

Module 12

Big Data as a Service using Google Cloud Platform

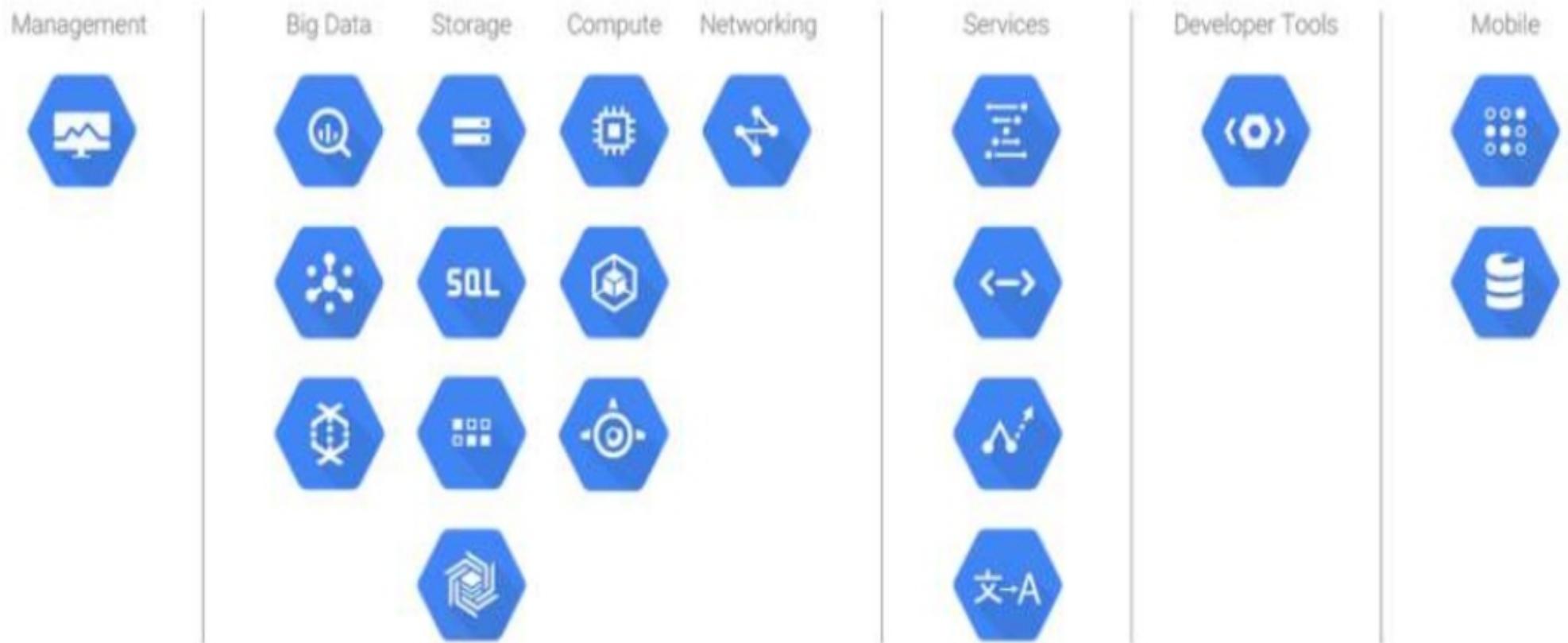
Thanachart Numnonda, Executive Director, IMC Institute

Mr.Aekanun Thongtae, Big Data Consultant, IMC Institute

Thanisa Numnonda, Faculty of Information Technology,

King Mongkut's Institute of Technology Ladkrabang

Google Cloud Platform Services



GCP Compute Services



COMPUTE ENGINE

Run large-scale workloads on virtual machines hosted on Google's infrastructure.



PREEMPTIBLE VMs

Preemptible VMs are a low cost choice for distributed and fault-tolerant workloads.



CUSTOM MACHINE TYPES

Create Compute Engine VMs with optimal amounts of vCPU and memory.



APP ENGINE

A platform for building scalable web apps and mobile backends.



CONTAINER ENGINE

Run Docker containers on Google's infrastructure, powered by Kubernetes.



CONTAINER REGISTRY

Fast, private Docker image storage on Google Cloud Platform.

GCP Storage Services



CLOUD STORAGE

Powerful, simple and cost effective object storage service with global edge-caching.



CLOUD STORAGE NEARLINE

A highly available, affordable solution for backup, archiving and disaster recovery.



CLOUD SQL

Store and manage data using a fully-managed, relational MySQL database.



CLOUD DATASTORE

A managed, NoSQL, schemaless database for storing non-relational data.



CLOUD BIGTABLE

Cloud Bigtable is a fast, fully managed, massively scalable NoSQL database service.

GCP Big Data Services



BIGQUERY

Analyze Big Data in the cloud. Run fast, SQL-like queries against petabytes of data in seconds.



CLOUD DATAFLOW

Dataflow is a real-time data processing service for batch and stream data processing.



CLOUD DATAPROC

Google Cloud Dataproc is a managed Spark and Hadoop service that is fast, easy to use, and low cost.



CLOUD DALAB

An easy to use interactive tool for large-scale data exploration, analysis and visualization.



CLOUD PUB/SUB

Connect your services with reliable, many-to-many, asynchronous messaging hosted on Google's infrastructure.

Google Data Studio 360

GAonGA SDX Report

File Edit View Insert Page Arrange Help

Jun 01, 2015 - Jun 30, 2015

Report Properties

BASIC DATA STYLE

Data Source: GAonGA SDX Data Source

Total Hits Processed: 937,857,670

Trending Users & Sessions

Trending Bounce Rate & Duration

Top Traffic Sources

Source	Users	Pages/Session	Avg. Session Duration
(direct)	243,164	1,003.15	00:26:10
plus.google.c...	215,475	1,091.49	00:26:10
(referral)	138,403	3,044.39	00:26:10
google.co...	73,094	5,091.42	00:26:10
support.goo...	73,080	4,705.17	00:26:10
google.it	64,360	4,373.99	00:26:10
google.co.jp	60,829	1,130.68	00:26:10
google.com.c...	56,792	1,780.74	00:26:10
newspaper	56,386	3,133.49	00:26:10

Device Category

Top Browsers

6

Source:



Cloud Datalab



Google Cloud Datalab Optimize sales with GA data-Completed (unreviewed)

Notebook Add Code Add Markdown Delete Move Up Move Down Run Clear Reset Session

```
%%sql --module conversations
select if(path = '/', 'home', 'product') as start,
       if(tx > 0, 'completed', 'abandoned') as outcome,
       count(*) as count from (
       select visitId, hits.page.pagePath as path, hits.hitNumber as hitNumber,
              sum(if(hits.page.pagePath == '/confirm.html', 1, 0)) within record as tx
         from [google.com:analytics-bigquery:LondonCycleHelmet.ga_sessions_28130918]
        order by visitStartTime, hitNumber)
    where hitNumber = 1
   group by start, outcome;
```

Navigation Help

Help for Python APIs
You can enter `class?` or `method?` within a code cell in the notebook to get help on a Python API.

For example, try `str?` to get help information on the built-in Python method to convert a value to its string representation.

Additional help topics and links are also available from the menu off the Help icon on the top of the page.

Docs and Samples
The [Datalab Guide](#) featuring documentation and sample notebooks is also a great way to check out how you can use Datalab.

Visualize paths taken

Sankey diagram makes it easier to see tabular data

Sankey diagram illustrating user conversion paths:

product → abandoned
count: 29

product → completed
count: 1

home → abandoned
count: 0

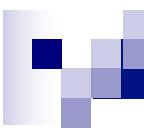
home → completed
count: 1

```
graph LR
    product((product)) --> abandoned1[abandoned  
count: 29]
    product --> completed1[completed  
count: 1]
    home((home)) --> abandoned2[abandoned  
count: 0]
    home --> completed2[completed  
count: 1]
```

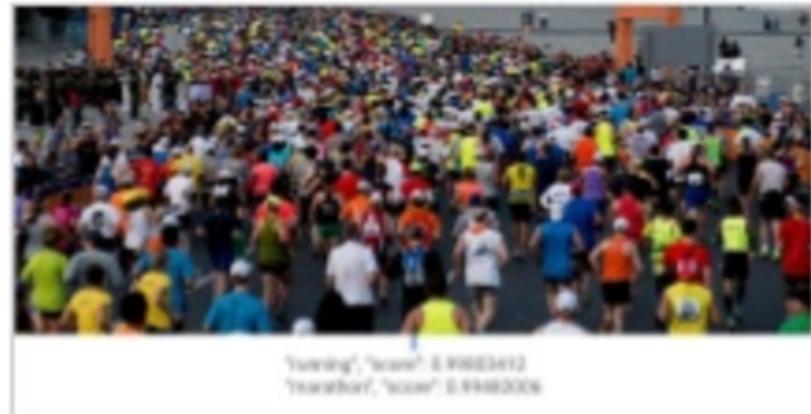
TensorFlow

- Deep Learning technology powering over 100 Google services
- Generalizable to vision, sound, text, video and other data
- Runs on CPUs or GPUs, desktop, server, or mobile computing platforms
- Available as Apache 2.0 OSS license

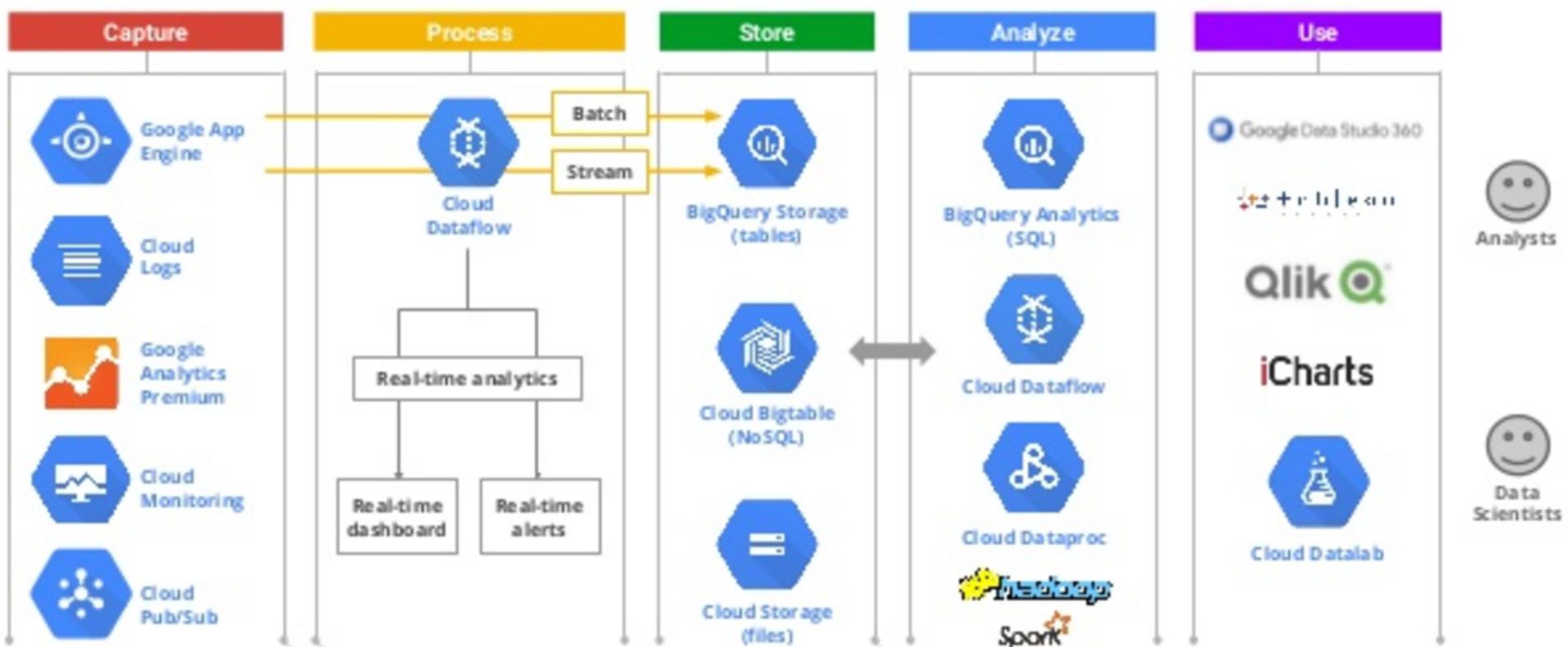




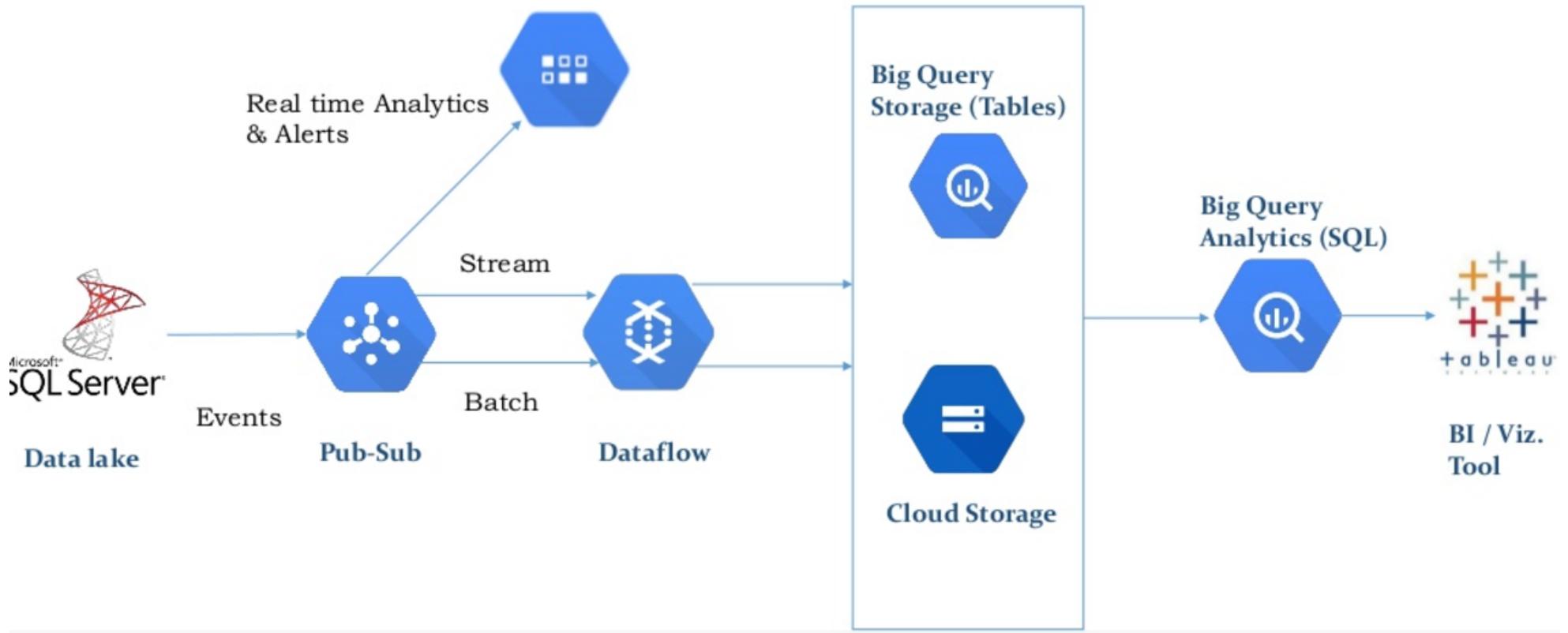
Cloud Vision API



Big Data Lifecycle



Big Data – Google Cloud Architecture



Hadoop v.s. Public Cloud Services

	Traditional	Google Cloud	AWS	Azure
Ingestion	Sqoop			
	Flume	DataFlow	Kinesis	Event Hub
	Kafka	Pub/Sub		
Storage	HDFS	Google Storage	S3	Azure Blob
	HBase/NoSQL	Big Table	DynamoDB	Storage Table
	RDBMS	Cloud SQL	RDS	Azure Database
Data warehouse	Teradata etc.		Redshift	Azure DW
Big Data Processing	Hadoop/Spark	DataProc	EMR	HD Insight
	Jupyter	DataLab		Visual Studio
	Hive	Big Query	Athena	Azure Functions
Machine Learning	Spark MLlib	ML Engine	Machine Learning	Azure ML
Data Visualization	Tableau etc.	Data Studio	QuickSight	Power BI



Google Cloud Shell

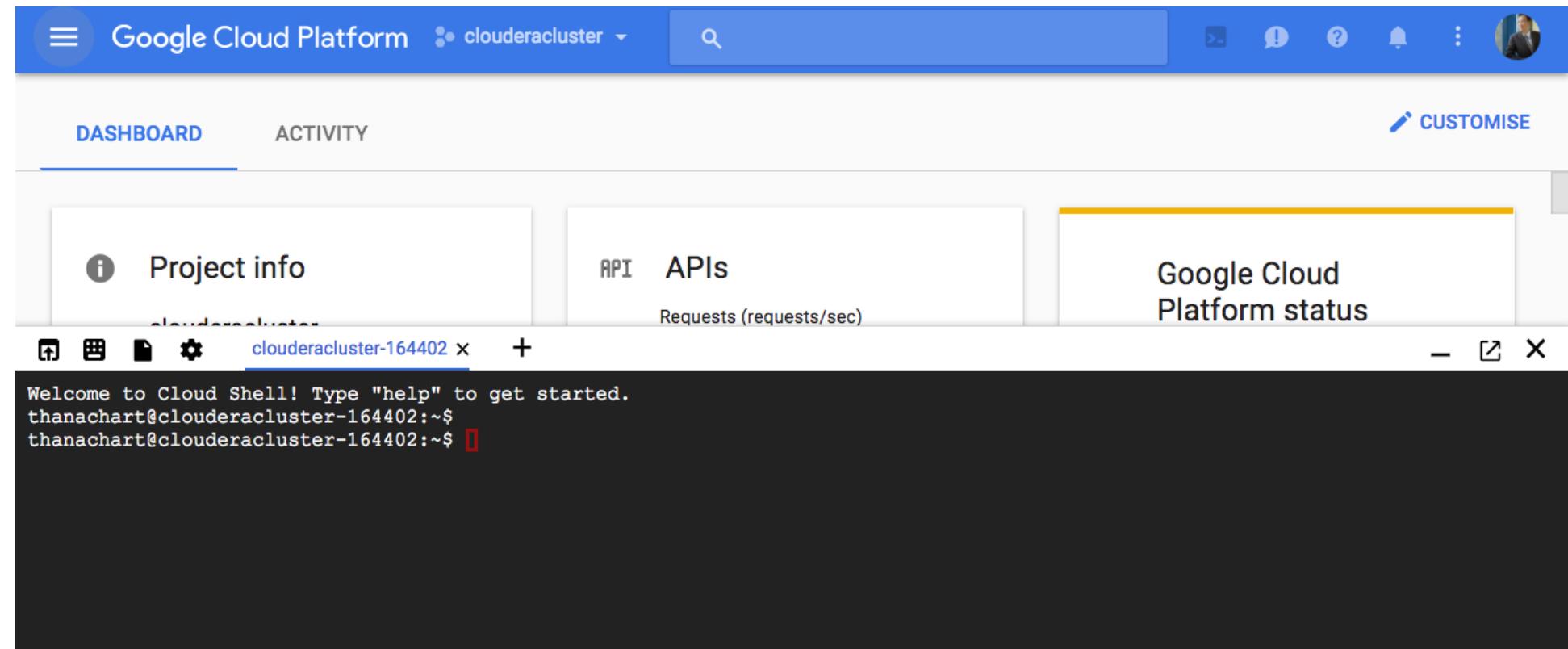
Google Cloud Shell is a shell environment for managing resources hosted on Google Cloud Platform. It is a micro VM and it is free of charge.



