slido

Join at

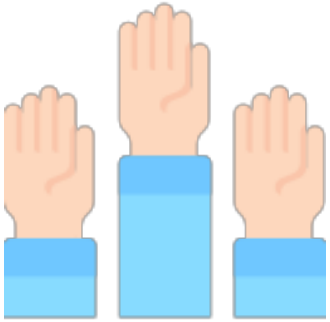
slido.com

#3557 928

ข้อตกลงการเรียน: จดได้ ถามได้

หาที่จดไว้ ไม่ลืมแน่นอน

ระหว่างเรียน ถามคำถามได้ตลอดเวลาใน slido.com



ตอนจบแต่ละ Section จะมีเวลาให้ถามคำถาม

Data TH - Data Science ชิลชิล

@datasciencechill · ⭐ 5 รีวิว 40 รายการ · เว็บไซต์เพื่อการศึกษา

ส่งข้อความ

สวัสดี! มีอะไรให้เราช่วยไหม

หน้าหลัก | งานกิจกรรม | รีวิว | วิดีโอ | เพิ่มเติม

ถูกใจแล้ว

Module agenda

Data Lakes and Data Warehouses

1. Database

เหมาะกับการเก็บ **Structured Data** หรือ **Unstructured Data** ที่ต้องการ เขียนและเข้าถึง อย่างรวดเร็ว เช่น Website, Application มือถือ, ระบบต่าง ๆ ที่มีผู้ใช้ ฯลฯ

แบ่งออกเป็น 2 ประเภทใหญ่ ๆ

1.1) SQL Database
สำหรับเก็บข้อมูลเป็นตาราง (Structured Data)

1.2) NoSQL Database
สำหรับเก็บข้อมูลแบบ Semi-Structured Data

DataTH

1.1 SQL Database

หรือ RDBMS (Relational Database Management System)

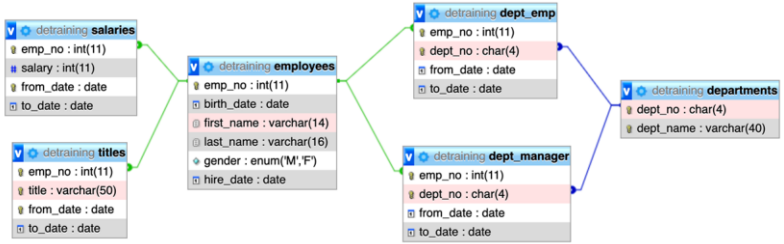
คือ ฐานข้อมูลสำหรับเก็บ **Structured Data** สามารถดึงข้อมูลด้วย **SQL**

เก็บข้อมูลเป็นตาราง

- **Table** (ตารางข้อมูล)
- **Schema** (โครงสร้างตาราง)
- **Relation** (ความสัมพันธ์ระหว่างตาราง)
- **Primary Key** (key ที่ไม่ซ้ำกัน เพื่อใช้อ้างอิงในแต่ละแถว)

Name	Type	Collation	Attributes	Null	Default
emp_no	int(11)			No	*None*
birth_date	date			No	*None*
first_name	varchar(14)	utf8mb4_0900_ai_ci		No	*None*
last_name	varchar(16)	utf8mb4_0900_ai_ci		No	*None*
gender	enum('M', 'F')	utf8mb4_0900_ai_ci		No	*None*
hire_date	date			No	*None*

ตัวอย่าง Database Schema



ตัวอย่าง Database Relation

PostgreSQL MySQL Microsoft SQL Server ORACLE

DataTH

1.2 NoSQL Database

ฐานข้อมูลอีกประเภท ที่เกิดขึ้นมาสำหรับ **Semi-Structured Data**

NoSQL = Not Only SQL หมายถึงว่า Database แบบ NoSQL บางตัวก็อ่าน SQL ได้

เช่น MongoDB หรือ Neo4J

Document stores

Key-value stores

Wide column stores

Graph DBMS

Database Models ยังมีอีกมากมาย

- Relational DBMS
- Key-value stores
- Document stores
- Graph DBMS
- Time Series DBMS
- Object oriented DBMS
- Search engines
- RDF stores (triplestore)
- Multivalue DBMS
- Wide column stores
- Native XML DBMS
- Event Stores
- Content stores
- Navigational DBMS

2. Data Warehouse

เหมาะกับการเก็บ**ข้อมูลประเภท Structured data** ที่มีขนาดใหญ่ ใช้พื้นที่จัดเก็บที่มีปริมาณมาก ข้อมูลที่เก็บมักจะเป็นข้อมูลในอดีต (historical data) ที่ไม่มีการเปลี่ยนแปลง

