb/google-http-client-1.20.0-cMWyQcNhqOYw3b6m58ER6g.jar
      7712  2016-06-23T13:46:30Z  gs://bbbbbbbbbb/google-http-client-jackson-1.15.0-rc-vfQ3gwRcTlzvFg2iSDwdtw.jar
      6720  2016-06-23T13:46:15Z  gs://bbbbbbbbbb/google-http-client-jackson2-1.20.0-w-ZUJ8lWn0zedD2Y_4n25g.jar
      7093  2016-06-23T13:46:29Z  gs://bbbbbbbbbb/google-http-client-protobuf-1.15.0-rc-BIGyHqYkps1j4mCKTyAXZg.jar
```

After lab - The Run Configuration



Create, manage, and run configurations

Found staging location gs://bbbbbbbbbb

Name: PipelineSample_StarterPipeline

Runner:

○ DirectPipelineRunner Runs the pipeline on the local machine.

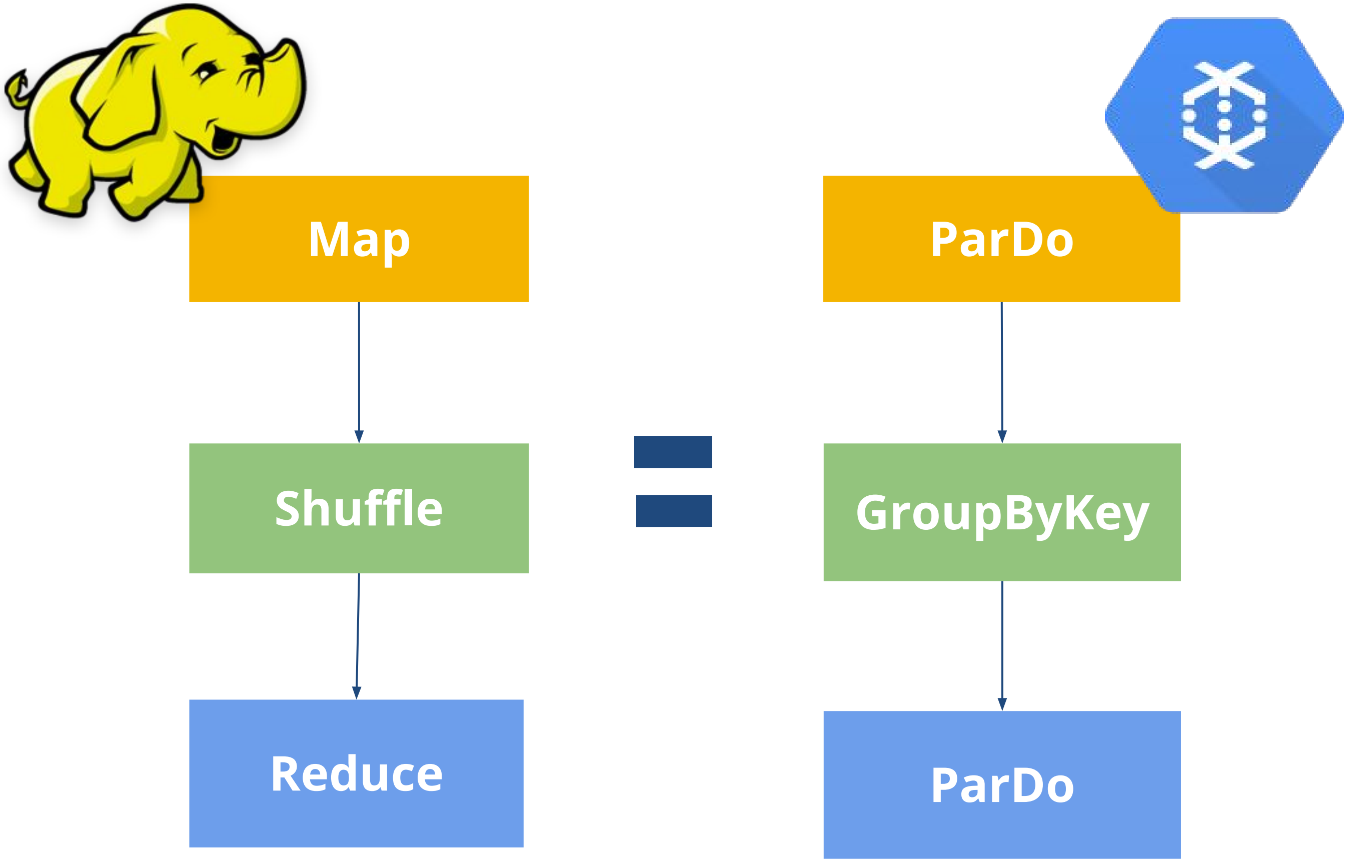
● DataflowPipelineRunner Runs the pipeline remotely on the Dataflow service and detach.

○ BlockingDataflowPipelineRunner Runs the pipeline remotely on the Dataflow service and wait for it to complete.

☑ Use default Dataflow options

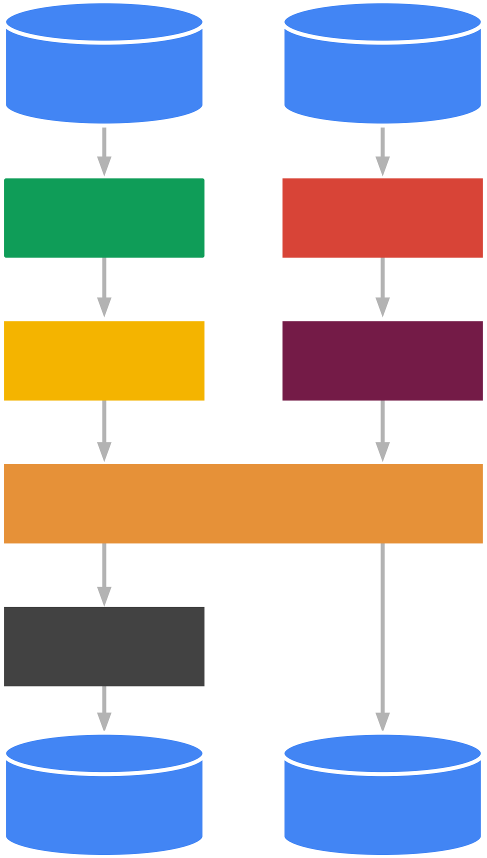
Dataflow in Batch Mode

What we do in Big Data process...