Launching Cloud Shell

- Select or create a Cloud Platform project.
- Click the Activate Google Cloud Shell button at the top of the Google Cloud Platform Console.
- A Cloud Shell session opens inside a new frame at the bottom of the console and displays a command-line prompt.

The image shows the Google Cloud Platform (GCP) dashboard. At the top, there's a blue header bar with the GCP logo, the project name "clouderacluster", a search bar, and several status icons. Below the header is a navigation menu on the left containing links like Home, API Manager, Billing, Cloud Launcher, Support, IAM & Admin, App Engine, and Compute Engine. The main area is the "DASHBOARD", which features two main cards: "Project info" (showing the project name "clouderacluster", Project ID "clouderacluster-164402", and number "No. 984926510584") and "Resources" (showing Cloud Storage with 3 buckets). To the right of these cards is a "CUSTOMISE" section with a "Google Cloud Shell" button, which is highlighted with a red arrow and the text "Google Cloud Shell". Below this is a chart titled "APIs" showing "Requests (requests/sec)" over time, with a single data point at 16 May, 10:44 showing 0.0133 requests.





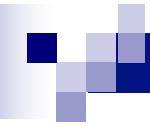
Google Cloud Storage

Google Cloud Storage is unified object storage for developers and enterprises, from live data serving to data analytics/ML to data archiving. It provides a unified offering across the availability spectrum: from live data tapped by today's most demanding applications, to cloud archival solutions Nearline and Coldline. Featuring a consistent API, latency, and speed across storage classes



Cloud Storage

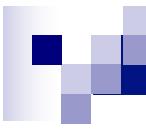
- Multi-Regional or Regional storage
- Nearline and Coldline storage solutions
- Offers a unified product offering with consistent access APIs across the entire range of storage classes
- With no minimum fees and a pay-per-use model,



Cloud Storage

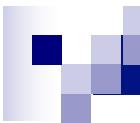
General Pricing

Multi-Regional Storage (per GB per Month)	Regional Storage (per GB per Month)	Nearline Storage (per GB per Month)	Coldline Storage (per GB per Month)
\$0.026	\$0.02	\$0.01	\$0.007



Lab I: Importing/Exporting

data to Google Cloud Storage



Launch Google Cloud console

The screenshot shows the Google Cloud Platform homepage. At the top, there's a navigation bar with links for "Why Google", "Products", "Solutions", "Launcher", "Pricing", and a "CONTACT SALES" button. To the right of the navigation is a search bar with a magnifying glass icon and the word "Search". Further right is a user profile icon showing a person's face. Below the navigation bar, the main heading "Build What's Next" is displayed in large, bold letters. Underneath it, the tagline "Make an impact with Google Cloud Platform" is shown. Two descriptive paragraphs follow: "Build and host applications and websites, store data, and analyze data on Google's scalable infrastructure." and "Develop faster, better software while staying fully compliant with industry standards." At the bottom left is a dark button with the text "GO TO CONSOLE" in white. A red arrow points from the text "GO TO CONSOLE" down towards the button. To the right of the button is another "CONTACT SALES" button.

Google Cloud Platform

Why Google Products Solutions Launcher Pricing > CONTACT SALES

Build What's Next

Make an impact with Google Cloud Platform

Build and host applications and websites, store data, and analyze data on Google's scalable infrastructure.

Develop faster, better software while staying fully compliant with industry standards.

GO TO CONSOLE

CONTACT SALES

Select Storage

The screenshot shows the Google Cloud Platform storage interface. At the top, there's a blue header bar with the 'Google Cloud Platform' logo, a dropdown for 'clouderacluster', a search bar, and various status icons. Below the header is a toolbar with 'Browser', 'CREATE BUCKET', 'REFRESH', 'DELETE', and 'SHOW INFO PANEL' buttons. On the left, a sidebar menu is open under the 'Menu' icon. It lists several services: Home, Networking, STORAGE (BigTable, Datastore, Storage, SQL, Spanner), and STACKDRIVER. The 'Storage' item is highlighted with a red arrow. A dropdown menu for 'Storage' is open, showing options: 'Browser' (which is selected and highlighted in grey), 'Transfer', and 'Settings'. The main content area shows a table of buckets:

Name	Default storage class	Location	Labels
dataproc-36c094d4-2796-4b03-85c9-7e9a90a99654-us	Multi-Regional	US	
dataproc-d2f44deb-3c45-4cc1-a5e4-329036679c87-asia-southeast1	Regional	ASIA-SOUTHEAST1	
imcinstitute	Multi-Regional	ASIA	
	Regional	ASIA-SOUTHEAST1	
	Regional	EUROPE-WEST1	

Click on Create Bucket

The screenshot shows the Google Cloud Platform Storage Browser interface. On the left, there's a sidebar with 'Storage' selected, showing options for 'Browser', 'Transfer', and 'Settings'. The main area is titled 'Browser' and contains a 'CREATE BUCKET' button with a red arrow pointing to it. Below the button is a search bar labeled 'Filter by prefix...'. A table lists five existing buckets: 'dataproc-36c094d4-2796-4b03-85c9-7e9a90a99654-us', 'dataproc-d2f44deb-3c45-4cc1-a5e4-329036679c87-asia-southeast1', 'waris', and 'waris2'. Each bucket entry includes columns for 'Name', 'Default storage class', 'Location', and 'Labels'.

Name	Default storage class	Location	Labels
dataproc-36c094d4-2796-4b03-85c9-7e9a90a99654-us	Multi-Regional	US	⋮
dataproc-d2f44deb-3c45-4cc1-a5e4-329036679c87-asia-southeast1	Regional	ASIA-SOUTHEAST1	⋮
waris	Regional	ASIA-SOUTHEAST1	⋮
waris2	Regional	EUROPE-WEST1	⋮

Choose a globally unique bucket name

The screenshot shows the Google Cloud Platform interface for creating a new storage bucket. The top navigation bar includes the 'Google Cloud Platform' logo, a dropdown for 'clouderacluster', a search bar, and various status icons. The left sidebar is titled 'Storage' and lists 'Browser', 'Transfer', and 'Settings'. The main content area is titled 'Create a bucket' with a back arrow. It features a 'Name' field containing 'imcinstitute', a 'Default storage class' section with options for 'Multi-Regional', 'Regional' (which is selected), 'Nearline', and 'Coldline', and a note about privacy.

Storage

Create a bucket

Name ?
Must be unique across Cloud Storage. Privacy: Do not include sensitive information in your bucket name. Others can discover your bucket name if it matches a name they're trying to use.

imcinstitute

Default storage class ?
[Find out about pricing](#)

Multi-Regional
Use to stream videos and host hot web content.
Best for data accessed frequently around the world.

Regional
Use to store data and run data analytics.
Best for data accessed frequently in one part of the world.

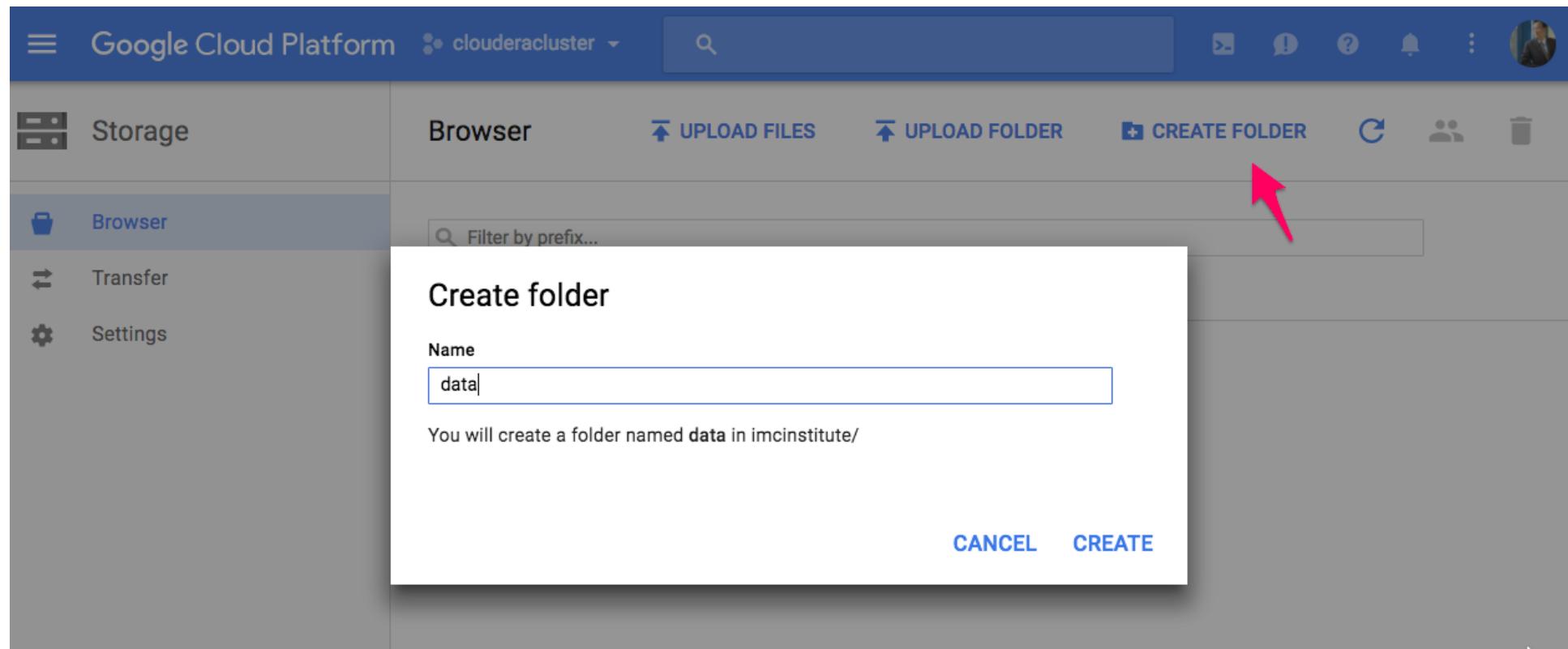
Nearline
Use to store rarely accessed documents.
Best for data accessed less than once per month.

Coldline
Use to store very rarely accessed documents.
Best for data accessed less than once per year.

Review objects in the bucket

The screenshot shows the Google Cloud Platform Storage Browser interface. The top navigation bar includes the Google Cloud Platform logo, the project name "clouderacluster", a search bar, and various notification and user icons. On the left, a sidebar menu lists "Storage", "Browser" (which is selected and highlighted in blue), "Transfer", and "Settings". The main content area is titled "Browser" and contains buttons for "UPLOAD FILES", "UPLOAD FOLDER", and "CREATE FOLDER". A "Filter by prefix..." input field is present. The path "Buckets / imcinstigate" is shown. A message at the bottom states "There are no objects in this bucket."

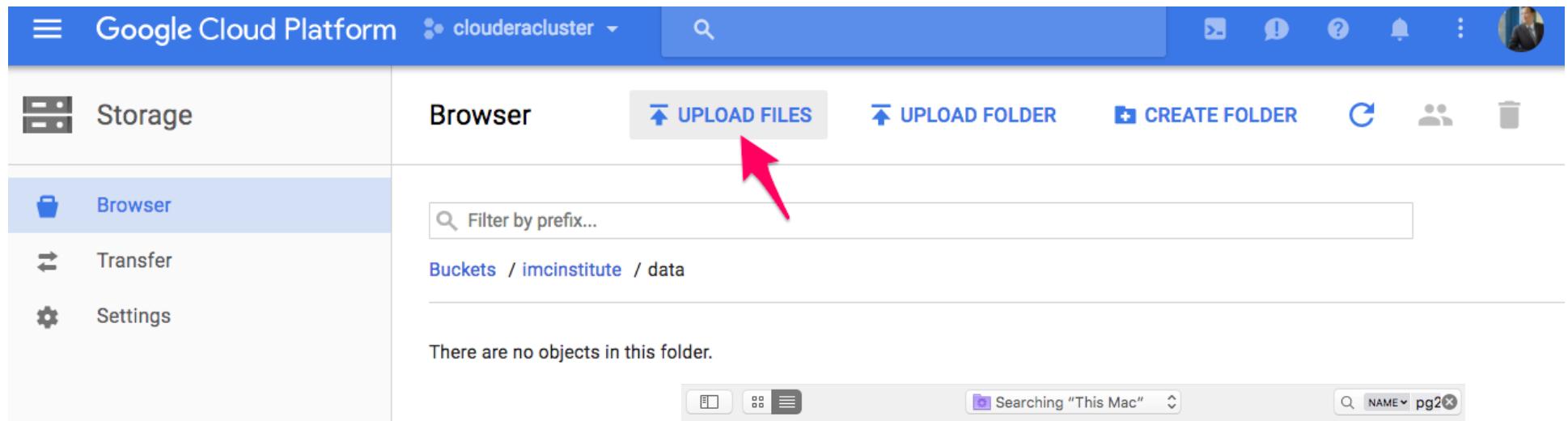
Create a new folder name as: **data**



The screenshot shows the Google Cloud Platform Storage Browser interface. The top navigation bar includes the Google Cloud Platform logo, the project name "clouderacluster", a search bar, and various status icons. On the left, a sidebar menu lists "Storage", "Browser" (which is selected and highlighted in blue), "Transfer", and "Settings". The main content area is titled "Browser" and features "UPLOAD FILES", "UPLOAD FOLDER", and "CREATE FOLDER" buttons. A "Filter by prefix..." input field is present. The "Buckets" section shows a single entry: "imcinstigate / imcinstigate". A table below lists a single folder named "data/".

Name	Size	Type	Storage class	Last modified	Sha
data/	-	Folder	-	-	

Upload a local file to the cloud storage



The screenshot shows the Google Cloud Platform Storage Browser interface. The left sidebar has 'Storage' selected. The main area shows a 'Browser' tab with 'UPLOAD FILES' highlighted by a red arrow. Below it is a search bar labeled 'Filter by prefix...' and a breadcrumb navigation 'Buckets / imcinstigate / data'. A message says 'There are no objects in this folder.' Below the interface is a Mac OS X Finder window showing a list of files in the 'Downloads' folder, including 'pg2600.txt' and 'pg26000.txt' from 2015.

