

# **PART 3**

## **DATA INGESTION - AIRBYTE**

# Setup Environment & requirements.txt



Hal yang pertama dilakukan untuk mengerjakan tugas adalah clone repository yang sudah disediakan

**git clone** <https://github.com/Immersive-DataEngineer-Resource/ingestion-data.git>

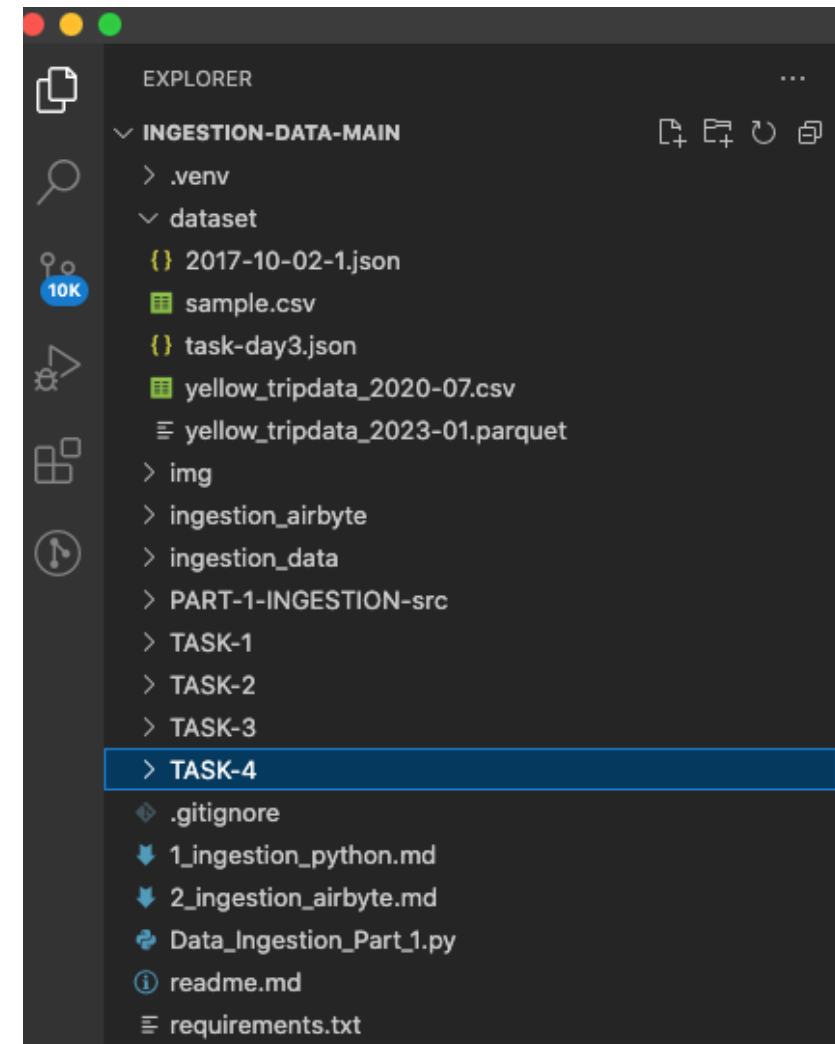
Bisa dilihat pada gambar disamping jika virtual environment sudah terinstall maka setelah itu bisa menggunakan perintah

**source .venv/bin/activate**

Mengaktifkan virtual environment sehingga semua paket Python yang diinstal atau digunakan adalah milik lingkungan terisolasi ini, yang menghindari dampak pada instalasi Python global.

Daftar requirements.txt mencakup paket-paket yang diperlukan untuk

**pip install -r requirements.**



# Deploy Postgresql, Citus and Airbyte Locally via Docker-Compose



```
(.venv) wartadi@Wartadis-MacBook-Pro ingestion_airbyte %
```

Pastikan environment sudah active dan di me direktori ingestion\_airbyte sebelum menjalankan perintah berikutnya

