

Nama: Farhan Riyandi

Mentor: Bilal Benefit

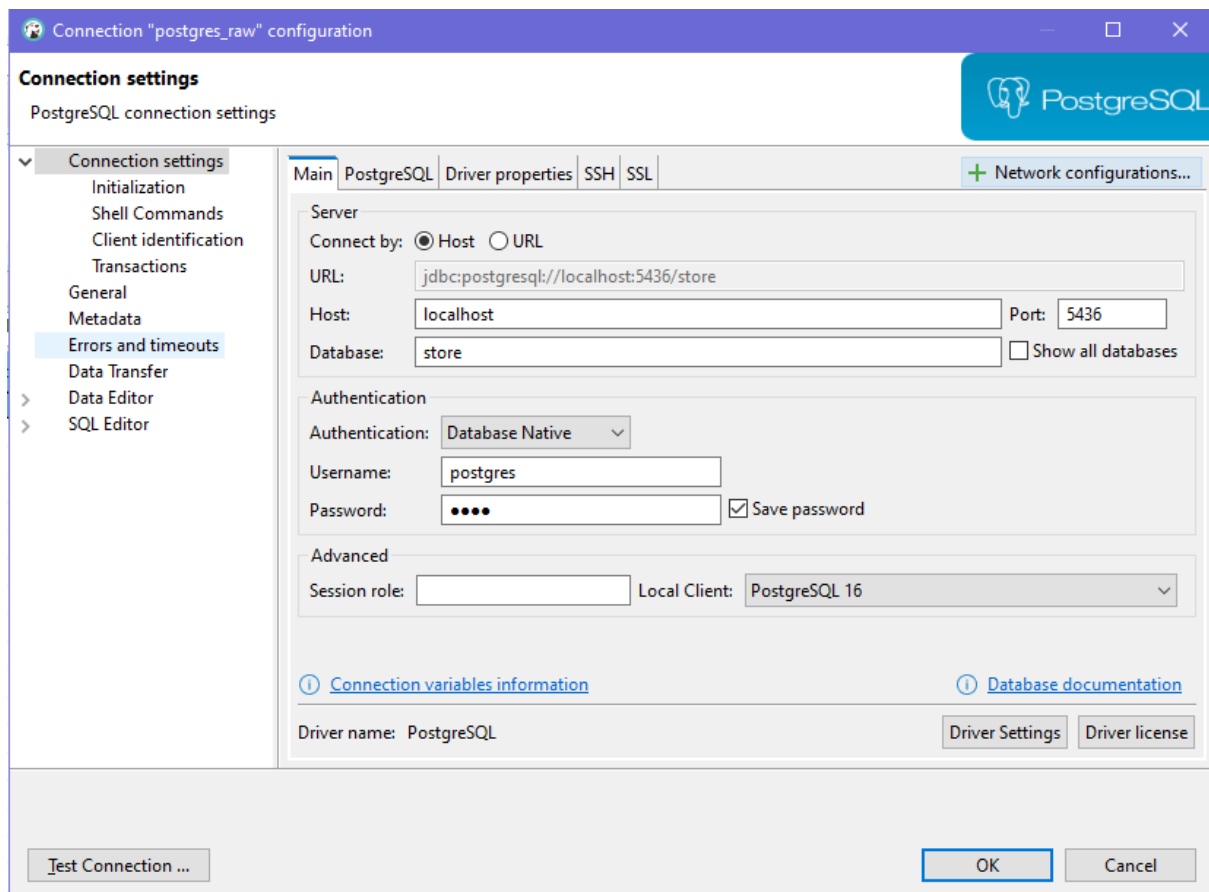
Data Engineer Batch 4

## Ingest from API to Postgresql

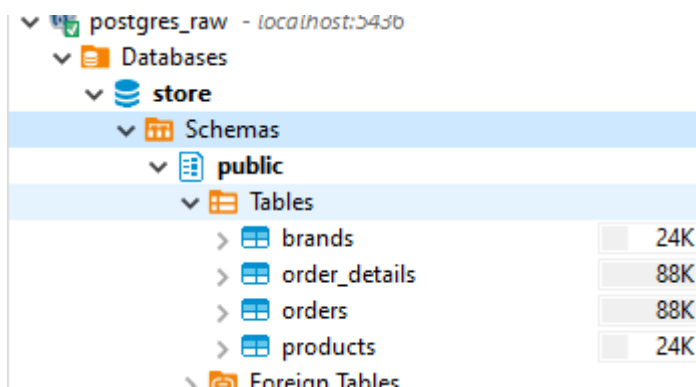
Jalankan docker compose yang menjalankan container airbyte

```
E:\Data_Ingestion_Tugas\ingestion-data\ingestion_airbyte>docker compose -f docker-compose.yml up -d
time="2024-07-28T11:41:19+07:00" level=warning msg="The \"DEPLOYMENT_MODE\" variable is not set. Defaulting to a blank string."
time="2024-07-28T11:41:19+07:00" level=warning msg="The \"APPLY_FIELD_SELECTION\" variable is not set. Defaulting to a blank string."
time="2024-07-28T11:41:19+07:00" level=warning msg="The \"DEPLOYMENT_MODE\" variable is not set. Defaulting to a blank string."
time="2024-07-28T11:41:19+07:00" level=warning msg="The \"FIELD_SELECTION_WORKSPACES\" variable is not set. Defaulting to a blank string."
time="2024-07-28T11:41:19+07:00" level=warning msg="The \"JOB_ERROR_REPORTING_SENTRY_DSN\" variable is not set. Defaulting to a blank string."
time="2024-07-28T11:41:19+07:00" level=warning msg="The \"LAUNCHDARKLY_KEY\" variable is not set. Defaulting to a blank string."
time="2024-07-28T11:41:19+07:00" level=warning msg="The \"LOG_CONNECTOR_MESSAGES\" variable is not set. Defaulting to a blank string."
time="2024-07-28T11:41:19+07:00" level=warning msg="The \"SECRET_PERSISTENCE\" variable is not set. Defaulting to a blank string."
time="2024-07-28T11:41:19+07:00" level=warning msg="The \"GITHUB_STORE_BRANCH\" variable is not set. Defaulting to a blank string."
time="2024-07-28T11:41:19+07:00" level=warning msg="The \"JOB_ERROR_REPORTING_SENTRY_DSN\" variable is not set. Defaulting to a blank string."
