

Ali Murtadho

linkedin : <https://www.linkedin.com/in/ali-murtadho>

telegram : t.me/alimurtadho_id

github : [@alimurtadho](https://github.com/alimurtadho)

medium : medium.com/@dho_aldho

GIMME



CODE

Outlines

- Introduction to ETL
- Extract
- Load



If you could be anywhere
in the world right now,
where would you be?

How Integrate Data to Data Warehouse?



ETL is a data integration process that involves extracting data from various sources, transforming it into a structured format, and loading it into a target system for analysis or storage.



- **In the extraction phase,** data is pulled from various sources, such as databases, files, or web services.
- **In the transformation phase,** the extracted data is cleaned, filtered, aggregated, or joined to prepare it for loading into the target system.
- **In the loading phase,** the transformed data is inserted or updated into the target system, such as a data warehouse or a database.

Where is ETL used?

Data Migration Scheme

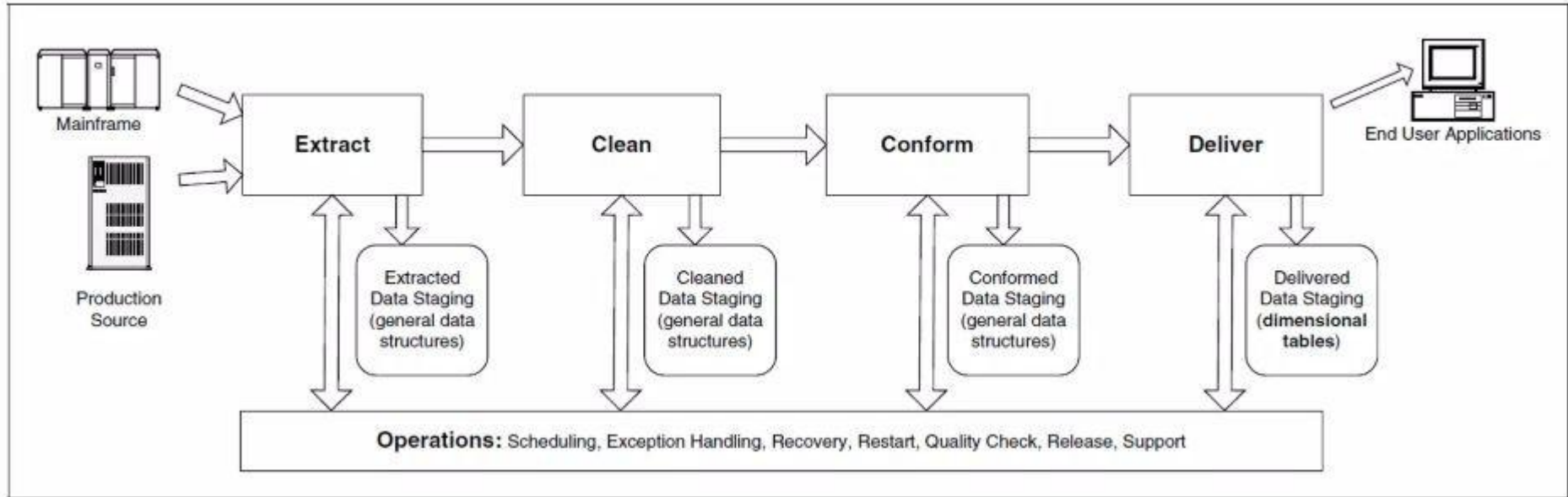




Benefits of ETL Process

- **Improved data quality:** Ensures that the data in the data warehouse is accurate, complete, and up-to-date.
- **Better data integration:** Helps to integrate data from multiple sources and systems, making it more accessible and usable.
- **Increased data security:** Help to improve data security by controlling access to the data warehouse and ensuring that only authorized users can access the data.
- **Improved scalability:** Help to improve scalability by providing a way to manage and analyze large amounts of data.
- **Increased automation:** Automate and simplify the ETL process, reducing the time and effort required to load and update data in the warehouse.

How to Implement ETL System



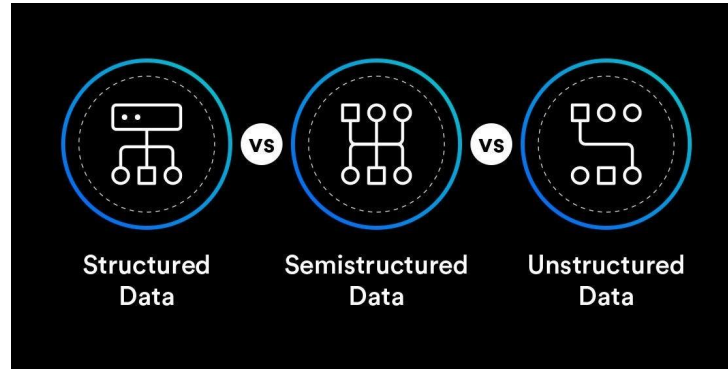


How to Implement ETL System

- **Define requirements:** Clearly identify the data sources, target system, and desired outcomes.
- **Design the ETL process:** Map out the extraction, transformation, and loading stages, including data cleansing and formatting steps.
- **Choose ETL tools:** Select appropriate software tools to automate and manage the ETL process.
- **Implement and monitor the system:** Develop and deploy the ETL process, followed by ongoing monitoring and maintenance.



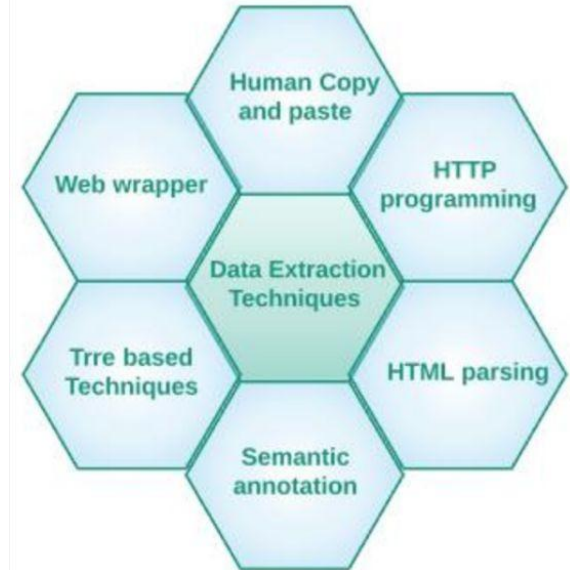
Extract



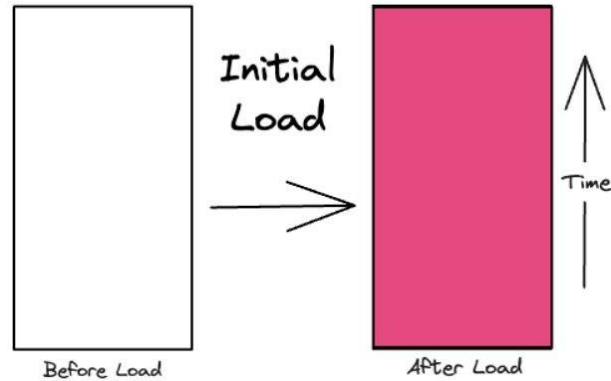
- Structured Data: defined format and organized in a row-column structure, stored in databases, spreadsheets.
- Semi-structured Data: not predefined, but somewhat organized, stored formats like XML, JSON, CSV.
- Unstructured Data: no predefined structure and can't be organized in any particular way, stored in files, multimedia files, emails.
- Metadata: data that describes other data, includes data dictionaries, database schemata, and catalogues.

Extraction Techniques

- Extraction techniques are methods used to retrieve data from various sources for use in the ETL process.
- Common extraction techniques include:
 - **Full extraction:** Extracts all data from a source at a specific point in time.
 - **Incremental extraction:** Extracts only the data that has changed since the last extraction.
 - **Real-time extraction:** Extracts data continuously as it is generated by the source.

