



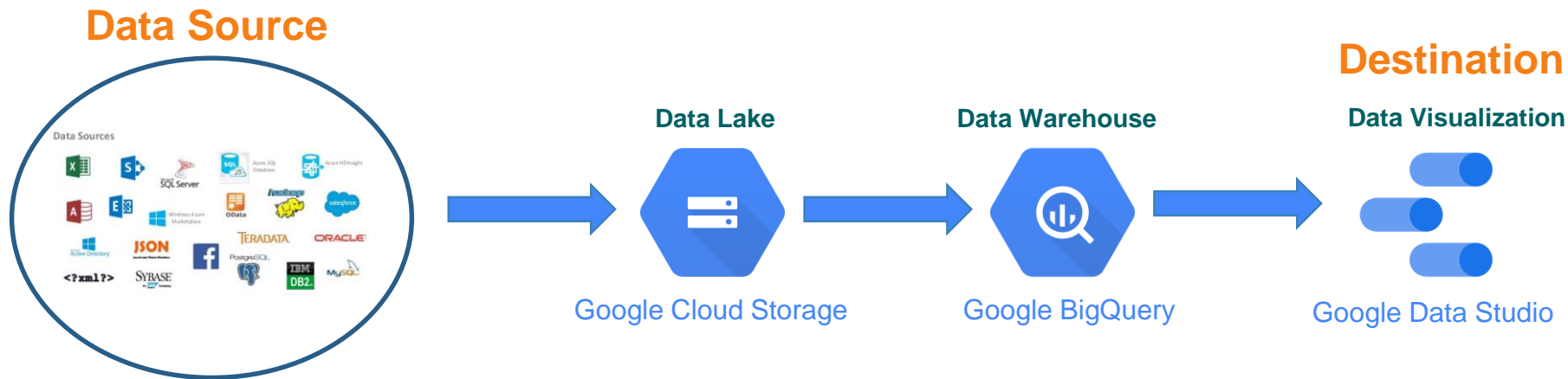
Building a data pipeline on Google Cloud Platform

พิชิตชัย พิมพ์โคตร

- Data pipeline
- Data pipeline Orchestration
- Apache Airflow
 - What is Apache Airflow
 - Airflow DAG definition file example
- Google Cloud Storage
 - Google Cloud Platform
 - Google Cloud Storage
- Workshop#1 upload data to google storage

- Google Cloud Composer
 - Google Cloud Composer
- Workshop#2 Automated Data Pipeline with Airflow
 - Workshop Hello world!
 - Workshop ingest from database and stored on data lake(google storage)
- Google BigQuery
 - Google BigQuery
- Workshop#3 Create BigQuery data warehouse and import data to BigQuery
- Workshop#4 Automated load data to BigQuery
- Exercise

- Data pipeline คือกระบวนการ หรือขั้นตอนในการ “ย้ายข้อมูลจากต้นทาง (Data Source) ไปยังปลายทาง (Destination)”



Technology Stack Overview

Data Lake



Amazon S3 Google Cloud Storage Azure Blob Storage

Data Warehouse



Google BigQuery



Amazon Redshift



Azure Synapse



Snowflake

Data Processing



Apache Hive

Apache Spark

Apache Beam



Amazon Glue



Cloud Dataflow



Cloud Dataproc

Data Pipeline



Azure Data Factory



Apache Airflow

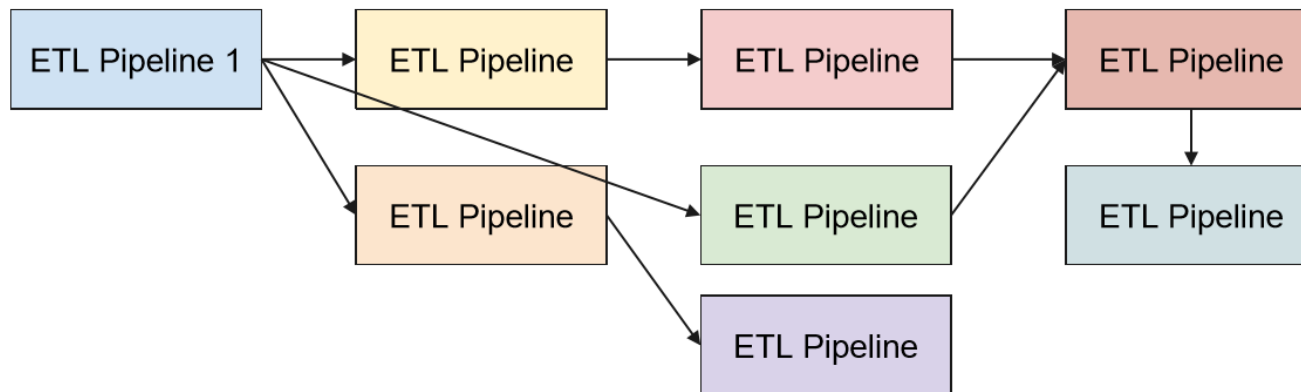


Apache Oozie



Luigi

- Data Pipeline Orchestration เป็นการจัดการ Pipeline ต่างๆให้เป็นระเบียบ เช่นการจัดคิว การ monitor การทำงานของ Pipeline ตั้งแต่ต้นจนจบ



- Oozie – เป็นตัวที่มาพร้อมๆ กับ Hadoop เลย เขียน DAG ด้วย XML
- Luigi – สร้างโดย Spotify เขียน DAG ด้วย Python
- Azkaban – สร้างโดย Linkedin เขียน DAG ด้วย YAML
- Airflow – สร้างโดย Airbnb เขียน DAG ด้วย Python

- พัฒนาโดย บริษัท Airbnb เป็นเครื่องมือที่ใช้จัดการ Task งานต่างๆ จะต้องเขียน Configuration เป็น Python Code โดย Workflow การทำงานจะเป็นแบบเป็นกราฟ DAG (Directed Acyclic Graph)

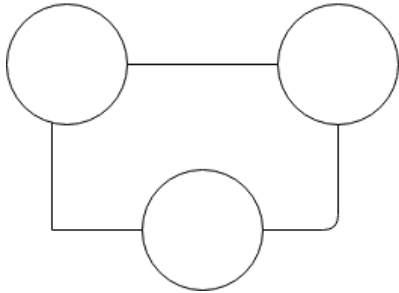


<https://airflow.apache.org/docs/stable/>

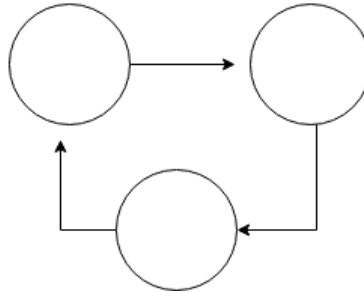
Directed Acyclic Graph (DAG)

- Directed Acyclic Graph (DAG) คือกราฟที่มีหัวลูกศรหรือทิศทางจากจุดหนึ่งไปอีกจุดหนึ่ง โดยไม่สามารถวนกลับมาที่จุดเดิมได้

