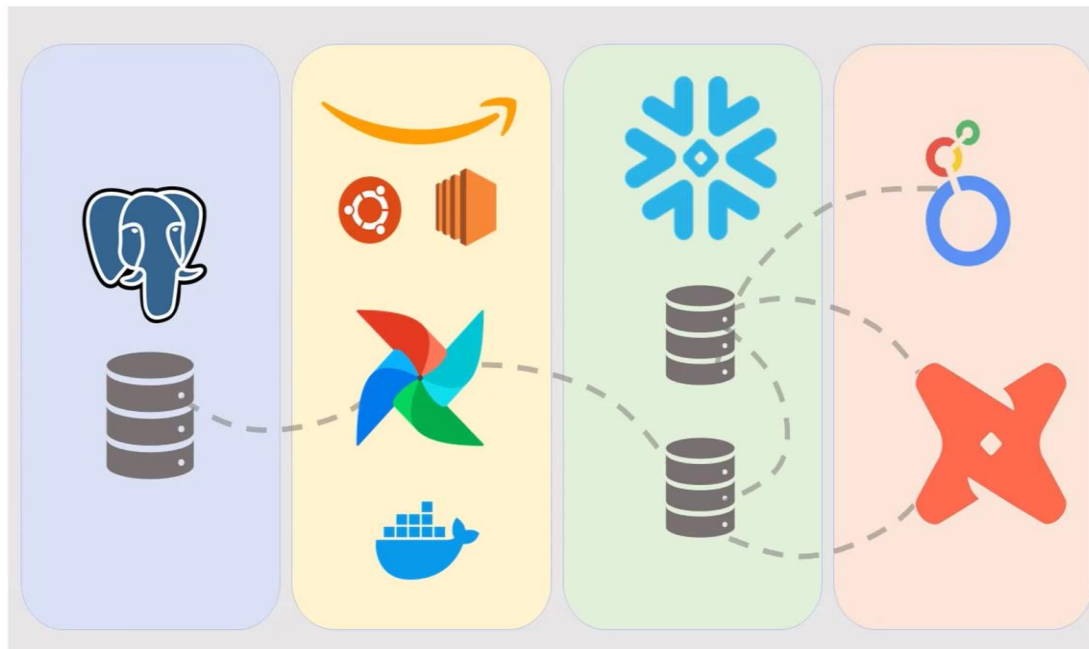


# Pipeline de ELT com DBT



# Projeto: Analytics Engineer (vendas)

Linguagem: Python

BD: PostgreSQL

Cloud: AWS EC2

SO: Linux

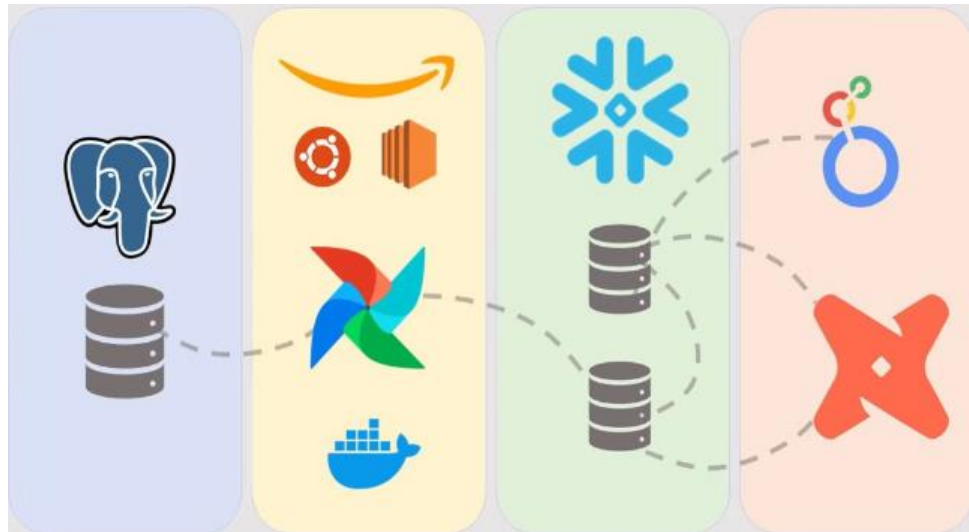
Container: Docker

Orquestrador: Airflow

ETL: dbt

DW: Snowflake

Dataviz: Looker Studio



# Etapas

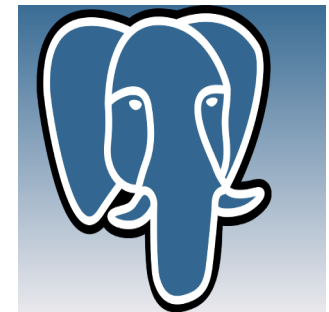
1. Conectar e explorar o BD de vendas;
2. Criar conta AWS ;
3. Criar VM Linux ( EC2);
4. Instalar e configurar o Docker;
5. Instalar e configurar o Apache Airflow (stage e teste de carga);
6. Configurar o Snowflake (criar BD, schemas, WH, Tabelas e etc)
7. Configurar DBT (criar modelos, jobs, testes);
8. Criar dashboards no Looker Studio;

# Etapa 1 - Conectar e explorar o BD de vendas;

```
/home/renatopatricio/.nushlogin file.  
renatopatricio@Renato:~$ psql -h 159.223.187.110 -U etlreadonly -d novadrive  
Password for user etlreadonly:  
psql (14.11 (Ubuntu 14.11-0ubuntu0.22.04.1), server 12.18 (Ubuntu 12.18-0ubuntu0.20.04.1))  
SSL connection (protocol: TLSv1.3, cipher: TLS_AES_256_GCM_SHA384, bits: 256, compression: off)  
Type "help" for help.  
  
novadrive=> SELECT schema_name FROM information_schema.schemata;  
      schema_name  
-----  
pg_catalog  
public  
information_schema  
(3 rows)
```



```
novadrive=> SELECT  
      pg_database.datname,  
      pg_size_pretty(pg_database_size(pg_database.datname)) AS size  
FROM pg_database;  
      datname | size  
-----+-----  
postgres     | 51 MB  
template1    | 7953 kB  
template0    | 7809 kB  
novadrive     | 22 MB  
novadrivebank | 8569 kB  
(5 rows)
```