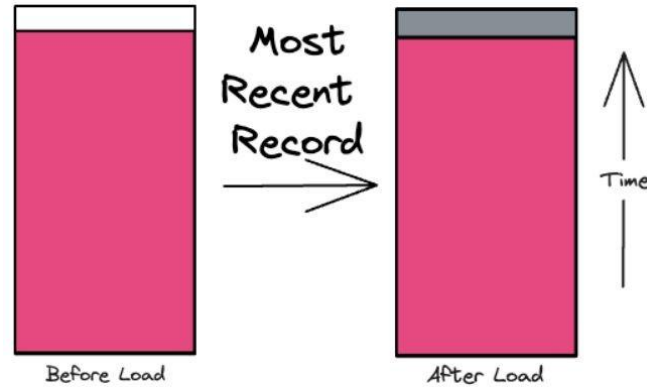


Extract Method: Full Refresh



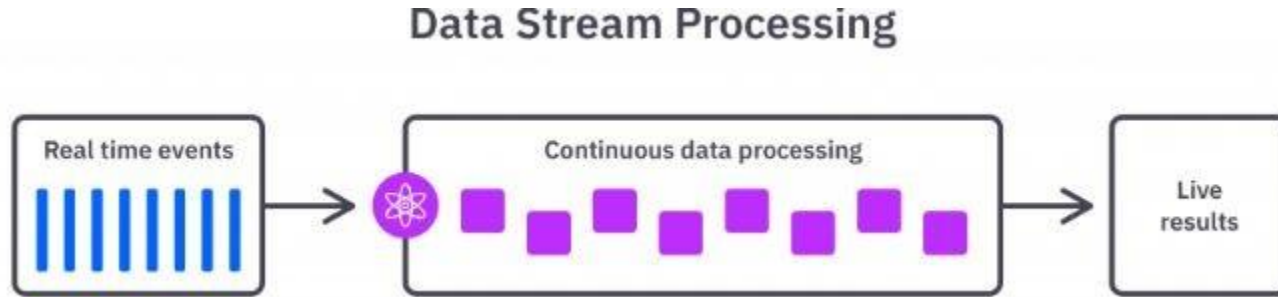
- Full refresh is a data extraction method that involves deleting all the data in a database system and reloading the data from the source system.
- This method is usually employed in data warehousing and data migration scenarios.
- This method provides accurate data and is often used for regulatory or financial reporting in organizations.

Extract Method: Incremental



- Incremental is a method of extraction the changes made to a database without affecting the existing data.
- This method is typically used when only new data or modifications need to be loaded into the database system.
- Incremental loading is faster than full refresh since it only processes the changes made to the data.

Extract Method: Streaming



- Streaming is a method of extraction data in a continuous and real-time manner.
- It is best suited for systems that require high-speed data processing or require up-to-date information.
- The architecture required for streaming is different than traditional batch loading, and it enables databases to be updated in real-time.

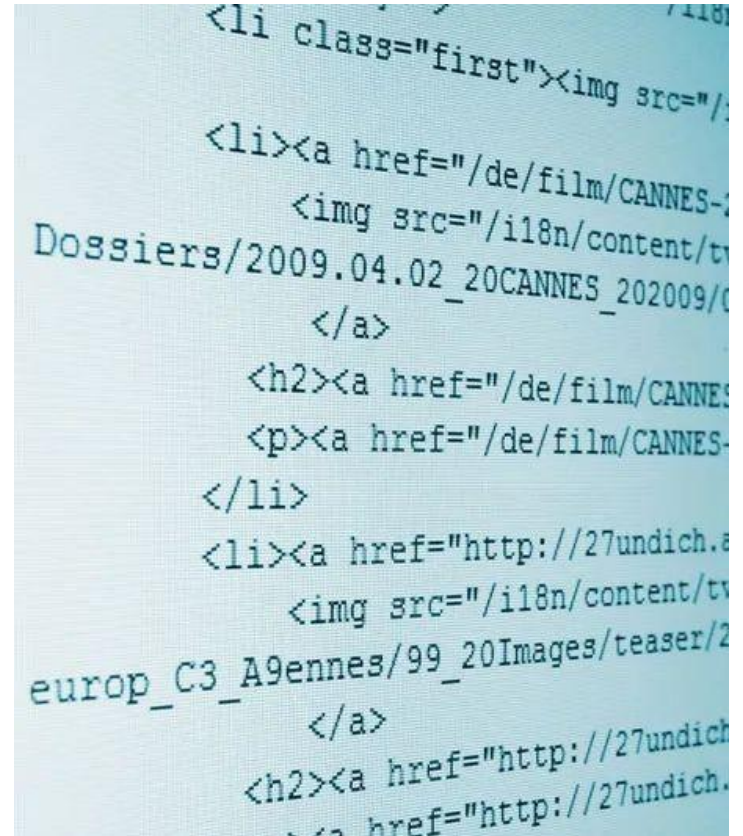
Choosing the Right Extraction Technique

- **Data volume:** For large datasets, incremental or real-time extraction might be more efficient.
- **Data update frequency:** For frequently changing data, incremental or real-time extraction is preferred.
- **Downtime tolerance:** If downtime for the source system is critical, full extraction might be less suitable.
- **Resource constraints:** Consider the processing power and storage capacity available for the chosen technique.



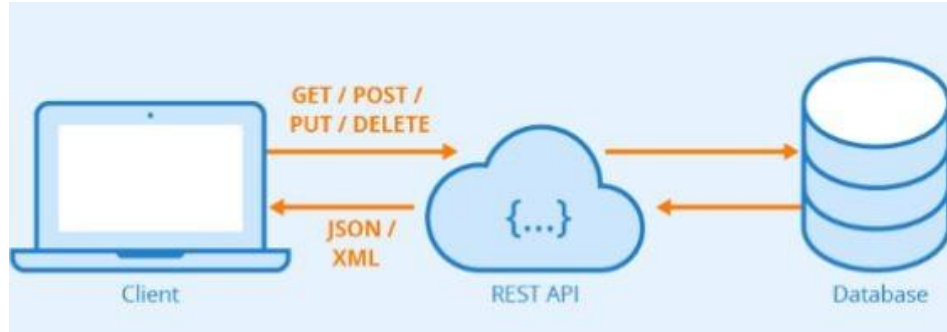
Extraction Technique: Web Scraping

- Web scraping is the process of extracting data from websites by automated means.
- Web scraping tools can extract data from various sources, including HTML, XML, and JSON.
- Data can be retrieved from different types of websites, including e-commerce websites, news websites, social media platforms, and more.
- Web scraping can be done using various programming languages, including Python, R, and JavaScript.



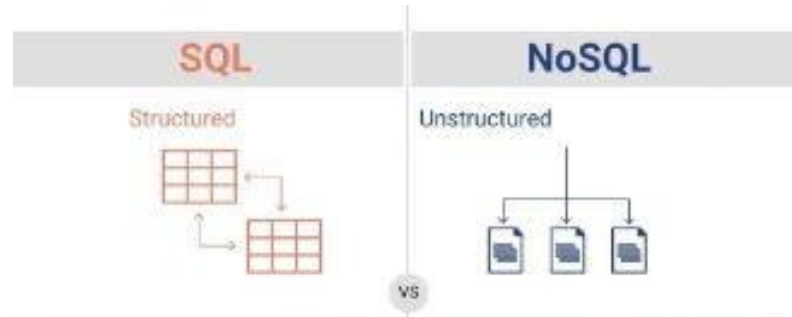
```
<li class="first">  
    <h2><a href="/de/film/CANNES  
    <p><a href="/de/film/CANNES-  
</li>  
<li><a href="http://27undich.e  
    <img src="/i18n/content/t  
europ_C3_A9ennes/99_20Images/teaser/2  
    </a>  
    <h2><a href="http://27undich  
    <a href="http://27undich.
```

Extraction Technique: API



- API stands for Application Programming Interface, which is a set of protocols used to build software applications.
- APIs allow developers to access data and functionality from other applications or services.
- Data can be retrieved in different formats, including JSON, XML, or CSV.
- APIs are used to extract data from various sources, including social media platforms, weather services, e-commerce websites, and more.

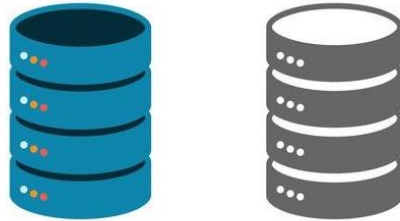
Extraction Technique: Query



- Querying is the process of retrieving data from databases using SQL (Structured Query Language) or NoSQL (Not Only SQL) databases.
- SQL is used to extract data from relational databases, including MySQL, PostgreSQL, Oracle, and more.
- NoSQL databases, including MongoDB, Cassandra, and Couchbase, are used to extract data from non-relational databases.
- Querying is used to extract data from various sources, including transactional databases, log files, and more.



Load



- The initial load is the first-time transfer of all relevant data from the source system(s) to the target system.
- It establishes the baseline for future data analysis and reporting.
- This process can be time-consuming and resource-intensive, especially for large datasets.



- The delta load focuses on transferring only the data that has changed since the last successful load.
- It's an efficient approach for frequently updated data sources.
- Requires mechanisms to identify and track changes in the source data.

Itu dia rangkaian proses ETL dari awal sampai akhir~

Tadi kita sempat bahas sekilas, bahwa penulisan query untuk ETL bisa kita lakukan dengan **coding manual, maupun menggunakan ETL tools.**

Apa sih bedanya? Yuk kita bahas!



Pertama, ETL dengan coding manual

Umumnya, cara ini dipilih oleh perusahaan yang belum memiliki banyak budget dan **baru mulai membuat data warehouse**.

Bahasa pemrograman yang biasa digunakan untuk mengatur perintah ETL adalah Bash, Python, Perl, C++, dan Java.

Meskipun demikian, cara ini juga punya kekurangan!



Mengandalkan coding manual untuk **jangka panjang** justru bisa menimbulkan **biaya operasional yang lebih besar**. Kenapa demikian?