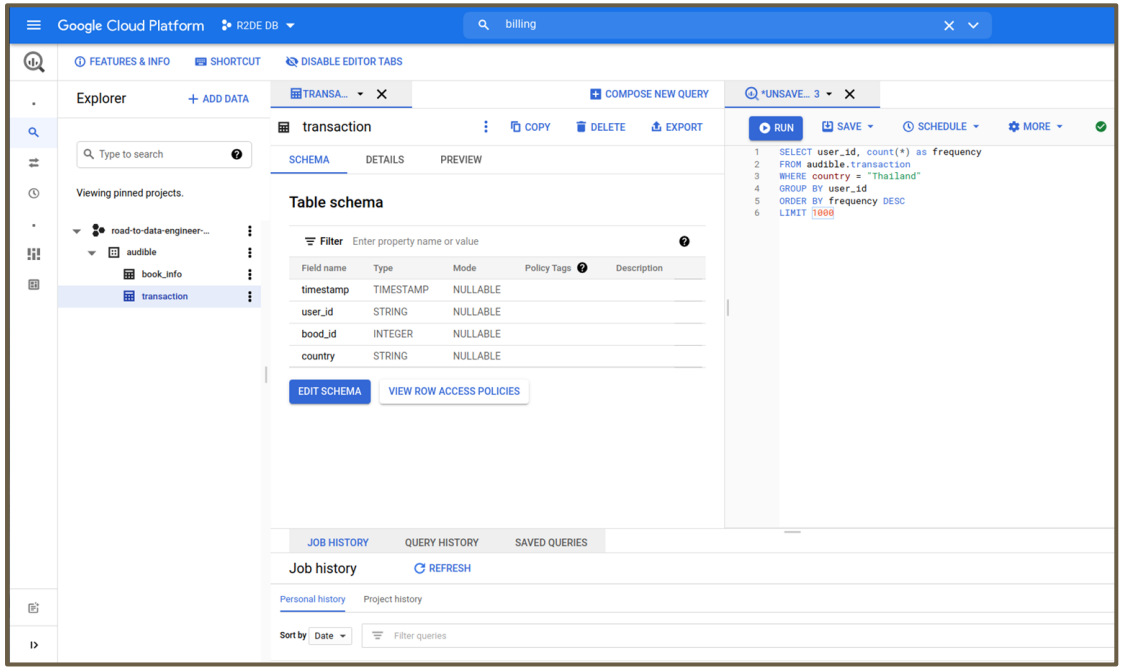
ตัวอย่างการนำไปใช้ เช่น

- การทำ Dashboard / Report หรือ การทำข้อมูลเพื่อประกอบการตัดสินใจ (Business Intelligence)
- การวิเคราะห์ข้อมูล Data Analytics หรือการนำข้อมูลไปเพื่อไปสร้างโมเดล

DataTH

ตัวอย่าง Data Warehouse

รูปตัวอย่างหน้าจอ (UI) ของ Google BigQuery



Apache Hive / Impala

Amazon Redshift

Google BigQuery

Azure Synapse

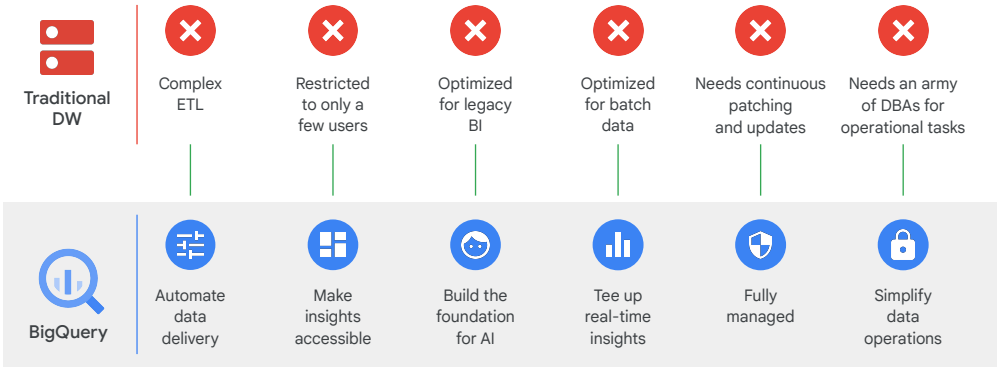
Snowflake

DataTH

Considerations when choosing a data warehouse

- Can it serve as a sink for both batch and streaming data pipelines?
- Can the data warehouse scale to meet my needs?
- How is the data organized, cataloged, and access controlled?
- Is the warehouse designed for performance?
- What level of maintenance is required by our engineering team?

BigQuery is a modern data warehouse that changes the conventional mode of data warehousing



Structured Data เก็บที่ไหนดี: [DB OLTP vs [DW] OLAP

OLTP (On-Line Transaction Processing) ⇒ Database

- ออกแบบมาสำหรับการเขียนและอัปเดตข้อมูล
- พบได้บ่อยใน application, เว็บไซต์

OLAP (On-Line Analytical Processing) ⇒ Data Warehouse

- ออกแบบมาสำหรับการอ่านข้อมูลเยอะ ๆ เพื่อนำไปวิเคราะห์ หรือทำ Data Visualization (Chapter 6)
- เหมาะกับการวิเคราะห์ข้อมูลที่ซับซ้อน โดย Data Analyst และ Data Scientist