time="2024-07-28T11:41:19+07:00" level=warning msg="The \"LAUNCHDARKLY_KEY\" variable is not set. Defaulting to a blank string."
time="2024-07-28T11:41:19+07:00" level=warning msg="The \"NEW_SCHEDULER\" variable is not set. Defaulting to a blank string."
time="2024-07-28T11:41:19+07:00" level=warning msg="The \"SECRET_PERSISTENCE\" variable is not set. Defaulting to a blank string."
time="2024-07-28T11:41:19+07:00" level=warning msg="The \"WORKER_ENVIRONMENT\" variable is not set. Defaulting to a blank string."
time="2024-07-28T11:41:19+07:00" level=warning msg="The \"DEPLOYMENT_MODE\" variable is not set. Defaulting to a blank string."
time="2024-07-28T11:41:19+07:00" level=warning msg="The \"TEMPORAL_HISTORY_RETENTION_IN_DAYS\" variable is not set. Defaulting to a blank string."
time="2024-07-28T11:41:19+07:00" level=warning msg="The \"UPDATE_DEFINITIONS_CRON_ENABLED\" variable is not set. Defaulting to a blank string."
time="2024-07-28T11:41:19+07:00" level=warning msg="The \"CDK_VERSION\" variable is not set. Defaulting to a blank string."
time="2024-07-28T11:41:19+07:00" level=warning msg="The \"DEPLOYMENT_MODE\" variable is not set. Defaulting to a blank string."
time="2024-07-28T11:41:19+07:00" level=warning msg="The \"PYTHON_VERSION\" variable is not set. Defaulting to a blank string."
time="2024-07-28T11:41:19+07:00" level=warning msg="E:\\Data_Ingestion_Tugas\\ingestion-data\\ingestion_airbyte\\docker-compose.yml: 'version' is obsolete"
[+] Running 15/19
  Network ingestion_airbyte_airbyte_internal Created                                0.0s
  Network ingestion_airbyte_postgres-network Created                               0.0s
```

