

## Solution Sprint - Fase 4

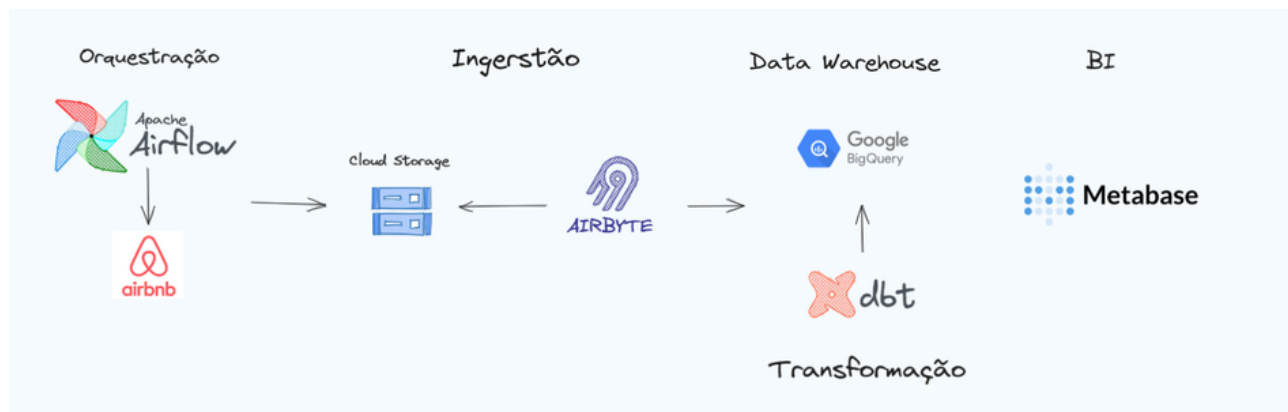
O desafio será construir uma plataforma de dados moderna (MDS) na nuvem e realizar integrações em real time para compor a solução. Para isso, vamos considerar as bases de dados do Inside Airbnb para o Rio de Janeiro e dados imobiliários sobre imóveis na cidade.

### Requisitos Técnicos: [🔗](#)

- a) No mínimo uma ferramenta utilizada para ingestão de dados;
- b) Uma camada de Storage, preferencialmente um Cloud Data Warehouse;
- c) Uma camada de transformação diretamente no Storage;
- d) No mínimo um dashboard ou report analisando os dados transformados;
- e) No mínimo um modelo analítico envolvendo técnicas de aprendizado de máquina;
- f) Uma solução para governança e catálogo de dados.

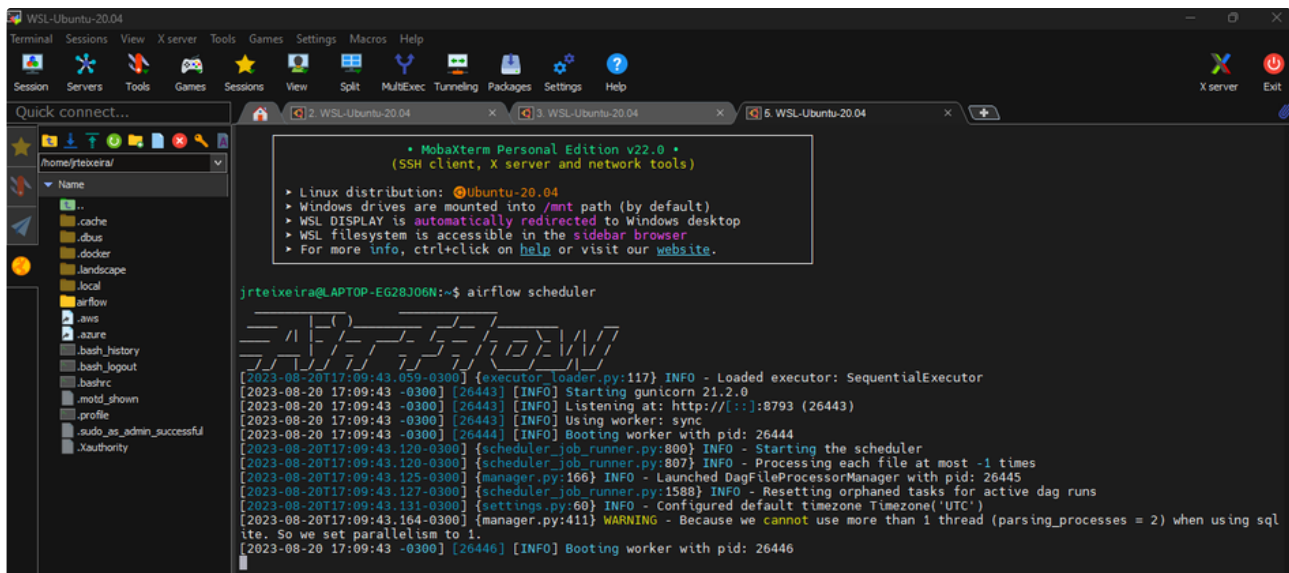
### Entregável 1 [🔗](#)