Run Postgresql, Citus and Airbyte locally via docker-compose

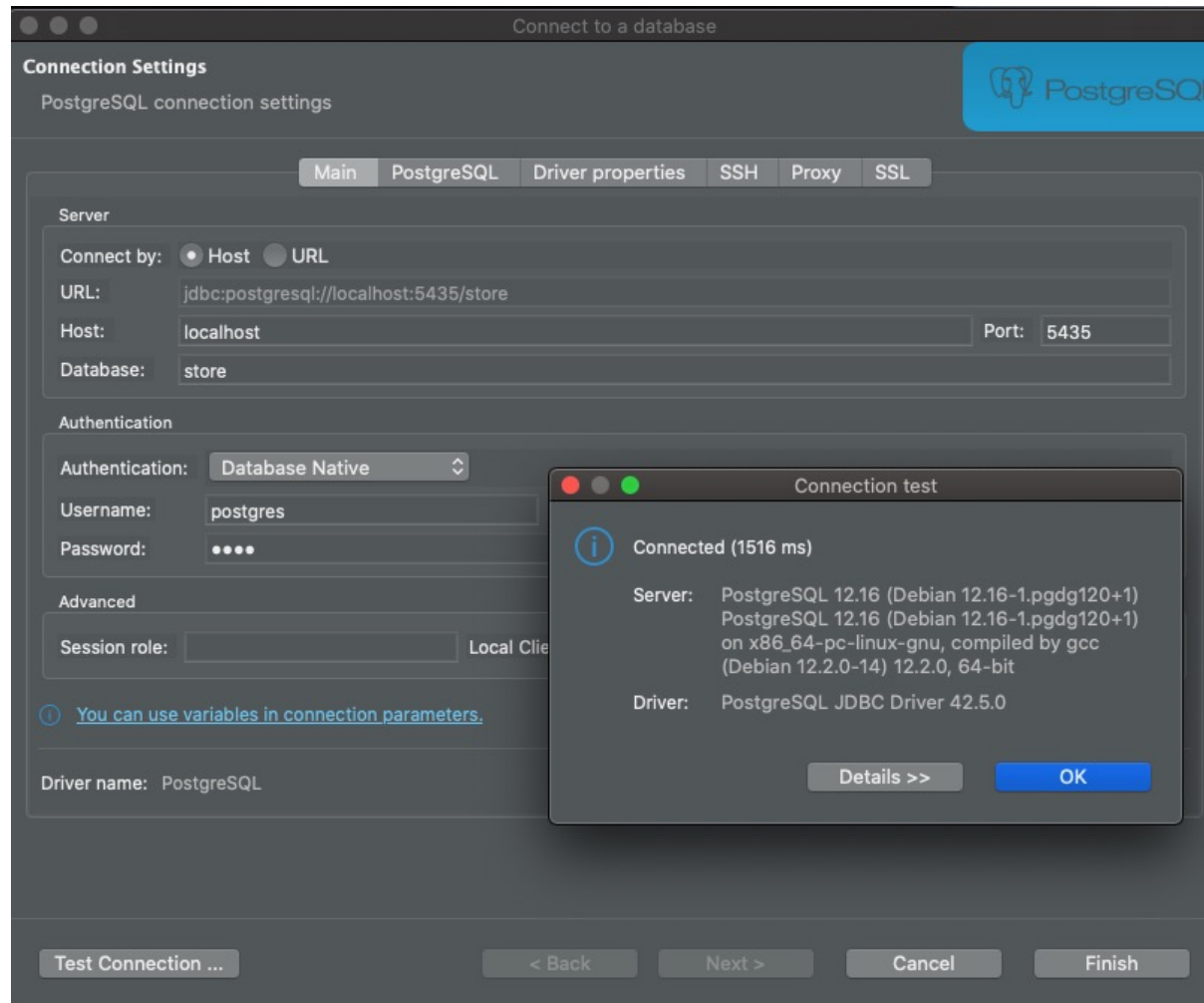
```
(.venv) wartadi@Wartadis-MacBook-Pro ingestion_airbyte % docker compose -f docker-compose.yml up -d
```

Berikut merukan hasil yang ditampilkan ketika menjalankan perintah diatas

```
[+] Running 19/19
# Network ingestion_airbyte_airbyte_internal Created 1.4s
# Network ingestion_airbyte_airbyte_public Created 1.0s
# Network ingestion_airbyte_default Created 0.6s
# Network ingestion_airbyte_postgres-network Created 0.7s
# Container init Exited 51.1s
# Container airbyte-temporal Started 42.9s
# Container airbyte-db Started 44.8s
# Container ingestion_airbyte Started 43.6s
# Container ingestion_airbyte_master Started 44.0s
# Container airbyte-bootloader Exited 310.2s
# Container ingestion_airbyte_manager Started 61.8s
# Container airbyte-connector-builder-server Started 316.0s
# Container airbyte-webapp Started 315.4s
# Container airbyte-worker Started 316.5s
# Container airbyte-cron Started 313.3s
# Container airbyte-server Started 315.5s
# Container airbyte-api-server Started 315.6s
# Container ingestion_airbyte-citus-worker-1 Started 75.9s
# Container airbyte-proxy Started 313.8s
(.venv) wartadi@Wartadis-MacBook-Pro ingestion_airbyte %
```

Create connection on DBeaver to Postgresql and Citus with these credentials:

Setelah Docker berjalan kita bisa membuat connection data base dengan postgreSQL

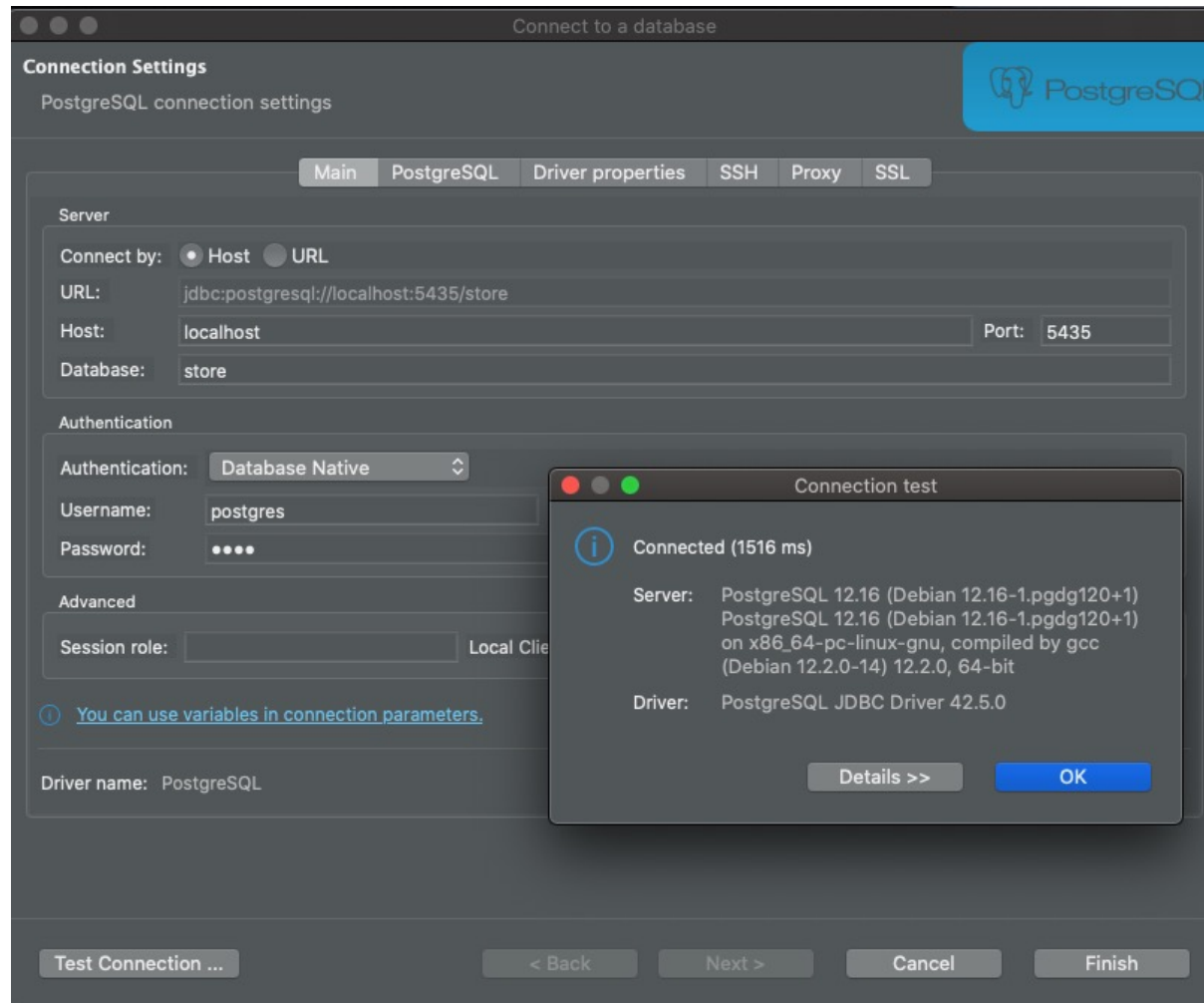


# Postgresql credential

- Host: localhost:5435
- Username: postgres
- Password: pass
- DB: store

Create connection on DBeaver to Postgresql and Citus with these credentials:

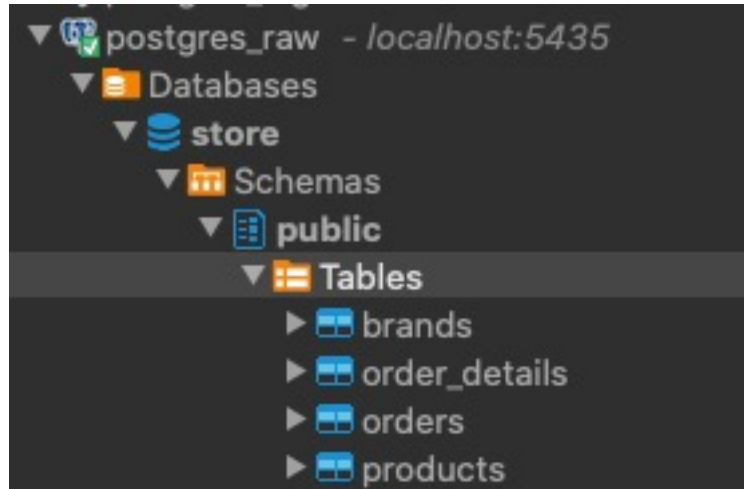
Setelah Docker berjalan kita bisa membuat connection data base dengan postgreSQL



# Citus credential

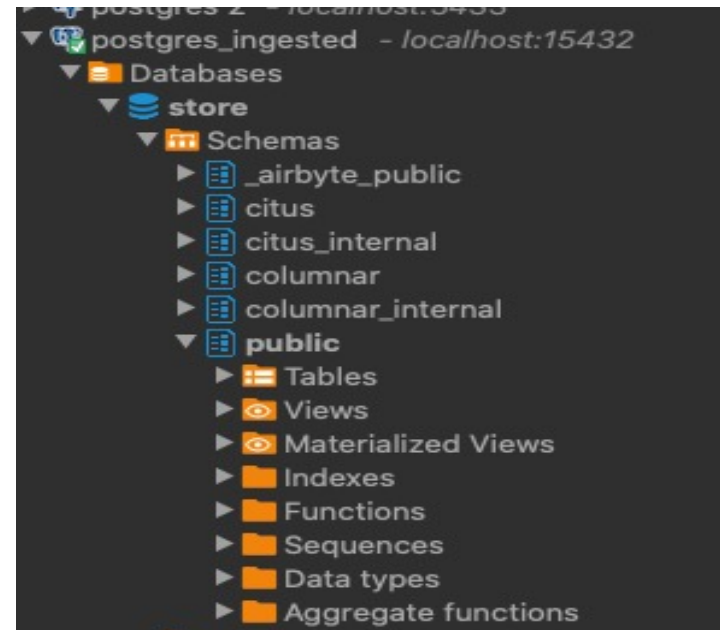
- Host: localhost:15432
- Username: posgres
- Password: pass
- DB: store