Date Modified	Size	Kind
3/24/2558 BE, 2:27 PM	3.3 MB	Plain Text
3/24/2558 BE, 2:27 PM	3.3 MB	Plain Text
12/23/2557 BE, 8:49 PM	3.3 MB	Plain Text

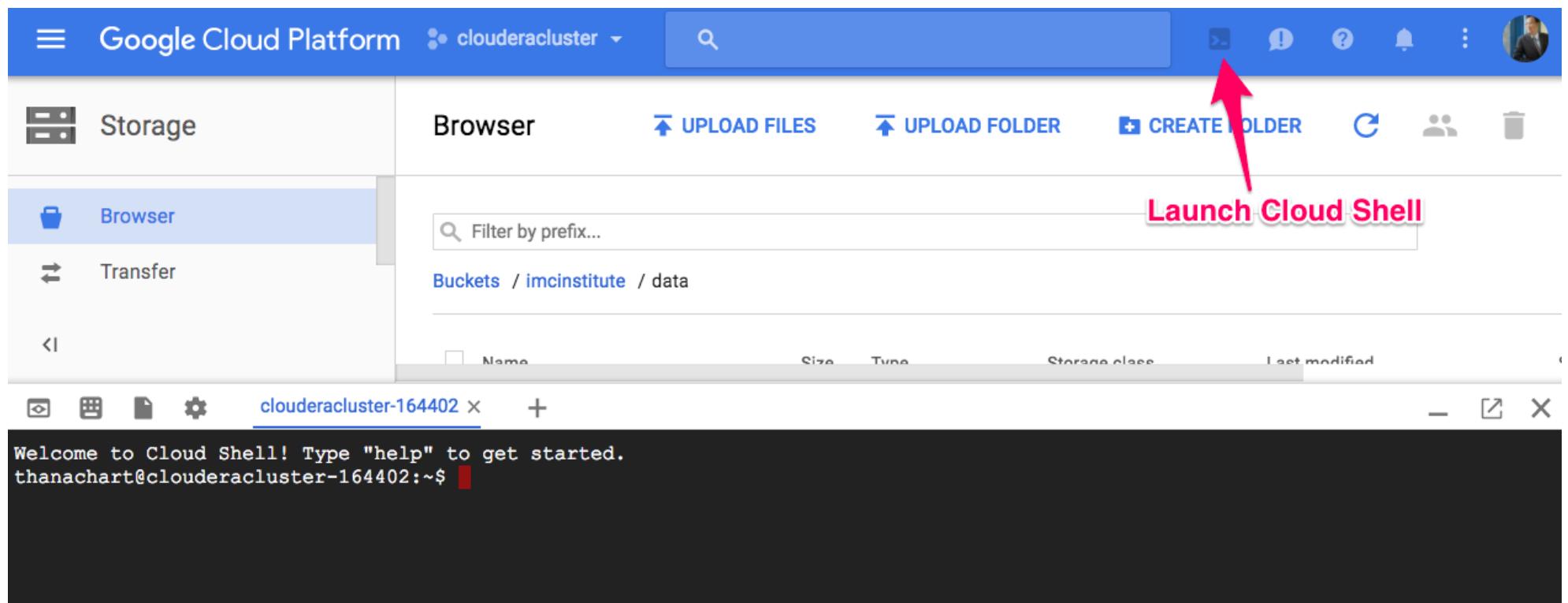
Format: All Files

Options Cancel Open

The screenshot shows the Google Cloud Platform Storage Browser interface. The left sidebar has 'Storage' selected, with 'Browser' highlighted. The main area shows a file named 'pg2600.txt' in the 'imcinstitute / data' bucket. The file is 3.14 MB, text/plain type, Regional storage class, and was last modified on 04/06/2017, 10:39.

Name	Size	Type	Storage class	Last modified
pg2600.txt	3.14 MB	text/plain	Regional	04/06/2017, 10:39

Upload a large file using Cloud shell



Download an example text file using wget command

```
$wget https://s3.amazonaws.com/imcbucket/input/pg2600.txt
```

```
Welcome to Cloud Shell! Type "help" to get started.  
thanachart@clouderacluster-164402:~$ wget https://s3.amazonaws.com/imcbucket/input/pg2600.txt  
--2017-06-04 10:47:06-- https://s3.amazonaws.com/imcbucket/input/pg2600.txt  
Resolving s3.amazonaws.com (s3.amazonaws.com) ... 52.216.1.3  
Connecting to s3.amazonaws.com (s3.amazonaws.com)|52.216.1.3|:443... connected.  
HTTP request sent, awaiting response... 200 OK  
Length: 3291648 (3.1M) [text/plain]  
Saving to: 'pg2600.txt'  
  
pg2600.txt          100%[=====>]  3.14M  402KB/s  in 19s  
  
2017-06-04 10:47:27 (165 KB/s) - 'pg2600.txt' saved [3291648/3291648]  
thanachart@clouderacluster-164402:~$
```

Upload Data to Google cloud storage using gsutil command

```
$gsutil cp pg2600.txt gs://<YOUR-BUCKET>/
```

```
thanachart@clouderacluster-164402:~$ gsutil cp pg2600.txt gs://imcinstitute/
Copying file://pg2600.txt [Content-Type=text/plain]...
- [1 files] [ 3.1 MiB/ 3.1 MiB]
Operation completed over 1 objects/3.1 MiB.
thanachart@clouderacluster-164402:~$
```

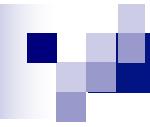
The screenshot shows the Google Cloud Platform Storage Browser interface. The top navigation bar includes the Google Cloud Platform logo, the project name "clouderacluster", a search bar, and various status icons. On the left, a sidebar menu is open, showing "Storage" selected, along with "Browser" (which is highlighted in blue), "Transfer", and "Settings". The main content area is titled "Browser" and displays a table of files. At the top of the table are buttons for "UPLOAD FILES" and "UPLOAD FOLDER", and icons for creating a new folder, refreshing, sharing, and deleting. A search bar labeled "Filter by prefix..." is also present. The table lists one file: "pg2600.txt", which is 3.14 MB in size, of type text/plain, and stored in the "Regional" storage class. It was last modified on 04/06/2017, 10:39.

<input type="checkbox"/>	Name	Size	Type	Storage class	Last modified
<input type="checkbox"/>	pg2600.txt	3.14 MB	text/plain	Regional	04/06/2017, 10:39



Google Cloud DataProc

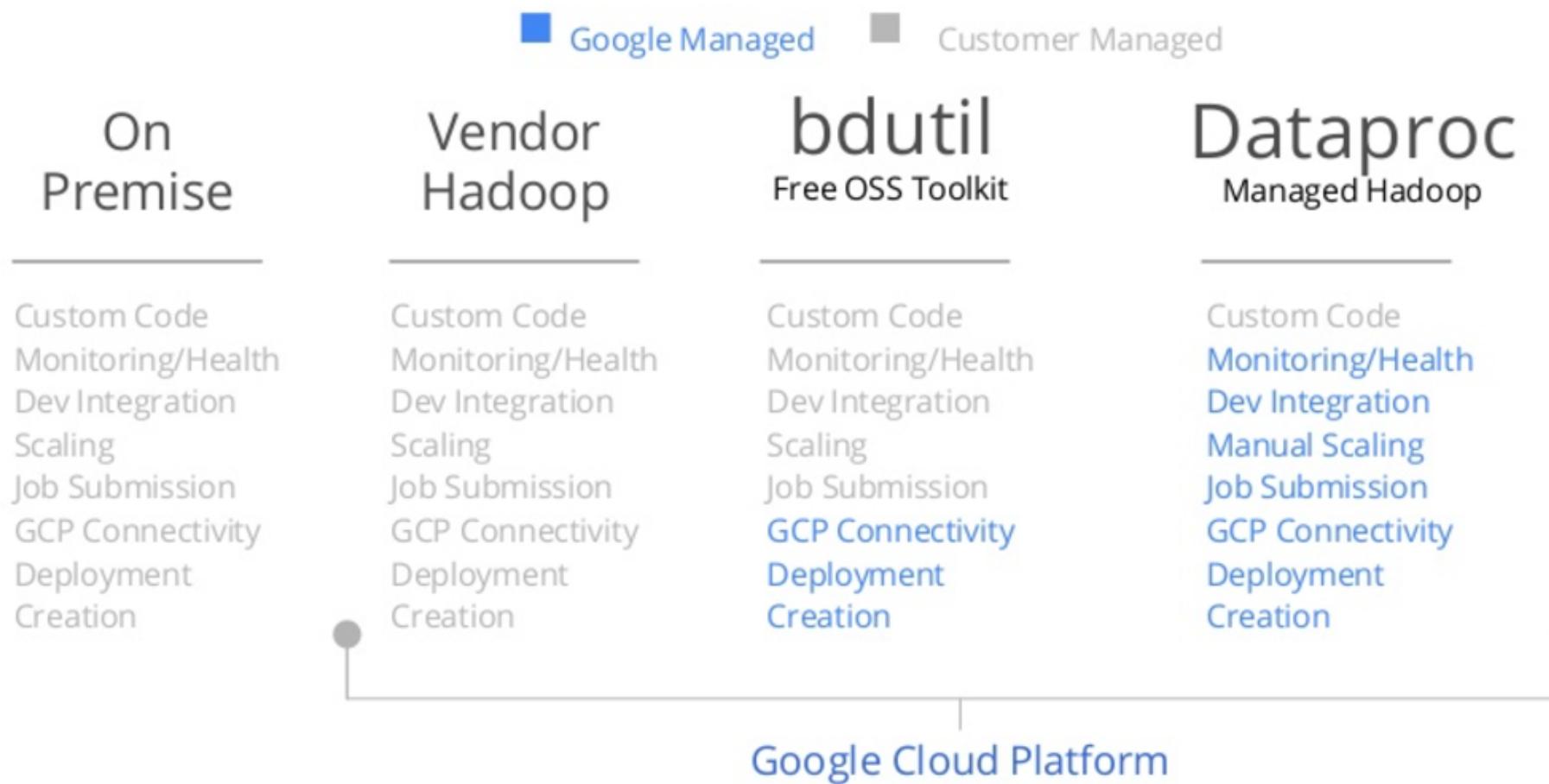
Google Cloud DataProc is a managed Spark and Hadoop service which is fast, easy to use, and low cost



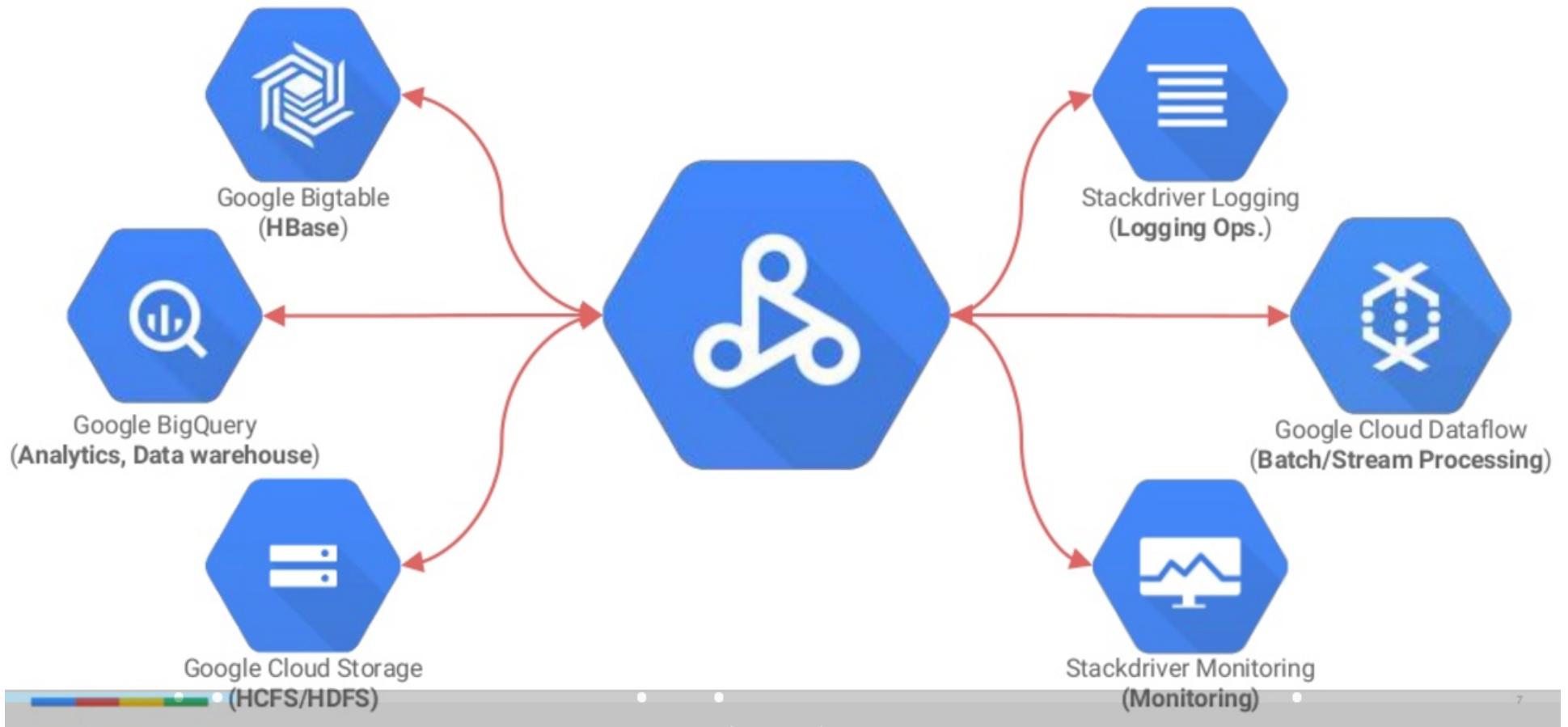
Google Cloud Dataproc

- Managed Hadoop & Spark, an Apache Hadoop, Apache Spark, Apache Pig, and Apache Hive service, to easily process big datasets at low cost.
- Fast & Scalable Data Processing
- Affordable Pricing
- Open Source Ecosystem