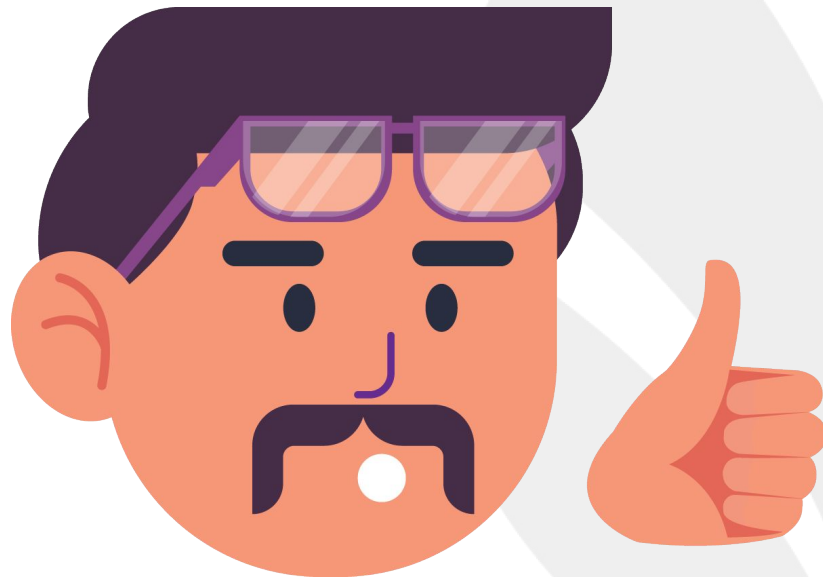
Alasannya, proses coding manual butuh **maintenance yang lebih rumit**, sehingga untuk perubahan atau penyesuaian kecil saja butuh usaha yang berat dan waktu yang lama.



Selain itu, bayangkan jika **data yang harus diproses jumlahnya sangat banyak.**

Pasti butuh waktu yang sangat lama dan melelahkan!

Nah, untungnya, kita bisa memanfaatkan apa yang disebut dengan **ETL Tools.**



Opsi kedua, ETL dengan Data Integration Tools

Banyak perusahaan yang menggunakan tools skala besar seperti [Informatica PowerCenter](#) atau [IBM DataStage](#).

Tapi, banyak juga yang menggunakan tools lain seperti Fivetran, Denodo, Xplenty, atau bahkan yang open source seperti Pentaho atau Talend.

Lebih lengkapnya, klik [link daftar ETL tools](#) berikut ini ya!

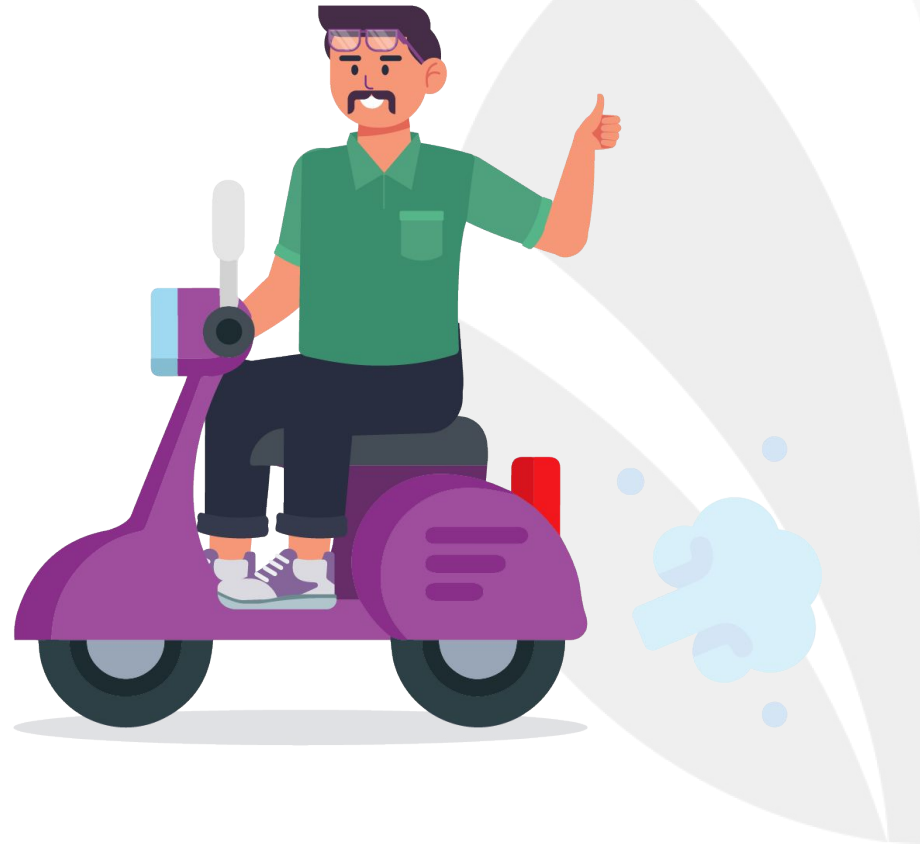


Banyak bisanya~

Sebenarnya, ETL tools ini berbasis logika yang sama dengan SQL.

Bedanya, tools ini **punya segudang kelebihan** yang membantu memudahkan proses ETL.

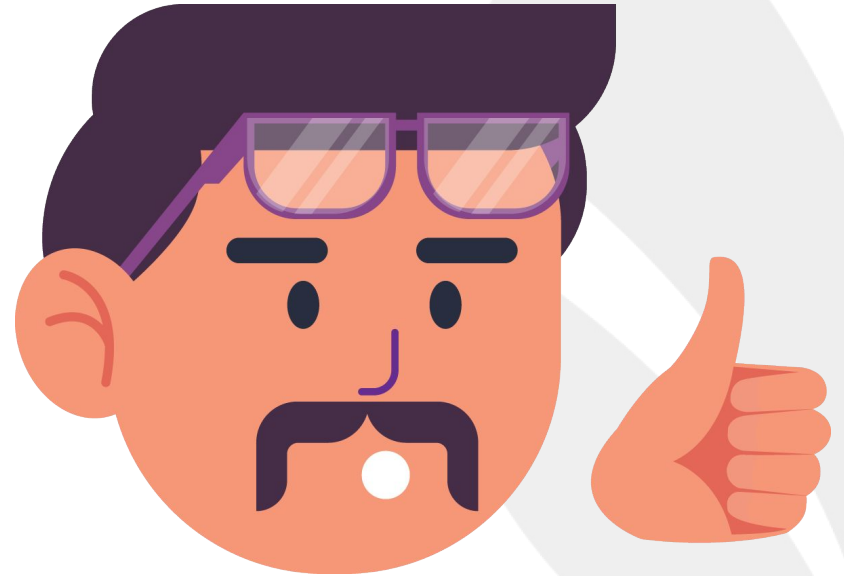
Yuk kita simak apa saja kelebihannya!



1 Performa yang kuat

Banyak ETL tools menyediakan engine dan hardware tersendiri yang kuat, serta dilengkapi fitur [parallel processing](#) yang membantu proses ETL semakin cepat.

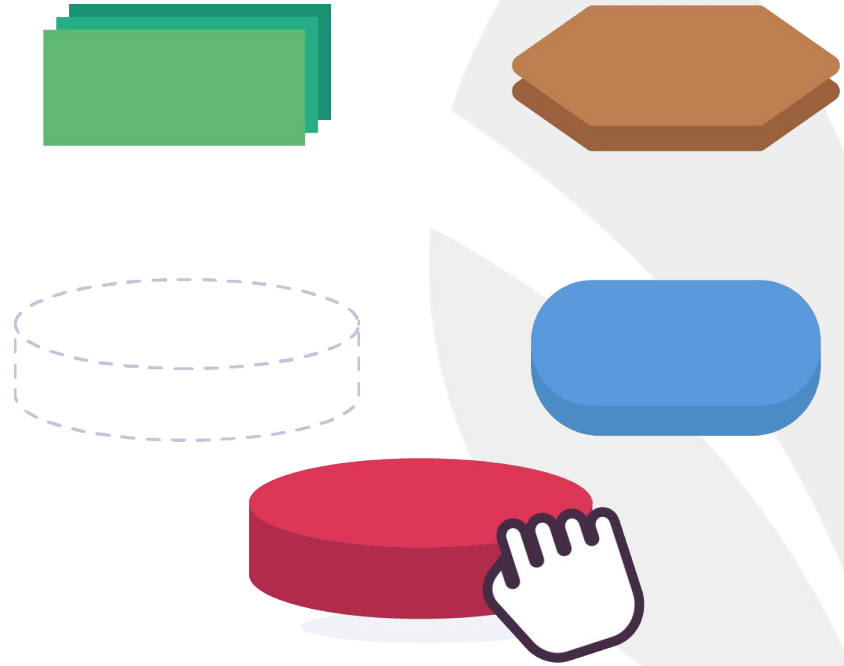
Jika kita hanya menggunakan coding, proses-proses tersebut dilakukan langsung oleh data warehouse, sehingga biasanya lambat dan dapat menghambat performa data warehouse itu sendiri.