#### Arquitetura projetada: [🔗](#)

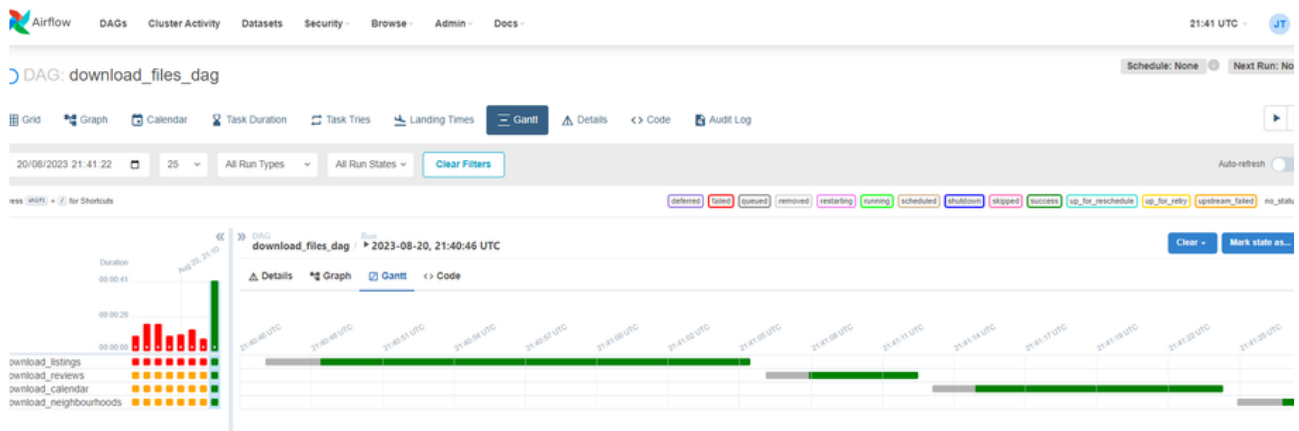


#### Orquestração: Airflow [🔗](#)

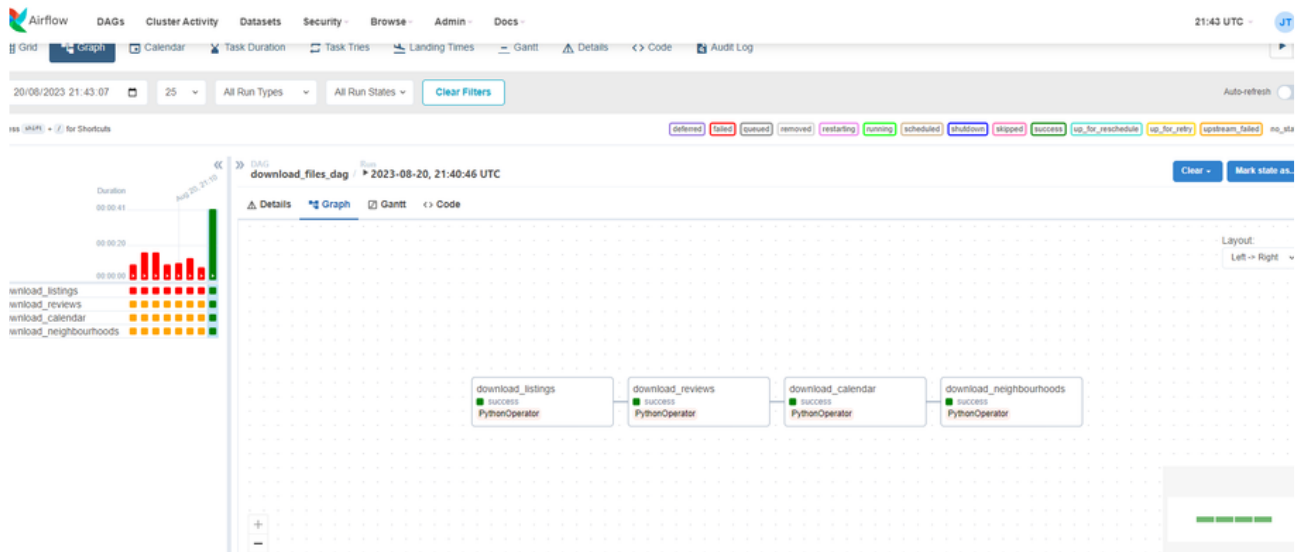
#### Evidências:



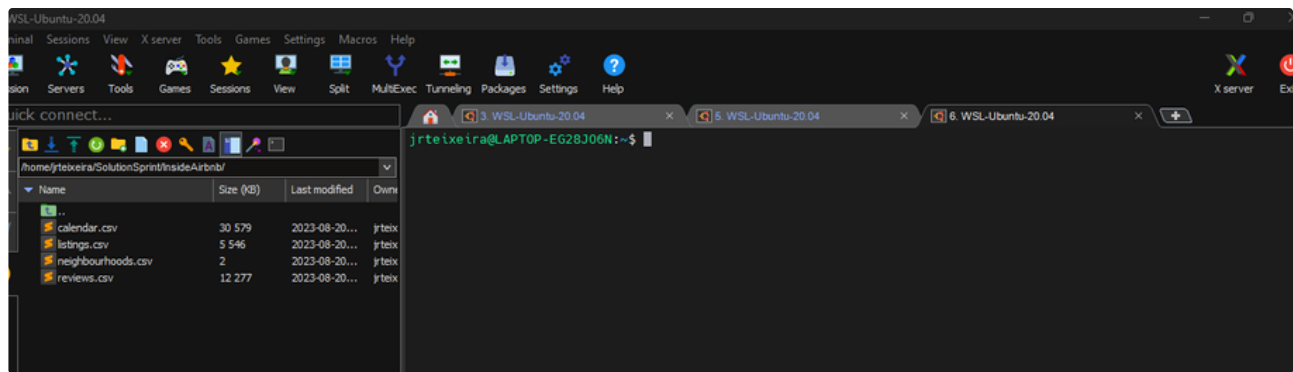
Start Airflow



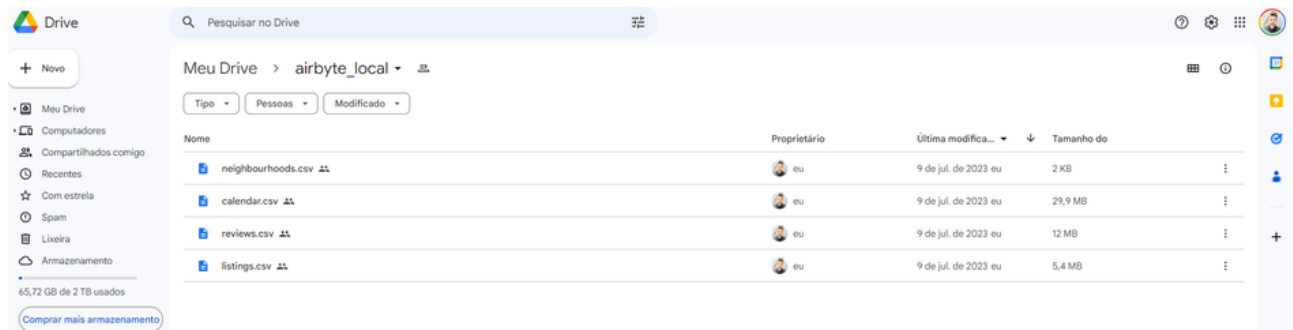
Airflow Tasks



Airflow Tasks

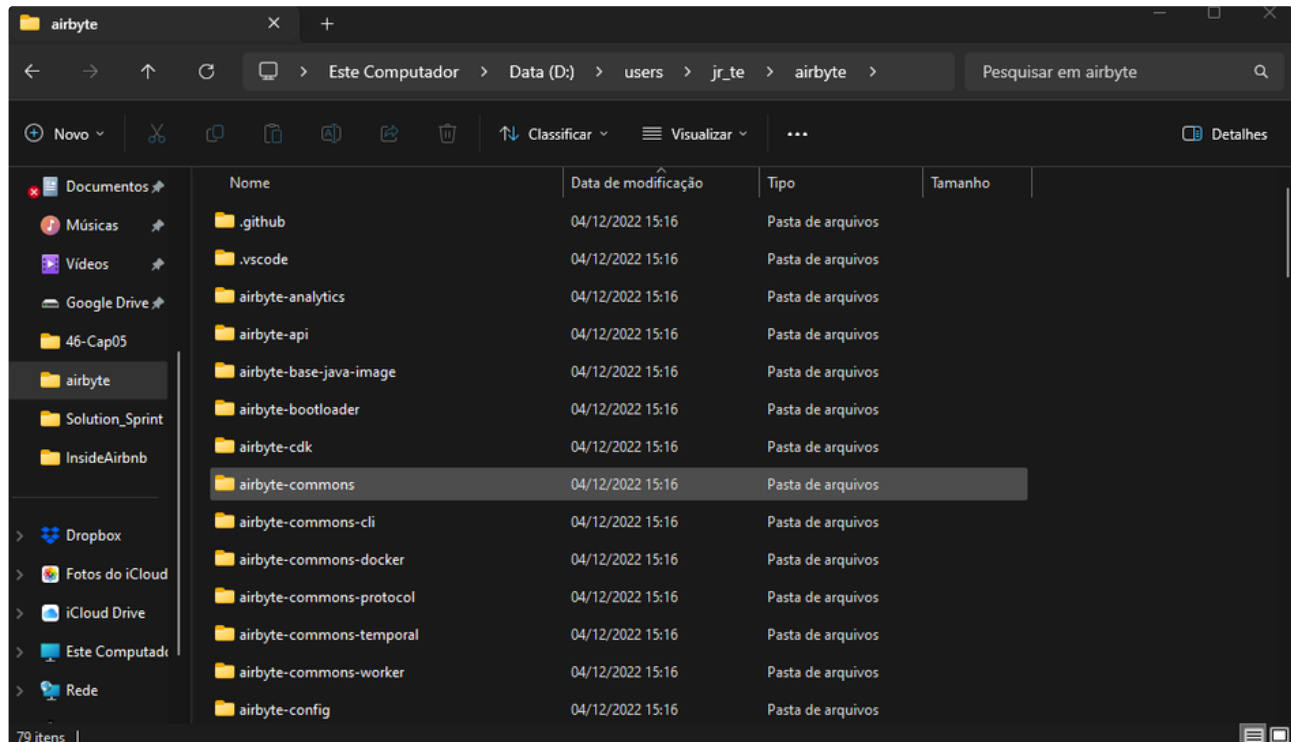


Arquivos listados no diretório local