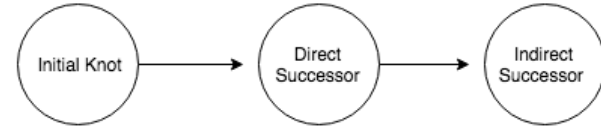
Graph



Directed Graph

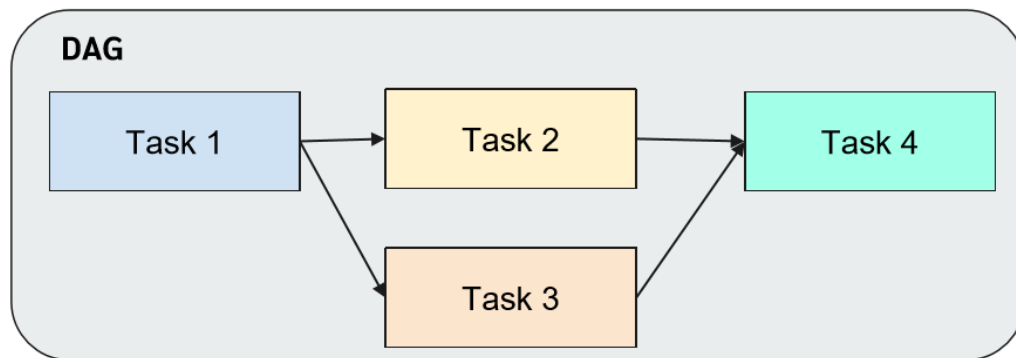


Directed Acyclic Graph



- ส่วนประกอบต่างๆ ของ DAG ใน Airflow

DAG ประกอบไปด้วย Task มาเชื่อมต่อกัน แต่ละ Task จะเป็น Operator ที่มีความสามารถต่างๆ



- Airflow DAG definition file
 1. Importing Modules
 2. Default Arguments
 3. Instantiate a DAG
 4. Tasks
 5. Setting up Dependencies

■ Airflow DAG definition file example

```
# The DAG object; we'll need this to instantiate a DAG
from airflow import DAG
# Operators; we need this to operate!
from airflow.operators.bash_operator import BashOperator
from airflow.utils.dates import days_ago
```

1

Importing Modules

```
default_args = {
    'owner': 'airflow',
    'depends_on_past': False,
    'start_date': days_ago(2),
    'email': ['airflow@example.com'],
    'email_on_failure': False,
    'email_on_retry': False,
    'retries': 1,
    'retry_delay': timedelta(minutes=5),
```

2

Default Arguments

<https://airflow.apache.org/docs/stable/tutorial.html#it-s-a-dag-definition-file>

■ Airflow DAG definition file example

```
dag = DAG(  
    'tutorial',  
    default_args=default_args,  
    description='A simple tutorial DAG',  
    schedule_interval=timedelta(days=1),  
)
```

3

Instantiate a DAG

```
t1 = BashOperator(  
    task_id='print_date',  
    bash_command='date',  
    dag=dag,  
)  
  
t2 = BashOperator(  
    task_id='sleep',  
    depends_on_past=False,  
    bash_command='sleep 5',  
    retries=3,  
    dag=dag,  
)
```

4

Tasks

<https://airflow.apache.org/docs/stable/tutorial.html#it-s-a-dag-definition-file>

- Airflow DAG definition file example

```
# The bit shift operator can also be  
# used to chain operations:  
t1 >> t2
```

5

Setting up Dependencies

<https://airflow.apache.org/docs/stable/tutorial.html#it-s-a-dag-definition-file>

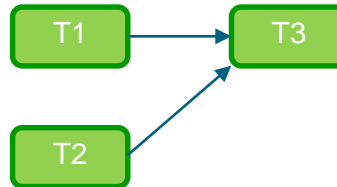
Setting up Dependencies



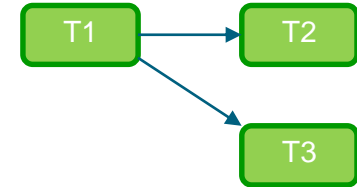
- $T1 \gg T2$
- or $T2 \ll T1$
- or `T1.set_downstream(T2)`
- or `T2.set_upstream(T1)`



- $T1 \gg T2 \gg T3$

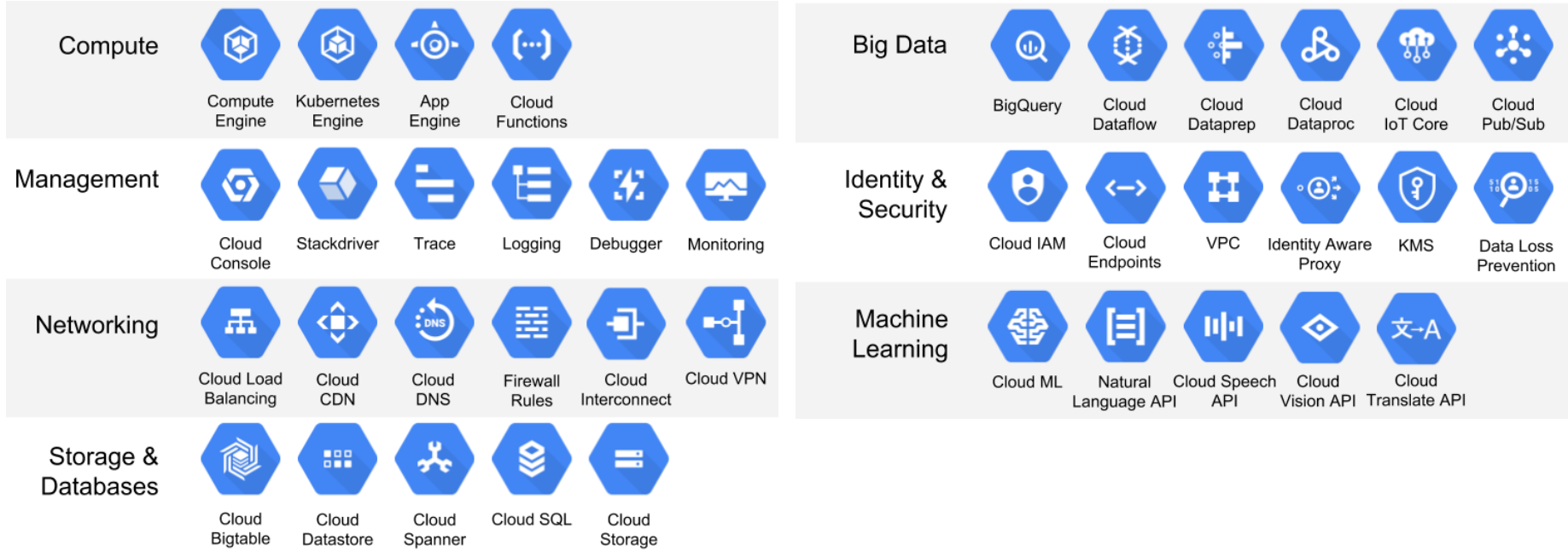


- $T1 \gg T3 \ll T2$



- $T1 \gg [T2, T3]$
- or $[T2, T3] \ll T1$
- or `T1.set_downstream([T2, T3])`

- Google Cloud Console : <https://cloud.google.com/>



<https://github.com/gregsrablins/google-cloud-4-words/blob/master/DarkPoster-medres.png>