Create connection on DBeaver to Postgresql and Citus with these credentials:



Database sudah terbuat dengan nama connection database kita rename menjadi postgres\_raw dengan table **brands**, **order\_details**, **Orders** dan **products**. Dimana data ini yang nantinya akan kita ingest melalui airbyte

Connection database dengan nama postgres\_ingested ini adalah connection yang nantinya akan digunakan sebagai destinasi setelah dilakukan proses ingestion melalui airbyte

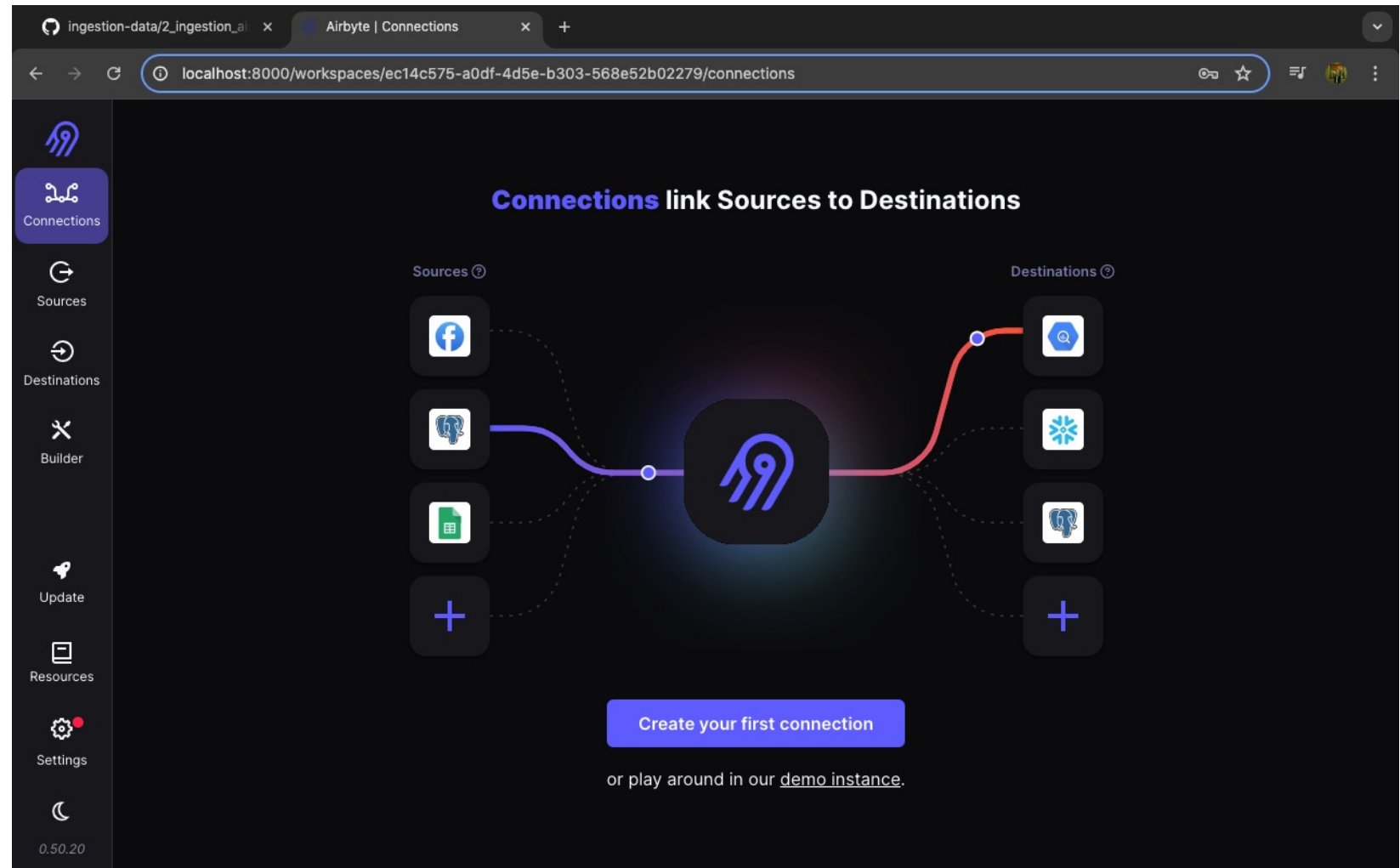


# Setup Connection in Airbyte

Buka Airbyte pada dashboard <http://localhost:8000/> di browser, login with

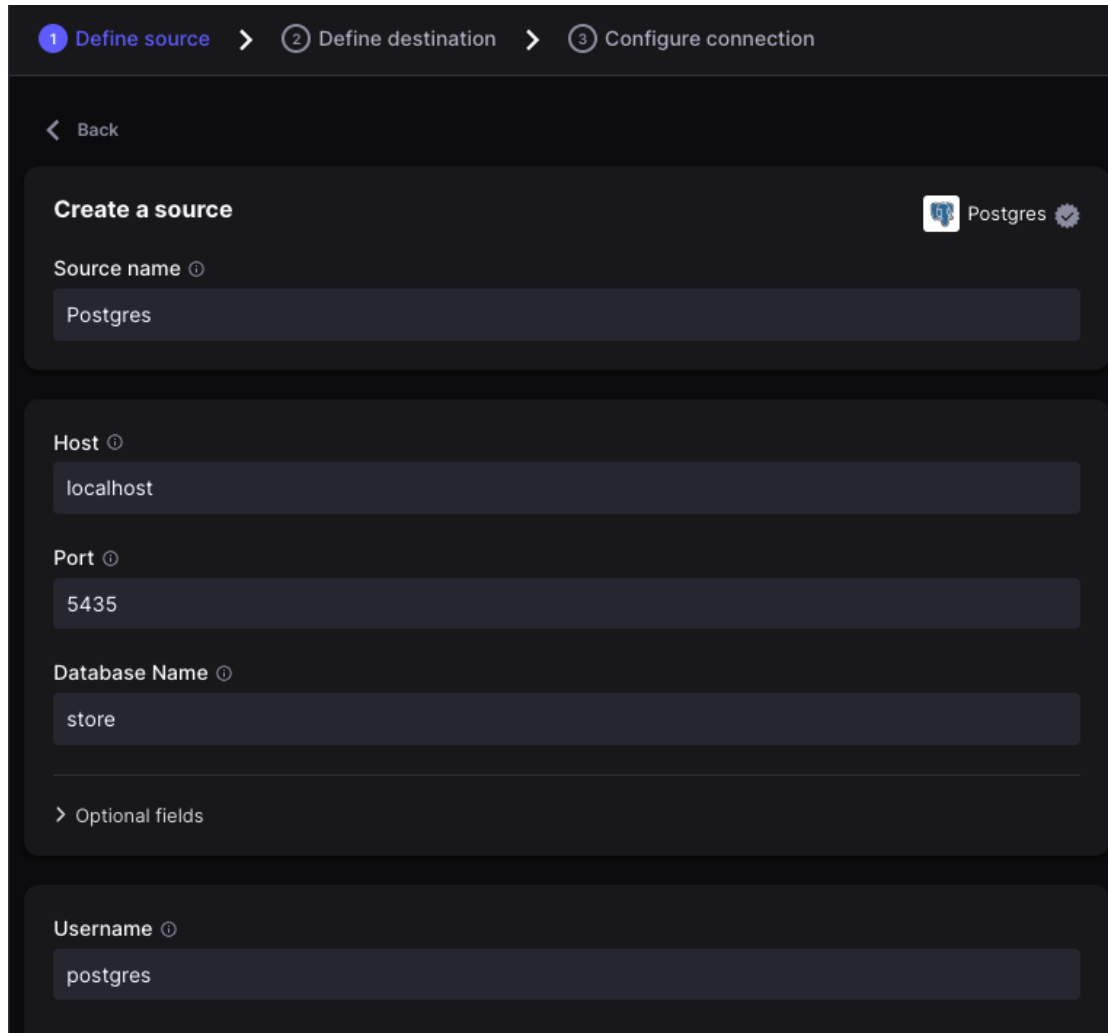
- Username: airbyte –
- Password: password

Lalu setelah itu membuat connection



# Ingest PostgreSQL to PostgreSQL – Define Source

Dalam kasus ini akan melakukan ingestion dari PostgreSQL to PostgreSQL.



The screenshot shows the Airbyte web interface for defining a new source. At the top, there are three steps: 1. Define source (active), 2. Define destination, and 3. Configure connection. Below the steps, there is a 'Back' button. The main section is titled 'Create a source' and features a 'Postgres' icon with a checkmark. The form contains the following fields:

- Source name**: postgres
- Host**: localhost
- Port**: 5435
- Database Name**: store
- Optional fields**: (collapsed)
- Username**: postgres

Didalam airbyte ada 3 konfigurasi untuk ingest data antara lain,  
Source : data yang akan dilakukan proses Ingestion  
Destination : Data yang akan disimpan setelah di lakukan ingestion  
Configursi Connection :

Source name : Postgres

Host: localhost:5435

Username: posgres

Password: pass

DB: store

Dalam memproses ingest pastikan bahwa browser yang digunakan sudah dihapus baik chache maupun cookie karena pada saat pelaksanaannya saya mendapatkan hal tersebut. Dan ketika hapus chache define