Microsoft SQL Server

MySQL

ORACLE

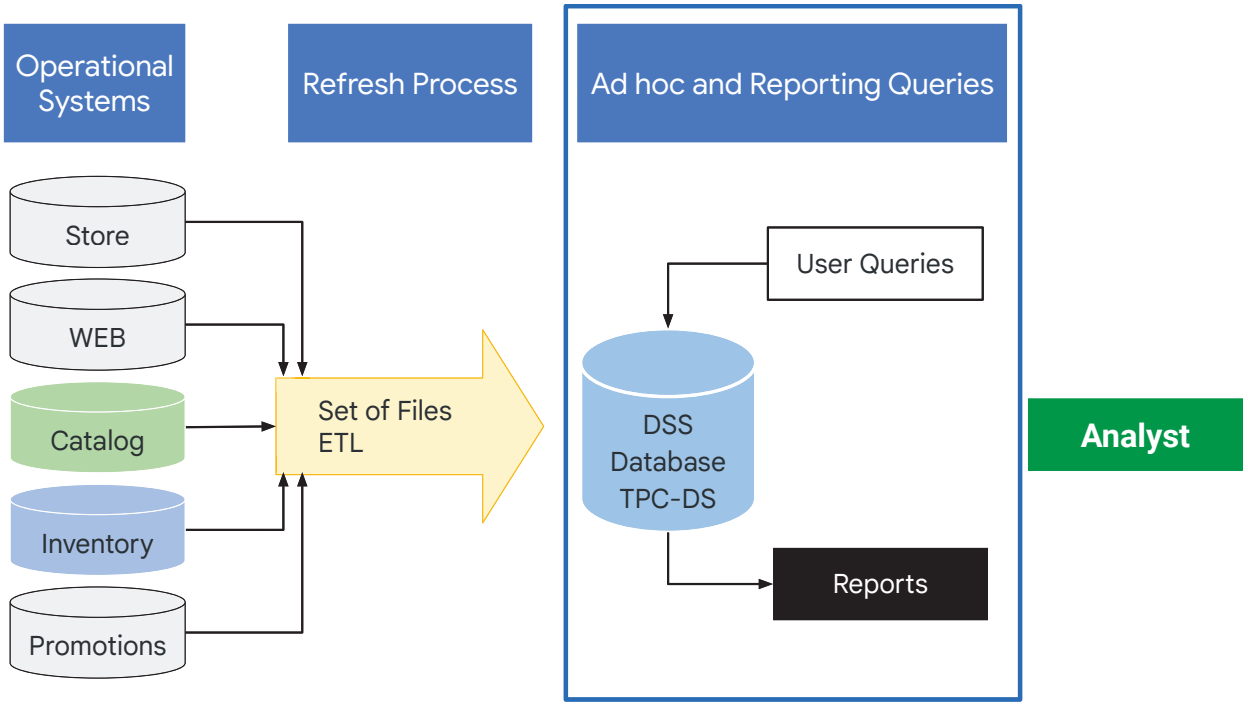
Oracle DB

HIVE

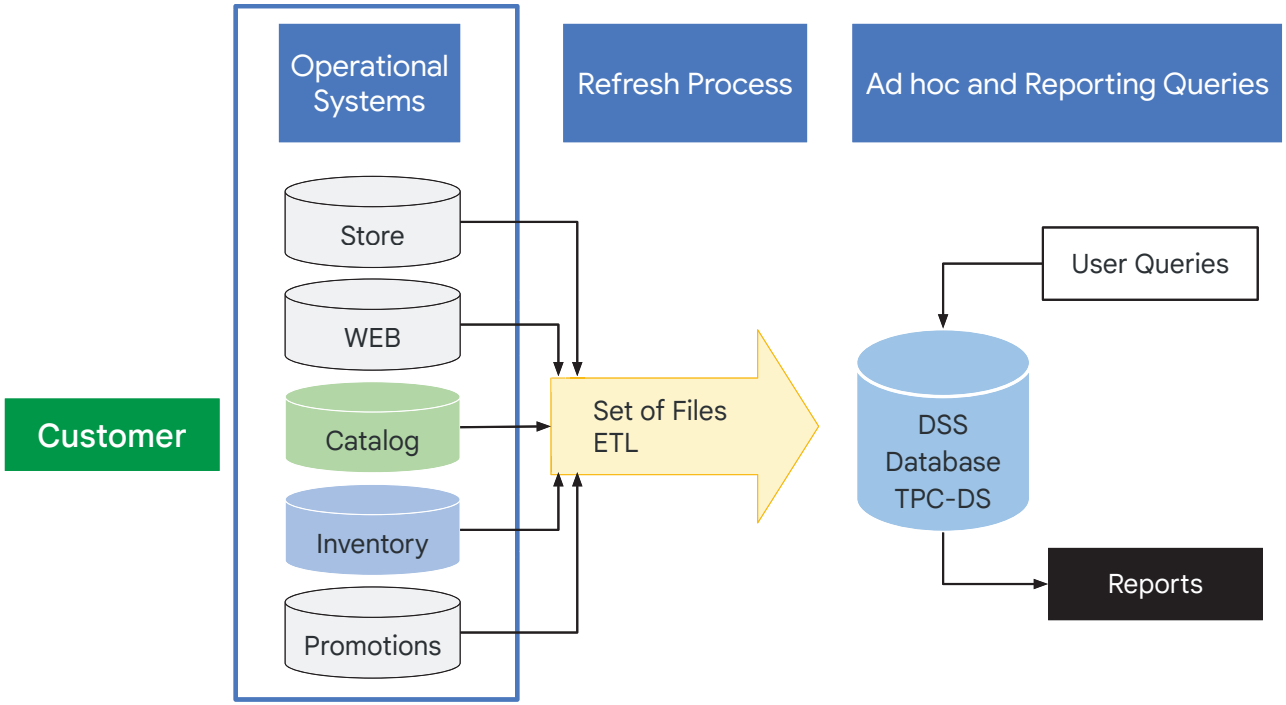
Apache Hive

DataTH

Analytical systems are 20% writes and 80% reads



Transactional systems are 80% writes and 20% reads



OLTP vs. OLAP

ONLINE TRANSACTION PROCESSING	ONLINE ANALYTICAL PROCESSING
Handles recent operational data	Handles all historical data
Size is smaller, typically ranging from 100 Mb to 10 Gb	Size is larger, typically ranging from 1 Tb to 100 Pb
Goal is to perform day-to-day operations	Goal is to make decisions from large data sources
Uses simple queries	Uses complex queries
Faster processing speeds	Slower processing speeds
Requires read/write operations	Requires only read operations

3. Data Lake

Data Lake เป็นที่เก็บข้อมูลขนาดใหญ่ รองรับข้อมูลทุกรูปแบบที่เป็นไฟล์