- บริการพื้นที่จัดเก็บข้อมูลแบบวัตถุ มีชื่อเรียกอย่างเป็นทางการว่า object storage ซึ่งสามารถเก็บข้อมูลได้หลากหลายประเภท เช่น Text, Image, Video เนื่องจากเป็นการเก็บข้อมูลบน Cloud จึงสามารถเข้าถึงข้อมูลได้จากทุกที่สามารถเพิ่มลด ขนาดพื้นที่ในการจัดเก็บข้อมูลได้ง่าย อีกทั้งยังสามารถกำหนดตำแหน่งทางภูมิศาสตร์ที่ใช้เก็บข้อมูลได้หลายที่เพื่อป้องกันเหตุการณ์ที่จะเป็นอันตรายต่อข้อมูล เช่น พายุ น้ำท่วม ไฟไหม้ เป็นต้น



Google Cloud Storage Option



Google Cloud Storage Option Source: <https://cloud.google.com/images/storage-options/flowchart.svg>

Google Cloud Storage Pricing

Singapore (asia-southeast1) ▼

Standard Storage
(per GB per Month)

\$0.020

Nearline Storage
(per GB per Month)

\$0.010

Coldline Storage
(per GB per Month)

\$0.005

Archive Storage
(per GB per Month)

\$0.0015

<https://cloud.google.com/storage/pricing>

Workshop#1 upload data to google storage

<https://cloud.google.com/storage/docs/how-to>

Create Bucket on Google Storage (console)

Google Cloud Platform My First Project Search products and resources

Storage Storage browser **1** CREATE BUCKET DELETE REFRESH SHOW INFO PANEL LEARN

Filter buckets

Bucket sorting and filtering are available in the Storage browser. Now you can filter your buckets by any value and sort by any column. DISMISS

<input type="checkbox"/>	Name ↑	Created	Location type	Location	Default storage class ?	Updated ?	Public access ?	Access control ?	Lifecycle rules ?
No rows to display									

Store and retrieve your data

Get started by creating a bucket – a container where you can organize and control access to your data and files in Cloud Storage.

1 CREATE BUCKET TAKE QUICKSTART

Create Bucket on Google Storage (console)

Google Cloud Platform

My First Project

Search products and resources

Storage

Browser

Monitoring

Transfer

Transfer for on-premises

Transfer Appliance

Settings

Create a bucket

Name your bucket

Pick a globally unique, permanent name. [Naming guidelines](#)

etltraining-bucket

Tip: Don't include any sensitive information

CONTINUE

Choose where to store your data

This permanent choice defines the geographic placement of your data and affects cost, performance, and availability. [Learn more](#)

Location type

☐ Region

Lowest latency within a single region

☐ Dual-region

High availability and low latency across 2 regions

☒ Multi-region

Highest availability across largest area

Location

us (multiple regions in United States)

CONTINUE

Monthly cost estimate

Enter values below to check this bucket's monthly cost. For guidance only. [Pricing details](#)

Storage and retrieval

Storage size

GB

\$0.026 per GB-month

Data retrieval size

GB

Free

Operations

Class A operations

per-month

\$0.005 per 1,000 ops

Class B operations

per-month

\$0.0004 per 1,000 ops

Availability SLA: 99.95%

Monthly cost: \$0.00

Currency: US Dollar (\$) ▼

2

ตั้งชื่อ bucket

3

กำหนด location
ของเครื่องที่ใช้เก็บข้อมูล


ประเมินค่าใช้จ่ายเบื้องต้น

CITE@DPU

22

- Choose a default storage class for your data

A storage class sets costs for storage, retrieval, and operations. Pick a default storage class based on how long you plan to store your data and how often it will be accessed. [Learn more](#)

- ☒ Standard 
Best for short-term storage and frequently accessed data
- ☐ Nearline
Best for backups and data accessed less than once a month
- ☐ Coldline
Best for disaster recovery and data accessed less than once a quarter
- ☐ Archive
Best for long-term digital preservation of data accessed less than once a year

CONTINUE

4 กำหนดประเภทของข้อมูลที่เก็บ

- Choose how to control access to objects

Access control

- ☐ Fine-grained
Specify access to individual objects by using object-level permissions (ACLs) in addition to your bucket-level permissions (IAM). [Learn more](#)
- ☒ Uniform
Ensure uniform access to all objects in the bucket by using only bucket-level permissions (IAM). This option becomes permanent after 90 days. [Learn more](#)

CONTINUE

5 กำหนดประเภทสิทธิ์การเข้าถึงข้อมูล

Create Bucket on Google Storage (console)

- Advanced settings (optional)

Encryption

- ☒ Google-managed key
No configuration required
- ☐ Customer-managed key
Manage via Google Cloud Key Management Service

6

กำหนดประเภทการเข้ารหัสข้อมูล

Retention policy

Set a retention policy to specify the minimum duration that this bucket's objects must be protected from deletion or modification after they're uploaded. You might set a policy to address industry-specific retention challenges. [Learn more](#)

- ☐ Set a retention policy

7

กำหนดระยะเวลาเมื่อข้อมูลถูกสั่งลบ

Labels

Labels are key:value pairs that allow you to group related buckets together or with other Cloud Platform resources. [Learn more](#)

+ ADD LABEL

8

CREATE

CANCEL

Create Bucket on Google Storage (command line)

Google Cloud Platform My First Project Search products and resources

Storage browser CREATE BUCKET DELETE REFRESH

Activate Cloud Shell

Filter buckets

Bucket sorting and filtering are available in the Storage browser. Now you can filter your buckets by any value and sort by any column.

<input type="checkbox"/>	Name ↑	Created	Location type	Location	Default storage class ?	Updated ?	Public access ?	Access control ?	
<input type="checkbox"/>	etttraining-buck...	Jul 9, 2020, 2:30:38 PM	Region	asia-southeast...	Standard	Jul 9, 2020, 2:30:38 PM	Not public	Uniform	

DISMISS

Create Bucket on Google Storage (command line)

Console gsutil Code samples REST APIs

Use the `gsutil mb` command:

```
gsutil mb gs://[BUCKET_NAME]/
```

Where:

- `[BUCKET_NAME]` is the name you want to give your bucket, subject to [naming requirements](#). For example, `my-bucket`.

Set the following optional flags to have greater control over the creation of your bucket:

- `-p` : Specify the project with which your bucket will be associated. For example, `my-project`.
- `-c` : Specify the default [storage class](#) of your bucket. For example, `NEARLINE`.
- `-l` : Specify the [location](#) of your bucket. For example, `US-EAST1`.
- `-b` : Enable [uniform bucket-level access](#) for your bucket.