# Ingest PostgreSQL to PostgreSQL – Define Destination



Create a destination Postgres ALPHA

**Alpha connectors** are in development and support is not provided. See our [documentation](#) for more details.

Destination name ⓘ  
Postgres

Host ⓘ  
localhost

Port ⓘ  
15432

DB Name ⓘ  
store

Default Schema ⓘ  
public

User ⓘ  
postgres

SSL modes ⓘ disable

SSH Tunnel Method ⓘ No Tunnel

Untuk Destination ini bisa disesuaikan dengan dengan konfigurasi citus

Host: localhost:15432

Username: posgres

Password: pass

DB: store

# Ingest PostgreSQL to PostgreSQL-Configure Connection



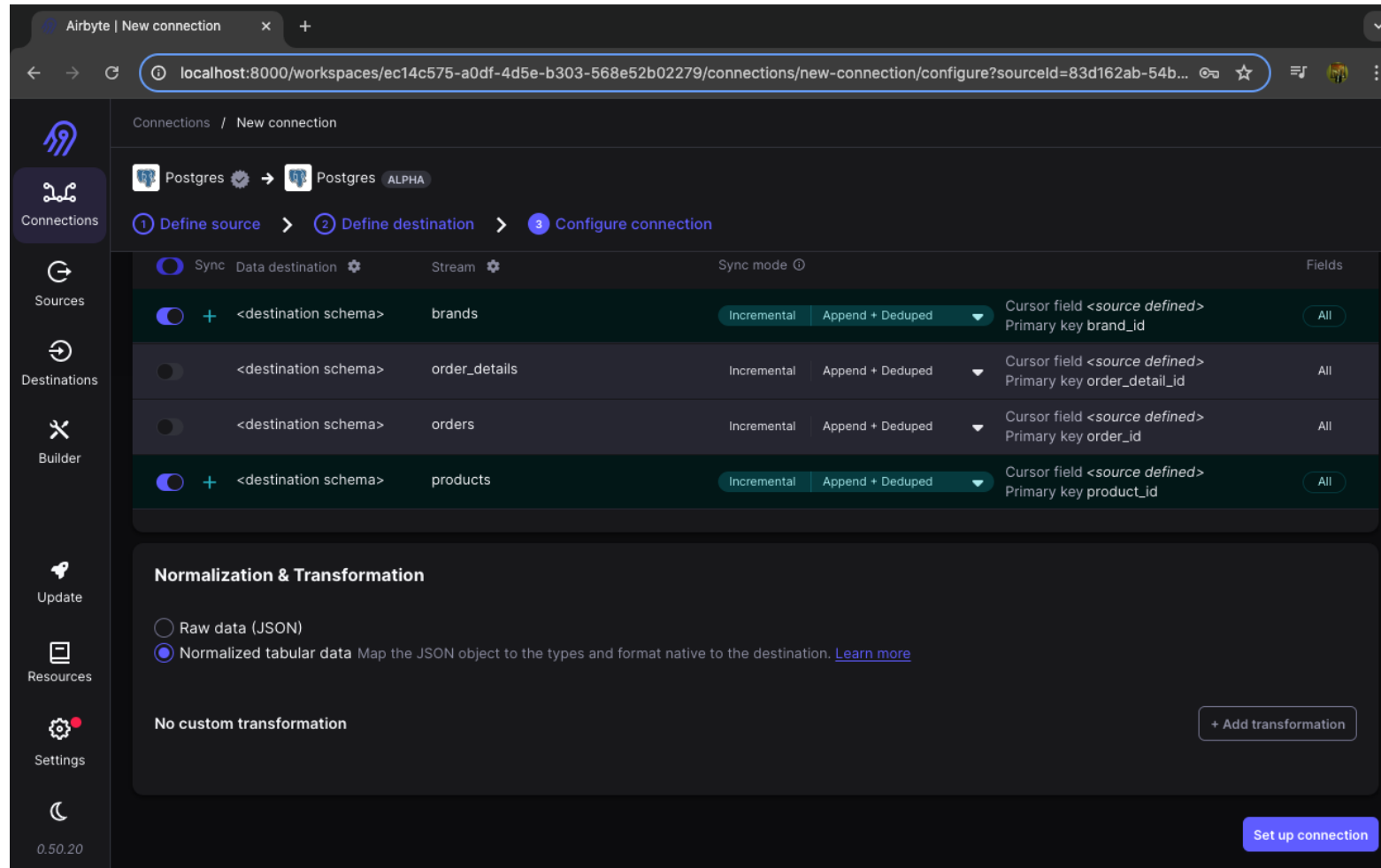
Setelah berhasil define source dan destination. Maka akan muncul dashboard seperti ini pada konfigurasi Airbyte. Untuk konfigurasi ini default sesuai yang dari airbyte. Ini yang nanti akan menghubungkan antara sources dan destination

A screenshot of the Airbyte web interface showing the 'New connection' configuration page. The browser address bar shows a localhost URL. The left sidebar contains navigation icons for Connections, Sources, Destinations, Builder, Update, Resources, and Settings. The main content area has a breadcrumb 'Connections / New connection' and a progress bar with three steps: '1 Define source', '2 Define destination', and '3 Configure connection'. The 'Configure connection' section is active and contains three main panels. The 'Connection' panel has a 'Connection name' field with the value 'Postgres -> Postgres'. The 'Configuration' panel includes 'Replication frequency' set to 'Every 24 hours', 'Destination Namespace' set to 'Destination default' with an 'Edit' button, 'Destination Stream Prefix' set to 'Mirror source name' with an 'Edit' button, and 'Detect and propagate schema changes' set to 'Ignore'. The bottom panel, 'Activate the streams you want to sync', features a search bar, a 'Refresh source schema' button, and a 'Hide disabled streams' toggle switch. The footer of the sidebar shows the version '0.50.20'.