② Menyediakan fitur drag-and-drop

ETL tools modern menyediakan **GUI (Graphical User Interface)** sehingga memudahkan seseorang mendesain proses ETL meskipun ia memiliki skill programming yang minimal.

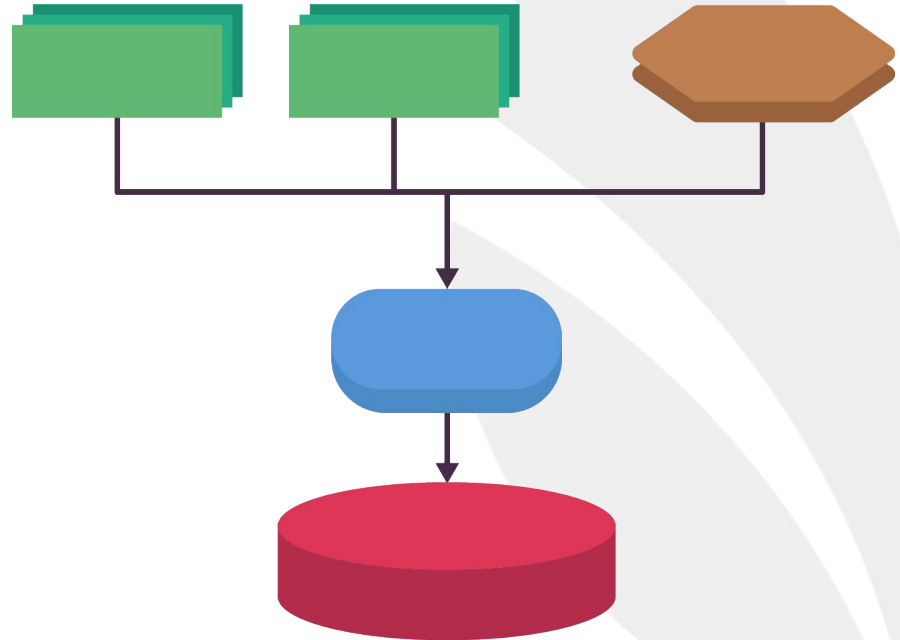
Daripada repot-repot coding, kita bisa tinggal drag-and-drop aja deh!



③ Menampilkan flow data secara visual

Masih berkaitan dengan **GUI**, dengan ETL tools kita bisa melihat desain dan flow dari ETL **dalam bentuk diagram-diagram yang mudah dimengerti.**

Dengan begini, kita bisa lebih mendapat gambaran untuk memahami logika di balik ETL tersebut, dan tentunya akan membantu meminimalisir kesalahan.



④ Menyediakan fitur data cleansing dan validasi

Banyak ETL tools yang menyediakan **pengaturan** atau **bantuan fitur validasi** untuk menjaga kualitas data yang diproses.

Contohnya adalah fitur-fitur yang membantu **pencegahan duplikasi, pemeriksaan kolom yang kosong**, dan sebagainya,

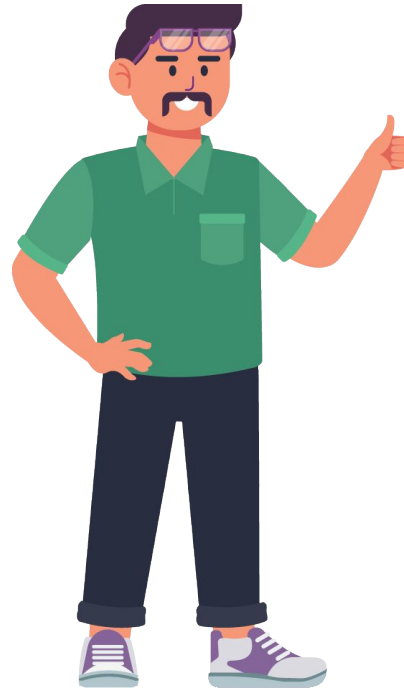


5 Mempercepat aktivitas operasional

Menggunakan tools dapat **mempersingkat banyak waktu** karena ada fitur monitoring atas pipeline yang berjalan.

ETL tools dapat menunjukkan di mana error terjadi, sekaligus memberikan rekomendasi untuk memperbaikinya.

Selain itu, melakukan perubahan/modifikasi terhadap pipeline yang sudah berjalan pun lebih mudah dan cepat dengan ETL tools~

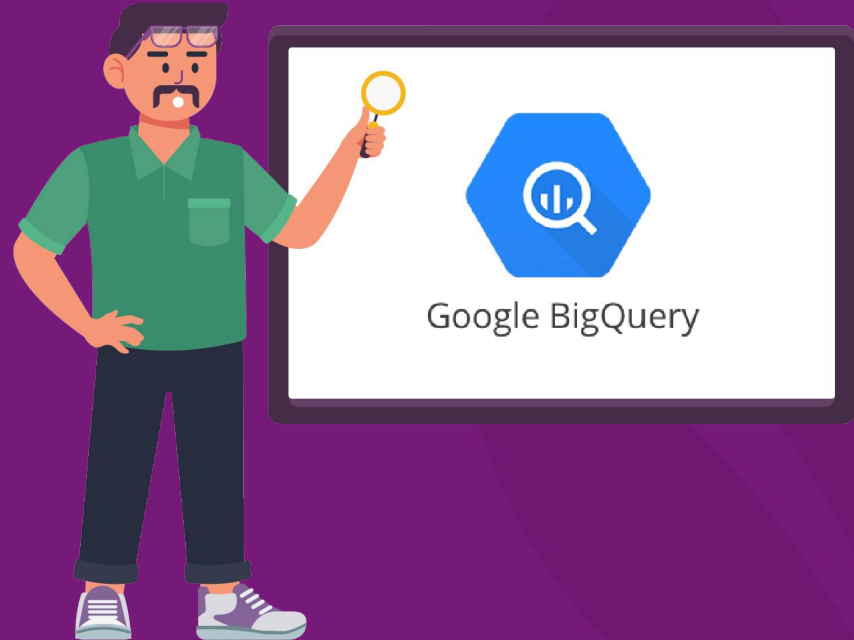


Tabel perbandingan ETL tools dan ETL manual

ETL dengan tools	ETL dengan coding manual
Performa lebih kuat	Hemat biaya jika untuk jangka pendek
Punya fitur drag and drop	Operasional jangka panjang jadi lebih boros biaya dibanding ETL tools
Flow data ditampilkan secara visual	Maintenance rumit
Menyediakan fitur data cleansing dan validasi	Operasi untuk data yang sangat banyak memakan waktu panjang
Aktivitas operasional lebih cepat	

Nah, biar kita makin kenal dengan ETL, Mas Gun bakal kasih tahu tutorial sederhana untuk melakukan **ETL dengan BigQuery** nih gengs!

Meskipun BigQuery sebenarnya adalah data warehouse, tetapi kita juga bisa melakukan ekstraksi data, transformasi data menggunakan SQL, dan menyimpan hasilnya ke langsung dalam tabel.



Masih ingat kan tentang BigQuery?

Sebagai bagian dari Google Cloud Platform (GCP), **BigQuery** adalah **data warehouse** berbasis **cloud** yang mampu memproses big data dengan efektif dan cepat.

Hmm, apa sih artinya “berbasis cloud”? Coba pilih jawaban yang menurutmu paling tepat!

- A** Memerlukan server fisik
- B** Tidak memerlukan server fisik
- C** Kantornya tinggi menembus awan

Jawabannya, B ya gengss!

Data warehouse berbasis cloud artinya kita tidak perlu menyiapkan server fisik~

A Memerlukan server fisik

B Tidak memerlukan server fisik

C Kantornya tinggi menembus awan

BigQuery adalah produk berbayar, tapi...

Kita tetap bisa pakai secara gratis dengan batas 10GB untuk penyimpanan data dan 1TB untuk pemrosesan data per bulan!

Untuk dapat menggunakan BigQuery, kita hanya perlu memiliki akun Google aktif saja.



Google
BigQuery

Buat mulai tutorialnya, klik [link berikut](#) ya!

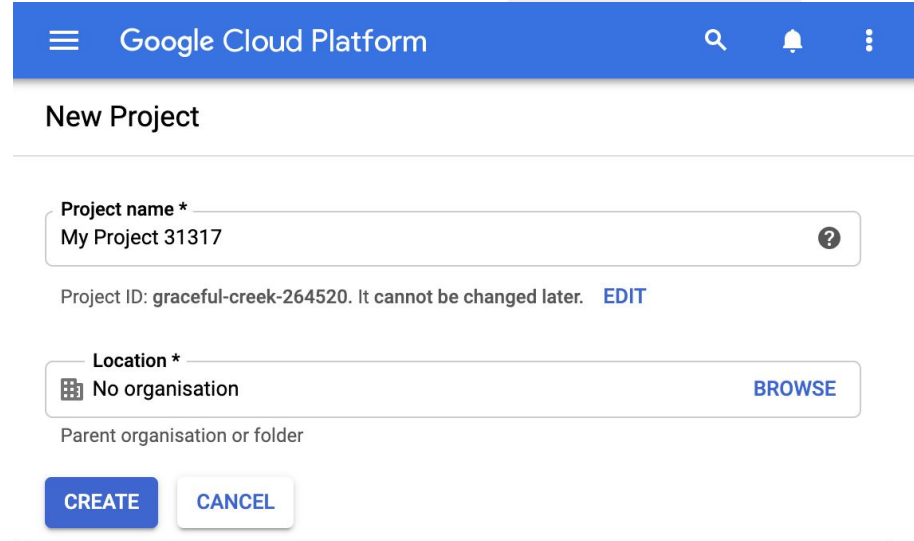
bit.ly/TutorialETL



Kamu juga bisa cek materi Chapter 5 Topic 2 lagi ya!