For example:

```
gsutil mb -p [PROJECT_ID] -c [STORAGE_CLASS] -l [BUCKET_LOCATION] -b on gs://[BUCKET_NAME]/
```

Create Bucket on Google Storage (command line)

The screenshot shows the Google Cloud Platform interface. The top navigation bar includes 'Google Cloud Platform', 'My First Project', a search bar, and various icons. The left sidebar lists navigation options: Storage, Monitoring, Transfer, Transfer for on-premises, Transfer Appliance, and Settings. The main area displays the 'Storage browser' with a table of buckets. A red box highlights the command 'gsutil mb gs://etltraining-bucket-1' in the Cloud Shell terminal, with a red circle containing the number '2' next to it.

Google Cloud Platform | My First Project | Search products and resources

Storage browser | CREATE BUCKET | DELETE | REFRESH | SHOW INFO PANEL

Filter buckets

Bucket sorting and filtering are available in the Storage browser. Now you can filter your buckets by any value and sort by any column. DISMISS

<input type="checkbox"/>	Name ↑	Created	Location type	Location	Default storage class ?	Updated ?	Public access ?	Access control ?
<input type="checkbox"/>	etltraining-buck...	Jul 9, 2020, 2:30:38 PM	Region	asia-southeast...	Standard	Jul 9, 2020, 2:30:38 PM	Not public	Uniform

CLOUD SHELL Terminal (bubbly-jigsaw-253814) x +

Welcome to Cloud Shell! Type "help" to get started.
Your Cloud Platform project in this session is set to **bubbly-jigsaw-253814**.
Use "gcloud config set project PROJECT_ID" to change to a different project.

pichitchai.pim@cloudshell:~ (bubbly-jigsaw-253814)\$ gsutil mb gs://etltraining-bucket-1

Edit & Delete Bucket

Google Cloud Platform My First Project Search products and resources

Storage Storage browser CREATE BUCKET DELETE REFRESH SHOW INFO PANEL

Filter buckets

Bucket sorting and filtering are available in the Storage browser. Now you can filter your buckets by any value and sort by any column. DISMISS

Name	Created	Location type	Location	Default storage class	Updated	Public access	Access control
etltraining-bucket	Jul 9, 2020, 2:30:38 PM	Region	asia-southeast...	Standard	Jul 9, 2020, 2:30:38 PM	Not public	Uniform
<input checked="" type="checkbox"/> etltraining-bucket...	Jul 9, 2020, 2:38:59 PM	Multi-region	us (multiple re...	Standard	Jul 9, 2020, 2:38:59 PM	Subject to object ACLs	Fine-grained

- Edit bucket permissions
- Edit labels
- Edit default storage class
- Delete bucket
- Export to Cloud Pub/Sub
- Process with Cloud Functions
- Scan with Cloud Data Loss Prevention

Upload data to Google Storage (console)

Google Cloud Platform My First Project Search products and resources

Storage Storage browser CREATE BUCKET DELETE REFRESH SHOW INFO PANEL

Filter buckets

Bucket sorting and filtering are available in the Storage browser. Now you can filter your buckets by any value and sort by any column. DISMISS

<input type="checkbox"/>	Name ↑	Created	Location type	Location	Default storage class ?	Updated ?	Public access ?	Access control ?
<input type="checkbox"/>	etitraining-bucket	Jul 9, 2020, 2:30:38 PM	Region	asia-southeast...	Standard	Jul 9, 2020, 2:30:38 PM	Not public	Uniform
<input type="checkbox"/>	etitraining-bu...	Jul 9, 2020, 2:38:59 PM	Multi-region	us (multiple re...	Standard	Jul 9, 2020, 2:38:59 PM	Subject to object ACLs	Fine-grained

1

เลือก Bucket ที่จะใช้เก็บข้อมูล

Upload data to Google Storage (console)

Google Cloud Platform My First Project Search products and resources

Storage

Bucket details EDIT BUCKET REFRESH BUCKET

etltraining-bucket

Objects Overview Permissions Bucket Lock

Upload files Upload folder Create folder Manage holds Delete

Filter by prefix...

Buckets / etltraining-bucket

There are no live objects in this bucket. If you have object versioning enabled, this bucket may contain noncurrent versions of objects, which aren't visible in the console. You can list noncurrent objects by using the [gsutil command line](#) or the [APIs](#).

Drop files here
or use the upload button

สามารถลาก file ที่จะ upload มาวางได้

Upload data to Google Storage (command line)

Google Cloud Platform My First Project Search products and resources

Storage browser CREATE BUCKET DELETE REFRESH

Filter buckets

Bucket sorting and filtering are available in the Storage browser. Now you can filter your buckets by any value and sort by any column.

Name	Created	Location type	Location	Default storage class	Updated	Public access	Access control
etttraining-buck...	Jul 9, 2020, 2:30:38 PM	Region	asia-southeast...	Standard	Jul 9, 2020, 2:30:38 PM	Not public	Uniform

Activate Cloud Shell

DISMISS

Synopsis

```
gsutil cp [OPTION]... src_url dst_url  
gsutil cp [OPTION]... src_url... dst_url  
gsutil cp [OPTION]... -I dst_url
```



Description

The `gsutil cp` command allows you to copy data between your local file system and the cloud, copy data within the cloud, and copy data between cloud storage providers. For example, to upload all text files from the local directory to a bucket you could do:

```
gsutil cp *.txt gs://my-bucket
```



Upload data to Google Storage (command line)

Google Cloud Platform My First Project Search products and resources

Storage

Bucket details EDIT BUCKET REFRESH BUCKET

etltraining-bucket

Objects Overview Permissions Bucket Lock

Upload files Upload folder Create folder Manage holds Delete

Filter by prefix...

Buckets / etltraining-bucket

Name	Size	Type	Storage class	Last modified	Public access	Encryption	Retention expiration date	Holds
Online Retail.csv	43.47 MB	application/vnd.ms-excel	Standard	7/9/20, 3:02:50 PM UTC+7	Not public	Google-managed key	-	None

Cloud Shell Terminal (bubbly-jigsaw-253814)

```
Welcome to Cloud Shell! Type "help" to get started.
Your Cloud Platform project in this session is set to bubbly-jigsaw-253814.
Use "gcloud config set project [PROJECT_ID]" to change to a different project.
pichitchai_pim@cloudshell:~ (bubbly-jigsaw-253814)$ ls
dagrest.py ddd.jpg hello.py README-cloudshell.txt
pichitchai_pim@cloudshell:~ (bubbly-jigsaw-253814)$ gsutil cp ddd.jpg gs://etltraining-bucket
Copying file:///ddd.jpg [Content-Type=image/jpeg]...
/ [1 files][104.7 KiB/104.7 KiB]
Operation completed over 1 objects/104.7 KiB.
pichitchai_pim@cloudshell:~ (bubbly-jigsaw-253814)$
```