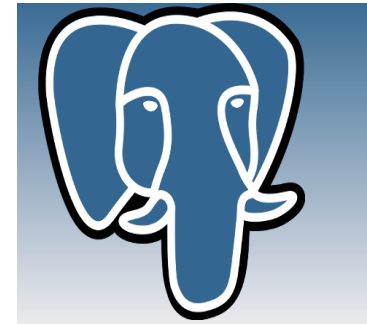
# Etapa 1 - Conectar e explorar o BD de vendas;

```
novadrive=> \dt+
```

List of relations							
Schema	Name	Type	Owner	Persistence	Access method	Size	Description
public	ciudades	table	postgres	permanent	heap	8192 bytes	
public	clientes	table	postgres	permanent	heap	6208 kB	
public	concessionarias	table	postgres	permanent	heap	8192 bytes	
public	estados	table	postgres	permanent	heap	8192 bytes	
public	veiculos	table	postgres	permanent	heap	8192 bytes	
public	vendas	table	postgres	permanent	heap	4736 kB	
public	vendedores	table	postgres	permanent	heap	8192 bytes	

(7 rows)

```
novadrive=> SELECT sum(pg_relation_size(c.oid)) AS total_size
FROM pg_class c
WHERE c.relname IN ('ciudades', 'clientes', 'concessionarias', 'estados', 'veiculos', 'vendas', 'vendedores');
total_size
-----
11173888
(1 row)
```



# Criação e configuração da VM/Linux

EC2 > Instances > i-074a37cde33d92cf4

**Instance summary for i-074a37cde33d92cf4 (pipeline\_vendas)** info

Updated less than a minute ago

Connect

Instance ID i-074a37cde33d92cf4 (pipeline_vendas)	Public IPv4 address -	Private IPv4 addresses 172.31.61.131
IPv6 address -	Instance state Stopped	Public IPv4 DNS -
Hostname type IP name: ip-172-31-61-131.ec2.internal	Private IP DNS name (IPv4 only) ip-172-31-61-131.ec2.internal	Elastic IP addresses -
Answer private resource DNS name IPv4 (A)	Instance type t2.large	AWS Compute Optimizer finding Opt-in to AWS Compute Optimizer for recomm
Auto-assigned IP address -	VPC ID vpc-046b555a2acc717f2	Auto Scaling Group name -
IAM Role -	Subnet ID subnet-07bac8a31ae4cd9d	
IMDSv2 Required		