Buat koneksi postgres

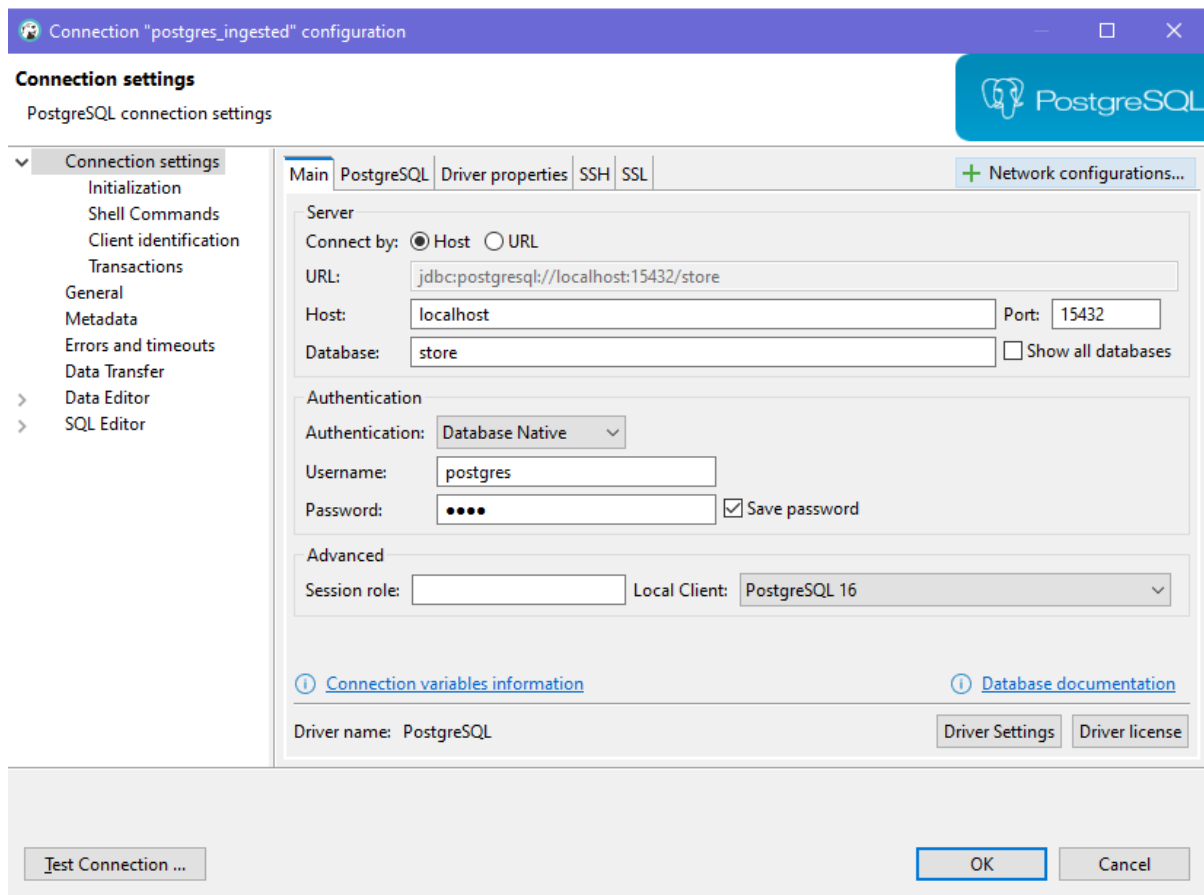


Dari store diubah Namanya menjadi postgres\_raw

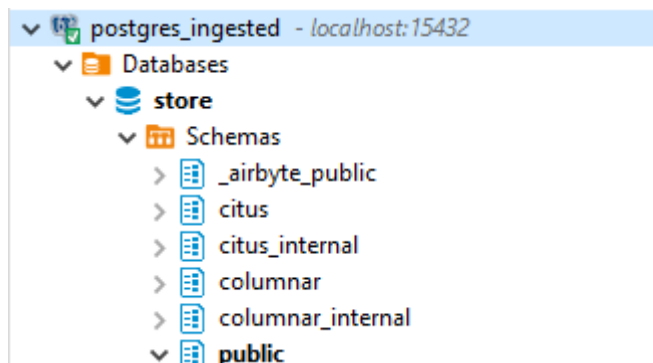


Bisa dilihat diatas sudah terdapat table yang sudah tersedia.