Langkah pertama, kita akan diminta untuk login menggunakan akun Google yang kita punya.

Lalu, ikuti prosesnya sampai diminta untuk membuat **project baru (new project)**.



The screenshot shows the 'New Project' page in the Google Cloud Platform console. At the top is a blue header with the Google Cloud Platform logo, a search icon, a notification bell, and a menu icon. Below the header, the title 'New Project' is displayed. The main form contains two input fields. The first field, labeled 'Project name *', has the text 'My Project 31317' and a help icon. Below it, the 'Project ID' is shown as 'graceful-creek-264520' with a note that it cannot be changed later and an 'EDIT' link. The second field, labeled 'Location *', has a dropdown menu showing 'No organisation' and a 'BROWSE' button. Below this field, the text 'Parent organisation or folder' is visible. At the bottom of the form are two buttons: 'CREATE' and 'CANCEL'.

Google Cloud Platform

New Project

Project name *
My Project 31317

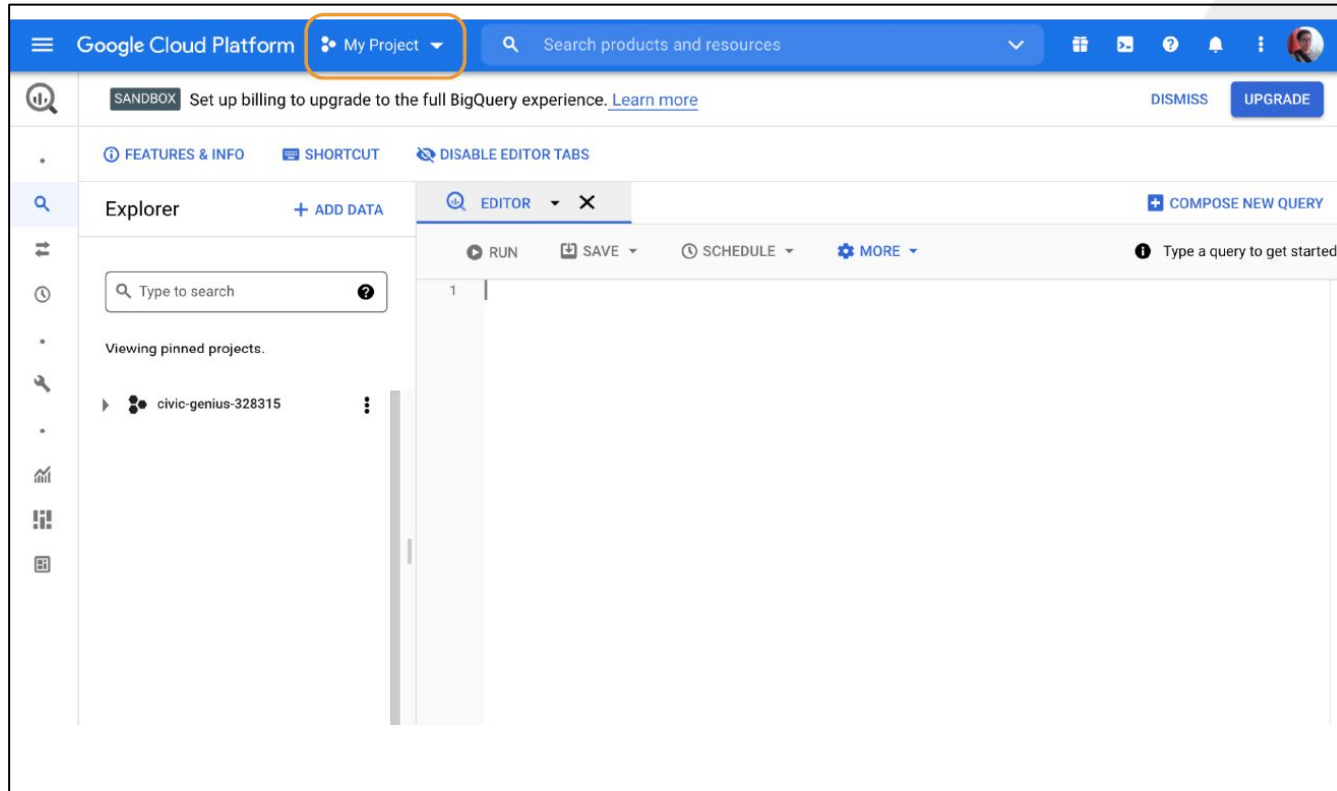
Project ID: graceful-creek-264520. It cannot be changed later. [EDIT](#)

Location *
No organisation [BROWSE](#)

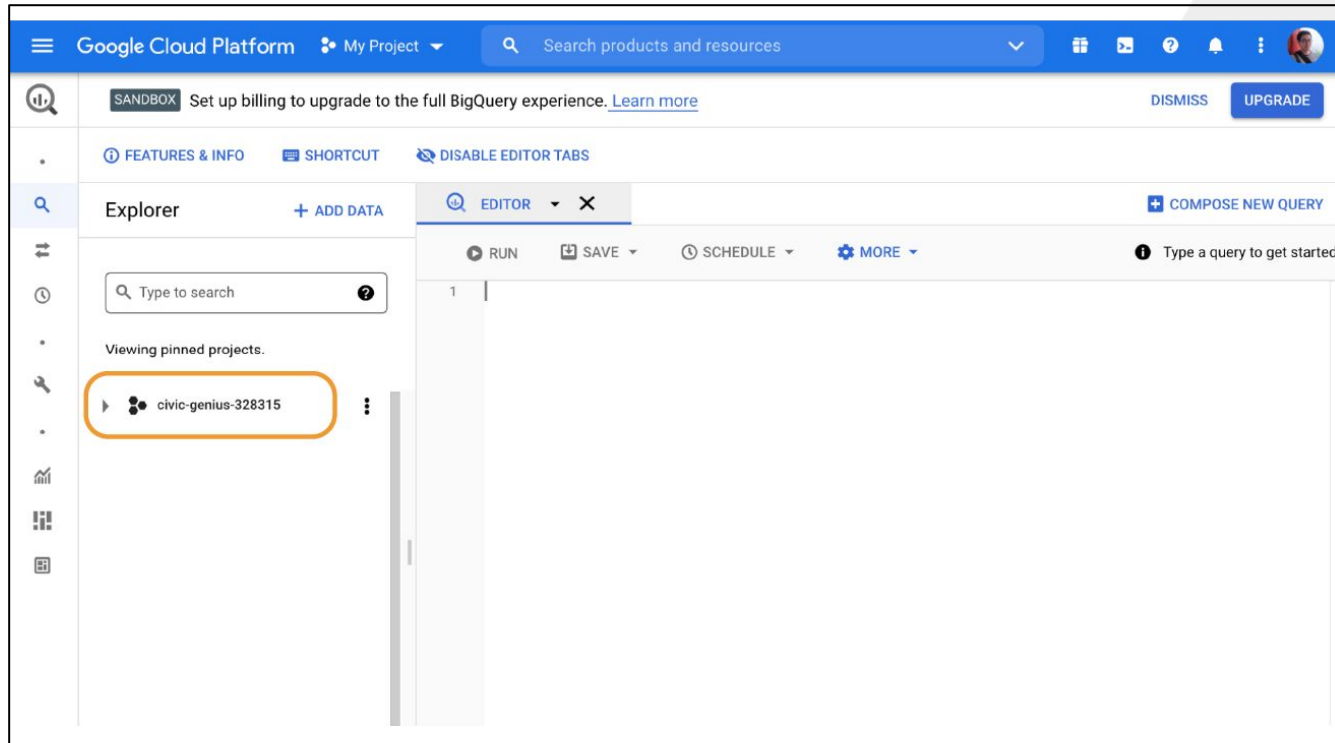
Parent organisation or folder

[CREATE](#) [CANCEL](#)

Setelah membuat project baru, tampilannya akan seperti ini. BigQuery siap digunakan!



Ini adalah **project id** yang diberikan oleh Google secara otomatis.
Setiap orang akan mendapatkan id yang berbeda-beda.