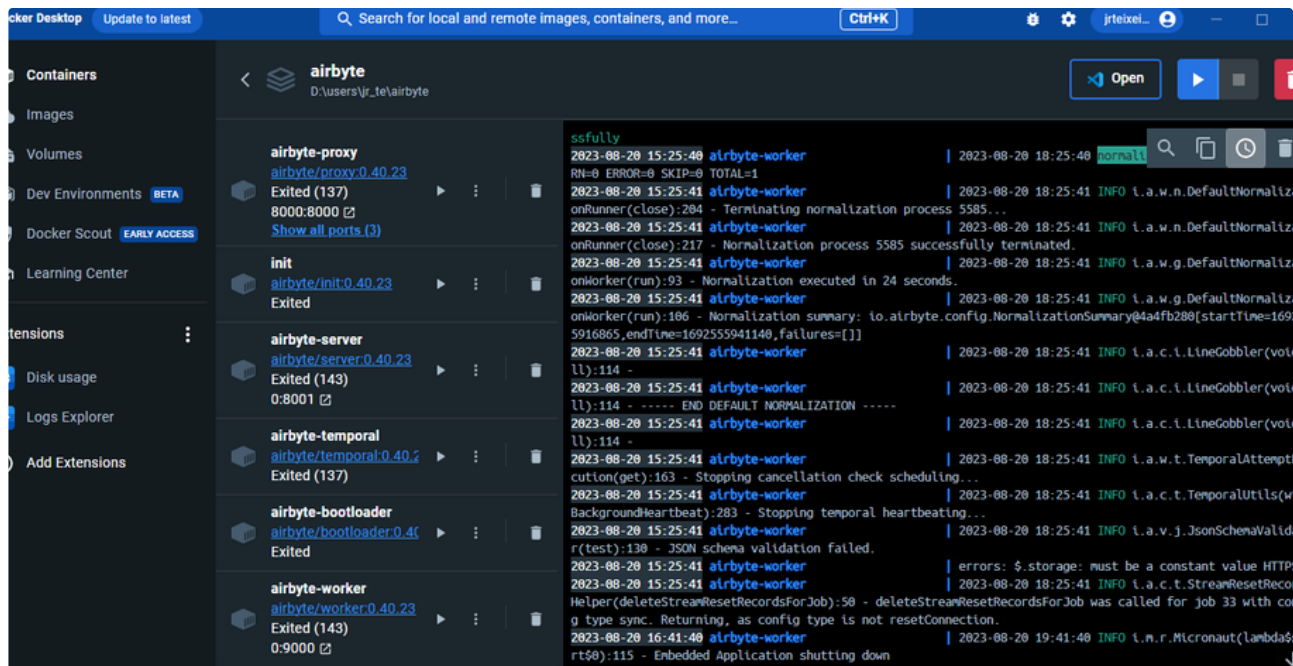


GDrive

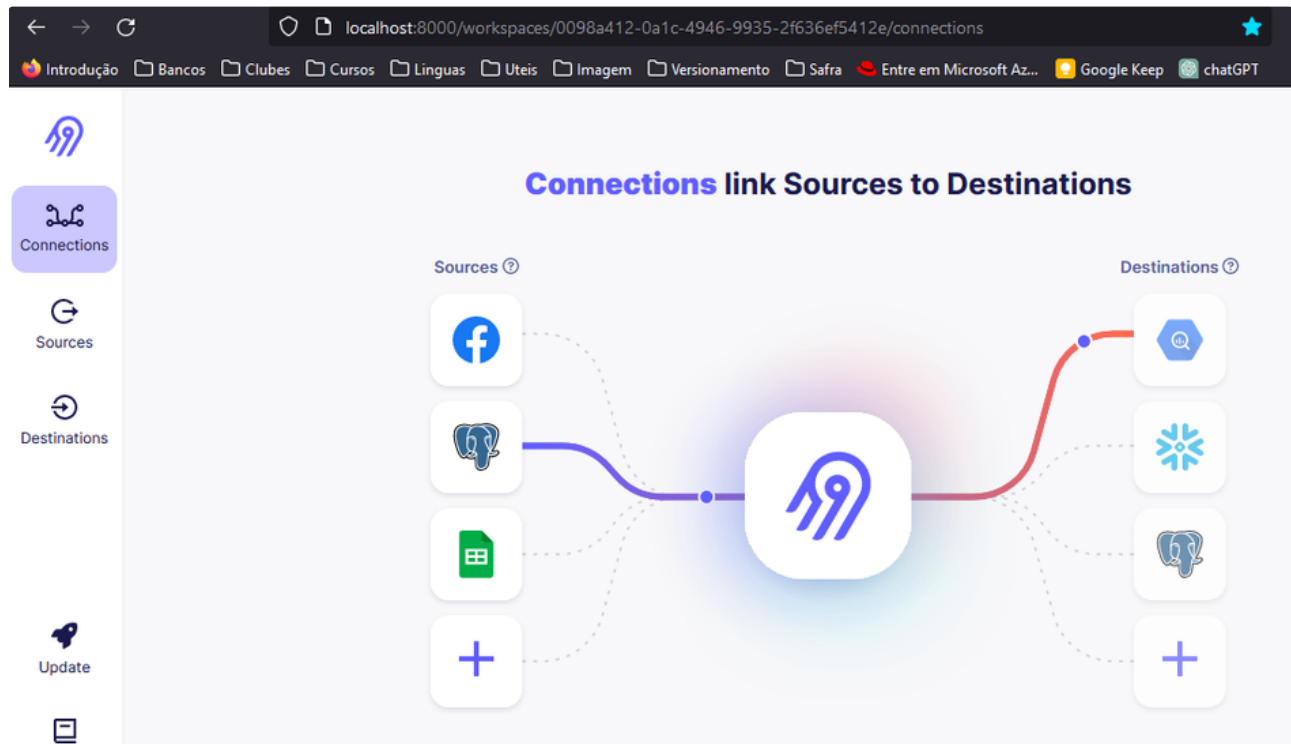
## Ingestão: Airbyte [↗](#)



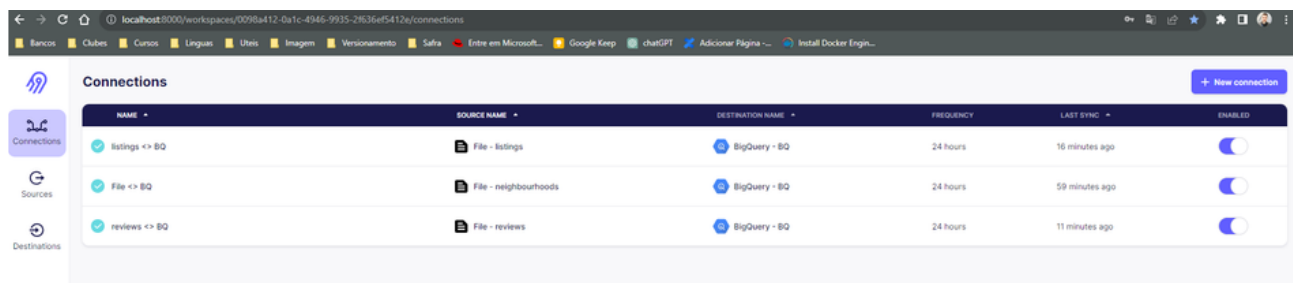
Git Clone do projeto Airbyte para diretório Local