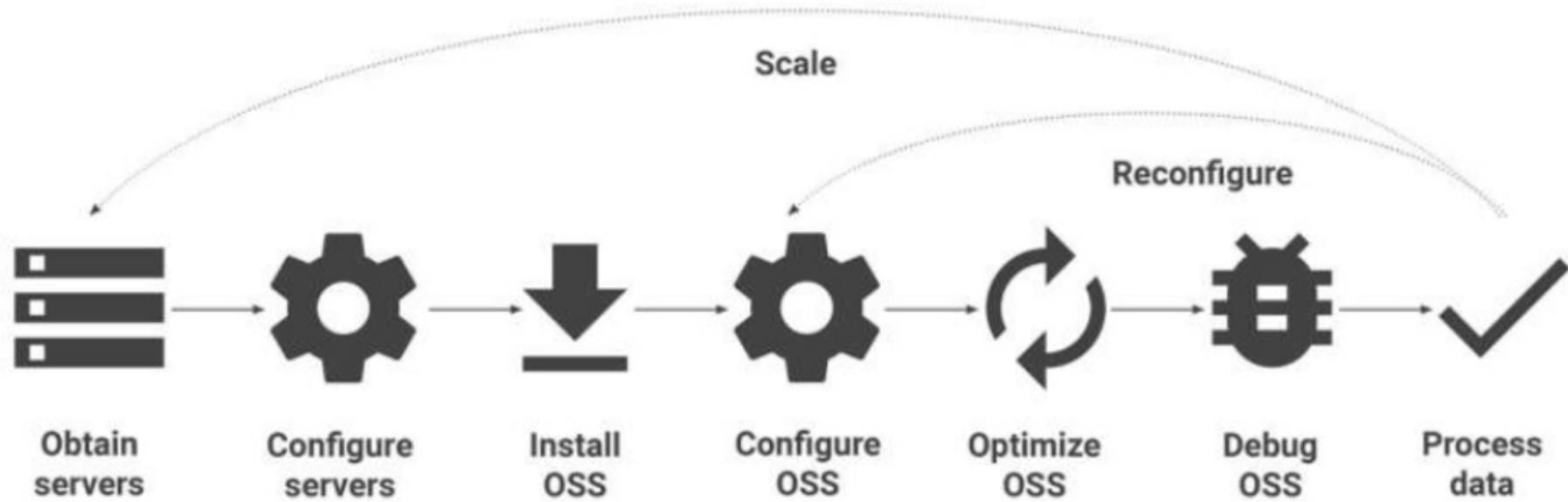
Running Hadoop on Google Cloud



Where Cloud Dataproc fits into GCP

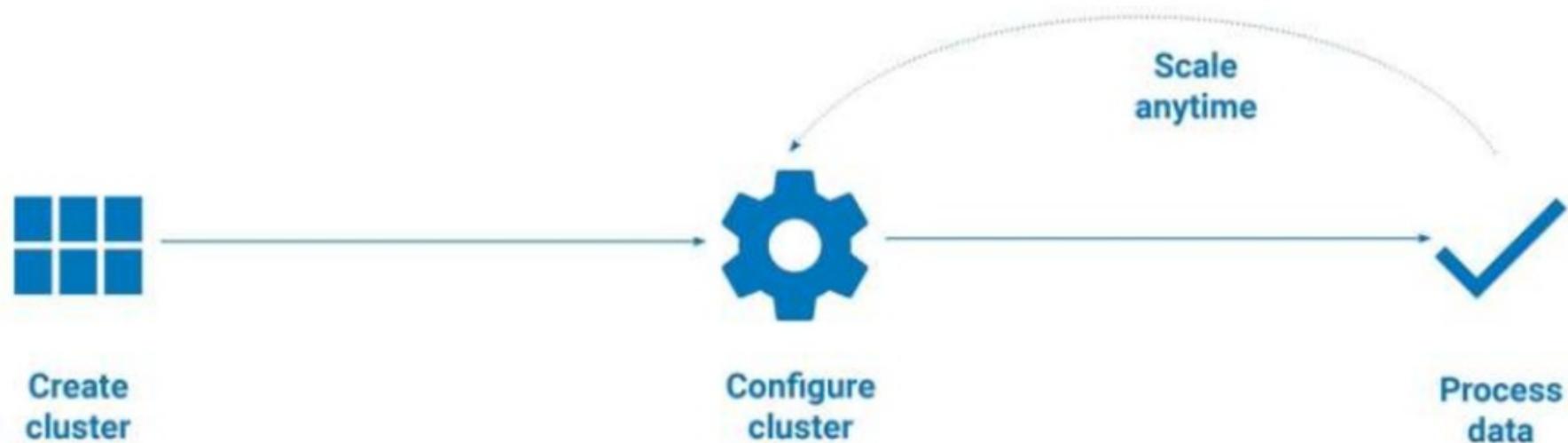


Traditional Spark and Hadoop clusters

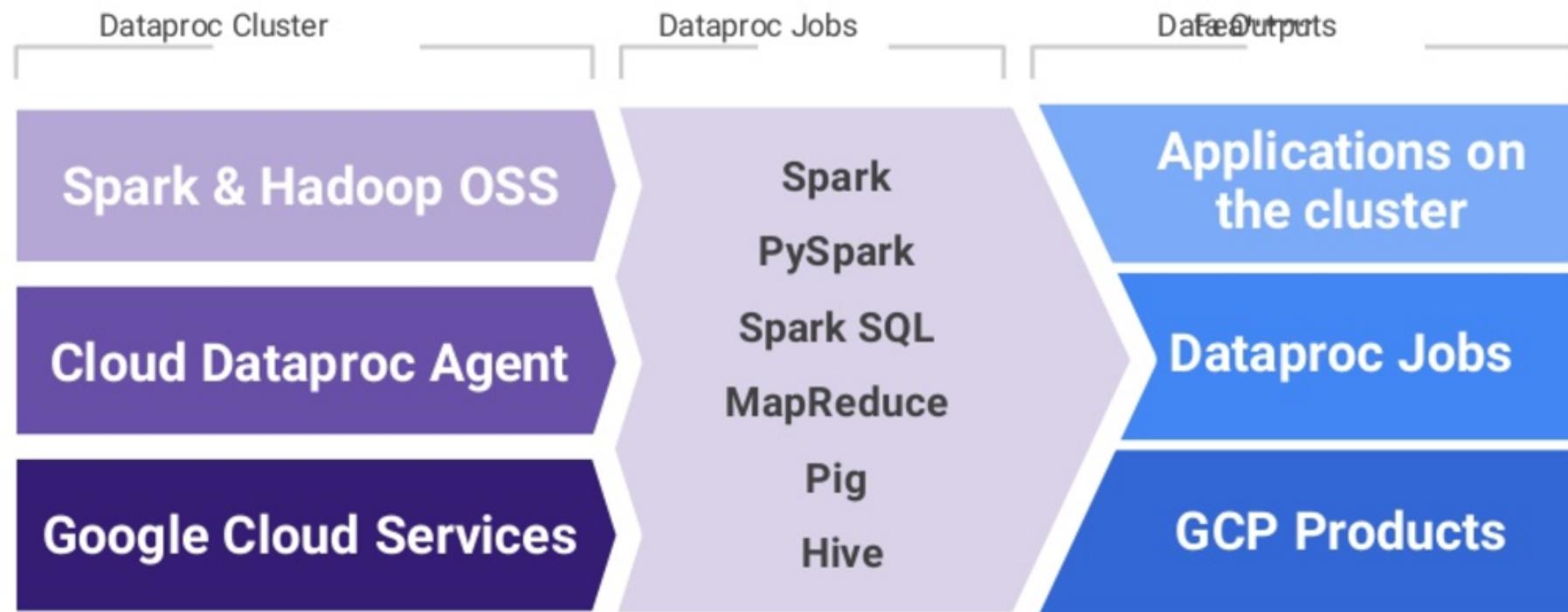


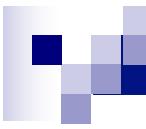


Google Cloud Dataproc



Google Cloud Dataproc - under the hood





Lab II: Launch Dataproc

select Dataproc >> Clusters

The screenshot shows the Google Cloud Platform dashboard. On the left, there's a sidebar with various services: Home, Pins appear here, Endpoints, BIG DATA (BigQuery, Pub/Sub), Dataproc, Dataflow, ML Engine, and Genomics. The 'Dataproc' item is highlighted with a red box and has a red arrow pointing to the 'Clusters' link in its dropdown menu. The main area displays an API requests chart and a section for Google Cloud Platform status and Billing.

Google Cloud Platform

Pins appear here

Home

Endpoints

BIG DATA

BigQuery

Pub/Sub

Dataproc

Dataflow

ML Engine

Genomics

clouderaclust...

API Requests (requests/sec)

0.04
0.03
0.02
0.01

09:36 6 Jun, 10:36

Requests: 0.04

Clusters

Jobs

Go to APIs overview

CUSTOMISE

Google Cloud Platform status

All services normal

Go to Cloud status dashboard

Billing

\$0.42

Approximate charges so far this month

View detailed charges

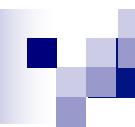
click Create cluster

The screenshot shows the Google Cloud Platform interface for Cloud Dataproc. The top navigation bar includes the Google Cloud logo, 'Google Cloud Platform' text, a dropdown menu showing 'clouderaclust...', a search bar, and various status icons. On the left, a sidebar for 'Cloud Dataproc' lists 'Clusters' (which is selected and highlighted in blue) and 'Jobs'. The main content area is titled 'Clusters' and contains a box with the heading 'Cloud Dataproc Clusters'. It describes the service as letting you provision Apache Hadoop clusters and connect to underlying analytic data stores. Below this text is a call-to-action button labeled 'Create cluster'.



Using the following configuration

- Name the cluster.
- Choose the region
- Change the machine type of both the Master and the Worker nodes to n1-standard-2
- Select disk size to 80 GB



Google Cloud Platform clouderaclust... ⚙️ 🔍 ⚡ ⚡ ⚡ ⚡ ⚡

Cloud Dataproc Create a cluster

Name [?](#)
thanachart-dataproc

Zone [?](#)
us-central1-a

Master node
Contains the YARN Resource Manager, HDFS NameNode and all job drivers

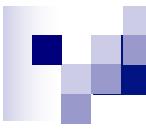
Machine type [?](#) Cluster mode [?](#)
n1-standard-2 (2 vCPU, 7.50 GB ...) Standard (1 master, N workers)

Primary disk size (minimum 10 GB) [?](#)
80 GB

Worker nodes
Each contains a YARN NodeManager and a HDFS DataNode.
The HDFS replication factor is 2.

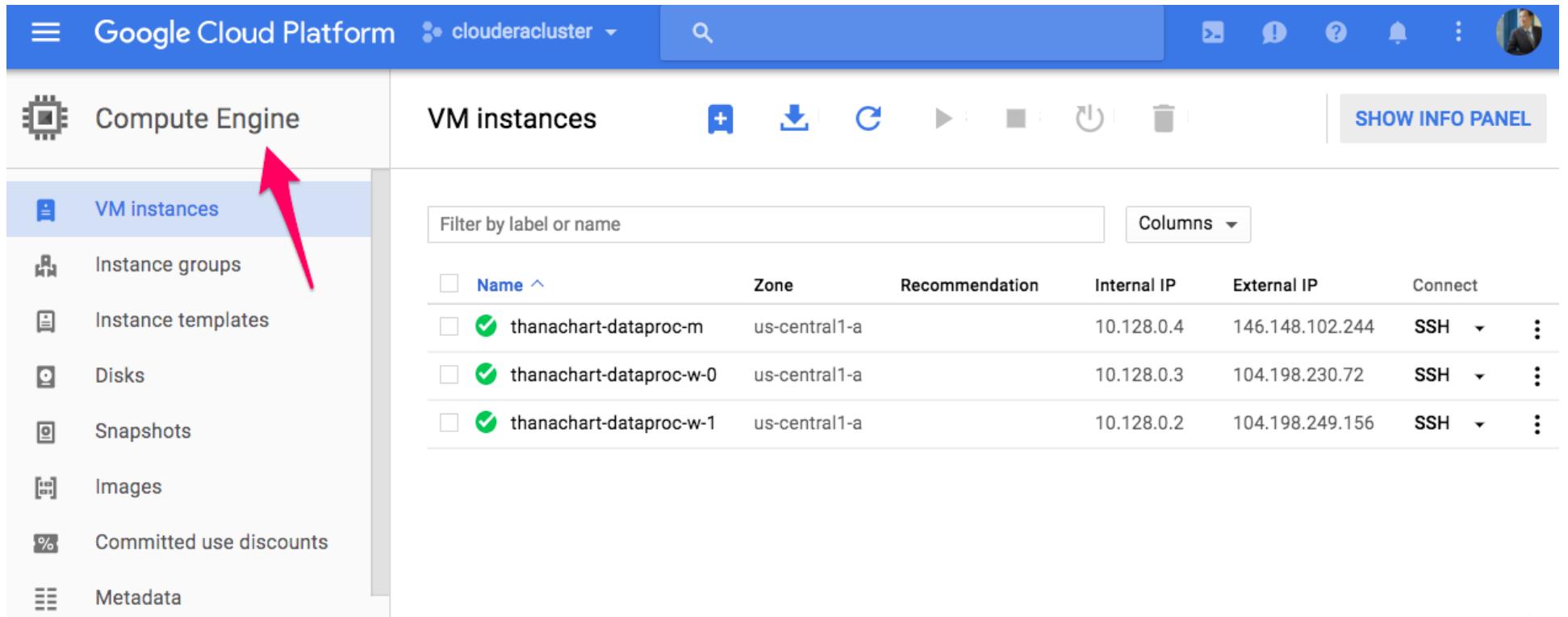
Machine type [?](#) Nodes (minimum 2) [?](#)
n1-standard-2 (2 vCPU, 7.50 GB ...) 2

Primary disk size (minimum 10 GB) [?](#) Local SSDs (0-8) [?](#)



Lab III: Running a Pig script

Select the Dataproc master instance from Compute engine



The screenshot shows the Google Cloud Platform Compute Engine interface. The left sidebar is titled "Compute Engine" and contains the following menu items:

- VM instances (selected, highlighted in blue)
- Instance groups
- Instance templates
- Disks
- Snapshots
- Images
- Committed use discounts
- Metadata

A red arrow points to the "VM instances" link in the sidebar.

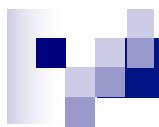
The main content area is titled "VM instances" and displays a table of running instances. The table includes columns for Name, Zone, Recommendation, Internal IP, External IP, and Connect (SSH). The instances listed are:

Name	Zone	Recommendation	Internal IP	External IP	Connect
thanachart-dataproc-m	us-central1-a		10.128.0.4	146.148.102.244	SSH
thanachart-dataproc-w-0	us-central1-a		10.128.0.3	104.198.230.72	SSH
thanachart-dataproc-w-1	us-central1-a		10.128.0.2	104.198.249.156	SSH

Start SSH from the master node

The screenshot shows the Google Cloud Platform Compute Engine interface. The left sidebar is collapsed. The main area displays a list of VM instances under the 'VM instances' tab. A red box highlights the first instance, 'thanachart-dataproc-m', which is checked. To the right of the instance details, there is a 'Connect' button with a dropdown arrow. A red arrow points to this button. A context menu is open over the 'Connect' button, listing four options: 'Open in browser window' (with a red arrow pointing to it), 'Open in browser window on custom port', 'View gcloud command', and 'Use another SSH client'. The 'Connect' button has a blue border.

Name	Zone	Internal IP	External IP	Connect
thanachart-dataproc-m	us-central1-a	10.128.0.4	146.148.102.244	SSH
thanachart-dataproc-w-0	us-central1-a	10.128.0.5	146.148.102.245	⋮
thanachart-dataproc-w-1	us-central1-a	10.128.0.6	146.148.102.246	⋮



```
Secure | https://ssh.cloud.google.com/projects/clouderacluster-164402/zones/us-central1-a/instances/thanachart-dataproc-m... ⓘ
Connected, host fingerprint: ssh-rsa 2048 3D:A2:32:CA:44:E3:41:31:12:6E:06:6B:C7:D3
:10:DD:CB:65:00:E3

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/*copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
thanachart@thanachart-dataproc-m:~$
```

Move Data from Google cloud storage to HDFS

```
$ gsutil cp gs://<YOUR-BUCKET>/pg2600.txt .
$ hadoop fs -mkdir /user/test
$ hadoop fs -put pg2600.txt /user/test
```

```
thanachart@thanachart-dataproc-m:~$ gsutil cp gs://imcinstitute/pg2600.txt .
Copying gs://imcinstitute/pg2600.txt...
/ [1 files] [ 3.1 MiB/ 3.1 MiB]
Operation completed over 1 objects/3.1 MiB.
```

Starting Pig Command Line

```
thanachart@thanachart-dataproc-m:~$ pig -x mapreduce
```

```
thanachart@thanachart-dataproc-m:~$ pig -x mapreduce

17/06/06 04:07:12 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
17/06/06 04:07:12 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
17/06/06 04:07:12 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2017-06-06 04:07:12,981 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.
0 (r: unknown) compiled May 18 2017, 01:29:04
2017-06-06 04:07:12,981 [main] INFO org.apache.pig.Main - Logging error messages t
o: /home/thanachart/pig_1496722032980.log
2017-06-06 04:07:13,004 [main] INFO org.apache.pig.impl.util.Utils - Default bootu
p file /home/thanachart/.pigbootup not found
2017-06-06 04:07:13,480 [main] INFO org.apache.hadoop.conf.Configuration.deprecati
on - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2017-06-06 04:07:13,481 [main] INFO org.apache.hadoop.conf.Configuration.deprecati
on - fs.default.name is deprecated. Instead, use fs.defaultFS
2017-06-06 04:07:13,481 [main] INFO org.apache.pig.backend.hadoop.executionengine.
HExecutionEngine - Connecting to hadoop file system at: hdfs://thanachart-datapro
m
2017-06-06 04:07:13,511 [main] INFO com.google.cloud.hadoop.fs.gcs.GoogleHadoopFil
eSystemBase - GHFS version: 1.6.1-hadoop2
2017-06-06 04:07:14,092 [main] INFO org.apache.pig.PigServer - Pig Script ID for t
he session: PIG-default-bac626fc-7afe-48cd-9f24-86adc3c08c4a
2017-06-06 04:07:14,092 [main] WARN org.apache.pig.PigServer - ATS is disabled sin
ce yarn.timeline-service.enabled set to false
grunt>
```

Writing a Pig Script for wordcount

Suggestion: Before run below codes, please make sure you that the file pg2600.txt still exists

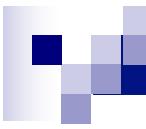
```
A = load '/user/test/*';
B = foreach A generate flatten(TOKENIZE((chararray)$0)) as word;
C = group B by word;
D = foreach C generate COUNT(B), group;
store D into '/user/test/wordcountPig';
```

```
2017-06-06 04:20:48,281 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at thanachart-dataproc-m/10.128.0.4:8032
2017-06-06 04:20:48,284 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-06-06 04:20:48,323 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at thanachart-dataproc-m/10.128.0.4:8032
2017-06-06 04:20:48,327 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-06-06 04:20:48,372 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at thanachart-dataproc-m/10.128.0.4:8032
2017-06-06 04:20:48,379 [main] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2017-06-06 04:20:48,424 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt> quit;
2017-06-06 04:21:21,248 [main] INFO org.apache.pig.Main - Pig script completed in 1 minute, 55 seconds and 805 milliseconds (115805 ms)
thanachart@thanachart-dataproc-m:~$
```

Viewing result in the HDFS

```
$ hadoop fs -ls /user/test/wordcountPig  
$ hadoop fs -cat /user/test/wordcountPig/part-r-00000
```

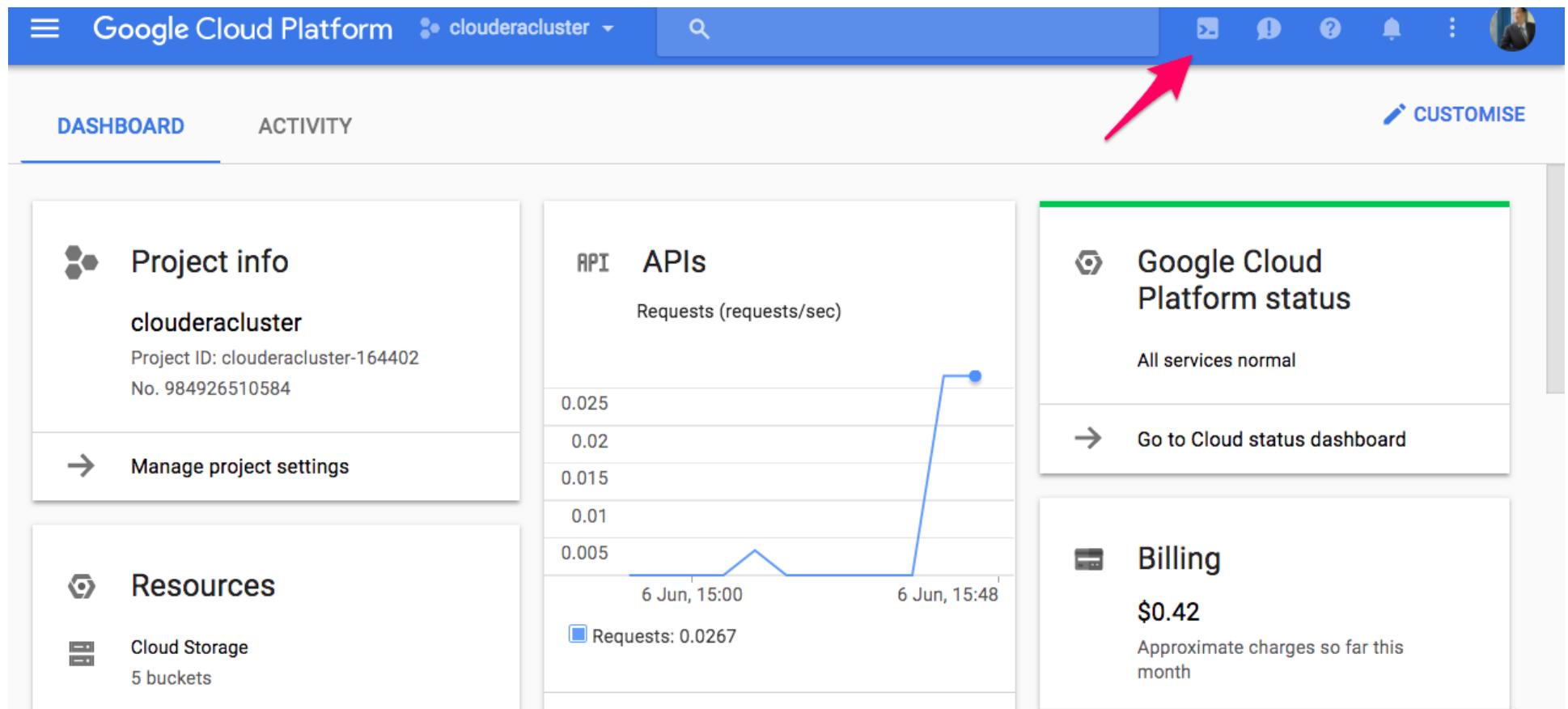
```
thanachart@thanachart-dataproc-m:~$ hadoop fs -ls /user/test/wordcountPig  
17/06/06 04:26:54 INFO gcs.GoogleHadoopFileSystemBase: GHFS version: 1.6.1-hadoop2  
Found 2 items  
-rw-r--r-- 2 thanachart hadoop 0 2017-06-06 04:20 /user/test/wordcountPig/_SUCCESS  
-rw-r--r-- 2 thanachart hadoop 353282 2017-06-06 04:20 /user/test/wordcountPig/part-r-00000  
thanachart@thanachart-dataproc-m:~$ hadoop fs -cat /user/test/wordcountPig/part-r-00000
```



Lab IV: Find top 10 best/worst airlines

Using Spark

Launch Dataproc using gcloud command: with Jupyter Notebook I) Launch Cloud Shell



The screenshot shows the Google Cloud Platform Dashboard for the project 'clouderacluster'. The dashboard includes sections for Project info, APIs, Resources, and Google Cloud Platform status. A red arrow points to the 'Cloud Shell' icon in the top right corner of the dashboard header.