คล้าย Harddrive ในเครื่องคอมพิวเตอร์ของเรา แต่มีระบบป้องกันข้อมูลสูญหายที่ดีกว่า

เก็บได้ทั้ง Structured Data, Semi-Structured Data และ Unstructured Data

Data Lake บางครั้งก็ถูกเรียกว่า object storage หรือ blob storage

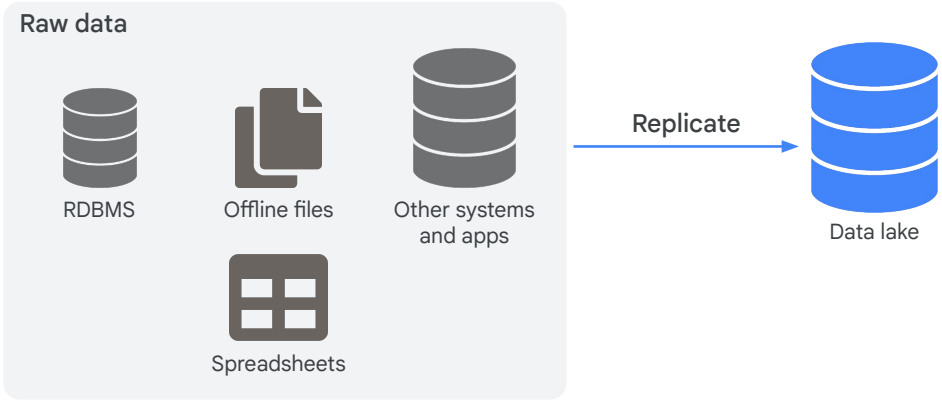
Apache Hadoop HDFS
Hadoop Distributed File System

Amazon S3

Google Cloud Storage

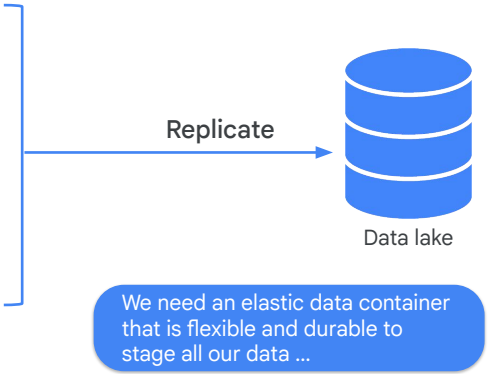
Azure Blob Storage

A data lake brings together data from across the enterprise into a single location

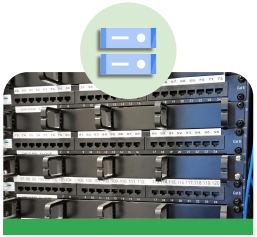


Key considerations when building a data lake

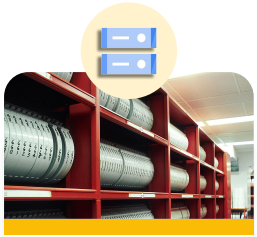
1. Can your data lake handle all the types of data you have?
2. Can it scale to meet the demand?
3. Does it support high-throughput ingestion?
4. Is there fine-grained access control to objects?
5. Can other tools connect easily?