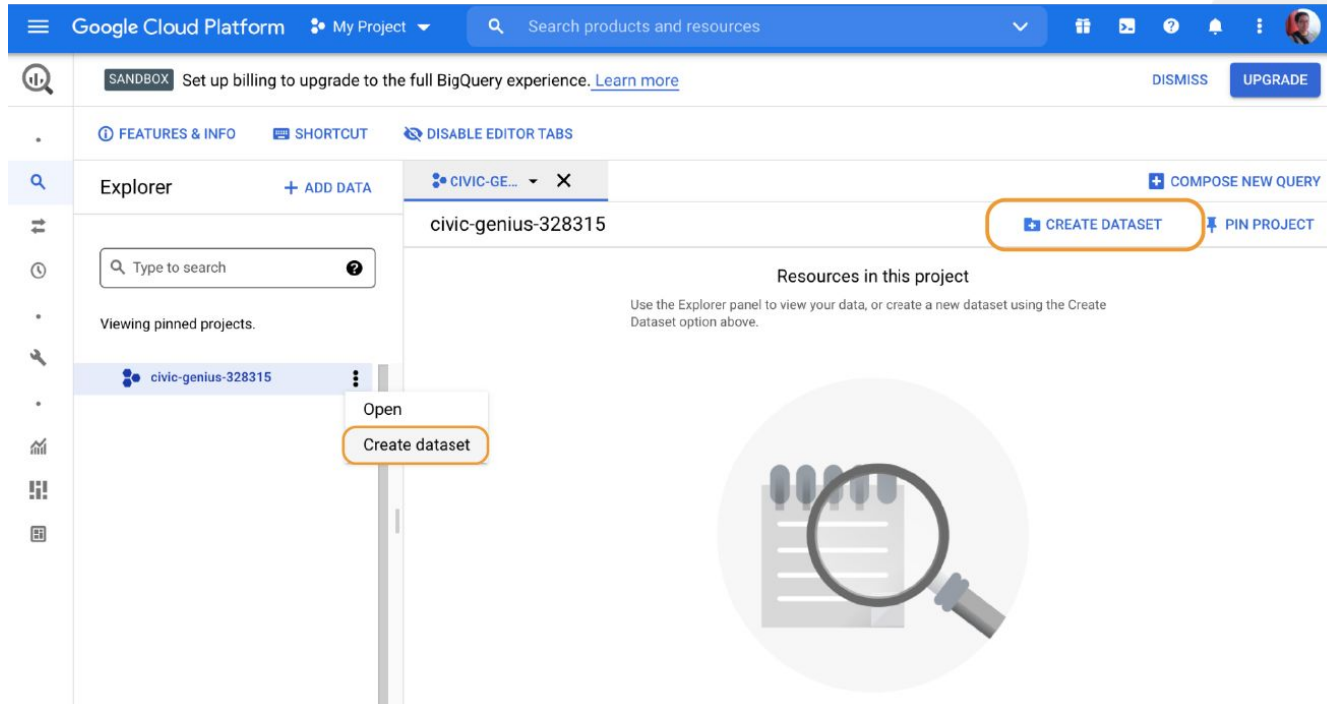
Selanjutnya, kita perlu **membuat dataset di dalam project.**

Dataset dalam BigQuery adalah tempat menyimpan dan mengelola tabel-tabel.



Google
BigQuery

Klik menu **Create Dataset** yang akan muncul kalau kita menekan **titik tiga di sebelah kanan project id** dalam panel **Explorer**. Atau, bisa juga pilih menu serupa yang ada di panel sebelah kanan.



Berikan id untuk dataset yang kita buat, lalu **pilih location** untuk tempat penyimpanan di cloud.

Create dataset

Letters, numbers, and underscores allowed

▼ ?

Default table expiration

☐ Enable table expiration ?

Days

Encryption

☒ Google-managed encryption key
No configuration required

☐ Customer-managed encryption key (CMEK)
Manage via Google Cloud Key Management Service

CREATE DATASET

CANCEL

Sebagai contoh, kita bisa memberikan id ``my_first_dataset`` seperti ini.

Pilih lokasi, lalu klik **CREATE DATASET**.

Create dataset

Dataset ID *

my_first_dataset

Letters, numbers, and underscores allowed

Data location

us-central1 (Iowa) ▼ ?

Default table expiration

☐ Enable table expiration ?

Default maximum table age

Days

Encryption

☒ Google-managed encryption key

No configuration required

☐ Customer-managed encryption key (CMEK)

Manage via Google Cloud Key Management Service

CREATE DATASET

CANCEL

Setelah berhasil, kita bisa menemukan dataset yang baru kita buat di dalam panel Explorer!

Explorer

[+ ADD DATA](#)

Viewing pinned projects.



civic-genius-328315



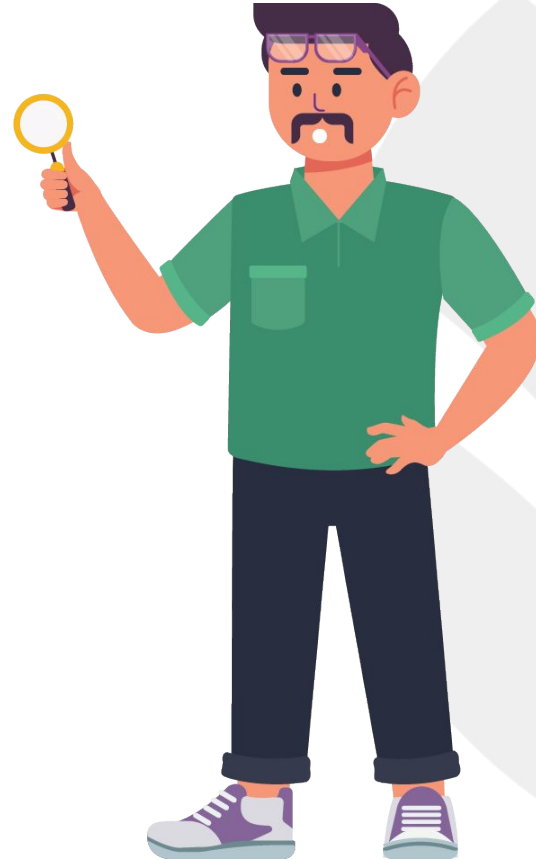
my_first_dataset



Nah, pada praktik ETL kali ini, Mas Gun udah menyiapkan dataset yang akan kita pakai latihan, yaitu data **Sample Superstore**.

Linknya bisa diakses di sini gengs!

[Link Sample Superstore](#)



Ini dia tampilan dari Sample Superstore yang akan kita pakai!

Sample - Superstore ☆ 📄 ☁

File Edit View Insert Format Data Tools Add-ons Help Last edit was on April 29

100% \$ % 0.00 123 Calibri 10 B I G A

A1	Order Date														
	A	B	C	D	E	F	G	H	I	J	K	L	M		
1	Order Date	Order ID	Category	Sub-Category	Product Name	Segment	Country	Region	State	City	Postal Code	Customer Name	Ship Mode	Ship	
2	2018-11-08	CA-2018-152156	Furniture	Bookcases	Bush Somerset Co	Consumer	United States	South	Kentucky	Henderson	42420	Claire Gute	Second Class		
3	2018-11-08	CA-2018-152156	Furniture	Chairs	Hon Deluxe Fabric	Consumer	United States	South	Kentucky	Henderson	42420	Claire Gute	Second Class		
4	2018-06-12	CA-2018-138688	Office Supplies	Labels	Self-Adhesive Add	Corporate	United States	West	California	Los Angeles	90036	Darrin Van Huff	Second Class		
5	2017-10-11	US-2017-108966	Furniture	Tables	Bretford CR4500 S	Consumer	United States	South	Florida	Fort Lauderdale	33311	Sean O'Donnell	Standard Class		
6	2017-10-11	US-2017-108966	Office Supplies	Storage	Eldon Fold 'N Roll	Consumer	United States	South	Florida	Fort Lauderdale	33311	Sean O'Donnell	Standard Class		
7	2016-06-09	CA-2016-115812	Furniture	Furnishings	Eldon Expressions	Consumer	United States	West	California	Los Angeles	90032	Brosina Hoffman	Standard Class		
8	2016-06-09	CA-2016-115812	Office Supplies	Art	Newell 322	Consumer	United States	West	California	Los Angeles	90032	Brosina Hoffman	Standard Class		
9	2016-06-09	CA-2016-115812	Technology	Phones	Mitel 5320 IP Pho	Consumer	United States	West	California	Los Angeles	90032	Brosina Hoffman	Standard Class		
10	2016-06-09	CA-2016-115812	Office Supplies	Binders	DXL Angle-View B	Consumer	United States	West	California	Los Angeles	90032	Brosina Hoffman	Standard Class		
11	2016-06-09	CA-2016-115812	Office Supplies	Appliances	Belkin F5C206VTE	Consumer	United States	West	California	Los Angeles	90032	Brosina Hoffman	Standard Class		
12	2016-06-09	CA-2016-115812	Furniture	Tables	Chromcraft Rectar	Consumer	United States	West	California	Los Angeles	90032	Brosina Hoffman	Standard Class		
13	2016-06-09	CA-2016-115812	Technology	Phones	Konitel 250 Confe	Consumer	United States	West	California	Los Angeles	90032	Brosina Hoffman	Standard Class		
14	2019-04-15	CA-2019-114412	Office Supplies	Paper	Xerox 1967	Consumer	United States	South	North Carolina	Concord	28027	Andrew Allen	Standard Class		
15	2018-12-05	CA-2018-161389	Office Supplies	Binders	Fellowes PB200 PI	Consumer	United States	West	Washington	Seattle	98103	Irene Maddox	Standard Class		
16	2017-11-22	US-2017-118983	Office Supplies	Appliances	Holmes Replacem	Home Office	United States	Central	Texas	Fort Worth	76106	Harold Pawlan	Standard Class		
17	2017-11-22	US-2017-118983	Office Supplies	Binders	Storex DuraTech R	Home Office	United States	Central	Texas	Fort Worth	76106	Harold Pawlan	Standard Class		
18	2016-11-11	CA-2016-105893	Office Supplies	Storage	Stur-D-Stor Shelv	Consumer	United States	Central	Wisconsin	Madison	53711	Pete Kriz	Standard Class		
19	2016-05-13	CA-2016-167164	Office Supplies	Storage	Fellowes Super St	Consumer	United States	West	Utah	West Jordan	84084	Alejandro Grove	Second Class		
20	2016-08-27	CA-2016-143336	Office Supplies	Art	Newell 341	Consumer	United States	West	California	San Francisco	94109	Zuschuss Donatell	Second Class		
21	2016-08-27	CA-2016-143336	Office Supplies	Art	Newell 341	Consumer	United States	West	California	San Francisco	94109	Zuschuss Donatell	Second Class		
22	2016-08-27	CA-2016-143336	Office Supplies	Art	Newell 341	Consumer	United States	West	California	San Francisco	94109	Zuschuss Donatell	Second Class		
23	2016-08-27	CA-2016-143336	Technology	Phones	Cisco SPA 501G IP	Consumer	United States	West	California	San Francisco	94109	Zuschuss Donatell	Second Class		
24	2016-08-27	CA-2016-143336	Technology	Phones	Cisco SPA 501G IP	Consumer	United States	West	California	San Francisco	94109	Zuschuss Donatell	Second Class		