Project info
clouderacluster
Project ID: clouderacluster-164402
No. 984926510584

API Requests (requests/sec)

Time	Requests (requests/sec)
6 Jun, 15:00	0.005
6 Jun, 15:08	0.005
6 Jun, 15:48	0.025

Requests: 0.0267

Resources
Cloud Storage
5 buckets

Google Cloud Platform status
All services normal
→ Go to Cloud status dashboard

Billing
\$0.42
Approximate charges so far this month

Upload a file to dataproc

```
$ wget https://s3.amazonaws.com/imcbucket/data/flights/2008.csv  
$ gsutil cp 2008.csv gs://<YOUR-BUCKET>/
```

```
thanachart@thanachart-dataproc-m:~$ wget https://s3.amazonaws.com/imcbucket/data/flights/2008.csv  
--2017-06-06 04:30:07-- https://s3.amazonaws.com/imcbucket/data/flights/2008.csv  
Resolving s3.amazonaws.com (s3.amazonaws.com) ... 54.231.81.91  
Connecting to s3.amazonaws.com (s3.amazonaws.com)|54.231.81.91|:443... connected.  
HTTP request sent, awaiting response... 200 OK  
Length: 689413344 (657M) [binary/octet-stream]  
Saving to: '2008.csv'  
  
2008.csv          100%[=====>] 657.48M 70.0MB/s   in 16s  
  
2017-06-06 04:30:23 (40.3 MB/s) - '2008.csv' saved [689413344/689413344]  
thanachart@thanachart-dataproc-m:~$ gsutil cp 2008.csv gs://imcinstiute/
```

Launch Dataproc using gcloud command with Jupyter Notebook

II) Type the following command

- gcloud dataproc clusters create datascience \
--zone us-central1-a \
--master-machine-type=n1-standard-2 \
--worker-machine-type=n1-standard-2 \
--initialization-actions \
gs://dataproc-initialization-actions/jupyter/jupyter.sh

```
thanachart@clouderacluster-164402:~$      --worker-machine-type=n1-standard-2 \  
>      --initialization-actions \  
>          gs://dataproc-initialization-actions/jupyter/jupyter.sh  
-bash: --worker-machine-type=n1-standard-2: command not found  
thanachart@clouderacluster-164402:~$ gcloud dataproc clusters create datascience      --zone us-central1-a      --master-machine-type  
=n1-standard-2      --worker-machine-type=n1-standard-2 \  
>      --initialization-actions \  
>          gs://dataproc-initialization-actions/jupyter/jupyter.sh  
Waiting on operation [projects/clouderacluster-164402/regions/global/operations/2ad165f0-c8c1-44cb-ab61-2d13f64077ce].  
Waiting for cluster creation operation...done.  
Created [https://dataproc.googleapis.com/v1/projects/clouderacluster-164402/regions/global/clusters/datascience].  
thanachart@clouderacluster-164402:~$
```

Google Cloud Platform

clouderacluster

Cloud Dataproc

Clusters

CREATE CLUSTER

REFRESH

DELETE

Clusters

Search clusters, press ?

Name	Zone	Total worker nodes	Cloud Storage staging bucket	Created	Status
datascience	us-central1-a	2	dataproc-36c094d4-2796-4b03-85c9-7e9a90a99654-us	6 Jun 2017, 15:57:07	Running

A red arrow points to the 'datascience' cluster row.

Google Cloud Platform

clouderacluster

Compute Engine

VM instances

SHOW INFO PANEL

VM instances

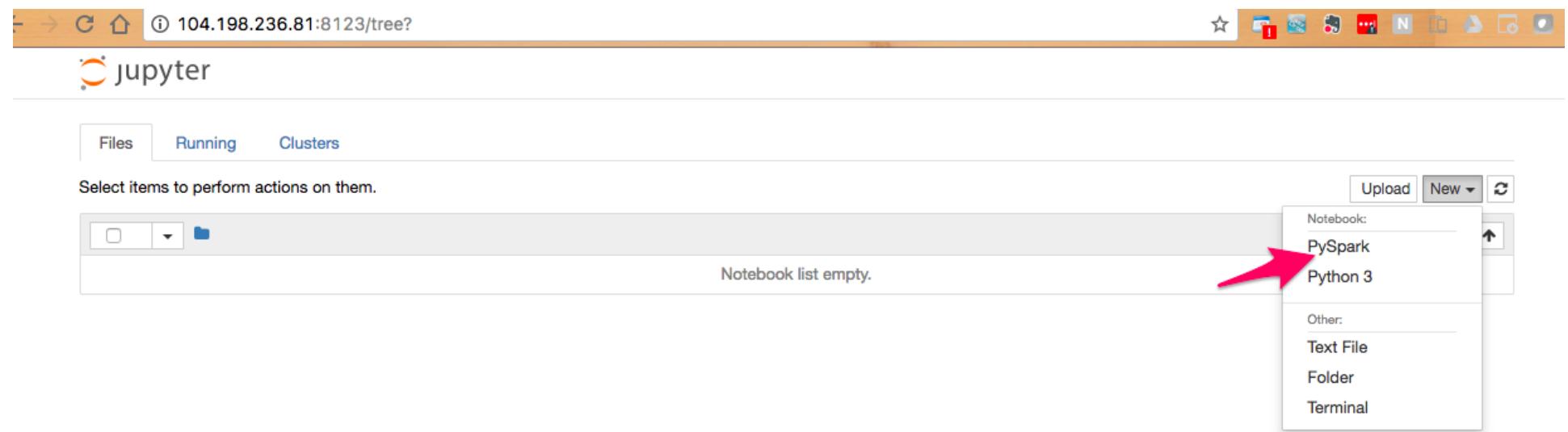
Filter by label or name

Columns ▾

Name	Zone	Recommendation	Internal IP	External IP	Connect
datascience-m	us-central1-a		10.128.0.2	104.198.236.81	SSH ▾
datascience-w-0	us-central1-a		10.128.0.3	130.211.171.250	SSH ▾
datascience-w-1	us-central1-a		10.128.0.4	104.155.180.252	SSH ▾

Lunch the Jupyter notebook

<<public ip>> :8123



Spark Program : Navigating Flight Data

```
>>> airline = sc.textFile("gs://<YOUR-BUCKET>/2008.csv")
>>> airline.take(2)
```

```
In [1]: airline = sc.textFile("gs://imcinstiute/2008.csv")
```

```
In [2]: airline.take(2)
```

```
Out[2]: ['Year,Month,DayOfMonth,DayOfWeek,DepTime,CRSDepTime,ArrTime,CRSArrTime,UniqueCarrier,FlightNum,TailNum,ActualElapsedTime,CRSElapsedTime,AirTime,ArrDelay,DepDelay,Origin,Dest,Distance,TaxiIn,TaxiOut,Cancelled,CancellationCode,Diverted,CarrierDelay,WeatherDelay,NASDelay,SecurityDelay,LateAircraftDelay',
'2008,1,3,4,2003,1955,2211,2225,WN,335,N712SW,128,150,116,-14,8,IAD,TPA,810,4,8,0,,0,NA,NA,NA,NA,NA']
```

Spark Program : Preparing Data

```
>>> header_line = airline.first()
>>> header_list = header_line.split(',')
>>> airline_no_header = airline.filter(lambda row: row != header_line)
>>> airline_no_header.first()
>>> def make_row(row):
...     row_list = row.split(',')
...     d = dict(zip(header_list, row_list))
...     return d
...
>>> airline_rows = airline_no_header.map(make_row)
>>> airline_rows.take(5)
```

```
In [3]: header_line = airline.first()
header_list = header_line.split(',')
airline_no_header = airline.filter(lambda row: row != header_line)
airline_no_header.first()
def make_row(row):
    row_list = row.split(',')
    d = dict(zip(header_list, row_list))
    return d
airline_rows = airline_no_header.map(make_row)
airline_rows.take(5)
```

```
Out[3]: [ {'ActualElapsedTime': '128',
   'AirTime': '116',
   'ArrDelay': '-14',
   'ArrTime': '2211',
   'CRSArrTime': '2225',
   'CRSDepTime': '1955',
   'CRSElapsedTime': '150',
   'CancellationCode': '',
   'Cancelled': '0',
   'CarrierDelay': 'NA',
   'DayOfWeek': '4',
   'DayofMonth': '3',
   'DepDelay': '8',
   'DepTime': '2003',
   'Dest': 'TPA',
   'Distance': '810',
   'Diverted': '0',
   'FlightNum': '335',
   'LateAircraftDelay': 'NA',
```

Spark Program : Define convert_float function

```
>>> def convert_float(value):
...     try:
...         x = float(value)
...         return x
...     except ValueError:
...         return 0
...
>>>
```

```
In [4]: def convert_float(value):
         try:
             x = float(value)
             return x
         except ValueError:
             return 0
```

Spark Program : Finding best/worst airline

```
>>> carrier_rdd = airline_rows.map(lambda  
row: (row['UniqueCarrier'],convert_float(row['ArrDelay'])))  
>>> carrier_rdd.take(2)
```

```
In [5]: carrier_rdd = airline_rows.map(lambda row:(row['UniqueCarrier'],convert_float(row['ArrDelay'])))  
carrier_rdd.take(2)  
  
Out[5]: [('WN', -14.0), ('WN', 2.0)]
```

Spark Program : Finding best/worst airlines

```
>>> mean_delays_dest =  
carrier_rdd.groupByKey().mapValues(lambda  
sum(delays)/len(delays))  
>>> mean_delays_dest.sortBy(lambda t:t[1],  
ascending=True).take(10)  
>>> mean_delays_dest.sortBy(lambda t:t[1],  
ascending=False).take(10)
```

```
In [11]: mean_delays_dest = carrier_rdd.groupByKey().mapValues(lambda delays: sum(delays.data)/len(delays.data))  
  
In [12]: mean_delays_dest.sortBy(lambda t:t[1], ascending=True).take(10)  
  
Out[12]: [(u'AO', -2.8708974358974357),  
(u'HA', 1.2518519716624075),  
(u'US', 2.800998260539828),  
(u'98', 3.987490846961191),  
(u'AS', 4.721360405553864),  
(u'WN', 5.115703380225903),  
(u'F9', 6.084135669681085),  
(u'OO', 6.43893863978179),  
(u'NW', 7.293465879672776),  
(u'DL', 7.716164635751918)]  
  
In [13]: mean_delays_dest.sortBy(lambda t:t[1], ascending=False).take(10)  
  
Out[13]: [(u'AA', 12.202853434950445),  
(u'OH', 11.404110178283158),  
(u'YV', 11.322566979170753),  
(u'UA', 11.001550560048052),  
(u'B6', 10.859381613638567),  
(u'CO', 10.809820575966226),  
(u'XE', 10.320298523403915),  
(u'EV', 10.00033146217589),  
(u'MQ', 9.496970610952266),  
(u'FL', 8.988157472371256)]
```

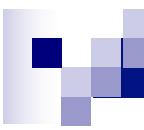
The screenshot shows the Google Cloud Platform Cloud Dataproc Clusters page. The left sidebar has 'Cloud Dataproc' selected under 'Clusters'. The main area shows a table of clusters. A red arrow points to the 'DELETE' button in the top right of the table header. The table data is as follows:

Name	Zone	Total worker nodes	Cloud Storage staging bucket	Created	Status
datascience	us-central1-a	2	datapro-36c094d4-2796-4b03-85c9-7e9a90a99654-us	6 Jun 2017, 15:57:07	Running



Google BigQuery

Google BigQuery is a fully-managed and cloud-based interactive query service for massive datasets. It's the externalization of Dremel, one of Google's core technologies



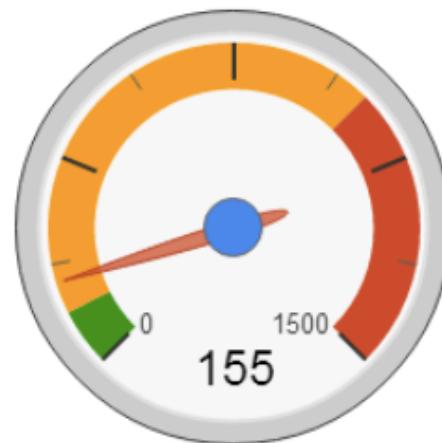
Why Google BigQuery ?



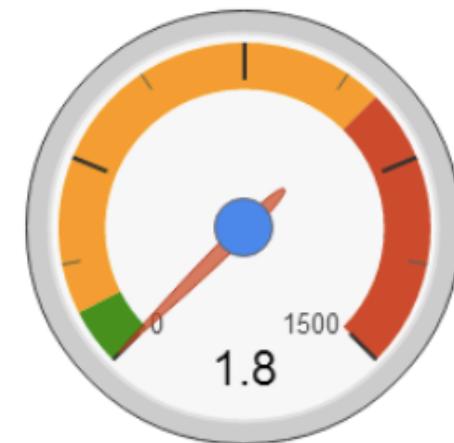
= 1.4 TB



Hadoop
(with Hive)



Amazon
Redshift



Google
BigQuery

On an average its within 8-10 seconds !!



Inside Google BigQuery

- BigQuery is based on **Dremel**, a technology pioneered by Google & extensively used within.
- It used **Columnar storage & multi-level execution trees** to achieve interactive performance for queries against multi-terabyte datasets.
- BigQuery's performance advantage comes from its parallel processing architecture.
- The query is processed by thousands of servers in a multi-level execution tree structure, with the final results aggregated at the root. BigQuery stores the data in a columnar format so that only data from the columns being queried are real.
- All this & more is now available as a publicly available service for any business or developer to use. This release made it possible for those outside of Google to utilize the power of Dremel for their Big Data processing requirements.

Features & Components

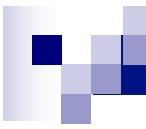
Features:

- Web GUI for BigQuery
- Affordable
- Run in Background
- Easy Data Importation
- Flexible (Addition of Columns, Native Support For Timestamp Type Of Data)
- REST API Support
- More than just Standard SQL

Components:

- Project
- Tables
- DataSets
- Jobs





Lab V : Query Bike data

Using Google BigQuery



OPEN DATA GIFT STORE ABOUT RESOURCES APP CONTACT LOGIN

SIGN UP HOW IT WORKS SUGGEST A STATION STATION MAP PRICING f t i

OPEN DATA

Here you'll find Bay Area Bike Share's trip data for public use. So whether you're a designer, developer, or just plain curious, feel free to download it and bring it to life!

THE DATA

Each trip is anonymized and includes:

- Bike number
- Trip start day and time
- Trip end day and time

YEAR 1 DATA

(August 2013 - August 2014)

YEAR 2 DATA

(September 2014 - August 2015)



Preparing a bike data from Cloud shell

```
$ wget https://s3.amazonaws.com/babs-open-data/babs_open_data_year_1.zip  
$ unzip babs_open_data_year_1.zip  
$ cd 201402_babs_open_data/  
$ gsutil cp 201402_trip_data.csv gs://<YOUR-BUCKET>/
```

```
thanachart@clouderacluster-164402:~$ ls  
201402_babs_open_data  babs_open_data_year_1.zip  pg2600.txt          training-data-analyst  
201408_babs_open_data  data-science-on-gcp        README-cloudshell.txt  
thanachart@clouderacluster-164402:~$ cd 201402_babs_open_data/  
thanachart@clouderacluster-164402:~/201402_babs_open_data$ gsutil cp 201402_trip_data.csv gs://imcinstitute/  
Copying file://201402_trip_data.csv [Content-Type=text/csv]...  
\ [1 files] [ 16.4 MiB / 16.4 MiB]  
Operation completed over 1 objects/16.4 MiB.  
thanachart@clouderacluster-164402:~/201402_babs_open_data$
```

Launch Google BigQuery

The screenshot shows the Google Cloud Platform Home page. On the left, there's a sidebar with sections for Home, Endpoints, and BIG DATA. Under BIG DATA, there are links for BigQuery, Pub/Sub, Dataproc, Dataflow, ML Engine, and Genomics. A red arrow points to the BigQuery link. The main content area features a chart titled "API APIs Requests (requests/sec)" showing a line graph with a single data series labeled "Requests: 0.0333". The x-axis shows two time points: "6 Jun, 17:30" and "6 Jun, 18:24". Below the chart is a button labeled "Go to APIs overview". To the right of the chart are two cards: "Google Cloud Platform status" (All services normal) and "Billing" (\$0.85, Approximate charges so far this month).