Pipeline

- A Direct Acyclic Graph of data processing transformations
- Can be submitted to the Dataflow Service for optimization and execution or executed on an alternate runner e.g. Spark
- May include multiple inputs and multiple outputs
- May encompass many logical MapReduce operations
- PCollections flow through the pipeline



Inputs & Outputs

- Read from standard Google Cloud Platform data sources
  - GCS, Pub/Sub, BigQuery, Datastore
- Write your own custom source by teaching Dataflow how to read it in parallel
  - Currently for bounded sources only
- Write to GCS, BigQuery, Pub/Sub
  - More coming…
- Can use a combination of text, JSON, XML, Avro formatted data

Your
Source/Sink
Here

PCollection

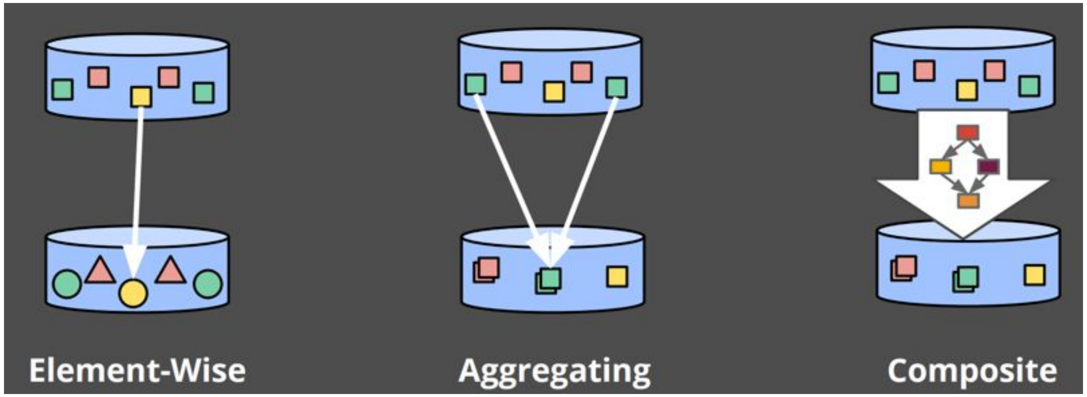
- A collection of data of type T in a pipeline
  - PCollection<K,V>
- Maybe be either ***bounded*** or ***unbounded*** in size
- Created by using a PTransform to:
  - Build from a java.util.Collection
  - Read from a backing data store
  - Transform an existing PCollection
- Often contain the key-value pairs using KV

```
{Seahawks, NFC, Champions, Seattle,
...}
```

```
{...,
 “NFC Champions #GreenBay”,
 “Green Bay #superbowl!”,
 ...
 “#GoHawks”,
  ...}
```

Transforms

- A step, or a processing operation that transforms data
    - convert format , group , filter data
- Type of Transforms
    - ParDo
    - GroupByKey
    - Combine
    - Flatten
        - Multiple PCollection objects that contain the same data type, you can merge them into a single logical PCollection using the Flatten transform



Pardo (Parallel do)

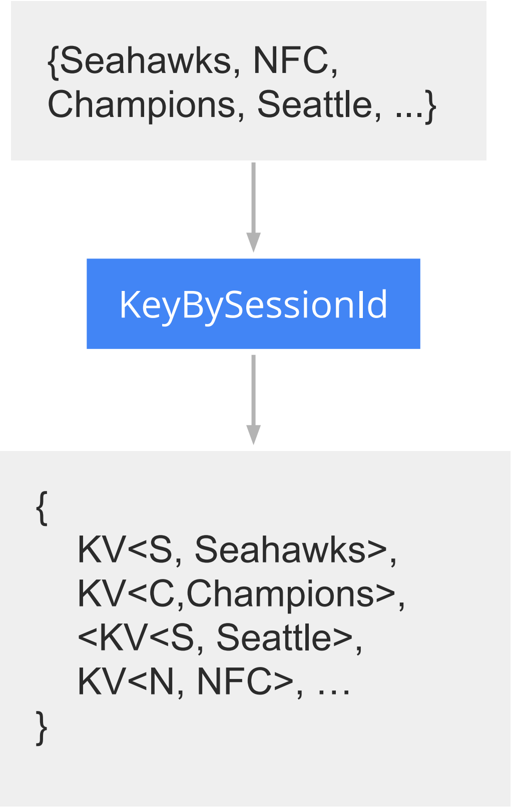
- Processes each element of a PCollection independently using a user-provided DoFn
- Corresponds to both the Map and Reduce phases in Hadoop i.e. ParDo->GBK->ParDo
- Useful for

  Filtering a data set.

  Formatting or converting the type of each element in a data set.

  Extracting parts of each element in a data set.

  Performing computations on each element in a data set.