Kemudian dibuat lagi koneksi yang satunya lagi



Berikut adalah hasil dari database koneksi diatas yang sudah dibuat



Dimana pada database tersebut masih kosong, tujuannya adalah memindahkan dari table postgres\_raw ke postgres\_ingested.

Kemudian buka link berikut: <http://localhost:8000/> . Kemudian login pada airbyte tersebut.

Sign in

http://localhost:8000

Username

airbyte

Password

Sign in

Cancel

Setelah login pada connection klik new connection

Connections

+ New connection

NAME	SOURCE NAME	DESTINATION NAME	FREQUENCY	LAST SYNC	ENABLED
File (CSV, JSON, Excel, Feather, Parquet) → Po...	File (CSV, JSON, Excel, Feather, Parquet) - File ...	Postgres - Postgres	Manual	11 days ago	<div>Sync now</div>
Postgres → Postgres	Postgres - Postgres	Postgres - Postgres	24 hours	2 hours ago	<div><div></div></div>

Pilihlah file (CSV, JSON, Excel, Feather, Parquet)

Select the type of source you want to connect

par

☒ Generally Available

☒ Beta

☐ Alpha

1 connector

File (CSV, JSON, Excel, Feather, Parquet)

Need to build your own source?

Use our low-code connector builder

+ Request a new connector

4 matching connectors are hidden by filters.

Show hidden results

Kemudian lakukan konfigurasi sebagai dibawah ini, jika sudah klik set up source

1 Define source > 2 Define destination > 3 Configure connection

< Back

Create a source

File (CSV, JSON, Excel, Feather, Parquet) ✓

Source name ⓘ

File (CSV, JSON, Excel, Feather, Parquet)

Dataset Name ⓘ

green-taxi-data

File Format ⓘ

parquet

Storage Provider ⓘ

HTTPS: Public Web

> Optional fields

URL ⓘ

https://d37ci6vzurychx.cloudfront.net/trip-data/green\_tripdata\_2023-01.parquet

> Optional fields

Set up source

Kemudian pada define destination pilih postgres

File (CSV, JSON, Excel, Feather, Parquet) →

1 Define source > 2 Define destination > 3 Configure connection

Define destination

Select an existing destination

Use a data destination that you've already set up in Airbyte

Set up a new destination

Configure a new destination from Airbyte's catalog of available connectors

Postgres

Postgres ALPHA

1 connection >

Pada configuration connection dilakukan konfigurasi sebagai berikut, kemudian klik set up connection

1 Define source > 2 Define destination > 3 Configure connection

Connection

Connection name ⓘFile (CSV, JSON, Excel, Feather, Parquet) ⌵

Configuration

Replication frequency ⓘManual ⌵

Destination Namespace ⓘDestination defaultEdit

Destination Stream Prefix ⓘOptionalMirror source nameEdit

Detect and propagate schema changes ⓘIgnore ⌵

Activate the streams you want to sync

Search stream name

Hide disabled streams

SyncData destination ⚙️Stream ⚙️Sync mode ⓘ

<destination schema>green-taxi-dataFull refreshOverwrite

Normalization & Transformation

Raw data (JSON)

Normalized tabular data

Map the JSON object to the types and format native to the destination. [Learn more](#)

No custom transformation

Klik sync now

File (CSV, JSON, Excel, Feather, Parquet) → PostgresEnabled

File (CSV, JSON, Excel, Feather, Parquet) → PostgresALPHA

StatusJob HistoryReplicationTransformationSettings

PendingReset your dataSync now

Enabled streams

Search

Status	Stream name	Last record loaded ⓘ
Pending	green-taxi-data	

Kemudian setelah berhasil

File (CSV, JSON, Excel, Feather, Parquet) → Postgres ALPHA

Status Job History Replication Transformation Settings

On time Next sync in a day Reset your data Sync now

Enabled streams

Status	Stream name	Last record loaded
On time	green-taxi-data	4 minutes ago

Kemudian berikut data green\_taxi\_data yang sudah masuk ke database

data\_parquet green\_taxi\_data \*mydb> Script-1

Properties Data ER Diagram postgres\_ingested Databases store Schemas public Tables green\_taxi\_data

green\_taxi\_data Enter a SQL expression to filter results (use Ctrl+Space)