Details | Status and alarms New | Monitoring | Security | Networking | Storage | Tags

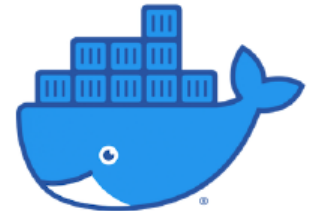
▼ Instance details info

Platform Ubuntu (Inferred)	AMI ID ami-080e1f13689e07408	Monitoring disabled
Platform details Linux/UNIX	AMI name ubuntu/images/hvm-ssd/ubuntu-jammy-22.04-amd64-server-20240301	Termination protection Disabled
Stop protection Disabled	Launch time Fri Apr 05 2024 22:57:16 GMT-0300 (Horário Padrão de Brasília) (14 days)	AMI location amazon/ubuntu/images/hvm-ssd/ubuntu-jam
Instance auto-recovery Default	Lifecycle normal	Stop-hibernate behavior Disabled
AMI Launch index 0	Key pair assigned at launch pipeline_vendas	State transition reason User initiated (2024-04-08 01:52:33 GMT)



# Instalação e configuração Docker/Airflow

```
ubuntu@ip-172-31-80-120:~$ sudo mkdir -m 0755 -p /etc/apt/keyrings
ubuntu@ip-172-31-80-120:~$ curl -fsSL https://download.docker.com/linux/ubuntu/gpg | sudo gpg --dearmor -o /usr/share/keyrings/docker-archive-keyring.gpg
ubuntu@ip-172-31-80-120:~$ echo "deb [arch=$(dpkg --print-architecture) signed-by=/usr/share/keyrings/docker-archive-keyring.gpg] https://download.docker.com/linux/ubuntu $(lsb_release -cs) stable" | sudo tee /etc/apt/sources.list.d/docker.list > /dev/null
ubuntu@ip-172-31-80-120:~$ sudo apt-get update
Hit:1 http://us-east-1.ec2.archive.ubuntu.com/ubuntu jammy InRelease
Hit:2 http://us-east-1.ec2.archive.ubuntu.com/ubuntu jammy-updates InRelease
Hit:3 http://us-east-1.ec2.archive.ubuntu.com/ubuntu jammy-backports InRelease
Get:4 https://download.docker.com/linux/ubuntu jammy InRelease [48.8 kB]
Hit:5 http://security.ubuntu.com/ubuntu jammy-security InRelease
Get:6 https://download.docker.com/linux/ubuntu jammy/stable amd64 Packages [26.1 kB]
Fetched 74.9 kB in 1s (136 kB/s)
Reading package lists... Done
ubuntu@ip-172-31-80-120:~$ sudo apt-get install docker-ce docker-ce-cli containerd.io docker-buildx-plugin docker-compose-plugin
```



```
ubuntu@ip-172-31-80-120:~$ curl -Lfo 'https://airflow.apache.org/docs/apache-airflow/stable/docker-compose.yaml'
% Total    % Received % Xferd Average Speed   Time    Time     Current
           Dload Upload   Total   Spent    Left   Speed
100 10940  100 10940    0     0  47729      0 --:--:-- --:--:-- --:--:-- 47772
ubuntu@ip-172-31-80-120:~$ ls
docker-compose.yaml
```



# Criação da DAG(Airflow)

Airflow DAGs Cluster Activity Datasets Security Browse Admin Docs

## DAGs

All 1 Active 1 Paused 0

Running 0 Failed 0

Filter DAGs by tag

Search DAGs

DAG	Owner	Runs	Schedule	Last Run	Next Run
postgres_to_snowflake	airflow	5	1 day, 0:00:00	2024-04-19, 18:50:26	2024-04-19, 00:00:00