Airbyte Docker



Página Inicia Airbyte



Conexões criadas

Sources / File

OverviewSettings

Source Settings

Source type

File

Source name

File

URL

gs://solutionsprint-fiap/neighbourhoods.csv

File Format

CSV

Storage Provider

GCS: Google Cloud Storage

Service Account JSON

Optional

{ "type": "service\_account", "project\_id": "dw-fiap-fase4", "private\_key\_id": "58f0b7856c"

Dataset Name

dw-fiap-fase4.fiap\_fase4

Reader Options

Optional

Save changes and test

Cancel

Retest source

Configuração Source

Connections

Sources

Destinations

Destinations / BQ

OverviewSettings

Destination Settings

Destination type

BigQuery

Destination name

BQ

Project ID

dw-fiap-fase4

Dataset Location

us-west1

Default Dataset ID

fiap\_fase4

Loading Method

Standard Inserts

Service Account Key JSON (Required for cloud, optional for open-source)

Optional

\*\*\*\*\*

Edit

Transformation Query Run Type

Optional

interactive

Google BigQuery Client Chunk Size

Optional

15

Configuração Destination

Google Cloud

dw-fiap-fase4

Pesquise (/) recursos, documentos, produtos e muito mais

Pesquisa

1

Detalhes do bucket

ATUALIZAR

SAIBA MAIS

solutionsprint-fiap

Local

us-west1 (Oregon)

Classe de armazenamento

Standard

Acesso público

Não público

Proteção

Nenhum

OBJETOS

CONFIGURAÇÃO

PERMISSÕES

PROTEÇÃO

CICLO DE VIDA

OBSERVABILIDADE

RELATÓRIOS DE INVENTÁRIO

Intervalos

solutionsprint-fiap

FAZER UPLOAD DE ARQUIVOS

CARREGAR PASTA

CRIAR PASTA

TRANSFERIR DADOS

GERENCIAR RETENÇÕES

FAZER O DOWNLOAD









EXCLUIR

Filtrar apenas pelo prefixo do nome

Filtro

Filtrar objetos e pastas

Mostrar dados excluídos

<input type="checkbox"/>	Nome	Tamanho	Tipo	Criado	Classe de armazenamento	Última modificação	Acesso público	Version history	Criptografia	Data de validade da retenção	Retenções	
<input type="checkbox"/>	 calendar.csv	29,9 MB	text/csv	14 de ago. de 2023 19:38:01	Standard	14 de ago. de 2023 19:38:01	Não público	—	Gerenciada pelo Google	—	Nenhum	
<input type="checkbox"/>	 listings.csv	5,4 MB	text/csv	14 de ago. de 2023 19:38:00	Standard	14 de ago. de 2023 19:38:00	Não público	—	Gerenciada pelo Google	—	Nenhum	
<input type="checkbox"/>	 neighborhoods.csv	2 KB	text/csv	14 de ago. de 2023 19:38:15	Standard	14 de ago. de 2023 19:38:15	Não público	—	Gerenciada pelo Google	—	Nenhum	
<input type="checkbox"/>	 reviews.csv	12 MB	text/csv	14 de ago. de 2023 19:38:27	Standard	14 de ago. de 2023 19:38:27	Não público	—	Gerenciada pelo Google	—	Nenhum	

Arquivos armazenados Bucket

Data WareHouse: Google BigQuery: [↗](#)

The screenshot shows the Google Cloud Explorer interface for the project 'dw-fiap-fase4'. The left sidebar contains navigation icons. The main area displays a search bar and a list of datasets under the 'fiap\_fase4' folder. The datasets listed are:

- Conexoes externas
- dbt\_amaevsky
- dbt\_fiap\_fase4
- fiap\_fase4
  - \_airbyte\_raw\_airbyte\_listings\_dw\_fiap\_fase4\_fiap\_fase4
  - \_airbyte\_raw\_airbyte\_neighborhoodsw\_dw\_fiap\_fase4\_fiap\_fase4
  - \_airbyte\_raw\_airbyte\_reviews\_dw\_fiap\_fase4\_fiap\_fase4
  - airbyte\_listings\_dw\_fiap\_fase4\_fiap\_fase4
  - airbyte\_neighborhoodsw\_dw\_fiap\_fase4\_fiap\_fase4
  - airbyte\_reviews\_dw\_fiap\_fase4\_fiap\_fase4

Datasets criados via Airflow

The screenshot shows the Google BigQuery Explorer and SQL Editor interface. The Explorer on the left shows the project structure, including the 'dw-fiap-fase4' project and its datasets. The SQL Editor on the right shows a query being executed:

```
1 SELECT * FROM `dw-fiap-fase4`.`fiap_fase4`.`vw_tabela_final` LIMIT 1000
```