		123 dolocationid	123 ratecodeid	123 fare_amount	123 congestion_surcharge	lpep_dropoff_datetime	123 vendorid	lpep_pickup_datetime	asc ehail_fee
1		143	1	14.9	2.75	2023-01-01 07:37:11.000 +0700	2	2023-01-01 07:26:10.000 +0700	[NULL]
2		43	1	10.7	0	2023-01-01 07:57:49.000 +0700	2	2023-01-01 07:51:03.000 +0700	[NULL]
3		179	1	7.2	0	2023-01-01 07:41:32.000 +0700	2	2023-01-01 07:35:12.000 +0700	[NULL]
4		238	1	6.5	0	2023-01-01 07:19:03.000 +0700	1	2023-01-01 07:13:14.000 +0700	[NULL]
5		74	1	6	0	2023-01-01 07:39:02.000 +0700	1	2023-01-01 07:33:04.000 +0700	[NULL]
6		262	1	17.7	2.75	2023-01-01 08:11:04.000 +0700	2	2023-01-01 07:53:31.000 +0700	[NULL]
7		45	1	19.1	2.75	2023-01-01 07:26:39.000 +0700	1	2023-01-01 07:09:14.000 +0700	[NULL]
8		75	1	14.2	0	2023-01-01 07:24:55.000 +0700	2	2023-01-01 07:11:58.000 +0700	[NULL]
9		166	1	7.2	0	2023-01-01 07:46:26.000 +0700	2	2023-01-01 07:41:29.000 +0700	[NULL]
10		140	1	24.7	2.75	2023-01-01 08:13:42.000 +0700	2	2023-01-01 07:50:32.000 +0700	[NULL]
11		234	1	26.8	2.75	2023-01-01 07:41:43.000 +0700	1	2023-01-01 07:16:12.000 +0700	[NULL]
12		140	1	11.4	2.75	2023-01-01 07:17:08.000 +0700	2	2023-01-01 07:08:43.000 +0700	[NULL]
13		148	1	30.3	2.75	2023-01-01 07:45:31.000 +0700	2	2023-01-01 07:26:32.000 +0700	[NULL]
14		255	1	17.7	0	2023-01-01 07:30:09.000 +0700	2	2023-01-01 07:18:35.000 +0700	[NULL]
15		186	1	35.9	2.75	2023-01-01 08:18:06.000 +0700	2	2023-01-01 07:39:32.000 +0700	[NULL]
16		210	1	32.5	0	2023-01-01 08:08:23.000 +0700	1	2023-01-01 07:49:34.000 +0700	[NULL]
17		129	5	15	0	2023-01-01 07:19:37.000 +0700	1	2023-01-01 07:10:45.000 +0700	[NULL]
18		68	1	44.3	2.75	2023-01-01 08:17:35.000 +0700	2	2023-01-01 07:35:11.000 +0700	[NULL]
19		260	1	12.8	0	2023-01-01 07:42:23.000 +0700	2	2023-01-01 07:31:06.000 +0700	[NULL]
20		75	1	8.6	0	2023-01-01 07:21:50.000 +0700	2	2023-01-01 07:14:37.000 +0700	[NULL]
21		74	1	7.2	0	2023-01-01 07:30:26.000 +0700	2	2023-01-01 07:26:21.000 +0700	[NULL]
22		42	1	5.8	0	2023-01-01 07:46:06.000 +0700	2	2023-01-01 07:41:56.000 +0700	[NULL]
23		24	1	11.4	0	2023-01-01 07:41:25.000 +0700	2	2023-01-01 07:32:02.000 +0700	[NULL]
24		42	1	7.2	0	2023-01-01 07:59:47.000 +0700	2	2023-01-01 07:55:13.000 +0700	[NULL]
25		244	1	16.3	0	2023-01-01 07:16:02.000 +0700	2	2023-01-01 07:01:31.000 +0700	[NULL]
26		146	1	19.8	0	2023-01-01 07:39:24.000 +0700	2	2023-01-01 07:22:31.000 +0700	[NULL]
27		135	1	14.2	0	2023-01-01 08:09:31.000 +0700	2	2023-01-01 07:57:28.000 +0700	[NULL]
28		173	1	19.1	0	2023-01-01 07:53:49.000 +0700	2	2023-01-01 07:35:21.000 +0700	[NULL]
29		216	1	18.4	0	2023-01-01 08:00:53.000 +0700	2	2023-01-01 07:52:52.000 +0700	[NULL]
30		142	1	13.5	2.75	2023-01-01 07:42:53.000 +0700	2	2023-01-01 07:32:56.000 +0700	[NULL]
31		238	1	8	2.75	2023-01-01 07:32:05.000 +0700	1	2023-01-01 07:24:01.000 +0700	[NULL]