Open Editor

Restart

Upload File

Download File

Safe Mode

2

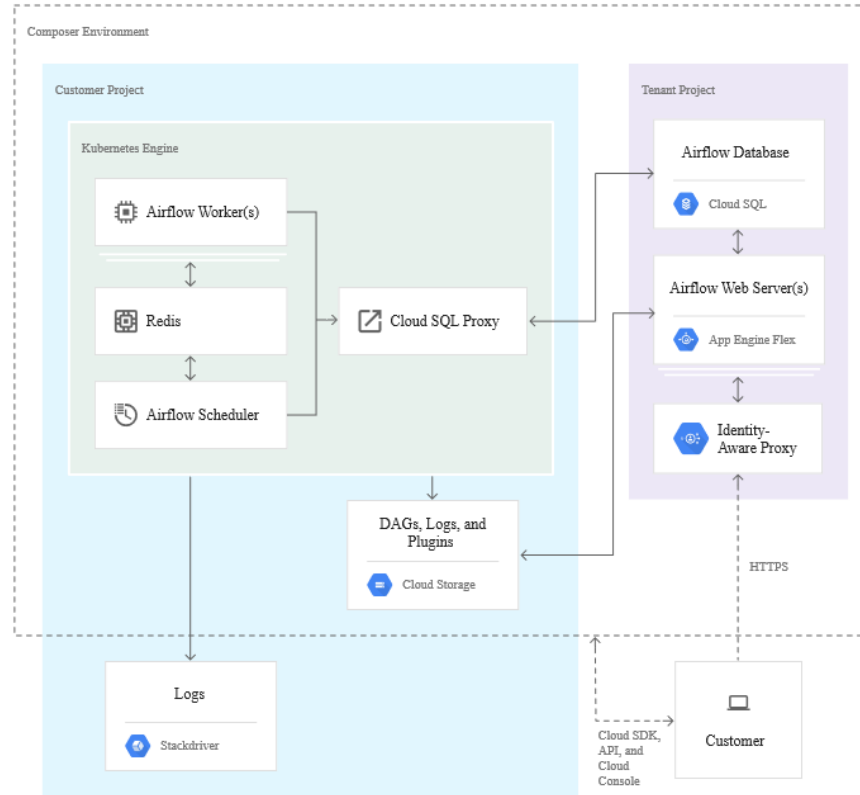
3

Copy file to cloud storage

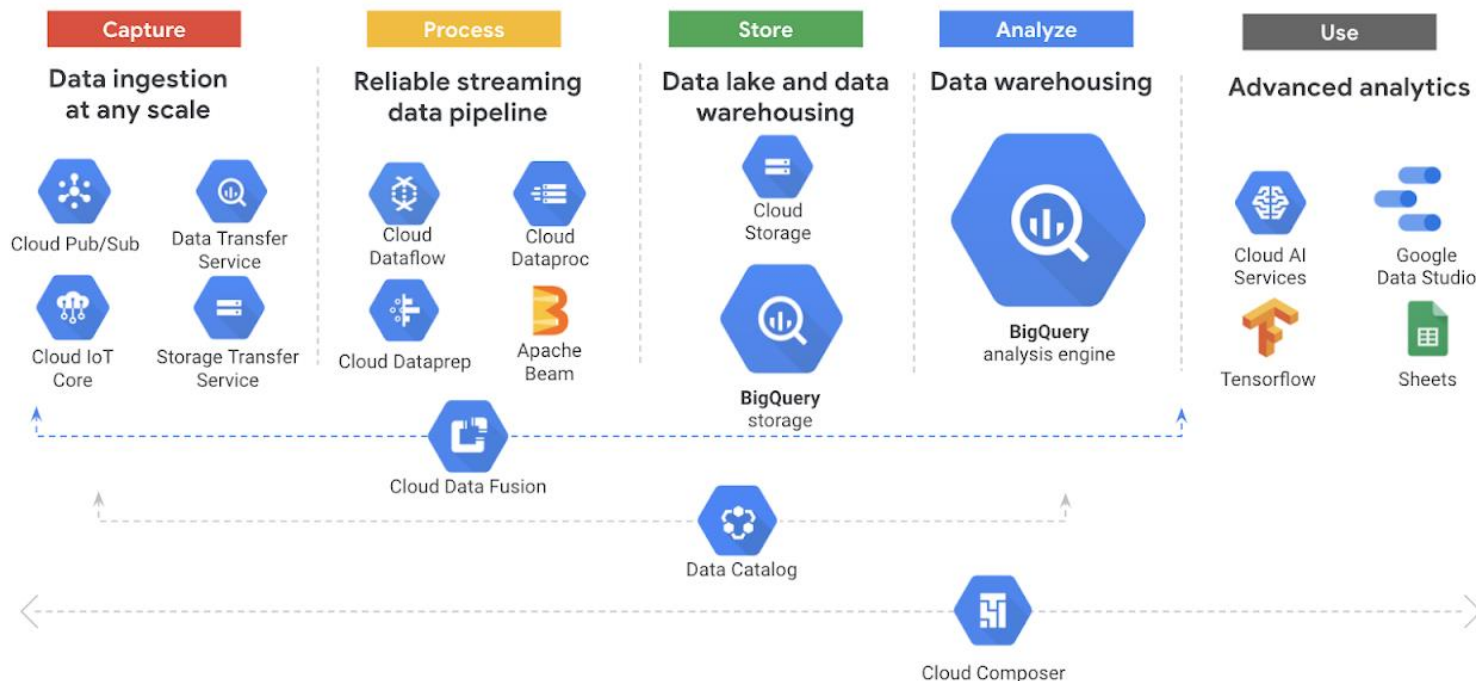
Upload file to cloud shell

- Cloud Composer คือ Service ที่เป็น Fully-managed ที่ไว้ควบคุมการทำงาน Workflow ต่างๆ ที่สร้างมาจาก **Apache Airflow** หน้าที่หลัก ๆ ก็คือ ทำ Schedule, Monitor เป็นต้น