# Ingest PostgreSQL to PostgreSQL-Configure Connection



Berikut merupakan tampilan table mana yang mau diambil dengan menonaktifkan seperti tombol on –off (Sync 1)



# Ingest PostgreSQL to PostgreSQL-Configure Connection



Setelah melakukan proses setup – connection makan akan muncul Status seperti pada gambar dibawah ini. Dimana dalam kasus ini masih pending . Klik Sync now untuk melanjutkan proses ingestion

The screenshot shows the Airbyte web interface in a dark theme. The browser address bar displays the URL: `localhost:8000/workspaces/ec14c575-a0df-4d5e-b303-568e52b02279/connections/fbc715c2-b4ff-4292-81b8-c3008edb55bb/status`. The left sidebar contains navigation icons for Connections, Sources, Destinations, Builder, Update, and Resources, with the 'Connections' icon highlighted. The main content area is titled 'Postgres → Postgres' and includes a toggle switch labeled 'Enabled' which is currently turned on. Below the title, there are tabs for 'Status', 'Job History', 'Replication', 'Transformation', and 'Settings', with 'Status' being the active tab. A large 'Pending' status card is displayed, featuring a 'Reset your data' button and a prominent blue 'Sync now' button. Underneath, the 'Enabled streams' section contains a search bar and a table with two entries:

Status	Stream name	Last record loaded
Pending	brands	
Pending	products	

# Ingest PostgreSQL to PostgreSQL-Configure Connection



Berikut merupakan tampilan jika sudah success melakukan sync dengan kata lain disini data sudah ke import ke connection postgres\_ingested.

A screenshot of the Airbyte web interface in a dark theme. The browser's address bar shows the URL: localhost:8000/workspaces/ec14c575-a0df-4d5e-b303-568e52b02279/connections/fbc715c2-b4ff-4292-81b8-c3008edb55bb/status. The left sidebar contains navigation icons for Connections, Sources, Destinations, Builder, Update, and Resources. The main content area is titled "Postgres → Postgres" and includes a toggle switch labeled "Enabled" which is turned on. Below this, there are tabs for Status, Job History, Replication, Transformation, and Settings. The "Status" tab is active, displaying a green checkmark icon and the text "On time". To the right of this status are two buttons: "Reset your data" and "Sync now". A section titled "Enabled streams" contains a search bar and a table with two rows of data.

Status	Stream name	Last record loaded
On time	brands	2 minutes ago
On time	products	2 minutes ago

# Ingest PostgreSQL to PostgreSQL-Configure Connection



Menambahkan table yang ingin di ingest dengan mode refresh, append pada table order\_details dan orders (Sync 2)

The screenshot displays the Alterra web interface for configuring a PostgreSQL to PostgreSQL connection. The left sidebar contains navigation icons for Connections, Sources, Destinations, Builder, Update, Resources, and Settings. The main panel is titled "Postgres → Postgres" and includes a toggle switch for "Enabled". Below the title are tabs for Status, Job History, Replication (selected), Transformation, and Settings. A "Refresh source schema" button is located in the top right of the Replication tab. The section "Activate the streams you want to sync" features a search bar and a "Hide disabled streams" toggle. A table lists the streams to be synced, with columns for Sync, Data destination, Stream, Sync mode, and Fields.

Sync	Data destination	Stream	Sync mode	Fields
<input type="checkbox"/>	<destination schema>	brands	Incremental Append + Deduped	Cursor field <source defined> Primary key brand_id All
<input checked="" type="checkbox"/>	+ <destination schema>	order_details	Full refresh Append	All
<input checked="" type="checkbox"/>	+ <destination schema>	orders	Full refresh Append	All
<input type="checkbox"/>	<destination schema>	products	Incremental Append + Deduped	Cursor field <source defined> Primary key product_id All

At the bottom right of the interface are "Cancel" and "Save changes" buttons.

# Ingest PostgreSQL to PostgreSQL-Configure Connection



Berikut merupakan tampilan jika sudah success melakukan sync dengan kata lain disini data sudah ke import ke connection postgres\_ingested.

The screenshot shows the Alterra web interface in a browser window. The address bar displays the URL: localhost:8000/workspaces/ec14c575-a0df-4d5e-b303-568e52b02279/connections/fbc715c2-b4ff-4292-81b8-c3008edb55bb/status. The interface has a dark theme. On the left is a sidebar with navigation icons for Connections, Sources, Destinations, Builder, Update, Resources, and Settings. The main content area is titled "Postgres → Postgres" and includes a toggle switch labeled "Enabled" which is turned on. Below this, there are tabs for Status, Job History, Replication, Transformation, and Settings. The "Status" tab is active, showing a green checkmark and the text "On time". To the right of this status are two buttons: "Reset your data" and "Sync now". Below the status bar is a section titled "Enabled streams" with a search bar. It contains a table with four rows of stream data.