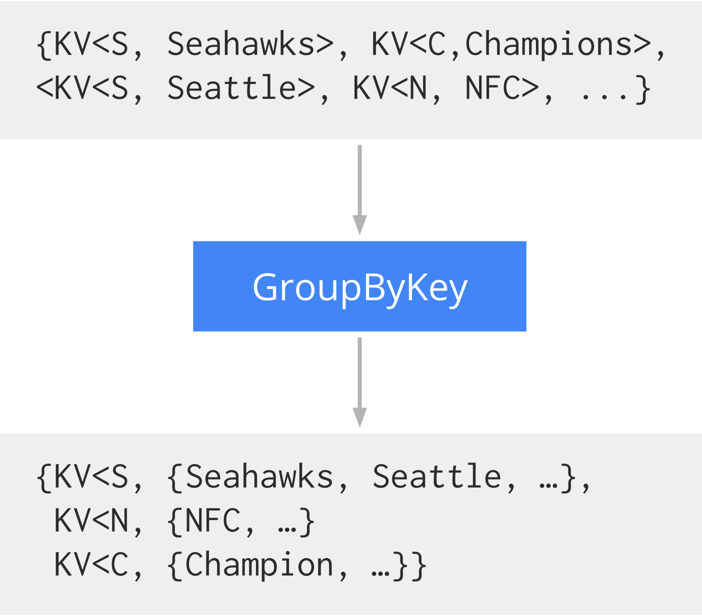
```
{Seahawks, NFC,
Champions, Seattle, ...}
```

```
{
   KV<S, Seahawks>,
   KV<C,Champions>,
   <KV<S, Seattle>,
   KV<N, NFC>, …
}
```

Group by key

- Takes a PCollection of key-value pairs and gathers up all values with the same key

- Corresponds to the shuffle phase in Hadoop

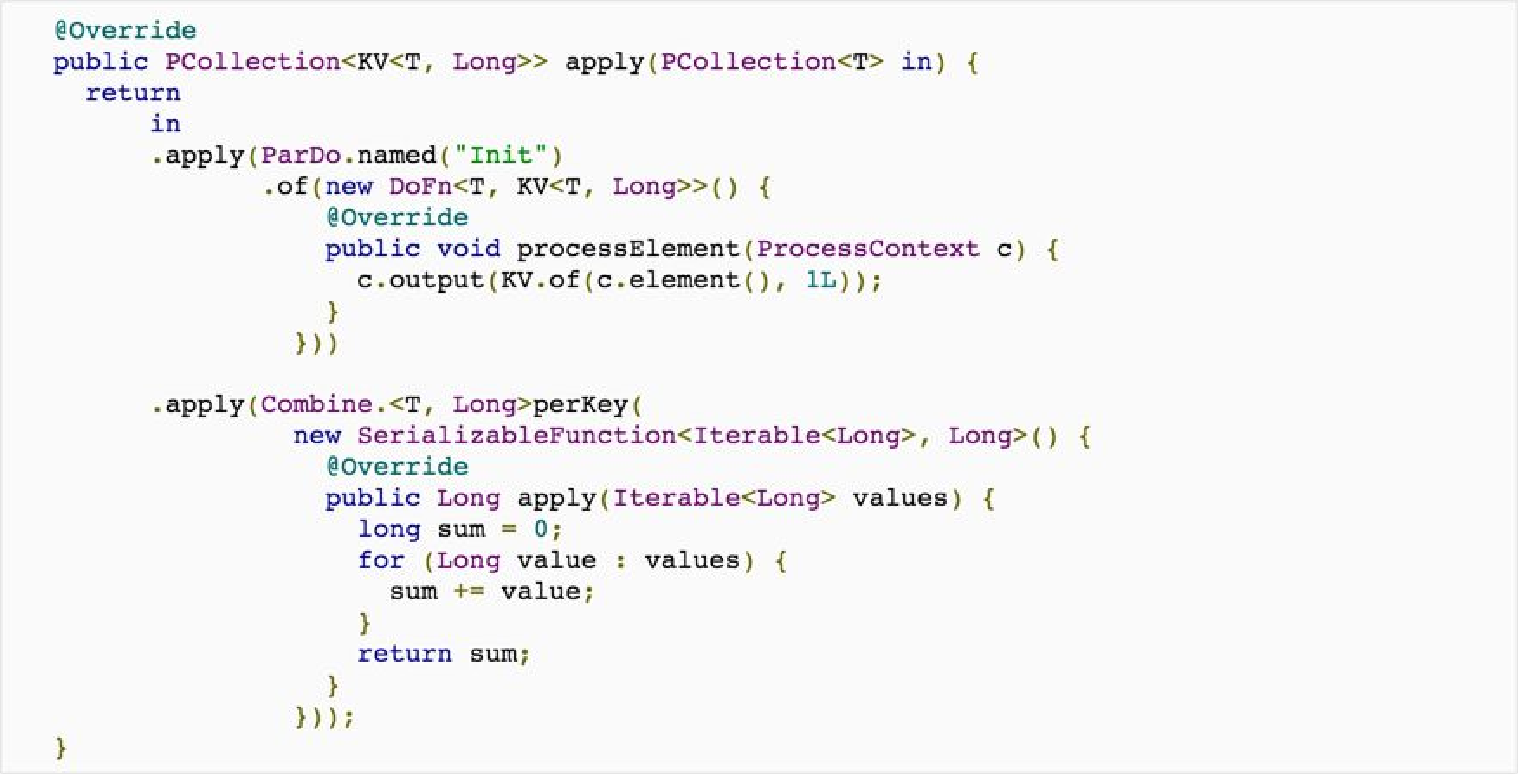


```
{KV<S, Seahawks>, KV<C,Champions>,
<KV<S, Seattle>, KV<N, NFC>, ...}
```

```
{KV<S, {Seahawks, Seattle, …},
 KV<N, {NFC, …}
 KV<C, {Champion, …}}
```

Wait a minute…
How do you do a GroupByKey on an unbounded PCollection?