Cloud Storage is designed for 99.999999999% annual durability



Backup



Replace/decommission infrastructure

Analytics and ML

Content storage and delivery

Quickly create buckets with Cloud Shell

```
gsutil mb gs://your-project-name
```

วิธีเลือก Data Storage: เลือกตามจุดประสงค์การใช้งาน

เก็บข้อมูลในระบบที่มีการใช้งานและเปลี่ยนแปลงตลอดเวลา

เช่น

- ฐานข้อมูลที่ใช้งานในธุรกิจ หรือ อุตสาหกรรม
- การใช้งานร่วมกับ application หรือเว็บไซต์

⇒ Database (DB)

เก็บข้อมูลเพื่อการวิเคราะห์ข้อมูล สำหรับข้อมูลที่ไม่มีการเปลี่ยนแปลงแล้ว

มักจะดึงข้อมูลจาก Database มาเก็บ เช่น

- การนำข้อมูลไปวิเคราะห์ หรือ สร้างโมเดล
- ใช้ข้อมูลเพื่อสร้าง BI dashboard / Visualization

⇒ Data Warehouse (DW)

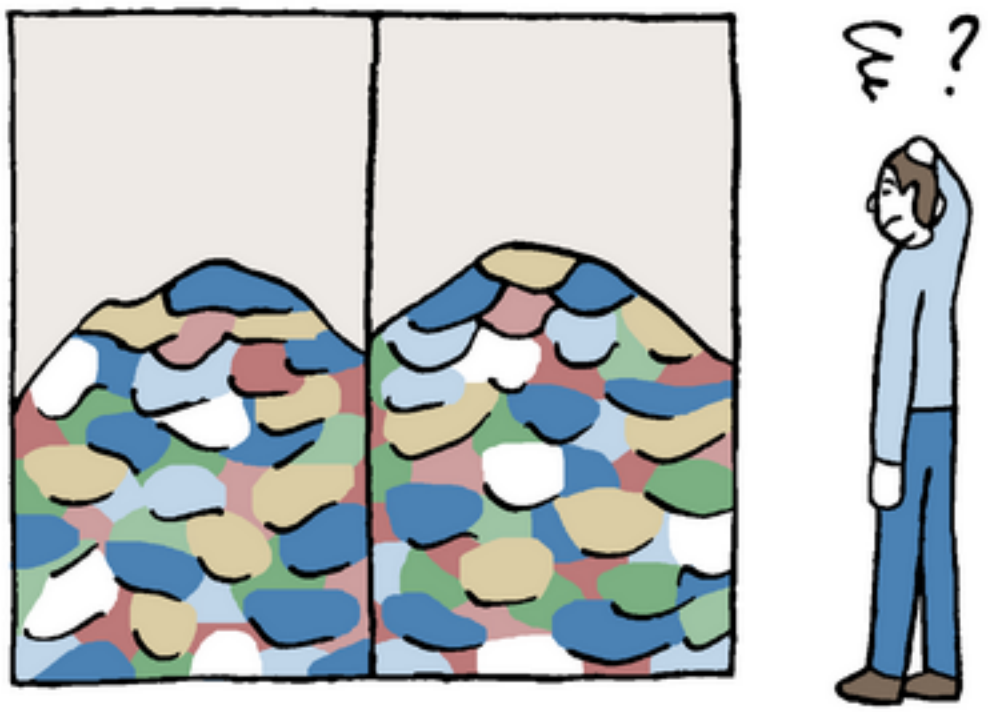
เก็บข้อมูล Unstructured Data เช่น รูปภาพ, ไฟล์เสียง, วีดิโอ

หรือการเก็บข้อมูลดิบ (raw data) ในรูปแบบของไฟล์ เช่น .txt, .csv ในราคาประหยัด

⇒ Data Lake (DL)

DATA WAREHOUSE

DATA LAKE

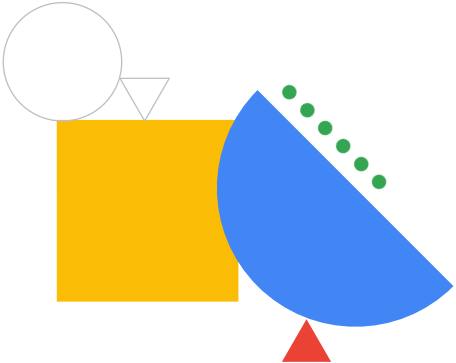


Dataedo /cartoon

Piotr@Dataedo

Google Cloud

Building a Data Warehouse



Module agenda

Introduction to BigQuery

A modern data warehouse

- Gigabytes to petabytes
- Serverless and no-ops, including ad hoc queries
- Ecosystem of visualization and reporting tools
- Ecosystem of ETL and data processing tools
- Up-to-the-minute data
- Machine learning
- Security and collaboration

BigQuery has many capabilities that make it an ideal data warehouse

- Interactive SQL queries over large datasets (petabytes) in seconds
- Serverless and no-ops, including ad hoc queries
- Ecosystem of visualization and reporting tools
- Ecosystem of ETL and data processing tools
- Up-to-the-minute data
- Machine learning
- Security and collaboration

BigQuery

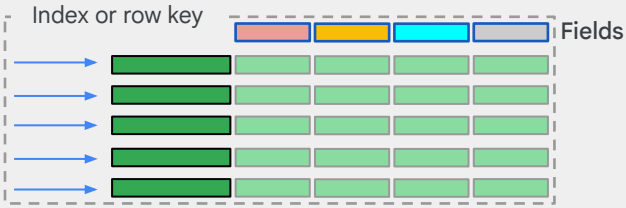
BigQuery is a serverless fully-managed service

- Data aging
- Storage management
- Fault recovery
- Query engine optimization
- Hardware
- Updates

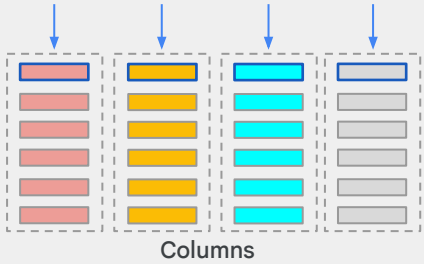
Free up real people-hours by not having to worry about common tasks.