≡ Google Cloud Platform clouderacluster ⚙️

Home

Pins appear here ⚙️

Endpoints

BIG DATA

BigQuery

Pub/Sub

Dataproc

Dataflow

ML Engine

Genomics

API APIs Requests (requests/sec)

0.08
0.06
0.04
0.02

6 Jun, 17:30 6 Jun, 18:24

Requests: 0.0333

Go to APIs overview

Google Cloud Platform status

All services normal

→ Go to Cloud status dashboard

Billing

\$0.85

Approximate charges so far this month

→ View detailed charges

Create new dataset

The screenshot shows the Google BigQuery web interface. On the left, there's a sidebar with 'COMPOSE QUERY' at the top, followed by 'Query History' and 'Job History'. Below that is a search bar labeled 'Filter by ID or label' with a question mark icon. A red arrow points to a dropdown menu that appears when you click the label field; it contains options: 'Create new dataset' (highlighted with a red arrow), 'Switch to project', and 'Refresh'. The main area is titled 'Queries' with tabs for 'Query History', 'Saved Queries', and 'Project Queries'. It shows a list of seven queries from May 14, all using the same SQL statement: 'SELECT commit, author, committer, repo_name FROM `bigquery-public-data.github_repos.commits` LIMIT 1000'. Each query has an 'Open Query' button and a date ('May 14'). At the bottom, a modal dialog is open for 'Create Dataset', asking for a 'Dataset ID' (set to 'demo_dataset') and 'Data location' (set to '(unspecified)'). It also has 'Data expiration' options: 'Never' (selected) or 'In [] days.'.

Create new table from the newly created dataset

The screenshot shows the Google BigQuery web interface. On the left, there's a sidebar with a 'COMPOSE QUERY' button, 'Query History', 'Job History', a 'Filter by ID or label' input, and a dropdown menu for 'clouderacluster'. Below that is a 'Public Datasets' section listing various datasets like 'bigquery-public-data:hacker_news', 'bigquery-public-data:noaa_gsod', etc. A red arrow points to the 'Share dataset' option in a context menu that appears when hovering over the 'clouderacluster' dropdown. The main area is titled 'Queries' and shows a list of 7 queries. The first query is highlighted with a red warning icon and the SQL command: 'SELECT commit,author, committer, repo_name FROM `bigquery-public-data.github_repos.commits` LIMIT 1000'. To the right of each query are 'Open Query' and 'May 14' buttons.

Query	Action	Date
SELECT commit,author, committer, repo_name FROM `bigquery-public-data.github_repos.commits` LIMIT 1000	Open Query	May 14
SELECT commit,author, committer, repo_name FROM `bigquery-public-data.github_repos.commits` LIMIT 1000	Open Query	May 14
SELECT commit,author, committer, repo_name FROM `bigquery-public-data.github_repos.commits` LIMIT 1000	Open Query	May 14
SELECT commit,author, committer, repo_name FROM `bigquery-public-data.github_repos.commits` LIMIT 1000	Open Query	May 14
SELECT commit,author, committer, repo_name FROM `bigquery-public-data.github_repos.commits` LIMIT 1000	Open Query	May 14
SELECT commit,author, committer, repo_name FROM `bigquery-public-data.github_repos.commits` LIMIT 1000	Open Query	May 14
SELECT commit,author, committer, repo_name FROM `bigquery-public-data.github_repos.commits` LIMIT 1000	Open Query	May 14

Select Source Data & Destination Table

Source Data

Create from source Create empty table

Repeat job [Select Previous Job](#) [?](#)

Location [Google Cloud Storage](#) [?](#)

File format [CSV](#) [?](#) [View Files](#)

Destination Table

Table name [demo_dataset](#) [?](#) [?](#)

Table type [Native table](#) [?](#)

Select Schema as follow

Schema Automatically detect [?](#)

Name	Type	Mode
	STRING	NULLABLE

[Add Field](#) [Edit as Text](#)

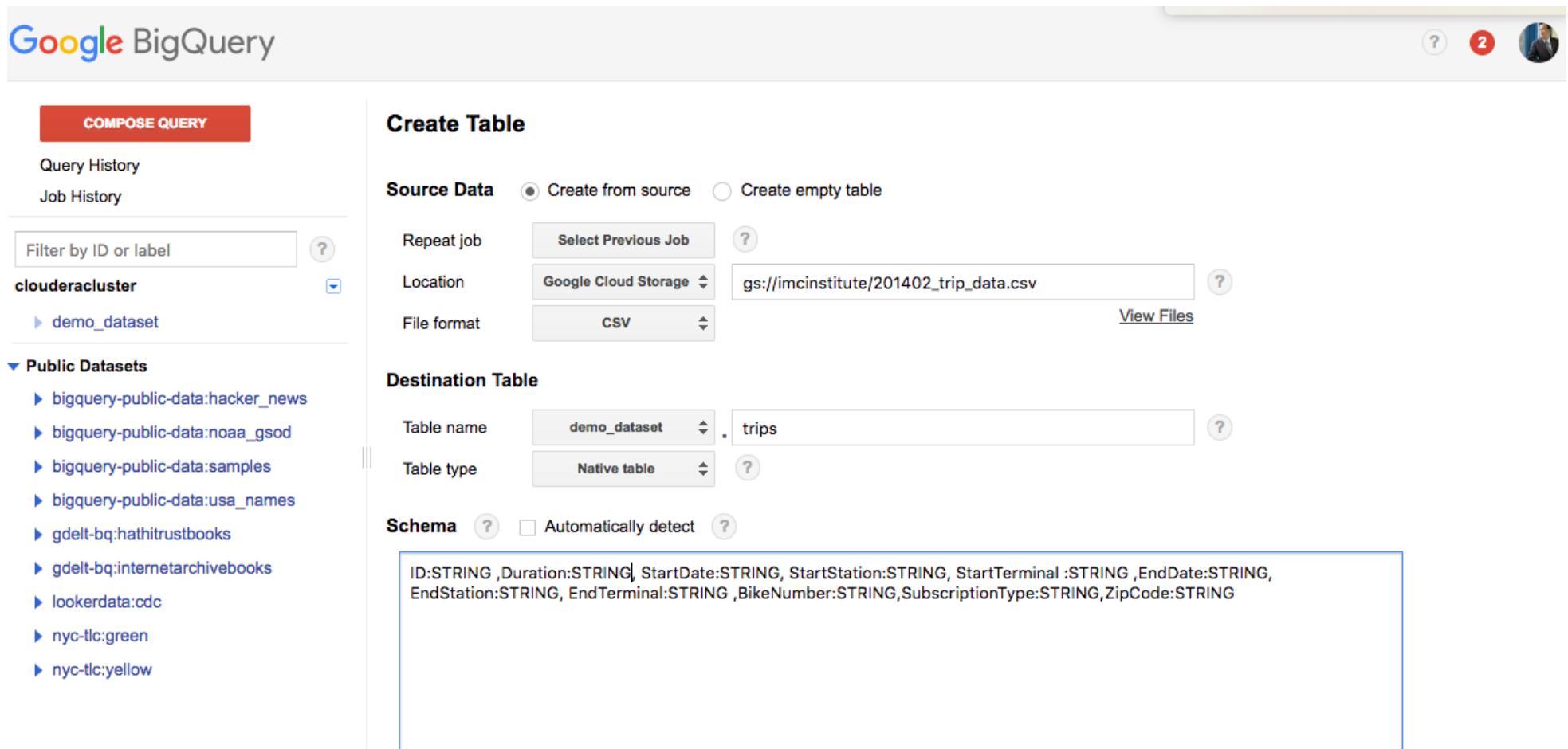
Schema [?](#) Automatically detect [?](#)

```
ID:STRING ,Duration:STRING, StartDate:STRING, StartStation:STRING, StartTerminal :STRING ,EndDate:STRING,  
EndStation:STRING, EndTerminal:STRING ,BikeNumber:STRING,SubscriptionType:STRING,ZipCode:STRING
```

id:STRING,

Duration:STRING,StartDate:STRING,StartStation:STRING,StartTerminal:STRING,EndDate:
STRING,EndStation:STRING,EndTerminal:STRING,BikeNumber:STRING,SubsriptionType:STRING,
ZipCode:STRING

Click Create Table



The screenshot shows the Google BigQuery web interface with the title "Create Table".

Left Sidebar:

- COMPOSE QUERY** button (red background)
- Query History
- Job History
- Filter by ID or label input field
- clouderacluster dataset dropdown:
 - demo_dataset
- Public Datasets** section:
 - bigquery-public-data:hacker_news
 - bigquery-public-data:noaa_gsod
 - bigquery-public-data:samples
 - bigquery-public-data:usa_names
 - gdelt-bq:hathitrustbooks
 - gdelt-bq:internetarchivebooks
 - lookerdata:cdc
 - nyc-tlc:green
 - nyc-tlc:yellow

Create Table Form:

Source Data: Create from source Create empty table

Repeat job: [Select Previous Job](#) [?](#)

Location: Google Cloud Storage [?](#) [View Files](#)

File format: CSV [?](#)

Destination Table:

Table name: demo_dataset [?](#) . trips [?](#)

Table type: Native table [?](#)

Schema: [?](#) Automatically detect [?](#)

```
ID:STRING ,Duration:STRING, StartDate:STRING, StartStation:STRING, StartTerminal :STRING ,EndDate:STRING,  
EndStation:STRING, EndTerminal:STRING ,BikeNumber:STRING,SubscriptionType:STRING,ZipCode:STRING
```

View Trips Table

The screenshot shows the Google BigQuery interface. On the left, there's a sidebar with a red "COMPOSE QUERY" button, "Query History", "Job History", a search bar, and a tree view of datasets and tables. The tree view shows "clouderacluster" expanded, with "demo_dataset" and "trips" under it. Below that, "Public Datasets" are listed. On the right, the main area is titled "Table Details: trips". It has tabs for "Schema", "Details", and "Preview". The "Schema" tab is selected, showing a table with 11 columns:

ID	STRING	NULLABLE	Describe this field...
Duration	STRING	NULLABLE	Describe this field...
StartDate	STRING	NULLABLE	Describe this field...
StartStation	STRING	NULLABLE	Describe this field...
StartTerminal	STRING	NULLABLE	Describe this field...
EndDate	STRING	NULLABLE	Describe this field...
EndStation	STRING	NULLABLE	Describe this field...
EndTerminal	STRING	NULLABLE	Describe this field...
BikeNumber	STRING	NULLABLE	Describe this field...
SubscriptionType	STRING	NULLABLE	Describe this field...
ZipCode	STRING	NULLABLE	Describe this field...

At the bottom of the schema table, there's a "Add New Fields" button. Above the schema table, there are buttons for "Query Table", "Copy Table", "Export Table", and "Delete Table". The top right corner shows a user profile icon with a red notification badge.

Query Trips Table

Google BigQuery

COMPOSE QUERY

Query History
Job History

Filter by ID or label ?

clouderacluster ▾
demo_dataset
trips

Public Datasets
▶ bigquery-public-data:hacker_news
▶ bigquery-public-data:noaa_gsod
▶ bigquery-public-data:samples
▶ bigquery-public-data:usa_names
▶ gdelt-bq:hathitrustbooks
▶ gdelt-bq:internetarchivebooks
▶ lookerdata:cdc
▶ nyc-tlc:green
▶ nyc-tlc:yellow

Table Details: trips

Schema Details Preview

ID	STRING	NULLABLE	Describe this field...
Duration	STRING	NULLABLE	Describe this field...
StartDate	STRING	NULLABLE	Describe this field...
StartStation	STRING	NULLABLE	Describe this field...
StartTerminal	STRING	NULLABLE	Describe this field...
EndDate	STRING	NULLABLE	Describe this field...
EndStation	STRING	NULLABLE	Describe this field...
EndTerminal	STRING	NULLABLE	Describe this field...
BikeNumber	STRING	NULLABLE	Describe this field...
SubscriptionType	STRING	NULLABLE	Describe this field...
ZipCode	STRING	NULLABLE	Describe this field...

Add New Fields

Query Table Copy Table Export Table Delete Table



Find the top 10 most popular start stations based on the trip data

```
SELECT startTerminal, startStation, COUNT(1) AS count FROM
[ <<Google Cloud Shell>> :demo_dataset.trips]
GROUP BY startTerminal, startStation ORDER BY count DESC LIMIT 10
```

The screenshot shows the Google BigQuery web interface. On the left, there's a sidebar with 'COMPOSE QUERY' and links to 'Query History', 'Job History', and datasets like 'clouderacluster' and 'Public Datasets'. The main area is titled 'New Query' and contains the SQL code shown above. Below the code, there are buttons for 'RUN QUERY', 'Save Query', 'Save View', 'Format Query', and 'Show Options'. A message says 'Query complete (3.6s elapsed, 4.10 MB processed)' with a green checkmark. At the bottom, there's a table with columns 'Row', 'startTerminal', 'startStation', and 'count', showing the top 6 results. There are also buttons to 'Download as CSV', 'Download as JSON', 'Save as Table', and 'Save to Google Sheets'. The bottom navigation includes 'Table' and 'JSON' tabs, and links for 'First', 'Prev', 'Rows 1 - 6 of 10', 'Next', and 'Last'.

Row	startTerminal	startStation	count
1	70	San Francisco Caltrain (Townsend at 4th)	9838
2	50	Harry Bridges Plaza (Ferry Building)	7343
3	60	Embarcadero at Sansome	6545
4	77	Market at Sansome	5922
5	55	Temporary Transbay Terminal (Howard at Beale)	5113
6	76	Market at 4th	5030

83



Google Cloud DataLab

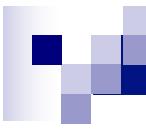
Cloud Datalab is a powerful interactive tool created to explore, analyze, transform and visualize data and build machine learning models on Google Cloud Platform. It runs on Google Compute Engine and connects to multiple cloud services easily so you can focus on your data science tasks.



Integrated & Open Source

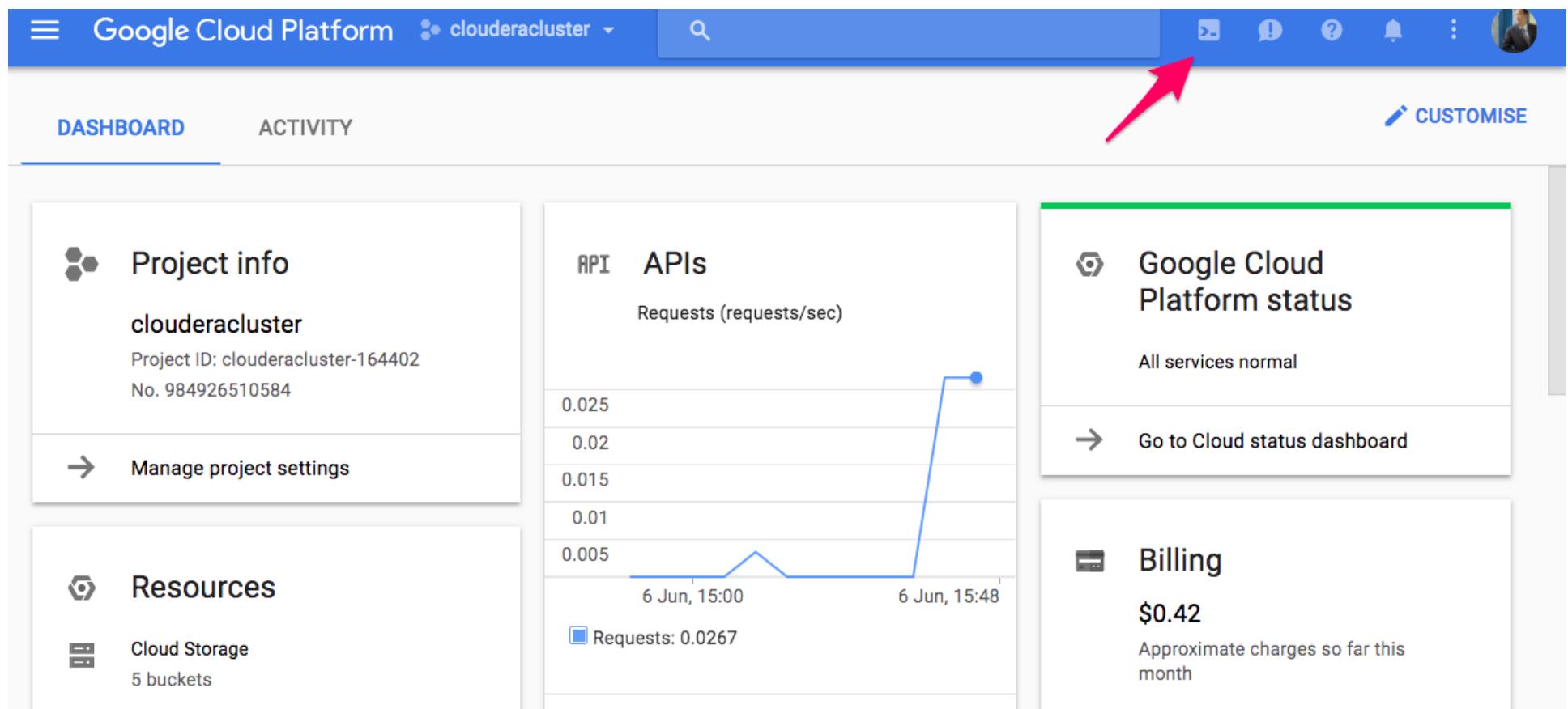
Cloud Datalab is **built on Jupyter (formerly IPython)**, which boasts a thriving ecosystem of modules and a robust knowledge base. Cloud Datalab **enables analysis of your data on Google BigQuery, Cloud Machine Learning Engine, Google Compute Engine, and Google Cloud Storage** using Python, SQL, and JavaScript (for BigQuery user-defined functions).

- Cloud Datalab simplifies data processing with Cloud BigQuery, Cloud Machine Learning Engine, Cloud Storage.
- Supports Python, SQL, and JavaScript
- Combines code, documentation, results, and visualizations together in an intuitive notebook format.
- Supports TensorFlow-based deep ML models in addition to scikit-learn.
- Based on Jupyter (formerly IPython) so you can use a large number of existing packages for statistics, machine learning etc..