Group by key sample



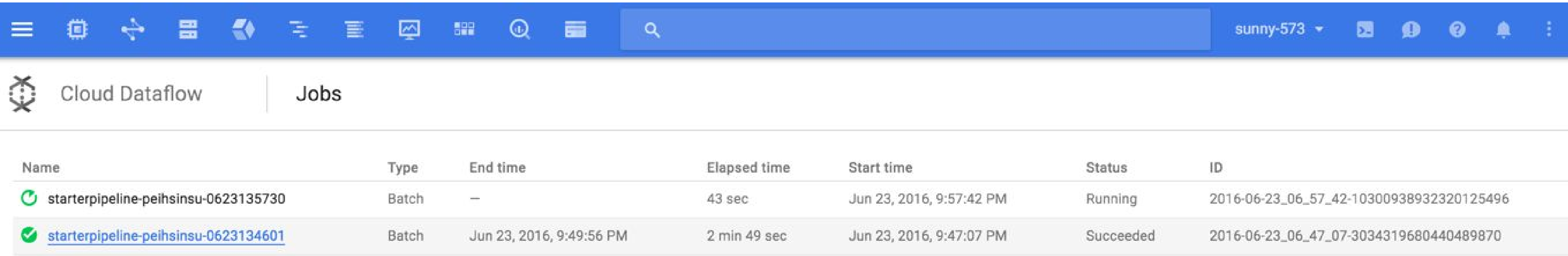
```
@Override
public PCollection<KV<T, Long>> apply(PCollection<T> in) {
  return
      in
      .apply(ParDo.named("Init")
             .of(new DoFn<T, KV<T, Long>>() {
                 @Override
                 public void processElement(ProcessContext c) {
                   c.output(KV.of(c.element(), 1L));
                 }
               }))

      .apply(Combine.<T, Long>perKey(
               new SerializableFunction<Iterable<Long>, Long>() {
                 @Override
                 public Long apply(Iterable<Long> values) {
                   long sum = 0;
                   for (Long value : values) {
                     sum += value;
                   }
                   return sum;
                 }
               }));
}
```

Lab 2: Deploy your first project to Google Cloud Platform

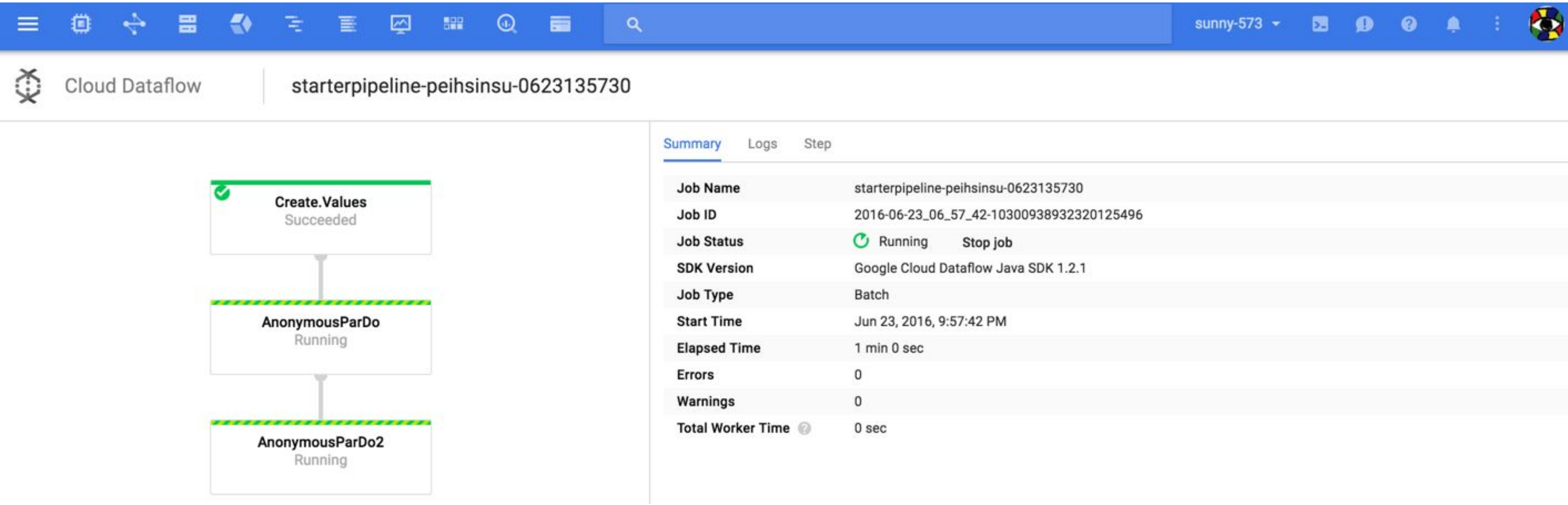
- Checking the Lab 1 project working well
- Deploy to cloud and watch the dataflow task dashboard

Dataflow Task Dashboard



Cloud Dataflow | Jobs