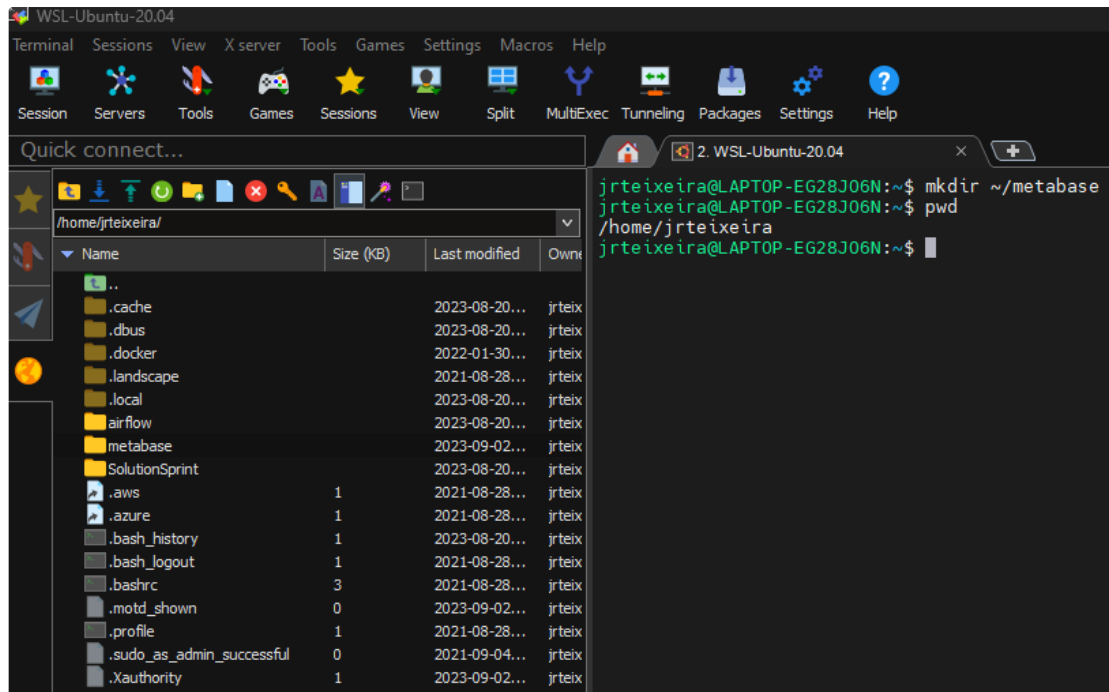
The results of the query are displayed in a table with the following columns: Linha, listing\_id, date, name, host\_id, host\_name, and neighbourhood\_group. The results show three rows of data:

Linha	listing_id	date	name	host_id	host_name	neighbourhood_group
1	17878	2014-11-10	Condo in Rio de Janeiro - • 4.70 · 2 bedrooms · 2 beds · 1 bath	68997	Matthias	null
2	17878	2020-06-19	Condo in Rio de Janeiro - • 4.70 · 2 bedrooms · 2 beds · 1 bath	68997	Matthias	null
3	17878	2013-06-17	Condo in Rio de Janeiro -	68997	Matthias	null

View Criada processo dbt

## Camada de Apresentação: Metabase

Ver arquivo procedimento de instalação metabase.txt

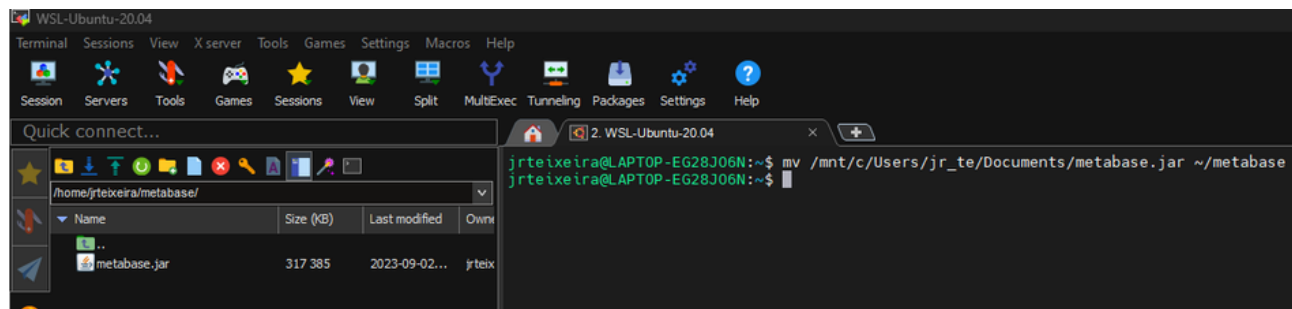


The screenshot shows a WSL terminal window with the following commands and output:

```
jrteixeira@LAPTOP-EG28J06N:~$ mkdir ~/metabase
jrteixeira@LAPTOP-EG28J06N:~$ pwd
/home/jrteixeira
jrteixeira@LAPTOP-EG28J06N:~$
```

Simultaneously, a file explorer window displays the contents of the /home/jrteixeira/ directory. The 'metabase' directory has been successfully created and is listed among other folders like .cache, .dbus, .docker, .landscape, .local, airflow, SolutionSprint, .aws, .azure, .bash\_history, .bash\_logout, .bashrc, .motd\_shown, .profile, .sudo\_as\_admin\_successful, and .Xauthority.