Lab VI: Launch Cloud DataLab

Launch Cloud Shell



The screenshot shows the Google Cloud Platform (GCP) dashboard for the project "clouderacluster". The dashboard includes sections for Project info, APIs, Resources, and Google Cloud Platform status.

- Project info:** clouderacluster, Project ID: clouderacluster-164402, No. 984926510584. Includes a link to "Manage project settings".
- APIs:** Requests (requests/sec) chart showing a sharp increase from ~0.005 to ~0.025 between 6 Jun, 15:00 and 6 Jun, 15:48. The chart also displays a legend entry: Requests: 0.0267.
- Resources:** Cloud Storage with 5 buckets.
- Google Cloud Platform status:** All services normal. Includes a link to "Go to Cloud status dashboard".
- Billing:** \$0.42, Approximate charges so far this month.

A red arrow points to the top right corner of the dashboard, where there are several icons: a gear for settings, a search bar, a user profile, and notification icons.

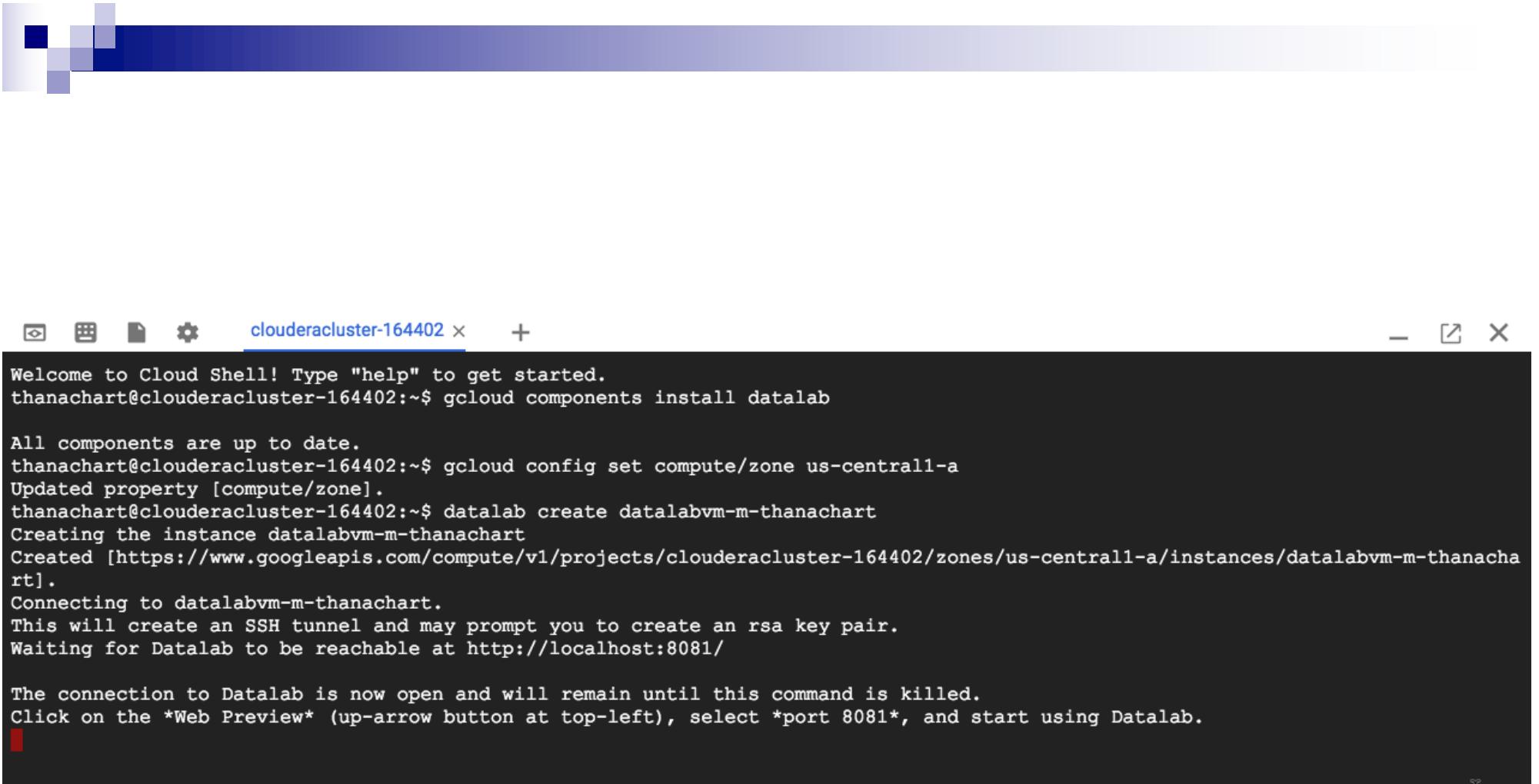
Launch DataLab



Docker on Compute Engine
Use GCP for processing
Notebooks on GCE disk
CloudShell ssh tunnel

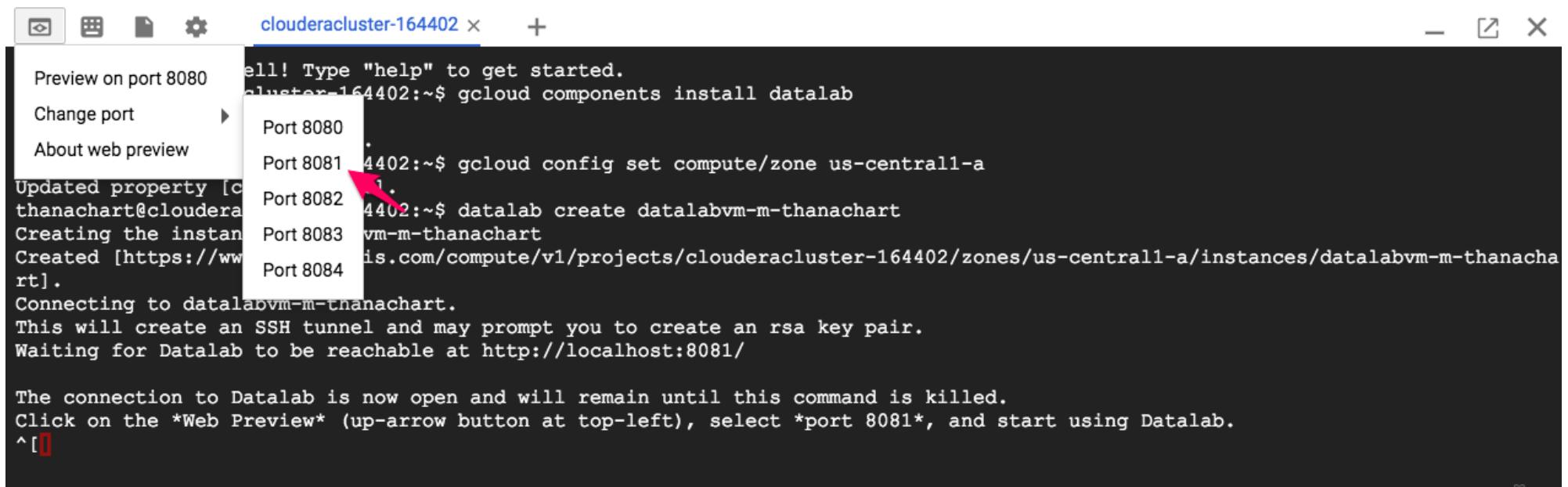
```
$ gcloud components install datalab  
$ gcloud config set compute/zone <ZONE>  
$ datalab create datalabvm-<USER>
```

```
#This will take several minutes. Wait for the message  
"You can now connect to Datalab at  
http://localhost:8081/"
```



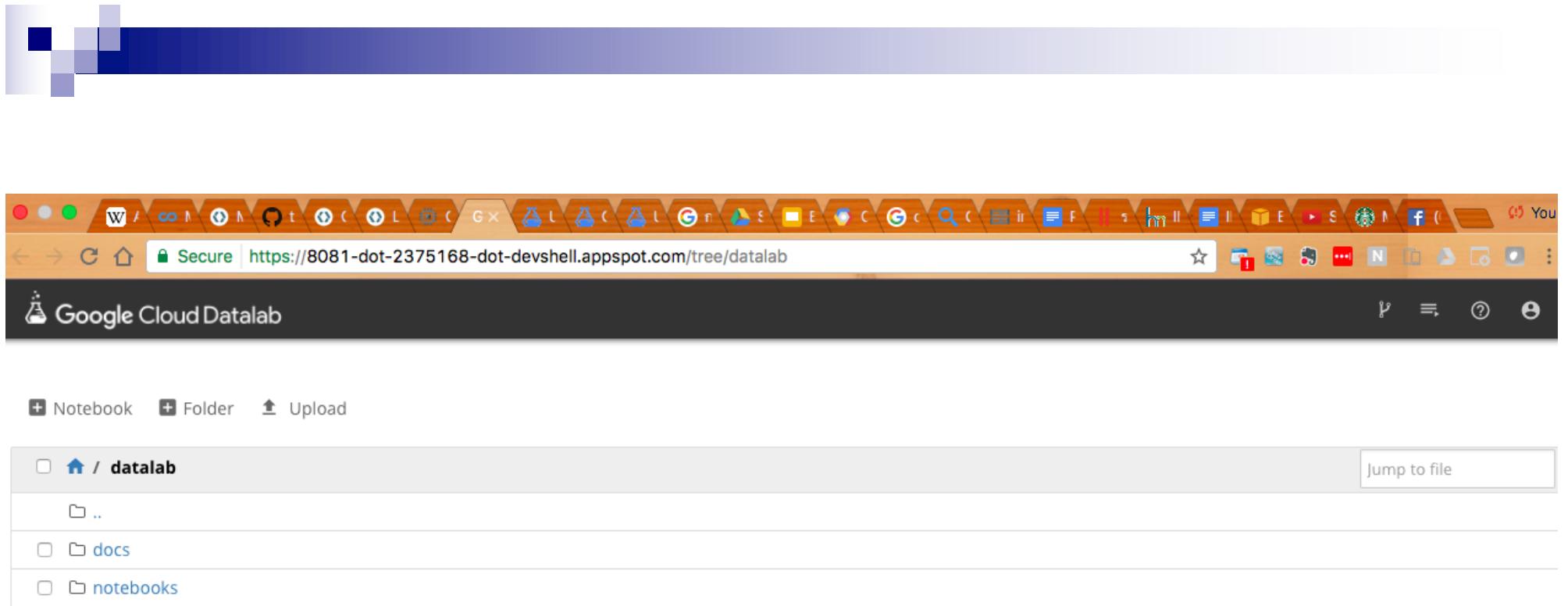
```
Welcome to Cloud Shell! Type "help" to get started.  
thanachart@clouderacluster-164402:~$ gcloud components install databricks  
  
All components are up to date.  
thanachart@clouderacluster-164402:~$ gcloud config set compute/zone us-central1-a  
Updated property [compute/zone].  
thanachart@clouderacluster-164402:~$ databricks create databricksvm-m-thanachart  
Creating the instance databricksvm-m-thanachart  
Created [https://www.googleapis.com/compute/v1/projects/clouderacluster-164402/zones/us-central1-a/instances/databricksvm-m-thanachart].  
Connecting to databricksvm-m-thanachart.  
This will create an SSH tunnel and may prompt you to create an rsa key pair.  
Waiting for Databricks to be reachable at http://localhost:8081/  
  
The connection to Databricks is now open and will remain until this command is killed.  
Click on the *Web Preview* (up-arrow button at top-left), select *port 8081*, and start using Databricks.
```

Launch Data Lab port 8081



The screenshot shows a terminal window titled "clouderacluster-164402". The window has a menu bar with icons for preview, change port, and about web preview. A dropdown menu is open under "About web preview" with options: "Port 8080", "Port 8081" (which is highlighted with a red arrow), "Port 8082", "Port 8083", and "Port 8084". The main terminal area displays the following command-line session:

```
hell! Type "help" to get started.  
cluster-164402:~$ gcloud components install datalab  
Port 8080  
. .  
Port 8081 4402:~$ gcloud config set compute/zone us-central1-a  
Port 8082 4402:~$ .  
Port 8083 4402:~$ datalab create datalabvm-m-thanachart  
Port 8084 vm-m-thanachart  
Created [https://www.googleapis.com/compute/v1/projects/clouderacluster-164402/zones/us-central1-a/instances/datalabvm-m-thanachart].  
Connecting to datalabvm-m-thanachart.  
This will create an SSH tunnel and may prompt you to create an rsa key pair.  
Waiting for Datalab to be reachable at http://localhost:8081/  
  
The connection to Datalab is now open and will remain until this command is killed.  
Click on the *Web Preview* (up-arrow button at top-left), select *port 8081*, and start using Datalab.  
^[[
```



Viewing running VM instance

The screenshot shows the Google Cloud Platform Compute Engine interface. The left sidebar is titled 'Compute Engine' and has a 'VM instances' section selected, indicated by a blue background. The main area is titled 'VM instances' and shows a table of running instances. A red arrow points to the first row of the table, which contains the following information:

Name	Zone	Recommendation	Internal IP	External IP	Connect
<input checked="" type="checkbox"/> datalabvm-m-thanachart	us-central1-a		10.128.0.2	146.148.97.217	SSH

Reconnecting to Datalab

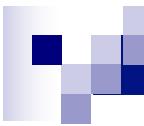
If the datalab command terminates (for example, if the laptop goes to sleep), reconnect to the VM using:

```
datalab connect databavm-<USER>
```

Cleanup (important)

When you are done using the Datalab VM, delete the instance using:

```
datalab delete databavm-<USER>
```



Google Machine Learning APIs

Google Cloud Machine Learning provides modern machine learning services, with pre-trained models and a service to generate your own tailored models. Major Google applications use Cloud Machine Learning, including Photos (image search), the Google app (voice search), Translate, and Inbox (Smart Reply).

Ready to use Machine Learning APIS



Cloud
Vision API



Cloud
Speech API



Cloud
Jobs API



Cloud
Translation
API



Cloud
Natural
Language API



Cloud
Video
Intelligence API



Coming
soon

Cloud Vision APIs

- Label / Face Detection
- Image Attributes
- Landmark Detection
- Logo Detection
- OCR Detection
- Safe Search Detection



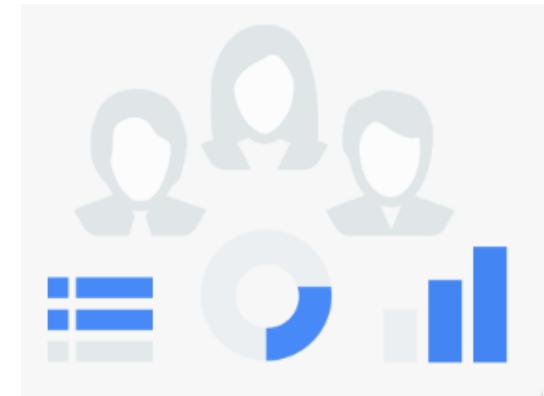
Cloud Speech API

- Audio → Text
- 80 languages & variants
- Audio uploaded on request and integrated GCS
- Result Streaming supported
- Noisy Environments & Context



Cloud Natural Language API

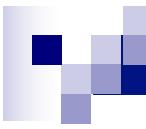
- Syntax Analysis
- Entity Recognition
- Sentiment Analysis
- Use in combination with Cloud Speech and Vision API (OCR)



Cloud Translation API

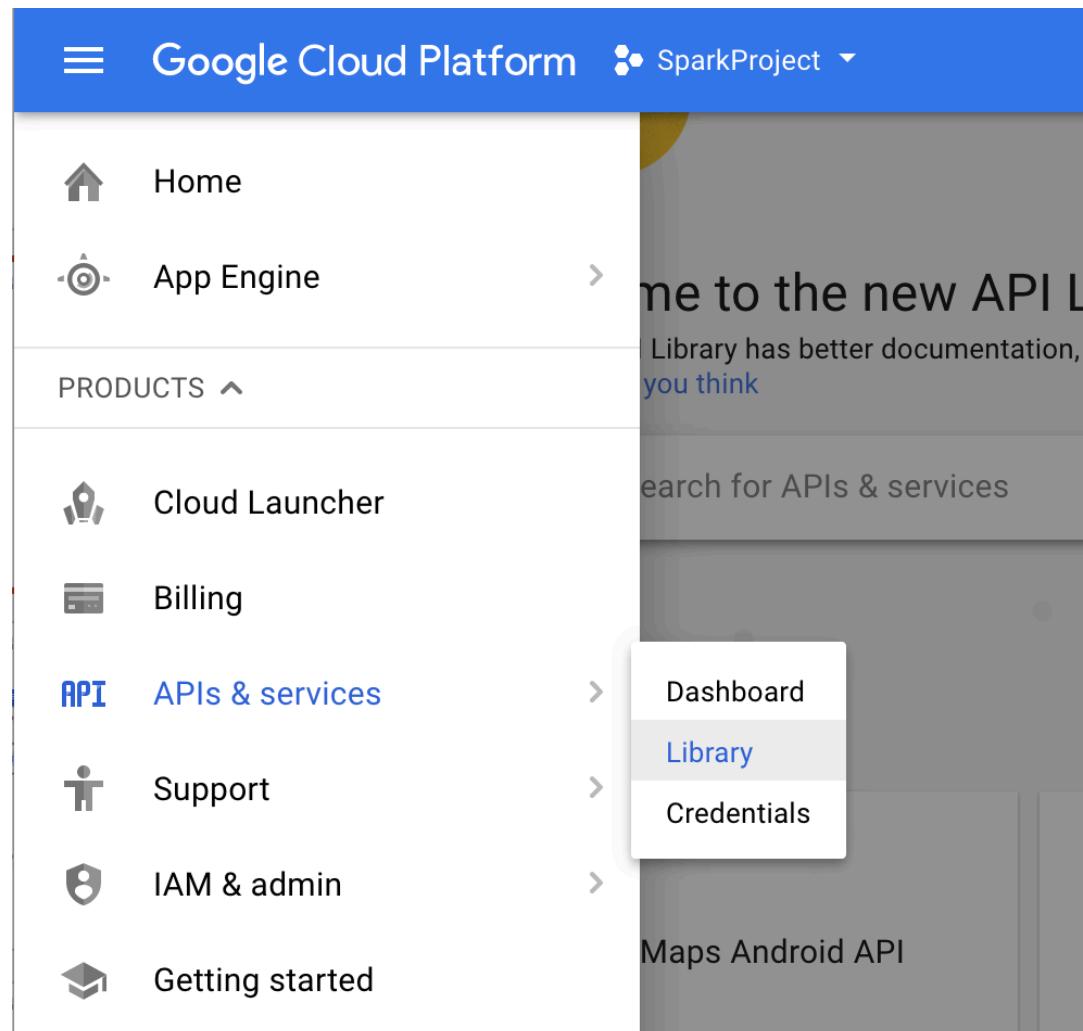
- Translate Many Languages
- Language Detection
- Supports more than 100 languages
- Learning continuously via Logs Analysis and Human Translation