Name	Type	End time	Elapsed time	Start time	Status	ID
starterpipeline-peihsinsu-0623135730	Batch	–	43 sec	Jun 23, 2016, 9:57:42 PM	Running	2016-06-23_06_57_42-10300938932320125496
starterpipeline-peihsinsu-0623134601	Batch	Jun 23, 2016, 9:49:56 PM	2 min 49 sec	Jun 23, 2016, 9:47:07 PM	Succeeded	2016-06-23_06_47_07-3034319680440489870



Cloud Dataflow | starterpipeline-peihsinsu-0623135730

Create.Values
Succeeded

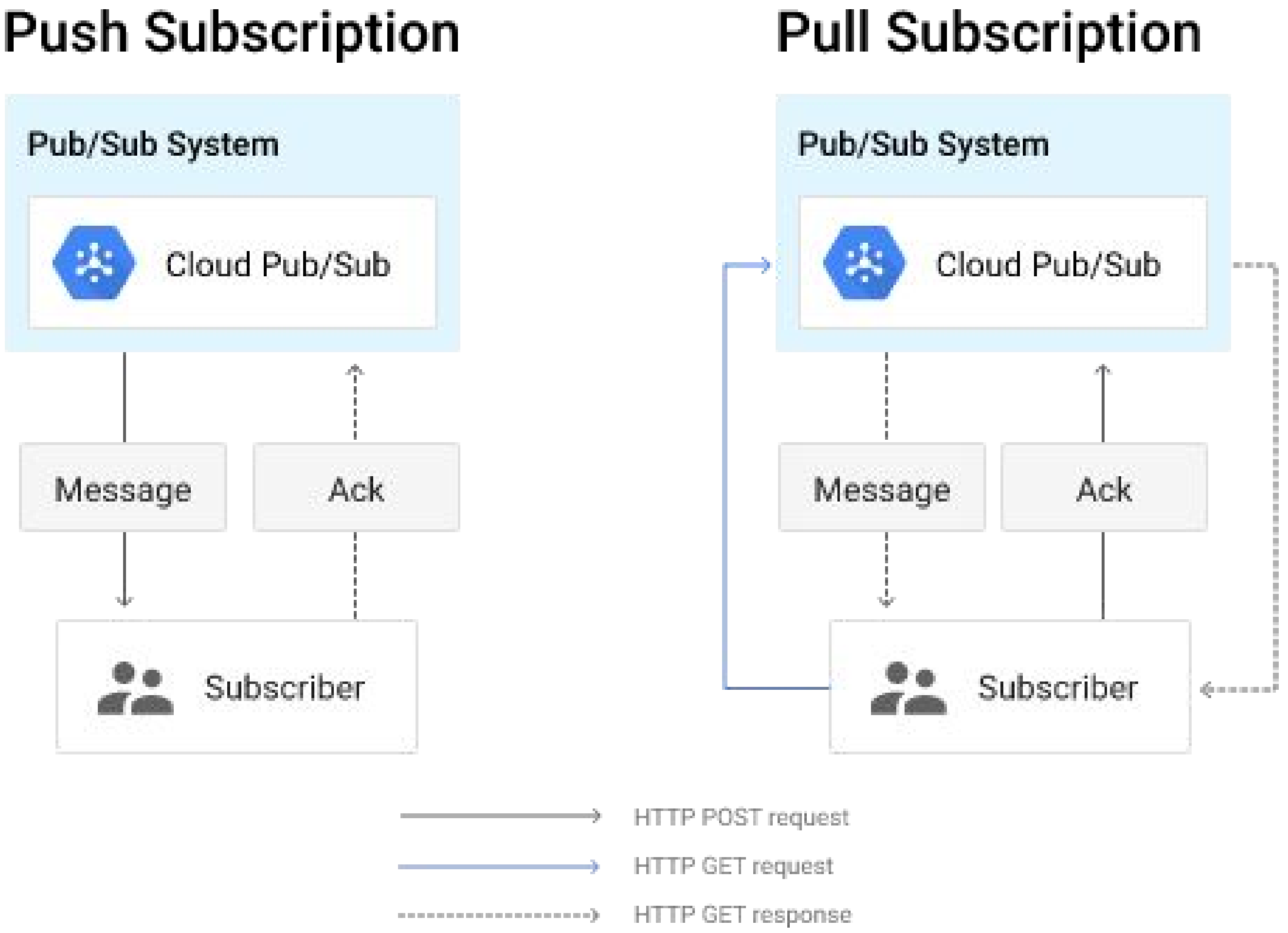
AnonymousParDo
Running

AnonymousParDo2
Running

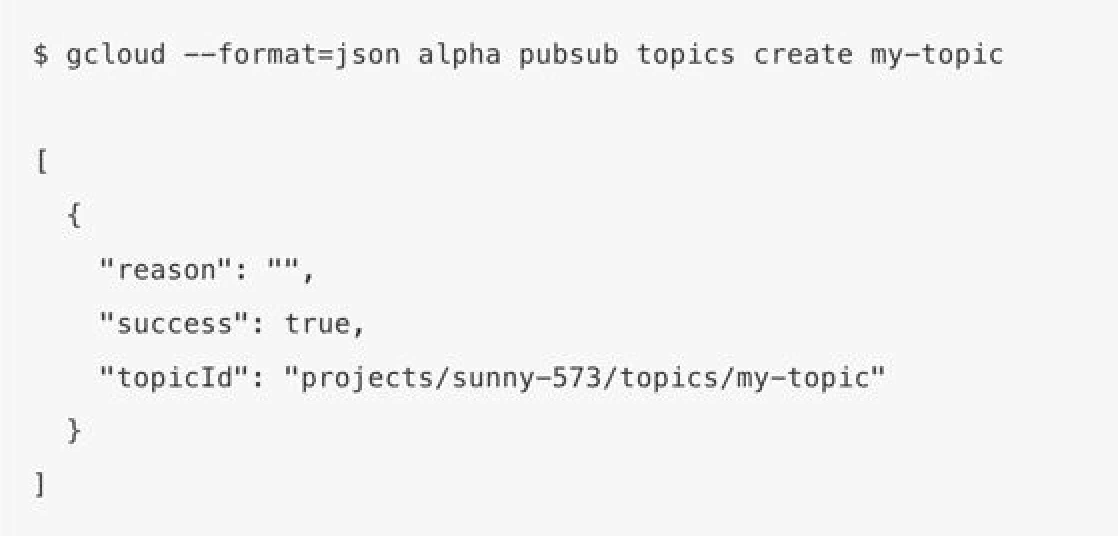
Job Name	starterpipeline-peihsinsu-0623135730
Job ID	2016-06-23_06_57_42-10300938932320125496
Job Status	Running Stop job
SDK Version	Google Cloud Dataflow Java SDK 1.2.1
Job Type	Batch
Start Time	Jun 23, 2016, 9:57:42 PM
Elapsed Time	1 min 0 sec
Errors	0
Warnings	0
Total Worker Time	0 sec

Dataflow in Streaming Mode

Pub/Sub working model

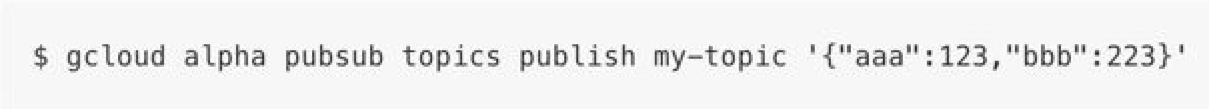


Pub/Sub Operations - Topics



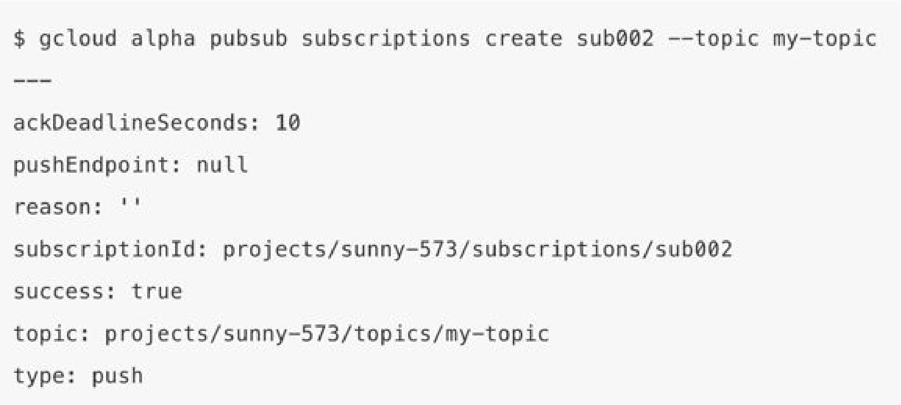
```
$ gcloud --format=json alpha pubsub topics create my-topic

[
  {