Google Cloud Composer Environment



Google Cloud Composer & Google Cloud Service



Google Cloud Composer Pricing

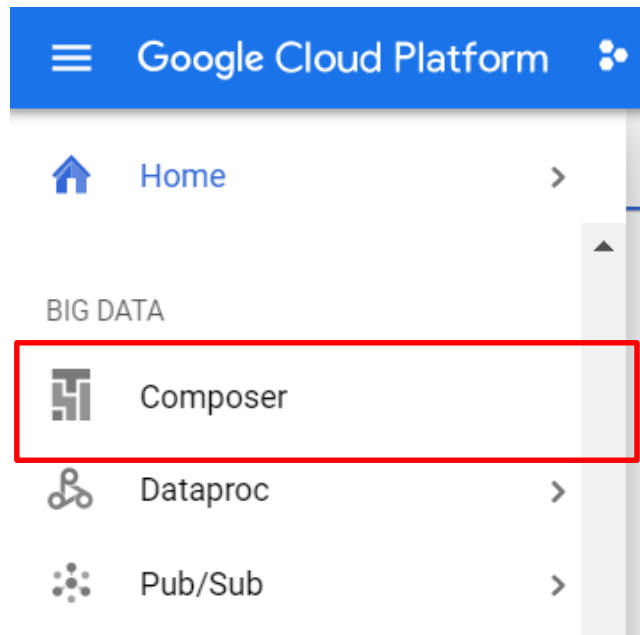
Tokyo (asia-northeast1) ▼

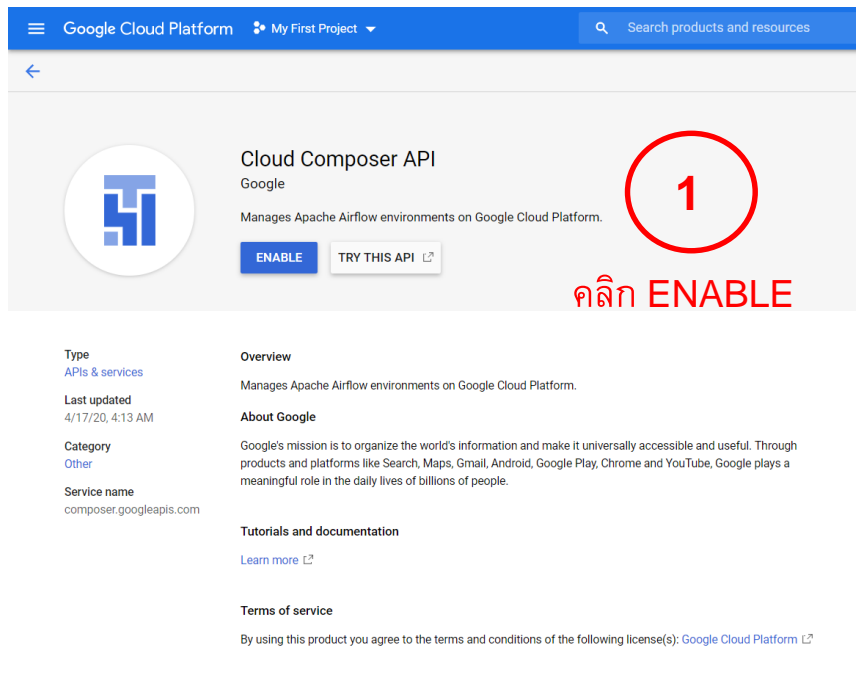
Item	Price (USD)
Web core hours	\$0.094 / vCPU hour
Database core hours	\$0.163 / vCPU hour
Web and database storage	\$0.354 per GB / month
Network egress	\$0.156 / GB

If you pay in a currency other than USD, the prices listed in your currency on [Cloud Platform SKUs](#) apply.

<https://cloud.google.com/composer/pricing?authuser=0&hl=ID>

- Google Cloud Platform Console (<https://console.cloud.google.com/>) >> ไปที่เมนู Composer (อยู่ในกลุ่ม BIG DATA)





Google Cloud Platform My First Project Search products and resources

Cloud Composer API
Google
Manages Apache Airflow environments on Google Cloud Platform.

ENABLE TRY THIS API

คลิก **ENABLE**

Type
APIs & services

Last updated
4/17/20, 4:13 AM

Category
Other

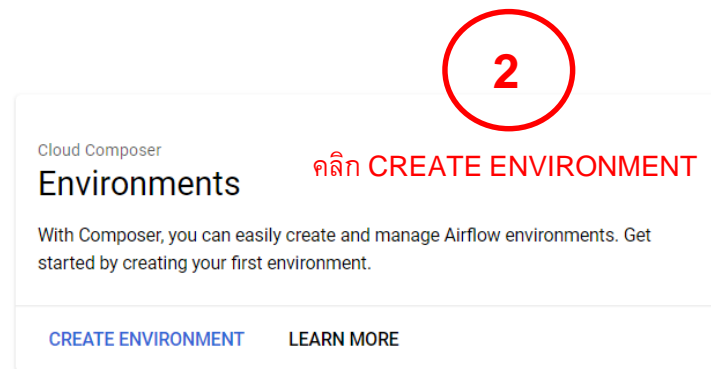
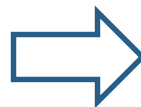
Service name
composer.googleapis.com

Overview
Manages Apache Airflow environments on Google Cloud Platform.

About Google
Google's mission is to organize the world's information and make it universally accessible and useful. Through products and platforms like Search, Maps, Gmail, Android, Google Play, Chrome and YouTube, Google plays a meaningful role in the daily lives of billions of people.

Tutorials and documentation
[Learn more](#)

Terms of service
By using this product you agree to the terms and conditions of the following license(s): [Google Cloud Platform](#)



Cloud Composer
Environments

คลิก **CREATE ENVIRONMENT**

With Composer, you can easily create and manage Airflow environments. Get started by creating your first environment.

CREATE ENVIRONMENT LEARN MORE

Google Cloud Platform My First Project

Composer Create environment

Name *
etltraining

3

Node configuration

The configuration information for the Google Kubernetes Engine nodes running the Airflow software.

Node count *
3

4

Location *
asia-northeast1

5

Zone
asia-northeast1-b

ตั้งชื่อต้องเป็นพิมพ์เล็ก ตัวเลข และ - และชื่อต้อง unique

กำหนดจำนวนเครื่อง ขั้นต่ำที่ 3 เครื่อง

กำหนด Location ของเครื่อง

Machine type
n1-standard-1

The Google Compute Engine machine type used for cluster instances. If unspecified, the machine type will default to 'n1-standard-1'. Must specify location before selecting machine type. [Learn more.](#)

6

กำหนด Spec เครื่องที่ใช้งาน

Disk size (GB)
20

The disk size in GB used for node VMs. Minimum is 20 GB. If unspecified, defaults to 100 GB. Cannot be updated.

7

กำหนด พื้นที่จัดเก็บข้อมูล

Python version
3

The Python version to use in the created environment.

8

กำหนด Version ของ Python เป็น Version 3

✓ NETWORKING, AIRFLOW CONFIG OVERRIDES, AND ADDITIONAL FEATURES.

CREATE

CANCEL

Create Cloud Composer

Google Cloud Platform

My First Project

Search products and resources

Composer

Environment details

REFRESH

DELETE

etltraining

This environment is running

MONITORING BETA

ENVIRONMENT CONFIGURATION

AIRFLOW CONFIGURATION OVERRIDES

ENVIRONMENT VARIABLES

LABELS

PYPI PACKAGES

Specify any required libraries from the Python Package Index (PyPI). If a library cannot be found, it will be removed.