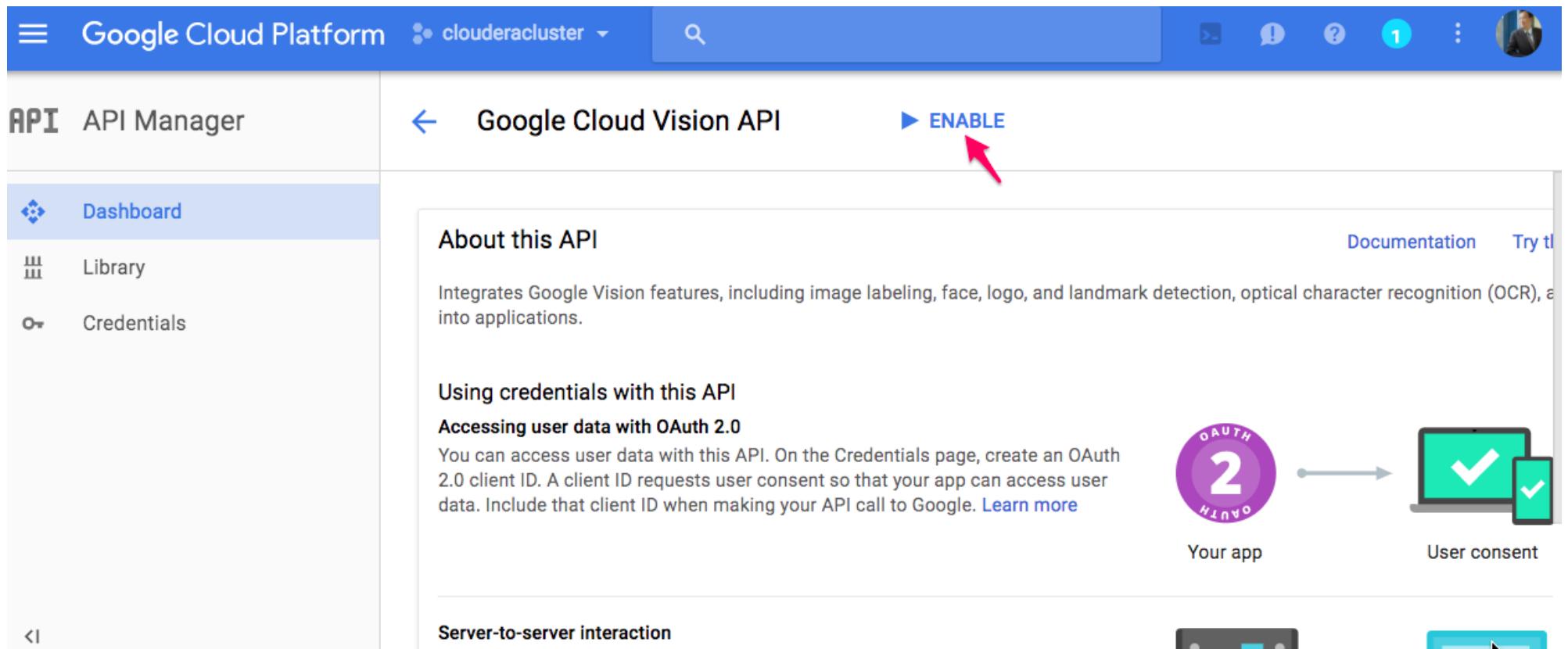


Lab VII: Using ML APIs

Select APIs & services -> Library



Enable Vision API



The screenshot shows the Google Cloud Platform API Manager interface. On the left, there's a sidebar with 'API Manager' selected. The main content area is titled 'Google Cloud Vision API'. At the top right of this area, there's a blue 'ENABLE' button with a red arrow pointing to it. Below the title, there's a section titled 'About this API' which describes the Vision API's features. To the right of this, there are 'Documentation' and 'Try it' links. Further down, there's a section about 'Using credentials with this API' and 'Accessing user data with OAuth 2.0'. It explains that you can create an OAuth 2.0 client ID and includes a link to learn more. To the right of this text, there's a diagram illustrating the OAuth 2.0 process: 'Your app' (represented by a laptop icon with a checkmark) has a double-headed arrow connecting to 'User consent' (represented by a smartphone icon with a checkmark). At the bottom, there's a section titled 'Server-to-server interaction'.

Enable Translation API

The screenshot shows the Google Cloud Platform API Manager interface. On the left, there's a sidebar with 'API Manager' selected. The main content area is titled 'Google Cloud Translation API'. A blue 'ENABLE' button is highlighted with a red arrow. Below it, there's a section titled 'About this API' with a description: 'The Google Cloud Translation API lets websites and programs integrate with Google Translate programmatically.' To the right of this are 'Documentation' and 'Try it' buttons. Further down, there's a section on 'Using credentials with this API' and 'Accessing user data with OAuth 2.0'. It explains how to create an OAuth 2.0 client ID and includes a link to 'Learn more'. To the right of this text is a diagram illustrating OAuth 2.0: 'Your app' (represented by a laptop icon with a checkmark) has an arrow pointing to 'User consent' (represented by a smartphone icon with a checkmark). At the bottom, there's a section on 'Server-to-server interaction' with a similar diagram showing a server icon connected to a cloud icon.

Get API key: Select Credential

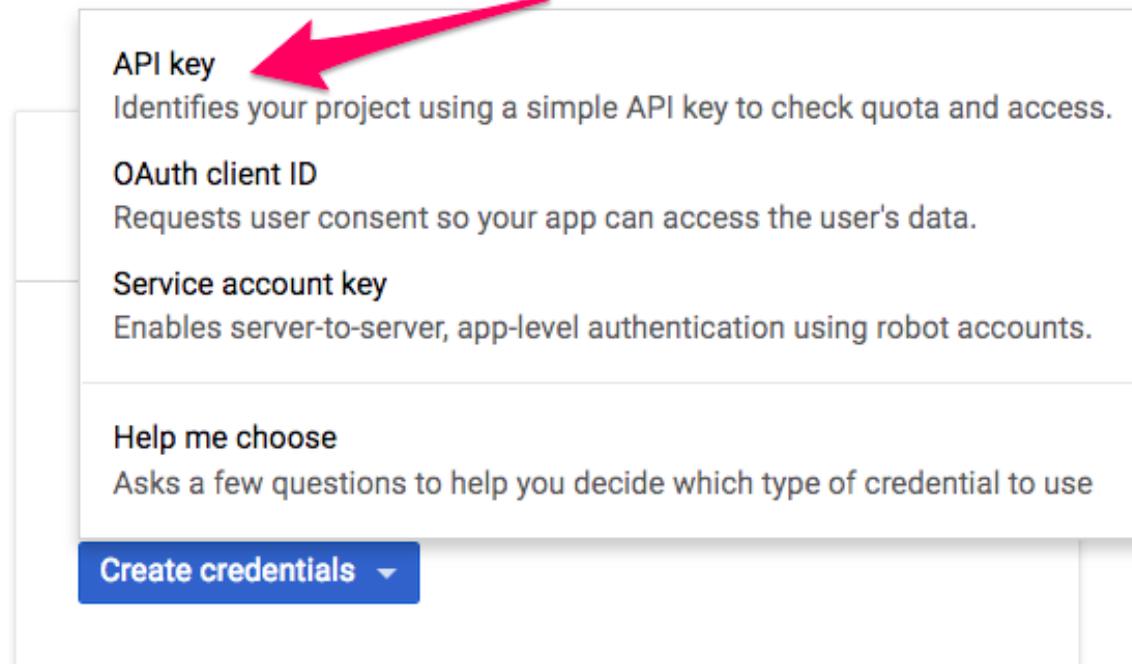
The screenshot shows the Google Cloud Platform API Manager interface. The left sidebar has 'API Manager' selected under 'API'. The main content area is titled 'Credentials' and contains the following text:

You need credentials to access APIs. [Enable the APIs that you plan to use](#) and then create the credentials that they require. Depending on the API, you need an API key, a service account or an OAuth 2.0 client ID. [Refer to the API documentation](#) for details.

A blue button labeled 'Create credentials ▾' is visible at the bottom of the main content area.

Click Create Credential, then API Key

Credentials



Copy API Key

API key created

Use this key in your application by passing it with the `key=API_KEY` parameter.

Your API key

AIzaSyCUT8K6oUDMG_id1WnLxHsJJ--xxbu7ZUM



Restrict your key to prevent unauthorised use in production.

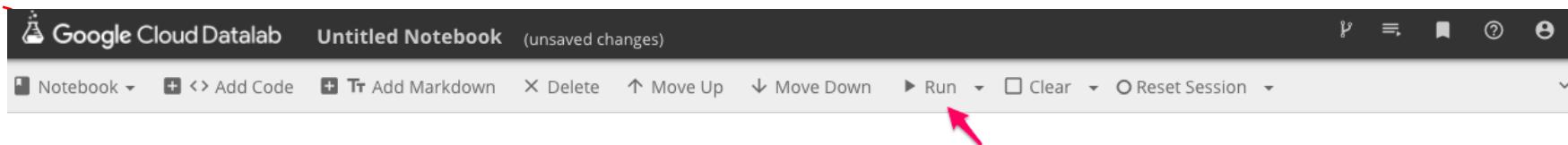
[CLOSE](#) [RESTRICT KEY](#)

Launch a new notebook



install the Python package

```
APIKEY="AIzaSyCUT8K6oUDMG_id1WnLxHsJJ--xxbu7ZUM"  
# Replace with your API key  
!pip install --upgrade google-api-python-client
```



```
APIKEY="AIzaSyCUT8K6oUDMG_id1WnLxHsJJ--xxbu7ZUM" # Replace with your API key  
!pip install --upgrade google-api-python-client  
  
Collecting google-api-python-client  
  Downloading google_api_python_client-1.6.2-py2.py3-none-any.whl (52kB)  
    100% #####| 61kB 2.2MB/s  
Requirement already up-to-date: six<2dev,>=1.6.1 in /usr/local/lib/python2.7/dist-packages (from google-api-python-client)  
Collecting httplib2<1dev,>=0.9.2 (from google-api-python-client)  
  Downloading httplib2-0.10.3.tar.gz (204kB)  
    100% #####| 204kB 3.0MB/s  
Collecting oauth2client<5.0.0dev,>=1.5.0 (from google-api-python-client)  
  Downloading oauth2client-4.1.1-py2.py3-none-any.whl (185kB)  
    100% #####| 194kB 3.5MB/s  
Collecting uritemplate<4dev,>=3.0.0 (from google-api-python-client)  
  Downloading uritemplate-3.0.0-py2.py3-none-any.whl  
Requirement already up-to-date: rsa>=3.1.4 in /usr/local/lib/python2.7/dist-packages (from oauth2client<5.0.0dev,>=1.5.0->google-api-python-client)  
Requirement already up-to-date: pyasn1>=0.1.7 in /usr/local/lib/python2.7/dist-packages (from oauth2client<5.0.0dev,>=1.5.0->google-api-python-client)  
Collecting pyasn1-modules>=0.0.5 (from oauth2client<5.0.0dev,>=1.5.0->google-api-python-client)  
  Downloading pyasn1_modules-0.9-py2.py3-none-any.whl (60kB)  
    100% #####| 61kB 6.0MB/s  
Building wheels for collected packages: httplib2  
  Running setup.py bdist_wheel for httplib2 ... - done  
  Stored in directory: /root/.cache/pip/wheels/ca/ac/5f/749651f7925b231103f5316cacca82a487810c22d30f011c0c  
Successfully built httplib2
```

Invoke Translation API

```
from googleapiclient.discovery import build  
service = build('translate', 'v2', developerKey=APIKEY)  
  
# use the service  
inputs = ['Today is a good day to learn Machine Learning ',  
          'It is an amazing technology', 'Great']  
outputs = service.translations().list(source='en',  
target='th', q=inputs).execute()  
  
# print outputs  
for input, output in zip(inputs, outputs['translations']):  
    print u"{0} -> {1}".format(input, output['translatedText'])
```

```
1 from googleapiclient.discovery import build
2 service = build('translate', 'v2', developerKey=APIKEY)
3 # use the service
4 inputs = ['Today is a good day to learn Machine Learning ', 'It is an amazing technology', 'Great']
5 outputs = service.translations().list(source='en', target='th', q=inputs).execute()
6 # print outputs
7 for input, output in zip(inputs, outputs['translations']):
8     print u"{} -> {}".format(input, output['translatedText'])
9 |
```

Today is a good day to learn Machine Learning -> วันนี้เป็นวันที่ดีในการเรียนรู้การเรียนรู้ด้วยเครื่อง
It is an amazing technology -> เป็นเทคโนโลยีที่น่าอัศจรรย์
Great -> ยิ่งใหญ่

Invoke Vision API

```
import base64
IMAGE="gs://cloud-training-demos/vision/sign2.jpg"
vservice = build('vision', 'v1', developerKey=APIKEY)
request = vservice.images().annotate(body={
    'requests': [
        {
            'image': {
                'source': {
                    'gcs_image_uri': IMAGE
                }
            },
            'features': [
                {
                    'type': 'TEXT_DETECTION',
                    'maxResults': 3,
                }
            ]
        },
    ]
})
responses = request.execute(num_retries=3)
print responses

foreigntext =
responses['responses'][0]['textAnnotations'][0]['description']
foreignlang =
responses['responses'][0]['textAnnotations'][0]['locale']
print foreignlang, foreigntext
```



```

import base64
IMAGE="gs://cloud-training-demos/vision/sign2.jpg"
vservice = build('vision', 'v1', developerKey=APIKEY)
request = vservice.images().annotate(body={
    'requests': [
        {
            'image': {
                'source': {
                    'gcs_image_uri': IMAGE
                }
            },
            'features': [
                {
                    'type': 'TEXT_DETECTION',
                    'maxResults': 3,
                }
            ]
        },
    ],
})
responses = request.execute(num_retries=3)
print responses

{u'responses': [{u'textAnnotations': [{u'locale': u'zh', u'description': u'\u8bf7\u60a8\u7231\u62a4\u548c\u4fdd\n\u62a4\u536b\u751f\u521b\u5efa\u4f18\n\u7f8e\u6c34\u73af\u5883\n', u'boundingPoly': {u'vertices': [{u'y': 103, u'x': 150}, {u'y': 103, u'x': 1084}, {u'y': 654, u'x': 1084}, {u'y': 654, u'x': 150}]}}}, {u'description': u'\u8bf7', u'boundingPoly': {u'vertices': [{u'y': 103, u'x': 178}, {u'y': 103, u'x': 322}, {u'y': 241, u'x': 322}, {u'y': 241, u'x': 178}]}}}, {u'description': u'\u60a8', u'boundingPoly': {u'vertices': [{u'y': 107, u'x': 327}, {u'y': 107, u'x': 471}, {u'y': 241, u'x': 471}, {u'y': 241, u'x': 327}]}}}, {u'description': u'\u7231\u62a4', u'boundingPoly': {u'vertices': [{u'y': 107, u'x': 481}, {u'y': 107, u'x': 774}, {u'y': 245, u'x': 774}, {u'y': 245, u'x': 481}]}}}, {u'description': u'\u548c', u'boundingPoly': {u'vertices': [{u'y': 110, u'x': 784}, {u'y': 110, u'x': 924}, {u'y': 245, u'x': 924}, {u'y': 245, u'x': 784}]}}}, {u'description': u'\u4fdd', u'boundingPoly': {u'vertices': [{u'y': 103, u'x': 921}, {u'y': 103, u'x': 1070}, {u'y': 227, u'x': 1070}, {u'y': 227, u'x': 921}]}}]
foreigntext = responses['responses'][0]['textAnnotations'][0]['description']
foreignlang = responses['responses'][0]['textAnnotations'][0]['locale']
print foreignlang, foreigntext

```

zh 请您爱护和保
护卫生创建优
美水环境

Translate Sign

```
inputs=[foreigntext]
outputs = service.translations().list(source=foreignlang,
target='th', q=inputs).execute()
# print outputs
for input, output in zip(inputs, outputs['translations']):
    print u"{0} -> {1}".format(input,
output['translatedText'])
```

```
inputs=[foreigntext]
outputs = service.translations().list(source=foreignlang, target='th', q=inputs).execute()
# print outputs
for input, output in zip(inputs, outputs['translations']):
    print u"{0} -> {1}".format(input, output['translatedText'])
```

請您爱护和保
护卫生创建优
美水环境
-> โปรดสร้างสภาพแวดล้อมการบ่อองกันน้ำที่สวยงามและการดูแลสุขภาพ