What makes BigQuery fast?

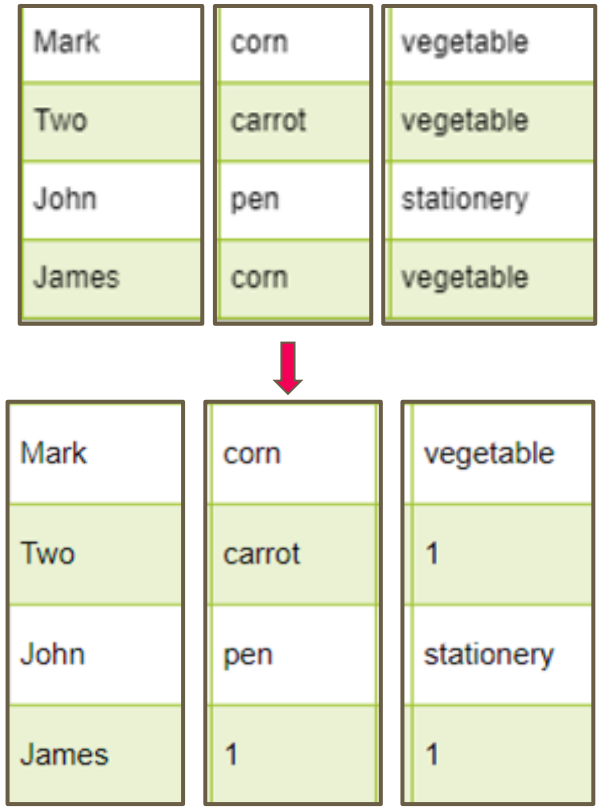
Relational Database tables are row-oriented



BigQuery tables are column-oriented



ใช้ Columnar Storage ช่วยให้พื้นที่เก็บข้อมูลน้อยลง



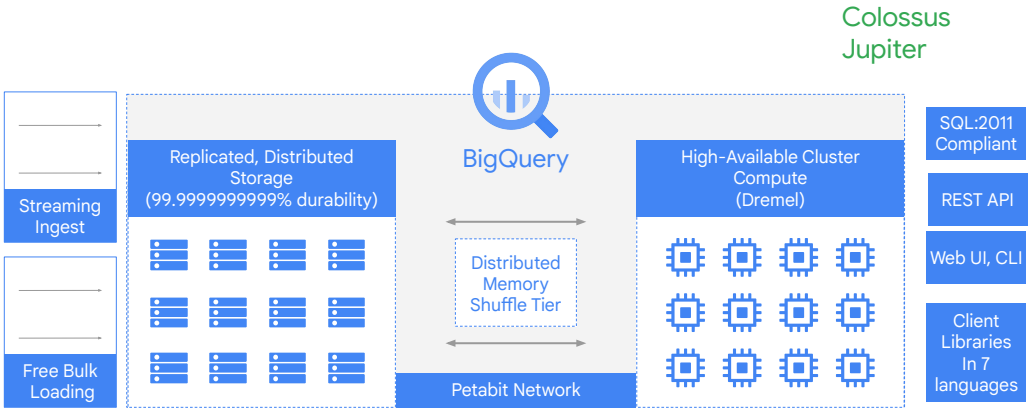
Data Warehouse ระบบใหม่ ๆ เช่น Google BigQuery, Amazon Redshift, Snowflake, Apache Cassandra, Apache HBase ฯลฯ ใช้การเก็บข้อมูลเป็นคอลัมน์

หรือเรียกว่า **Columnar Storage / Column-oriented Database**

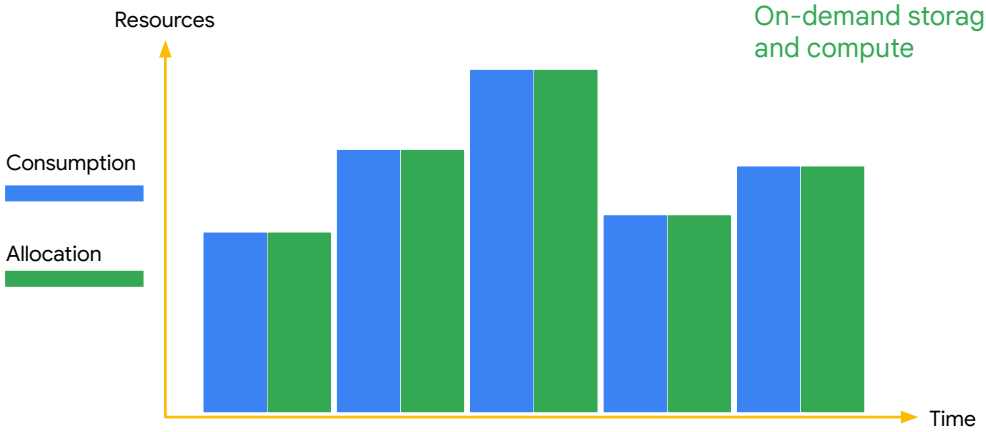
การเก็บค่าเป็น Column ข้อดี คือ ถ้าค่าในคอลัมน์เหมือนกัน ก็ไม่ต้องเก็บค่าซ้ำ ใช้วิธีอ่านจากค่าที่เคยเก็บแล้วได้เลย
(ป.ล. บางฐานข้อมูลอาจจะมีการแอบทำ Normalization เบื้องหลัง แต่ไอเดียเหมือนกัน)