PyPI packages

Package name	Extras and version
requests	for example: [extra]==1.1
pandas	for example: [extra]==1.1
pymysql	for example: [extra]==1.1

+ ADD PACKAGE

SAVE CANCEL

9

หลังจากติดตั้ง Cloud Composer เรียบร้อยแล้ว
ระบุ Packages ที่ใช้

Workshop#2 Automated Data Pipeline with Airflow

- Workshop Hello world!
- Workshop ingest from database and stored on data lake(google storage)

- Google BigQuery เป็น Data Warehouse ที่ตั้งอยู่บนโครงสร้างของ Google Cloud Platform (เป็น Serverless) สามารถทำงานกับข้อมูลขนาดใดๆ ตั้งแต่จาก Excel เล็กๆ จนถึง Big Data ขนาดหลาย Petabyte ได้ในเวลาอันสั้น

BigQuery เน้นสนับสนุนข้อมูลที่มีการเขียนเข้าเป็นหลัก ไม่เอื้อแก่การแก้ไขหรือลบข้อมูล (append-only tables) เท่าไหร่จึงเหมาะจะใช้เป็น Data Warehouse ([LDAP](#)) สำหรับเก็บข้อมูล

ที่ไม่ต้องถูกแก้ไขบ่อยนัก เช่น Event Logs, Analytical Data, หรือ Time Series Events Data แทนที่จะใช้เป็น Operational Database (OLTP) สำหรับเก็บข้อมูลแบบทั่วไป



Google Big Query

- ปริมาณข้อมูลทั้งหมดที่มีอยู่ โดยคิด \$0.02/GB ต่อเดือน
- ปริมาณของข้อมูลที่ถูกนำเข้า โดยคิด \$0.0136/200MB
- ปริมาณของข้อมูลทั้งหมดที่ถูกค้นหา โดยคิด \$6.75/TB

*** Storage : The first 10 GB per month is free.

*** Queries (analysis) : The first 1 TB of query data processed per month is free

ตัวอย่าง คำนวนการ Query ข้อมูล

Query ข้อมูลขนาด 500GB วันละ 50 ครั้ง = $(0.5 \times 6.75) \times 50 = 168.75$ USD
= $168.75 \times 31 = 5,231.25$ THB

<https://cloud.google.com/bigquery/pricing>

Singapore (asia-southeast1) Monthly		
Operation	Pricing	Details
Active storage	\$0.020 per GB	The first 10 GB is free each month. See Storage pricing for details.
Long-term storage	\$0.010 per GB	The first 10 GB is free each month. See Storage pricing for details.
BigQuery Storage API	Unavailable	The BigQuery Storage API is not included in the free tier .
Streaming Inserts	\$0.0136 per 200 MB	You are charged for rows that are successfully inserted. Individual rows are calculated using a 1 KB minimum size. See Streaming pricing for details.
Queries (on-demand)	\$6.75 per TB	First 1 TB per month is free, see On-demand pricing for details.
Queries (hourly Flex slots)	\$27.00 per 500 slots	You can purchase additional slots in 500 slot increments. See Flex slots pricing for details.
Queries (monthly flat-rate)	\$13,500 per 500 slots	You can purchase additional slots in 500 slot increments. See Monthly flat-rate pricing for details.
Queries (annual flat-rate)	\$11,475 per 500 slots	You can purchase additional slots in 500 slot increments. You are billed monthly. See Annual flat-rate pricing for details.

BigQuery ใช้ระบบ Columnar Storage บนเครื่อง Cluster หลายๆเครื่อง (Distributed File System) ซึ่งทำให้การอ่านและคำนวณข้อมูลมหาศาลสามารถทำขนานกัน สเกลได้ง่าย และทำได้เร็วมาก

แต่โครงสร้างดังกล่าวทำให้ BigQuery ไม่มีระบบ Indexing เหมือน Database/Data Warehouse ทั่วไป ทำให้การ Query ข้อมูลแต่ละครั้ง ไม่ว่าจะทำกับข้อมูลส่วนไหนก็ตาม ก็ต้องเริ่มต้นจากการค้นหาจากข้อมูลทั้งตาราง

จากข้อจำกัดดังกล่าว ทำให้แนวทางการลดค่าใช้จ่ายที่เกิดจาก BigQuery สามารถทำได้สามวิธีหลักๆ ดังนี้

- 1. Monitor ดูค่าใช้จ่ายที่ผ่านมาและสถิติการใช้งาน และใส่ลิมิตไว้คุมค่าใช้จ่ายไม่ให้ใช้เยอะเกินไป
- 2. ลดจำนวน Query ที่ใช้ลง
- 3. ลดขนาดข้อมูลที่ต้องใช้ในแต่ละ Query ลง

Workshop#3 Create BigQuery data warehouse and import data to BigQuery

Create BigQuery Dataset

The screenshot displays the Google Cloud Platform BigQuery interface. The top navigation bar includes the Google Cloud Platform logo, the project name 'My First Project', a search bar, and various utility icons. The left sidebar contains a list of navigation options: Query history, Saved queries, Job history, Transfers, Scheduled queries, Reservations, BI Engine, and Resources. The 'Resources' section is expanded, showing a search bar and a list of resources. The resource 'bubbly-jigsaw-253814' is highlighted with a red circle and the number 1. The main area is the 'Query editor', which is currently empty. Below the query editor, there is a toolbar with buttons for 'Save query', 'Save view', 'Schedule query', and 'More'. The bottom section of the interface shows a message: 'No datasets available. Use the controls above to create a dataset and start building out your Resources tree, or check your permissions for this project.' A red circle with the number 2 highlights the 'CREATE DATASET' button in the top right corner of the main area.

Create BigQuery Dataset

The screenshot shows the 'Create dataset' dialog in Google Cloud. It contains the following fields and options:

- Dataset ID:** A text input field containing 'etltraining'. It is circled in red with the number 3 and the text 'ตั้งชื่อ dataset' (Set dataset name).
- Data location (Optional):** A dropdown menu showing 'Singapore (asia-southeast1)'. It is circled in red with the number 4 and the text 'กำหนด location ของเครื่องที่เก็บข้อมูล' (Specify location of the device that stores data).
- Default table expiration:** Radio buttons for 'Never' (selected) and 'Number of days after table creation:'. The latter is circled in red with the number 5 and the text 'กำหนดระยะเวลา' (Specify duration).
- Encryption:** Radio buttons for 'Google-managed key' (selected), 'No configuration required', and 'Customer-managed key'. The 'Google-managed key' option is circled in red with the number 6 and the text 'กำหนดประเภทการเข้ารหัส' (Specify encryption type).
- Buttons:** 'Create dataset' and 'Cancel' buttons at the bottom. The 'Create dataset' button is circled in red with the number 7.