Status	Stream name	Last record loaded
On time	brands	3 minutes ago
On time	order_details	3 minutes ago
On time	orders	3 minutes ago
On time	products	3 minutes ago

# Connection – Sync Mode



Sync mode pada Airbyte mengacu pada cara data disalin dari sumber ke tujuan. Ada beberapa mode sinkronisasi yang dapat dipilih, tergantung pada kebutuhan dan karakteristik data Anda:

## 1. Full Refresh (Overwrite):

1. **Deskripsi:** Seluruh data dari sumber akan ditulis ulang ke tujuan setiap kali sinkronisasi dilakukan.
2. **Kapan digunakan:** Berguna saat data di sumber dapat dihapus dan digantikan sepenuhnya tanpa kehilangan informasi penting. Misalnya, jika data sumber adalah snapshot atau laporan yang tidak berubah, dan Anda ingin memastikan tujuan selalu mencerminkan keadaan terbaru dari sumber.

## 2. Full Refresh (Append):

1. **Deskripsi:** Seluruh data dari sumber akan ditambahkan ke data yang sudah ada di tujuan.
2. **Kapan digunakan:** Jika Anda ingin menambahkan data baru dari sumber ke dalam data yang sudah ada di tujuan, tanpa menghapus data lama. Ini bisa berguna untuk data yang bersifat kumulatif atau historis.

## 3. Incremental (Append):

1. **Deskripsi:** Hanya data yang baru atau yang telah diubah sejak sinkronisasi terakhir yang akan ditambahkan ke tujuan.
2. **Kapan digunakan:** Ideal untuk kasus di mana data sumber berubah secara berkala dan Anda hanya ingin memperbarui data di tujuan dengan perubahan terbaru, tanpa menimpa atau menambah seluruh data.

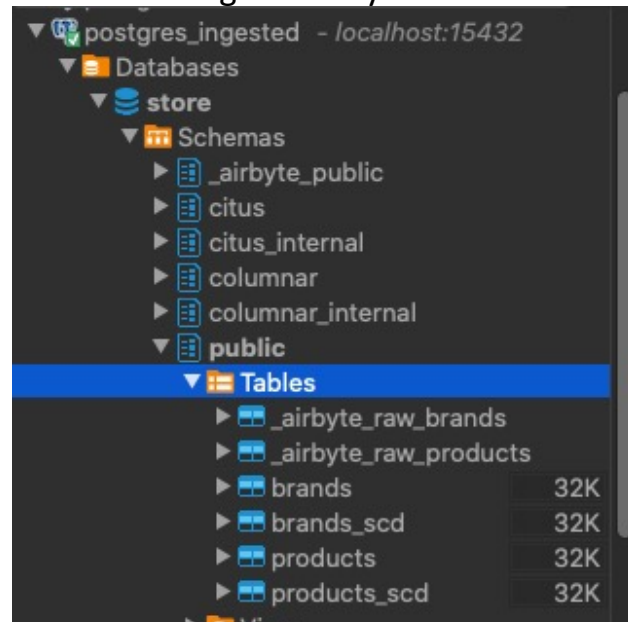
## 4. Incremental (Upsert):

1. **Deskripsi:** Data yang baru atau diubah dari sumber akan diperbarui atau ditambahkan di tujuan. Ini sering melibatkan pengidentifikasian dan penggabungan data berdasarkan kunci unik.
2. **Kapan digunakan:** Berguna saat data sumber dapat diperbarui secara teratur dan Anda ingin memastikan bahwa tujuan mencerminkan status terbaru dari data tanpa kehilangan data lama yang relevan.



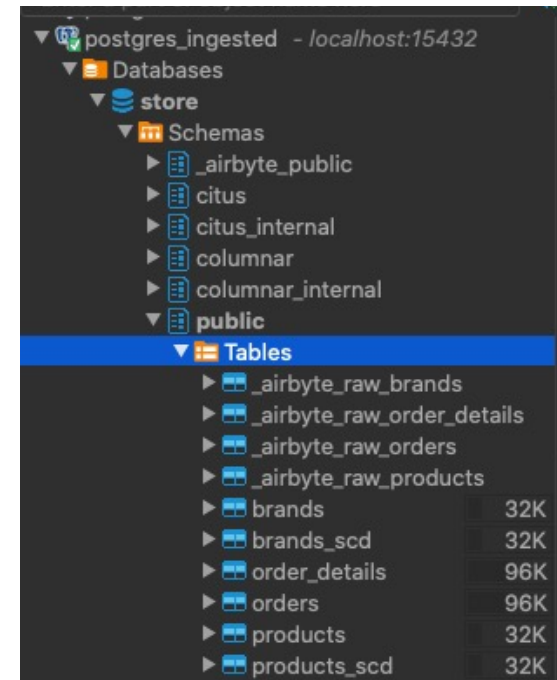
# Check Connection PostgreSQL (postgres\_ingested)

Ketika kita menggunakan sync mode incremental | full + Depuped, ada penambahan data scd yang otomatis tercreate ketika kita melakukan ingestion. Sync 1



Dengan SCD di Airbyte, organisasi dapat melacak perubahan data seiring waktu dan memastikan bahwa semua versi historis data tetap tersedia untuk analisis mendalam dan pelaporan historis.

Terkonfirmasi bahwa data sudah teringest pada postgres\_ingested connection database. Syn 2



THANK YOU 😊