Criando diretório metabase

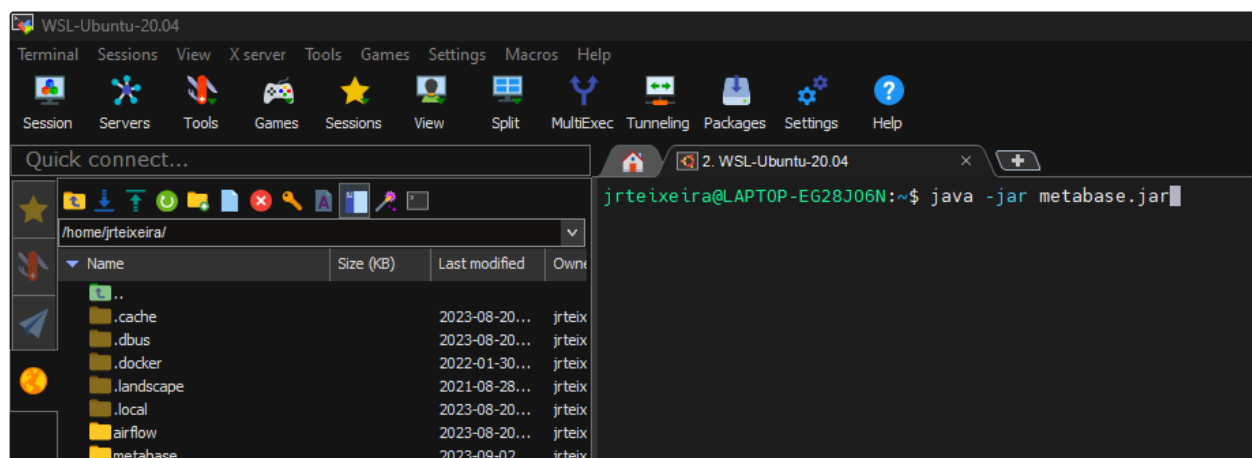


The screenshot shows a WSL terminal window with the following command and output:

```
jrteixeira@LAPTOP-EG28J06N:~$ mv /mnt/c/Users/jr_te/Documents/metabase.jar ~/metabase
jrteixeira@LAPTOP-EG28J06N:~$
```

The file explorer window shows the /home/jrteixeira/metabase/ directory, which now contains the 'metabase.jar' file (317,385 KB, last modified 2023-09-02).

Move arquivo JAR para novo diretório



The screenshot shows a WSL terminal window with the following command and output:

```
jrteixeira@LAPTOP-EG28J06N:~$ java -jar metabase.jar
```

The file explorer window shows the /home/jrteixeira/ directory, where the 'metabase' directory is visible.

Run JAR



```

jrteixeira@LAPTOP-EG2BJ06N:~/metabase$ echo $JAVA_HOME
/usr/lib/jvm/java-11-openjdk-amd64
jrteixeira@LAPTOP-EG2BJ06N:~/metabase$ java -jar metabase.jar
2023-09-02 16:17:05,609 INFO metabase.util :: Maximum memory available to JVM: 2.2 GB
2023-09-02 16:17:10,825 WARN db.env :: WARNING: Using Metabase with an H2 application database is not recommended for production deployments. For production deploy
ment use Postgres, MySQL, or MariaDB instead. If you decide to continue to use H2, please be sure to back up the database file regularly. For more information,
ase.com/docs/latest/operations-guide/migrating-from-h2.html
2023-09-02 16:17:11,324 INFO util.encryption :: Saved credentials encryption is DISABLED for this Metabase instance. @
For more information, see https://metabase.com/docs/latest/operations-guide/encrypting-database-details-at-rest.html
2023-09-02 16:17:28,027 INFO driver.impl :: Registered abstract driver :sql
2023-09-02 16:17:28,062 INFO driver.impl :: Registered abstract driver :sql-jdbc (parents: [:sql])
2023-09-02 16:17:28,079 INFO metabase.util :: Load driver :sql-jdbc took 308.4 ms
2023-09-02 16:17:28,081 INFO driver.impl :: Registered driver :h2 (parents: [:sql-jdbc])
2023-09-02 16:17:28,243 INFO driver.impl :: Registered driver :mysql (parents: [:sql-jdbc])
2023-09-02 16:17:28,363 INFO driver.impl :: Registered driver :postgres (parents: [:sql-jdbc])
2023-09-02 16:17:37,078 INFO metabase.core ::
Metabase v0.47.0 (682ec2d release-x.47.x)

Copyright © 2023 Metabase, Inc.

Metabase Enterprise Edition extensions are NOT PRESENT.
2023-09-02 16:17:37,108 INFO metabase.core :: Starting Metabase in STANDALONE mode
2023-09-02 16:17:37,302 INFO metabase.server :: Launching Embedded Jetty Webserver with config:
{:port 3000}

2023-09-02 16:17:37,461 INFO metabase.core :: Starting Metabase version v0.47.0 (682ec2d release-x.47.x) ...
2023-09-02 16:17:37,530 INFO metabase.core :: System info:
{"file.encoding" "UTF-8",
 "java.runtime.name" "OpenJDK Runtime Environment",
 "java.runtime.version" "11.0.20.1+1-post-Ubuntu-0ubuntu120.04",
 "java.vendor" "Ubuntu",
 "java.vendor.url" "https://ubuntu.com/",
 "java.version" "11.0.20.1",
 "java.vm.name" "OpenJDK 64-Bit Server VM",
 "java.vm.version" "11.0.20.1+1-post-Ubuntu-0ubuntu120.04",
 "os.name" "Linux",
 "os.version" "5.10.102.1-microsoft-standard-WSL2",
 "user.language" "en",
 "user.timezone" "America/Sao_Paulo"}

2023-09-02 16:17:37,535 INFO metabase.plugins :: Loading plugins in /home/jrteixeira/metabase/plugins...
2023-09-02 16:17:38,646 INFO util.files :: Extract file /modules/sparksql.metabase-driver.jar -> /home/jrteixeira/metabase/plugins/sparksql.metabase-driver.jar
2023-09-02 16:17:39,696 INFO util.files :: Extract file /modules/athena.metabase-driver.jar -> /home/jrteixeira/metabase/plugins/athena.metabase-driver.jar
2023-09-02 16:17:39,793 INFO util.files :: Extract file /modules/sqlserver.metabase-driver.jar -> /home/jrteixeira/metabase/plugins/sqlserver.metabase-driver.jar
2023-09-02 16:17:39,793 INFO util.files :: Extract file /modules/presto-jdbc.metabase-driver.jar -> /home/jrteixeira/metabase/plugins/presto-jdbc.metabase-driver.jar
2023-09-02 16:17:40,222 INFO util.files :: Extract file /modules/sqlite.metabase-driver.jar -> /home/jrteixeira/metabase/plugins/sqlite.metabase-driver.jar
2023-09-02 16:17:40,593 INFO util.files :: Extract file /modules/vertica.metabase-driver.jar -> /home/jrteixeira/metabase/plugins/vertica.metabase-driver.jar
2023-09-02 16:17:40,596 INFO util.files :: Extract file /modules/redshift.metabase-driver.jar -> /home/jrteixeira/metabase/plugins/redshift.metabase-driver.jar
2023-09-02 16:17:40,640 INFO util.files :: Extract file /modules/druid.metabase-driver.jar -> /home/jrteixeira/metabase/plugins/druid.metabase-driver.jar
2023-09-02 16:17:40,671 INFO util.files :: Extract file /modules/mongo.metabase-driver.jar -> /home/jrteixeira/metabase/plugins/mongo.metabase-driver.jar
2023-09-02 16:17:40,813 INFO util.files :: Extract file /modules/oracle.metabase-driver.jar -> /home/jrteixeira/metabase/plugins/oracle.metabase-driver.jar
2023-09-02 16:17:40,822 INFO util.files :: Extract file /modules/snowflake.metabase-driver.jar -> /home/jrteixeira/metabase/plugins/snowflake.metabase-driver.jar
2023-09-02 16:17:41,840 INFO util.files :: Extract file /modules/googleanalytics.metabase-driver.jar -> /home/jrteixeira/metabase/plugins/googleanalytics.metabase-d
2023-09-02 16:17:41,928 INFO util.files :: Extract file /modules/bigquery-cloud-sdk.metabase-driver.jar -> /home/jrteixeira/metabase/plugins/bigquery-cloud-sdk.meta
2023-09-02 16:17:43,439 DEBUG plugins.lazy-loaded-driver :: Registering lazy loading driver :athena...
2023-09-02 16:17:43,440 INFO driver.impl :: Registered driver :athena (parents: [:sql-jdbc])