Import data to BigQuery (console)

The screenshot displays the Google Cloud BigQuery console interface. On the left sidebar, under the 'Resources' section, the dataset 'bubbly-jigsaw-253814:etltraining' is listed. A red box and a red circle with the number '1' highlight this dataset name. The main panel shows the details for the 'bubbly-jigsaw-253814:etltraining' dataset. At the top of this panel, a red box and a red circle with the number '2' highlight the 'CREATE TABLE' button. Other buttons visible include 'COPY DATASET' and 'DELETE DATASET'. Below the buttons, the 'Description' and 'Labels' sections both show 'None'. The 'Dataset info' section contains a table with the following details:

Dataset ID	bubbly-jigsaw-253814:etltraining
Created	Jul 9, 2020, 1:30:12 PM
Default table expiration	Never
Last modified	Jul 9, 2020, 1:30:12 PM
Data location	asia-southeast1

Import data to BigQuery (console)

Create table

Source

Create table from: Select file from GCS bucket: Browse File format:

☐ Source Data Partitioning

Destination

☒ Search for a project ☐ Enter a project name

Project name: Dataset name: Table type:

Table name:

Schema

Auto detect ☒ Schema and input parameters

Partition and cluster settings

Partitioning:

3

เลือกข้อมูลต้นทาง

4

เลือก dataset และตั้งชื่อตาราง

5

กำหนด schema ของตาราง

Advanced options

Write preference:

Number of errors allowed: Unknown values: ☐ Ignore unknown values

Field delimiter: Custom field delimiter:

Header rows to skip: Quoted newlines: ☐ Allow quoted newlines Jagged rows: ☐ Allow jagged rows

Encryption

Data is encrypted automatically. Select an encryption key management solution.

☒ Google-managed key
No configuration required

☐ Customer-managed key
Manage via Google Cloud Key Management Service

Create table Cancel

6

ระบุเครื่องหมายที่คั่น
Columns ของข้อมูล

7

ระบุประเภทการเข้ารหัส

Import data to BigQuery Partition table (console)

The screenshot displays the Google Cloud BigQuery console interface. On the left, the 'Resources' sidebar shows a search bar and a list of datasets under the project 'bubbly-jigsaw-253814'. The dataset 'etltraining' is selected and highlighted with a red box and a red circle labeled '1'. The main panel shows the details for 'bubbly-jigsaw-253814:etltraining'. At the top of this panel, there are buttons for 'Run', 'Save query', 'Save view', 'Schedule query', and 'More'. To the right of these buttons, the 'CREATE TABLE' button is highlighted with a red box and a red circle labeled '2'. Below the buttons, there are sections for 'Description' (None), 'Labels' (None), and 'Dataset info'. The 'Dataset info' section contains a table with the following data:

Dataset ID	bubbly-jigsaw-253814:etltraining
Created	Jul 9, 2020, 1:30:12 PM
Default table expiration	Never
Last modified	Jul 9, 2020, 1:30:12 PM
Data location	asia-southeast1

Import data to BigQuery Partition table (console)

Create table

3

เลือกข้อมูลต้นทาง

Source

Create table from:

Google Cloud Storage ▼

Select file from GCS bucket: ?

✔ etltraining-bucket/Online Retail.csv

Browse

File format:

CSV ▼

☐ Source Data Partitioning

Destination

☒ Search for a project

☐ Enter a project name

4

เลือก dataset และตั้งชื่อตาราง

Project name

My First Project ▼

Dataset name

etltraining ▼

Table type ?

Native table ▼

Table name

online_retail_partition

Import data to BigQuery Partition table (console)

Schema

Auto detect

☐ Schema and input parameters

☐ Edit as text

Name	Type	Mode	
InvoiceNo	STRING	NULLABLE	×
StockCode	STRING	NULLABLE	×
Description	STRING	NULLABLE	×
Quantity	INTEGER	NULLABLE	×
InvoiceDate	TIMESTAMP	NULLABLE	×
UnitPrice	FLOAT	NULLABLE	×
CustomerID	INTEGER	NULLABLE	×
Country	STRING	NULLABLE	×
+ Add field			

5 กำหนด schema ของตาราง

Partition and cluster settings

Partitioning: ?

No partitioning

No partitioning

Partition by ingestion time

Partition by field

invoicedate

customerid

6 กำหนดการทำ Data Partition
เป็น Partition by field
เลือก Column ที่จะใช้กำหนด
Partition

Import data to BigQuery Partition table (console)

Partitioning type: ?

- ☒ By day
☐ By hour

7

กำหนด Partition Type

Clustering order (optional): ?

Clustering order determines the sort order of the data. Clustering can be used on both partitioned and non-partitioned tables.

Comma-separated list of fields to define clustering order (up to 4)

Advanced options ^

Write preference:

Write if empty

8

ระบุเครื่องหมายที่คั่น
Columns ของข้อมูล

Number of errors allowed: ?

0

Unknown values: ?

☐ Ignore unknown values

Field delimiter: ?

Custom

Custom field delimiter: ?

;

Header rows to skip: ?

1

Quoted newlines: ?

☐ Allow quoted newlines

Jagged rows: ?

☐ Allow jagged rows

Encryption

Data is encrypted automatically. Select an encryption key management solution.

☒ Google-managed key

No configuration required

☐ Customer-managed key

Manage via Google Cloud Key Management Service

9

ระบุประเภทการเข้ารหัส

Create table


Cancel

Query from Normal table

Unsaved query Edited + COMPOSE NEW QUERY HIDE EDITOR FULL SCREEN

```
1 SELECT * FROM bubbly-jigsaw-253814.etltraining.online_retail
2 where invoicedate ='2010-12-01'
```

Run Save query Save view Schedule query More


This query will process 46 MB when run. 

Query from Partition table

Unsaved query Edited + COMPOSE NEW QUERY HIDE EDITOR FULL SCREEN

```
1 SELECT * FROM bubbly-jigsaw-253814.etltraining.online_retail_partition
2 where invoicedate ='2010-12-01'
```

Run Save query Save view Schedule query More

This query will process 269.6 KB when run. 

Workshop#4 automated load data to BigQuery

Create with auto detection schema

```
'bq load --source_format=CSV --autodetect --skip_leading_rows=1 \  
[DATASET_ID].[TABLE_NAME] \  
gs://[GCS_BUCKET_NAME]/data/online_retail_result.csv'
```

Create partition table

```
'bq load \  
--source_format=CSV \  
--skip_leading_rows=1 \  
--schema  
InvoiceNo:STRING,StockCode:STRING,Description:STRING,Quantity:INTEGER,InvoiceDate:DATE,UnitPrice:FLOAT,CustomerID:  
FLOAT,Country:STRING,InvoiceTimestamp:TIMESTAMP,date:DATE,Rate:FLOAT,THBPrice:FLOAT \  
--time_partitioning_field=InvoiceDate \  
--time_partitioning_type=DAY \  
[DATASET_ID].[TABLE_NAME] \  
gs:// [GCS_BUCKET_NAME]/data/online_retail_result.csv'
```

1. อ่านข้อมูล “online_retail_from_result.csv” จาก data lake
2. แก้ไขข้อมูลฟิวส์ Rate = 31.0
3. คำนวณ THBPrice จาก Rate ใหม่
4. บันทึกข้อมูลใน data lake ชื่อ online_retail_result31.csv
5. สร้าง table ใน BigQuery โดย Import ข้อมูลจาก online_retail_result31.csv