» DAG  
postgres\_to\_snowflake

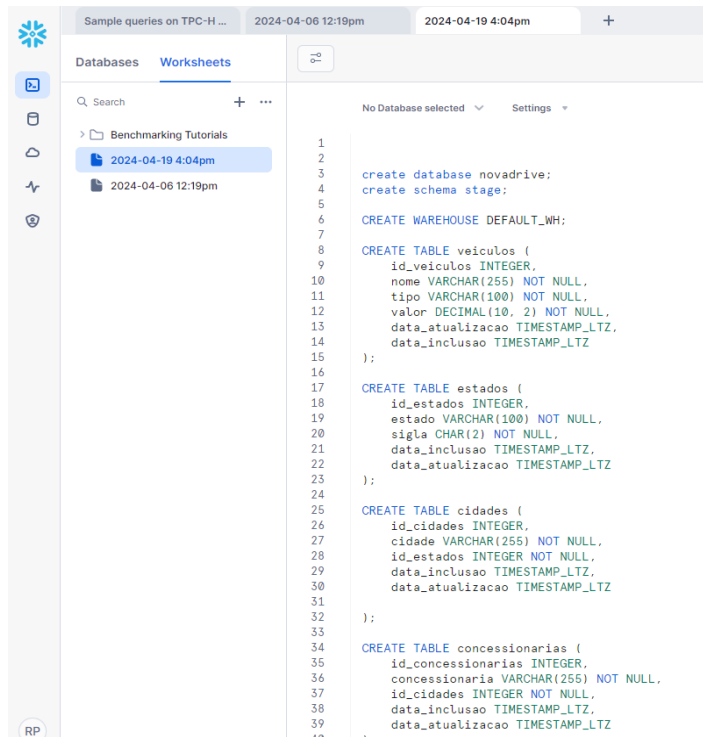
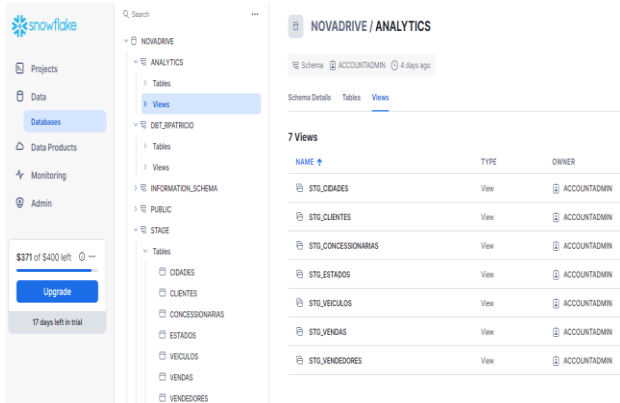
Details Graph Gantt Code

Parsed at: 2024-04-19, 18:52:43 UTC

```
1 from datetime import datetime, timedelta
2 from airflow.decorators import dag, task
3 from airflow.providers.postgres.hooks.postgres import PostgresHook
4 from airflow.providers.snowflake.hooks.snowflake import SnowflakeHook
5
6 default_args = {
7     'owner': 'airflow',
8     'depends_on_past': False,
9     'start_date': datetime(2024, 1, 1),
10    'email_on_failure': False,
11    'email_on_retry': False,
12    'retries': 0,
13    'retry_delay': timedelta(minutes=1),
14 }
15
16 @dag(
17     dag_id='postgres_to_snowflake',
18     default_args=default_args,
19     description='Load data incrementally from Postgres to Snowflake',
20     schedule_interval=timedelta(days=1),
21     catchup=False
```