DATATH

The data is physically stored in a redundant way separate from the compute cluster

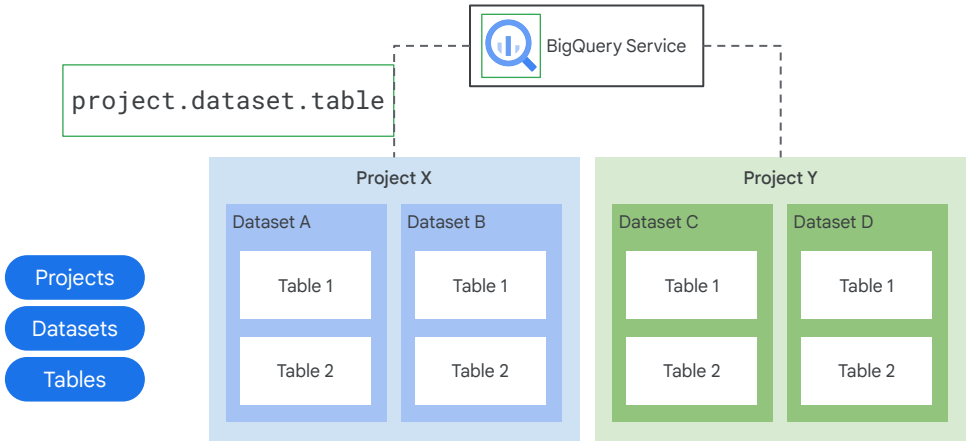


You don't need to provision resources before using BigQuery



Get Started with BigQuery

What are some reasons to structure your information?



Explorer

Type to search

Viewing resources.

SHOW STARRED ONLY

- poc-transl
  - Queries
  - Notebooks
  - External connections
  - billing_export
  - gcp_billing_ex...
  - poc_translated
    - arm_nut_p...
    - transaction...
    - translation...
    - translation...
    - translation...
    - translation...
  - translation_rep...

SUMMARY

Untitled | arm_nut_puk_quality_table

arm_nut_puk_quality_table

QUERY | SHARE | COPY | SNAPSHOT | DELETE | EXPORT

This is a partitioned table. Learn more

SCHEMA | DETAILS | PREVIEW | LINEAGE | DATA PROFILE | DATA QUALITY

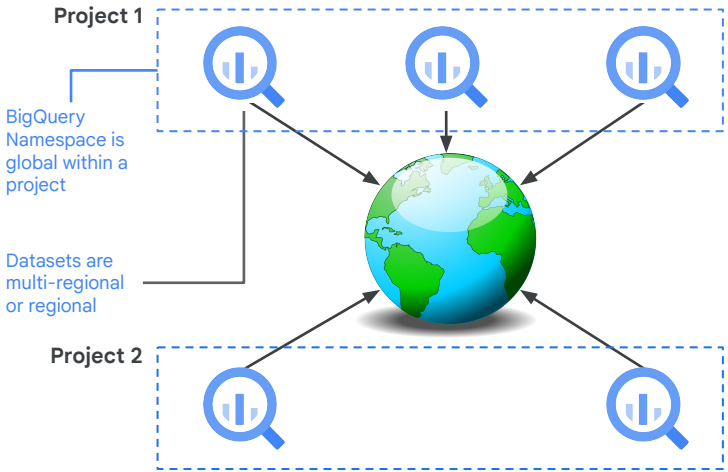
Filter Enter property name or value

Field name	Type	Mode	Key	Collation	Default Value	Policy Tags	Description
data_quality_scan	RECORD	NULLABLE	-	-	-	-	Data quality scan information.
data_source	RECORD	NULLABLE	-	-	-	-	The data source of the data scan.
data_quality_job_id	STRING	NULLABLE	-	-	-	-	Data quality scan job id.
data_quality_job_configuration	JSON	NULLABLE	-	-	-	-	Data quality job configuration.
job_labels	JSON	NULLABLE	-	-	-	-	The data scan job labels.
job_start_time	TIMESTAMP	NULLABLE	-	-	-	-	The start time of the data scan job.
job_end_time	TIMESTAMP	NULLABLE	-	-	-	-	The end time of the data scan job.
job_quality_result	RECORD	NULLABLE	-	-	-	-	The overall result of the data quality job.
job_dimension_result	JSON	NULLABLE	-	-	-	-	The dimension result of the data quality job.
job_rows_scanned	INTEGER	NULLABLE	-	-	-	-	The number of rows that have been scanned during this data scan job.
rule_name	STRING	NULLABLE	-	-	-	-	Data quality rule name.
rule_description	STRING	NULLABLE	-	-	-	-	Data quality rule description.
rule_type	STRING	NULLABLE	-	-	-	-	Data quality rule type.
rule_evaluation_type	STRING	NULLABLE	-	-	-	-	Data quality rule evaluation type.
rule_column	STRING	NULLABLE	-	-	-	-	The column name in the source table of the rule runs against.
rule_dimension	STRING	NULLABLE	-	-	-	-	Data quality rule dimension.