    "reason": "",
    "success": true,
    "topicId": "projects/sunny-573/topics/my-topic"
  }
]
```

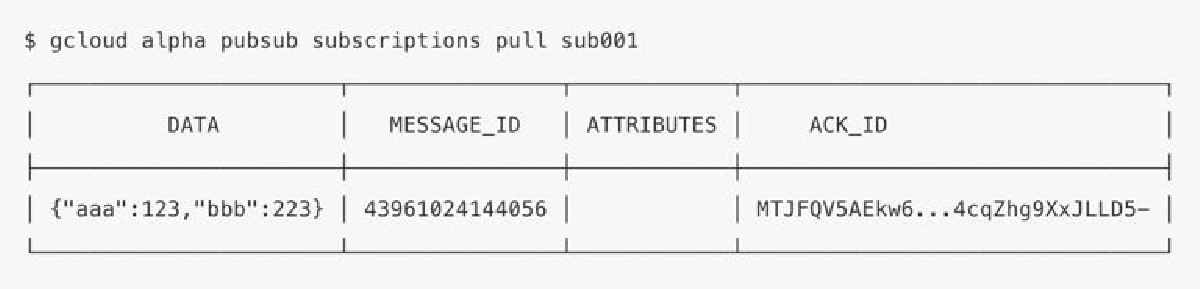


```
$ gcloud alpha pubsub topics publish my-topic '{"aaa":123,"bbb":223}'
```

Pub/Sub Operations - Subscriber

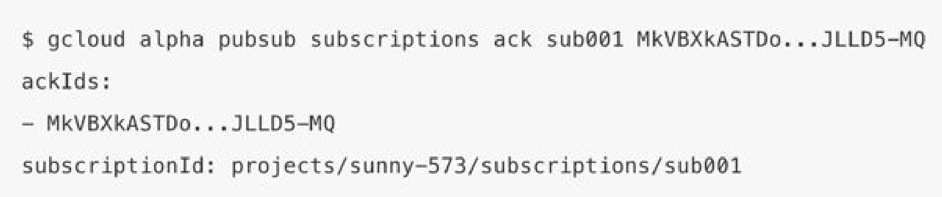


```
$ gcloud alpha pubsub subscriptions create sub002 --topic my-topic
---
ackDeadlineSeconds: 10
pushEndpoint: null
reason: ''
subscriptionId: projects/sunny-573/subscriptions/sub002
success: true
topic: projects/sunny-573/topics/my-topic
type: push
```



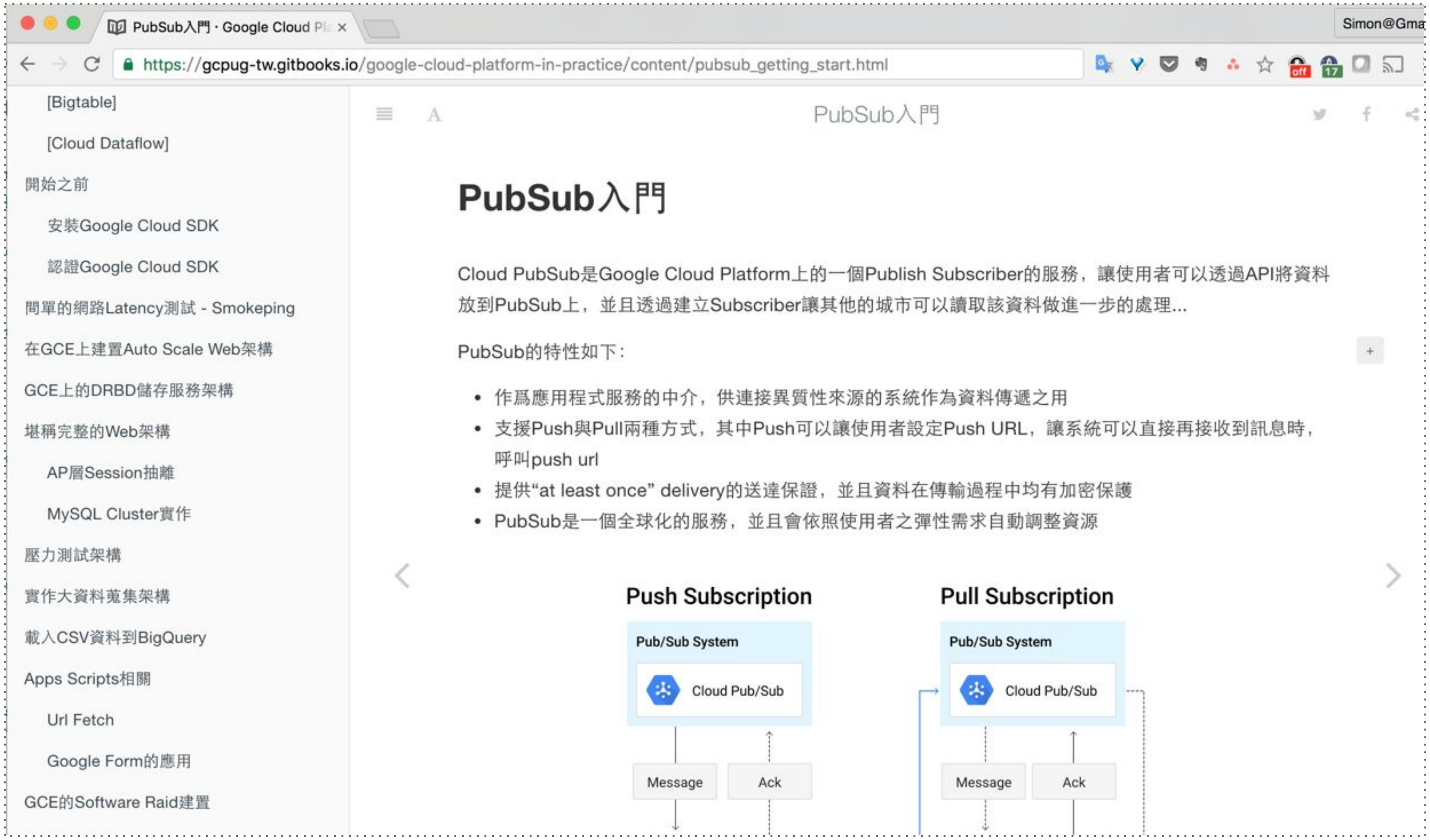
```
$ gcloud alpha pubsub subscriptions pull sub001

┌───────────────────────┬──────────────────┬────────────┬──────────────────────────────────┐
│         DATA          │    MESSAGE_ID    │ ATTRIBUTES │     ACK_ID                       │
├───────────────────────┼──────────────────┼────────────┼──────────────────────────────────┤
│ {"aaa":123,"bbb":223} │ 43961024144056   │            │ MTJFQV5AEkw6...4cqZhg9XxJLLD5-   │
└───────────────────────┴──────────────────┴────────────┴──────────────────────────────────┘