Sample - Superstore Explore

Tujuan dari praktik ETL kita kali ini adalah **memuat data Sample Superstore ke dalam BigQuery.**

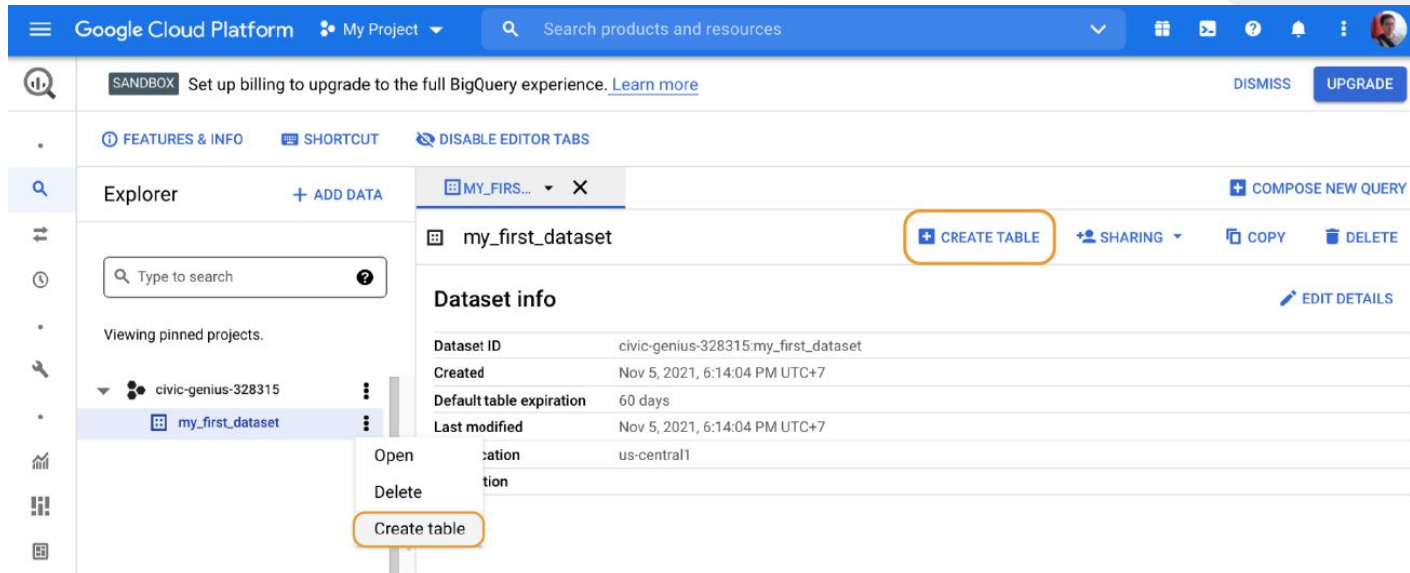
Artinya, data tersebut akan kita coba masukkan ke dalam tabel di dataset yang tadi sudah kita buat.

Yuk kita mulai langkah-langkahnya!



Langkah Pertama

Klik menu **Create Table** yang akan muncul kalau kamu menekan titik tiga di sebelah kanan `my_first_dataset`` dalam panel Explorer. Menu ini bisa diakses juga di panel sebelah kanan.



Di halaman ini, kita perlu menetapkan beberapa hal berikut:

1. **Source:** sumber data
2. **Destination:** nama tabel tujuan
3. **Schema:** skema dari data

Kita bahas satu per satu ya~

Create table

Source

Create table from
Empty table

Destination

Project *
civic-genius-328315 [BROWSE](#)

Dataset ID *
my_first_dataset

Table name *

Unicode letters, marks, numbers, connectors, dashes or spaces allowed.

Table type
Native table

Schema

☐ Edit as text

+

Partition and cluster settings

Partitioning
No partitioning

[CREATE TABLE](#) [CANCEL](#)

1 Sumber data

Data **Sample Superstore** yang mau kita pakai tersedia di Google Drive dan berbentuk Google Sheets, sehingga di sini kita pilih:

- Create table from: **Drive**
- File format: **Google Sheet**

Kemudian kita salin [link Sample Superstore](#) dan masukkan pada bagian **Select Drive URL**.

Source

Create table from	Drive
Select Drive URI *	https://docs.google.com/spreadsheets/d/1QoucX508yEay_Gbtk4B4xUJhWdvZdr0FOFa5nl9jEDI
File format	Google Sheet
Sheet range	

2 Nama tabel tujuan

Di sini kita beri nama tabel tujuan **`sample_superstore`**. Nama tabel yang dianjurkan adalah nama yang sebisa mungkin mewakili isi dari tabel itu.

Destination

Project *	civic-genius-328315	BROWSE
Dataset ID *	my_first_dataset	
Table name *	sample_superstore	
<small>Unicode letters, marks, numbers, connectors, dashes or spaces allowed.</small>		
Table type	External table	▼ ⓘ

3 Skema data

Skema data adalah **daftar setiap kolom** dan **masing-masing tipe datanya**.

Contohnya bisa kamu lihat pada gambar ya!

Field name	Type
Order_Date	DATE
Order_ID	STRING
Category	STRING
Sub_Category	STRING
Product_Name	STRING
Segment	STRING
Country	STRING
Region	STRING
State	STRING
City	STRING
Postal_Code	INTEGER
Customer_Name	STRING

Ada cara cepat menentukan skema data nih gengs!

BigQuery memiliki fitur **Auto detect** yang bisa kita gunakan agar **skema terdefiniskan secara otomatis**. Caranya, kita cukup klik pada bagian checkbox “Auto detect”.

Tapi perlu dicatat, fitur ini akan **mendeteksi baris pertama untuk dijadikan nama kolom**. Maka dari itu, kita perlu menetapkan **Header rows to skip = 1** agar skema terdeteksi dengan benar~

Schema

☒ Auto detect

i Schema will be automatically generated.