EDIT SCHEMA | VIEW ROW ACCESS POLICIES

Job history

BigQuery datasets can be regional or multi-regional



Access control is through IAM

Organization
Project
Dataset

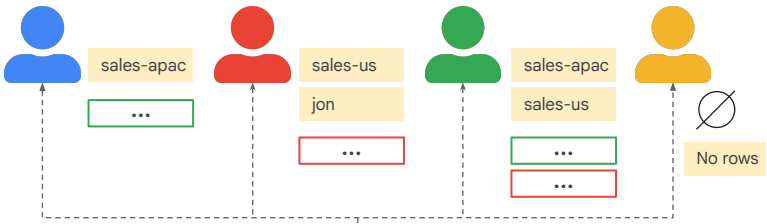
Access control is to the Dataset, Table, View, or Column

A user has to run jobs to query datasets

Row-level security in BigQuery

```
CREATE ROW ACCESS POLICY
    apac_filter
ON
    dataset1.table1
GRANT TO
("group:sales-apac@example.com")
FILTER USING
    (Region="APAC");
```

```
CREATE ROW ACCESS POLICY
    us_filter
ON
    dataset1.table1
GRANT TO
("group:sales-us@example.com",
 "user:jon@example.com")
FILTER USING
    (Region="US");
```



Partner	Contact	Country	Region
Example Customers Corp	alice@examplecustomers.com	Japan	APAC
Example Enterprise Group	bob@exampleenterprisegroup.com	Singapore	APAC
Example HighTouch Co.	carrie@examplehightouch.com	USA	US
Example Buyers Inc.	david@examplebuyersinc.com	USA	US

Load Data into BigQuery

The method you use to load data depends on how much transformation is needed

EL

Extract and **L**oad

ELT

Extract, **L**oad, and **T**ransform

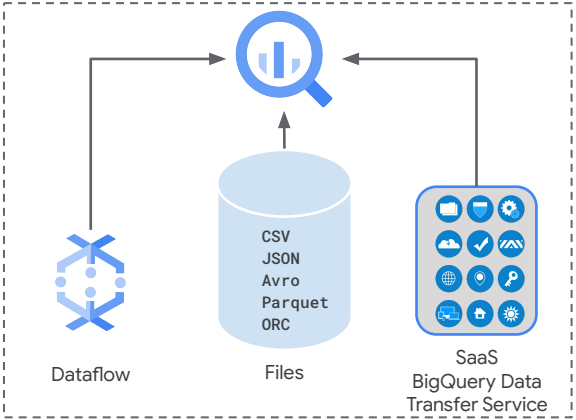
ETL

Extract, **T**ransform, and **L**oad

Batch load supports different file formats

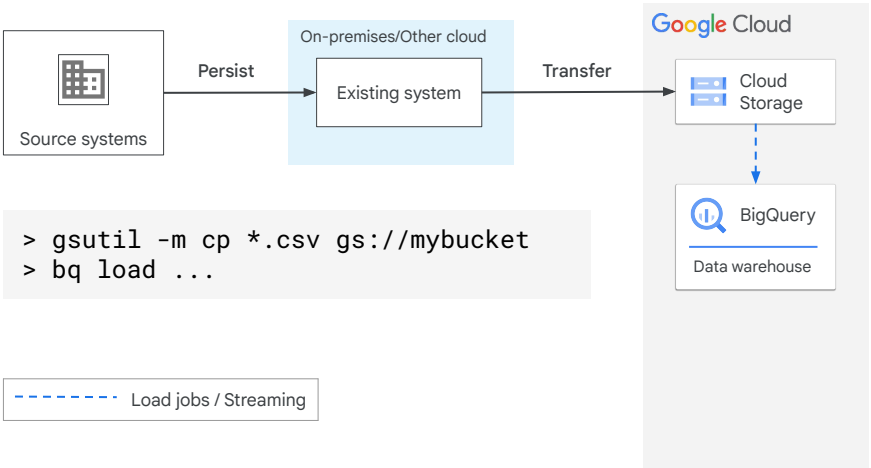
- CSV
- NEWLINE_DELIMITED_JSON
- AVRO
- DATASTORE_BACKUP
- PARQUET
- ORC

Most common is loading data into BigQuery tables (batch, periodic)



Loading data into BigQuery tables (batch, periodic) offers the best performance.

Loading data through Cloud Storage




```
> gsutil -m cp *.csv gs://mybucket
> bq load ...
```

Load jobs / Streaming

Lab Intro

Working with JSON and Array Data in BigQuery