```



```
$ gcloud alpha pubsub subscriptions ack sub001 MkVBXkASTDo...JLLD5-MQ
ackIds:
- MkVBXkASTDo...JLLD5-MQ
subscriptionId: projects/sunny-573/subscriptions/sub001
```

Simple Guide for Pub/Sub



PubSub入門 · Google Cloud Pl
Simon@Gma
https://gcpug-tw.gitbooks.io/google-cloud-platform-in-practice/content/pubsub_getting_start.html

PubSub入門

PubSub入門

Cloud PubSub是Google Cloud Platform上的一個Publish Subscriber的服務，讓使用者可以透過API將資料放到PubSub上，並且透過建立Subscriber讓其他的城市可以讀取該資料做進一步的處理...

PubSub的特性如下：

- 作爲應用程式服務的中介，供連接異質性來源的系統作為資料傳遞之用
- 支援Push與Pull兩種方式，其中Push可以讓使用者設定Push URL，讓系統可以直接再接收到訊息時，呼叫push url
- 提供“at least once” delivery的送達保證，並且資料在傳輸過程中均有加密保護
- PubSub是一個全球化的服務，並且會依照使用者之彈性需求自動調整資源

Ref: https://gcpug-tw.gitbooks.io/google-cloud-platform-in-practice/content/pubsub_getting_start.html

Dataflow with Cloud Pub/Sub Use Case

Globally redundant

Low latency (sub sec.)

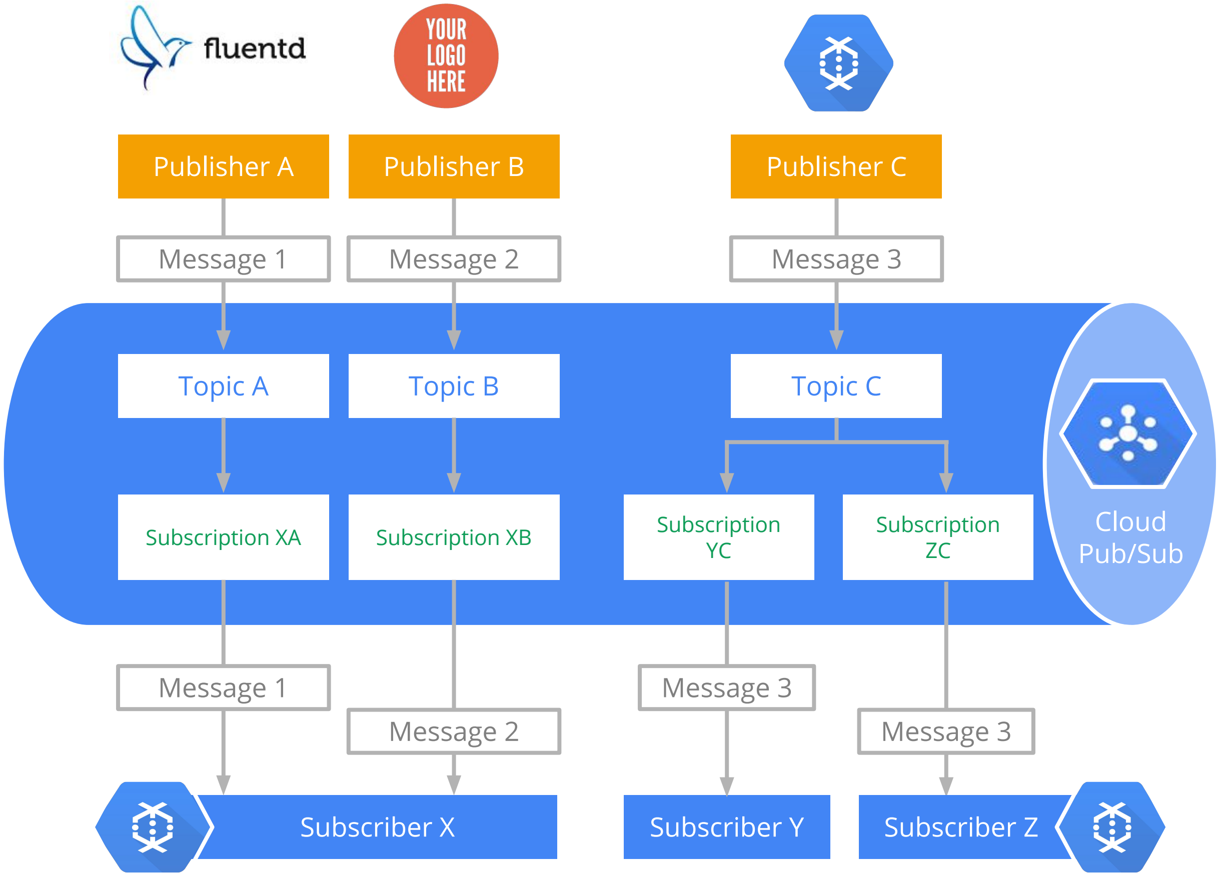
N to N coupling

Batched read/write

Push & Pull

Guaranteed Delivery

Auto expiration



Lab 3: Create a Streaming Dataflow model