Advanced options

☐ Unknown values **?**

Header rows to skip







1 **?**

Kalau sudah benar, tinggal klik **CREATE TABLE** dan tabel akan segera terlihat di panel Explorer setelah sukses terbuat! ✨

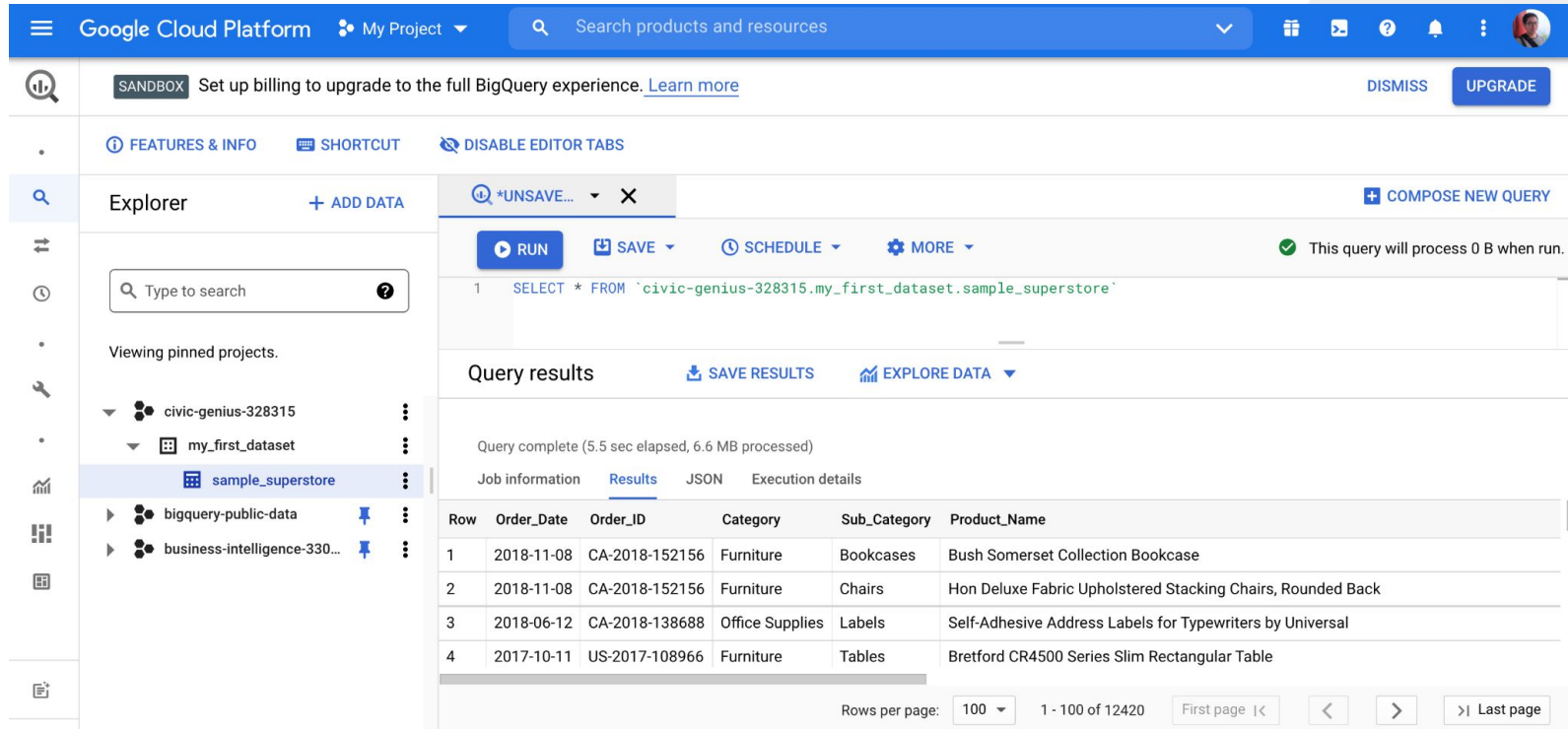
Explorer

[+ ADD DATA](#)

Viewing pinned projects.

- ▼  civic-genius-328315 
- ▼  my_first_dataset 
-  sample_superstore 

Kalau sudah, coba cek hasil ETL kamu dengan menjalankan query **SELECT**~



The screenshot shows the Google Cloud Platform BigQuery interface. The top navigation bar includes the Google Cloud Platform logo, 'My Project' dropdown, a search bar, and various utility icons. Below the navigation bar, there's a 'SANDBOX' banner with a message about upgrading to the full BigQuery experience. The main interface is divided into three sections: Explorer, Query Editor, and Query Results.

Explorer: Shows a list of projects and datasets. The selected project is 'civic-genius-328315', and the selected dataset is 'my_first_dataset'. The 'sample_superstore' table is highlighted.

Query Editor: Contains the SQL query: `SELECT * FROM `civic-genius-328315.my_first_dataset.sample_superstore``. The query is ready to be executed, with buttons for 'RUN', 'SAVE', 'SCHEDULE', and 'MORE'. A status message indicates: 'This query will process 0 B when run.'

Query Results: Shows the execution status: 'Query complete (5.5 sec elapsed, 6.6 MB processed)'. Below this, there's a table of results with columns: Row, Order_Date, Order_ID, Category, Sub_Category, and Product_Name.

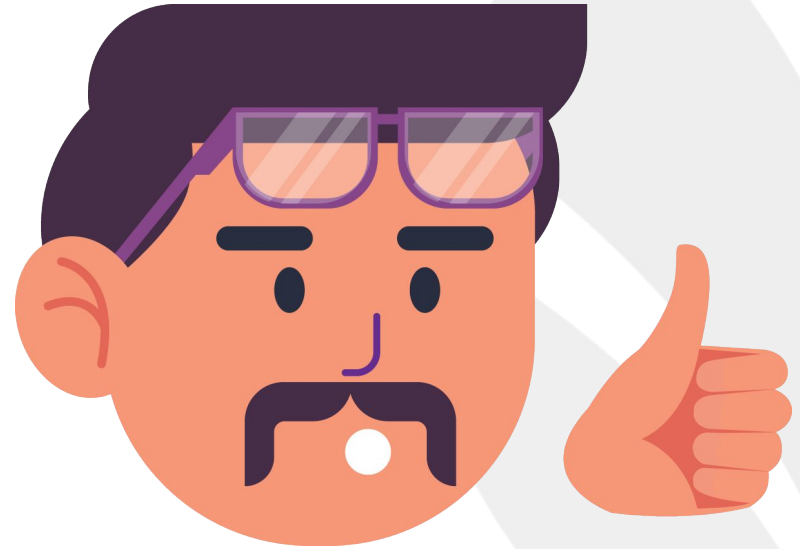
Row	Order_Date	Order_ID	Category	Sub_Category	Product_Name
1	2018-11-08	CA-2018-152156	Furniture	Bookcases	Bush Somerset Collection Bookcase
2	2018-11-08	CA-2018-152156	Furniture	Chairs	Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back
3	2018-06-12	CA-2018-138688	Office Supplies	Labels	Self-Adhesive Address Labels for Typewriters by Universal
4	2017-10-11	US-2017-108966	Furniture	Tables	Bretford CR4500 Series Slim Rectangular Table

At the bottom, there's a pagination bar showing 'Rows per page: 100', '1 - 100 of 12420', and navigation buttons for 'First page', '<', '>', and '>| Last page'.

Gampang kan gengs? 🧐

Di tutorial tadi, kita dimudahkan oleh fitur canggih BigQuery di mana kita hanya perlu klik beberapa tombol saja untuk melakukan ETL sederhana.

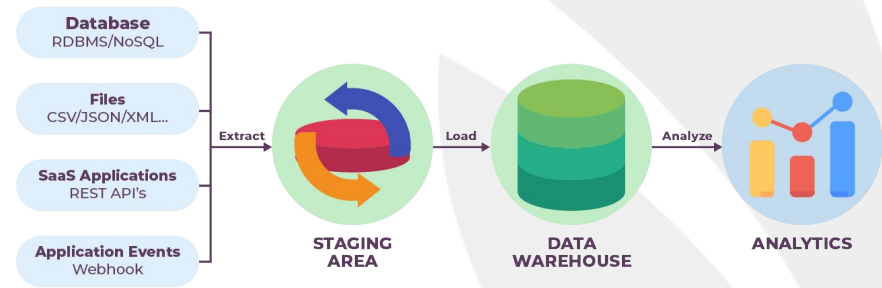
Perlu kamu catat, tadi kita hanya mempraktikkan tahapan **Extract** dan **Load** saja. Ini karena data pada Sample Superstore adalah data yang sudah rapi dan ditransformasi.



Pada kenyataannya nanti di proses ETL yang sesungguhnya, data biasanya belum tersedia secara rapi dan lengkap seperti Sample Superstore.

Selain itu, proses **Extract & Load** seperti tadi bisa jadi **perlu kita lakukan beberapa kali** apabila ada lebih dari satu sumber data mentah.

Kemudian setelah itu, baru kita akan perlu melakukan Transform menggunakan SQL, menyimpan hasilnya ke dalam tabel berbeda, untuk akhirnya dapat digunakan oleh user deh~



Saatnya pengerjaan studi kasus 🌟

Bukalah folder [berikut](#).

Di dalamnya terdapat 2 dataset. Tugasmu adalah melakukan proses ETL untuk membuat satu database baru yang bersih dan hanya berisi data negara Indonesia saja.

Gunakan Google BigQuery ya!

Itulah materi tentang **proses ETL beserta detail dari setiap tahapannya!**

Menurutmu, bagian mana dari ETL yang paling rumit? Mengapa? 🤔



Saatnya Quiz

installasi python windwos,
python mac

download vscode





Hands ON Session

Before We Start

Python

[Links Google](#)
[Colab](#)



Thank you