# Configuração do Snowflake



# Testando a DAG (Airlflow)



# Testando a carga incremental (Snowflake)

ACCOUNTADMIN

NOVADRIVE.STAGE Settings

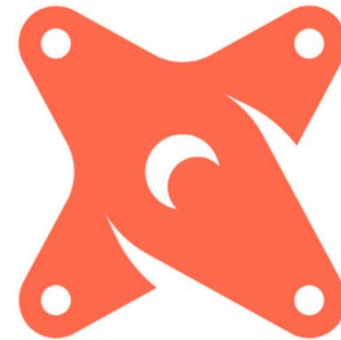
1 SHOW TABLES IN stage;  
2  
3  
4  
5

Results Chart

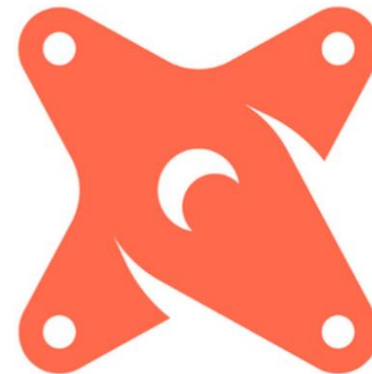
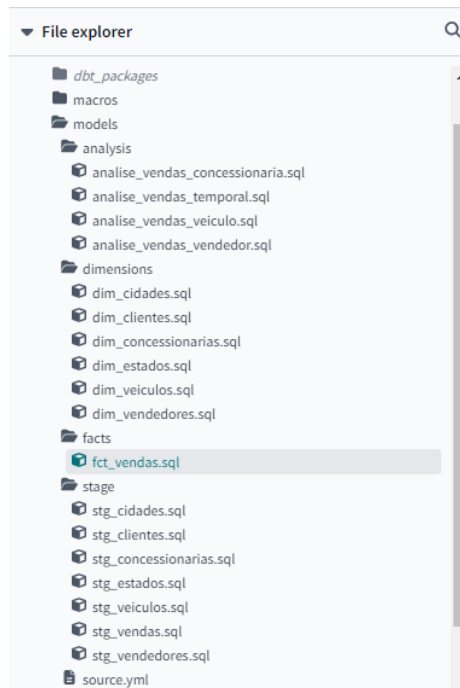
	created_on	name	database_name	schema_name	kind	comment	cluster_by	rows	bytes	owner
1	2024-04-06 08:29:09.768 -0700	CIDADES	NOVADRIVE	STAGE	TABLE			54	4608	ACCOUNTADMIN
2	2024-04-06 08:29:10.992 -0700	CLIENTES	NOVADRIVE	STAGE	TABLE			10862	317440	ACCOUNTADMIN
3	2024-04-06 08:29:10.138 -0700	CONCESSIONARIAS	NOVADRIVE	STAGE	TABLE			58	4608	ACCOUNTADMIN
4	2024-04-06 08:29:09.325 -0700	ESTADOS	NOVADRIVE	STAGE	TABLE			54	4608	ACCOUNTADMIN
5	2024-04-06 08:29:08.851 -0700	VEICULOS	NOVADRIVE	STAGE	TABLE			16	4608	ACCOUNTADMIN
6	2024-04-06 08:29:11.446 -0700	VENDAS	NOVADRIVE	STAGE	TABLE			10897	262144	ACCOUNTADMIN
7	2024-04-06 08:29:10.564 -0700	VENDEDORES	NOVADRIVE	STAGE	TABLE			106	3072	ACCOUNTADMIN

# Criação do arquivo source.yml (dbt)

```
source.yml x
models > source.yml
1 | version: 2
2 |
3 | sources:
4 |   - name: sources
5 |     database: NOVADRIVE
6 |     schema: STAGE
7 |     tables:
8 |       Generate model
9 |       - name: cidades
10 |      Generate model
11 |      - name: clientes
12 |      Generate model
13 |      - name: concessionarias
14 |      Generate model
15 |      - name: estados
16 |      Generate model
17 |      - name: veiculos
18 |      Generate model
19 |      - name: vendas
20 |      Generate model
21 |      - name: vendedores
```



# Criação dos modelos de consulta (dbt)



# Deploy do job de produção (dbt)

Deploy > Prod > Main.Job > Run #70403107385073

**Run #70403107385073** Success

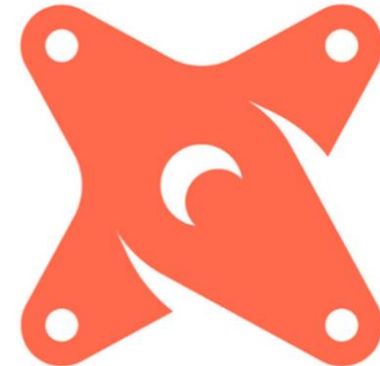
Triggered Mon, 15 Apr 2024 13:51:18 GMT

Trigger	Commit SHA	Environment	Run duration	Documentation
Triggered manually	#149755ea	Prod	48s	<a href="#">View</a>

**Run summary** | Lineage | Model timing | Artifacts

- Time in queue  
Prep time 4s
- Clone git repository
- Create profile from connection Snowflake
- Invoke dbt deps
- Invoke dbt build
- Invoke dbt docs generate

Finished Mon, 15 Apr 2024 13:52:11 GMT



# Dashboard das análises (Looker Studio)