## Ingest data from Postgresql to Postgresql Citus

Klik new connection

Connections						<a href="#">+ New connection</a>
NAME	SOURCE NAME	DESTINATION NAME	FREQUENCY	LAST SYNC	ENABLED	
File (CSV, JSON, Excel, Feather, Parquet) → Postgres	File (CSV, JSON, Excel, Feather, Parquet) - File (CSV, JSON, Excel, F...	Postgres - Postgres	24 hours	12 minutes ago	<input type="checkbox"/>	
File (CSV, JSON, Excel, Feather, Parquet) → Postgres	File (CSV, JSON, Excel, Feather, Parquet) - File (CSV, JSON, Excel, F...	Postgres - Postgres	Manual	11 days ago	<a href="#">Sync now</a>	
File (CSV, JSON, Excel, Feather, Parquet) → Postgres	File (CSV, JSON, Excel, Feather, Parquet) - File (CSV, JSON, Excel, F...	Postgres - Postgres	Manual		<a href="#">Sync now</a>	
Postgres → Postgres	Postgres - Postgres	Postgres - Postgres	24 hours	2 hours ago	<input type="checkbox"/>	

Pilih postgres pada define source

[1 Define source](#) > [2 Define destination](#) > [3 Configure connection](#)

**Define source**

☒ Select an existing source  
Use a data source that you've already set up in Airbyte

☐ Set up a new source  
Configure a new source from Airbyte's catalog of available connectors

File (CSV, JSON, Excel, Feather, Parquet)

File (CSV, JSON, Excel, Feather, Parquet)

3 connections >

File (CSV, JSON, Excel, Feather, Parquet)

File (CSV, JSON, Excel, Feather, Parquet)

>

Postgres

Postgres

1 connection >

Pada define destination pilih postgres

Postgres →

[1 Define source](#) > [2 Define destination](#) > [3 Configure connection](#)

**Define destination**

☒ Select an existing destination  
Use a data destination that you've already set up in Airbyte

☐ Set up a new destination  
Configure a new destination from Airbyte's catalog of available connectors

Postgres

Postgres ALPHA

4 connections >

Pada configuration diatur konfigurasi sebagai berikut kemudian klik set up connection



Postgres → Postgres **ALPHA**

1 Define source > 2 Define destination > 3 Configure connection

Destination Stream Prefix: Optional

Detect and propagate schema changes:

---

**Activate the streams you want to sync**

Search stream name

☐ Hide disabled streams

☒ Sync   Sync mode:

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

---

**Normalization & Transformation**

☐ Raw data (JSON)

☒ Normalized tabular data Map the JSON object to the types and format native to the destination. [Learn more](#)

No custom transformation

Status Job History Replication Transformation Settings

☒ On time

Enabled streams

Status	Stream name	Last record loaded
On time	brands	4 minutes ago
On time	order_details	4 minutes ago
On time	orders	4 minutes ago
On time	products	4 minutes ago

Kemudian hasilnya pada database postgres ingested terdapat beberapa tabel.

Schema	Table Name	Size
public	Tables	
	_airbyte_raw_brands	32K
	_airbyte_raw_green_taxi_data	
	_airbyte_raw_order_details	
	_airbyte_raw_orders	752K
	_airbyte_raw_products	32K
	brands	32K
	brands_scd	32K
	green_taxi_data	19M
	order_details	512K
	orders	480K
	products	32K
	products_scd	32K
	Foreign Tables	
	Views	