- Create PubSub topic
- Deploy Dataflow streaming sample
- Watch Dataflow task dashboard

After Dataflow

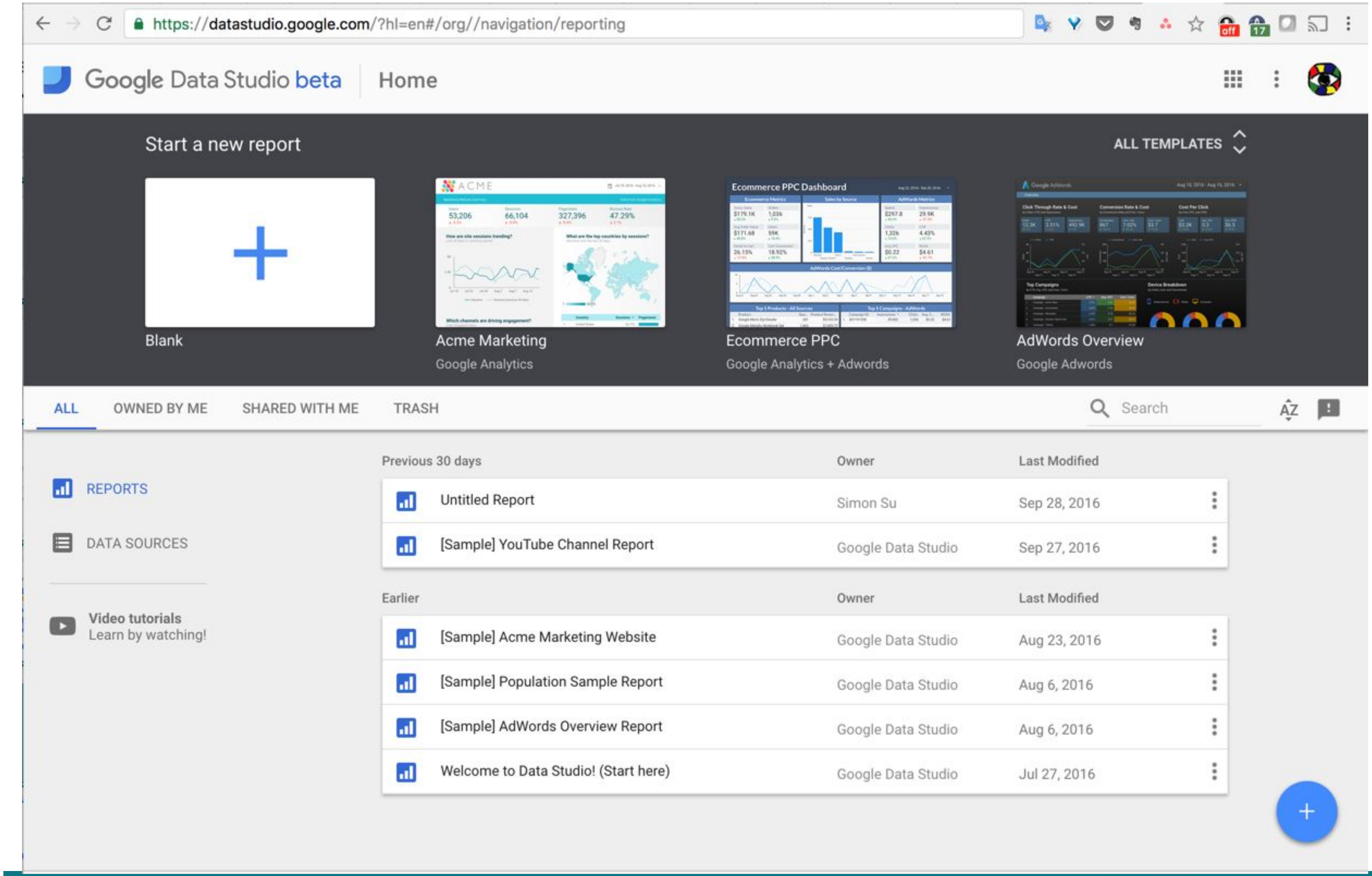
Datalab

An easy tool for analysis and report

BigQuery

An interactive analysis service

Google Data Studio



https://datastudio.google.com/?hl=en#/org//navigation/reporting

Google Data Studio beta | Home

Start a new report

ALL TEMPLATES

Blank

Acme Marketing
Google Analytics

Ecommerce PPC
Google Analytics + Adwords

AdWords Overview
Google Adwords

ALL | OWNED BY ME | SHARED WITH ME | TRASH | Search

REPORTS

DATA SOURCES

Video tutorials
Learn by watching!

Previous 30 days	Owner	Last Modified
Untitled Report	Simon Su	Sep 28, 2016
[Sample] YouTube Channel Report	Google Data Studio	Sep 27, 2016

Earlier	Owner	Last Modified
[Sample] Acme Marketing Website	Google Data Studio	Aug 23, 2016
[Sample] Population Sample Report	Google Data Studio	Aug 6, 2016
[Sample] AdWords Overview Report	Google Data Studio	Aug 6, 2016
Welcome to Data Studio! (Start here)	Google Data Studio	Jul 27, 2016

--- THANKS ---