```

localhost:3000/setup

Cursos

Linguas

Uteis

Imagem

Versionamento

Safra

Entre em Microsoft...

Google Keep

chatGPT

Adicionar Página ...

✓

Olá, Jose. Prazer em te conhecer!

3

Adicione seus dados

Você está pronto para começar a explorar seus dados? Adicione abaixo.

Não está pronto? Pule e dê uma olhada no nosso banco de dados de exemplo.

Busque por um banco de dados...

MySQL

PostgreSQL

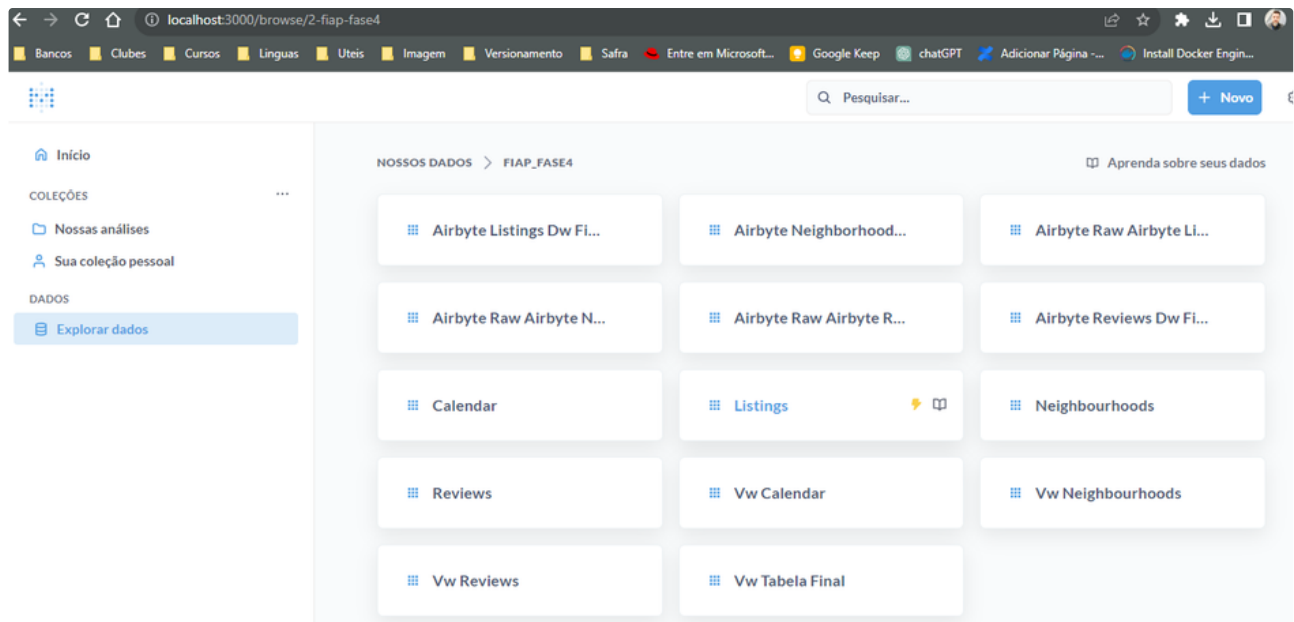
SQL Server

Amazon Redshift

BigQuery

Snowflake

Metabase: localhost:3000



Metabase conectado ao BigQuery

### Dicionário das variáveis

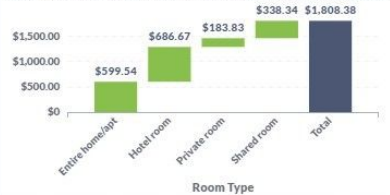
- `id` - número de identificação do imóvel
- `name` - Título do anúncio da propriedade
- `host_id` - número de identificação do proprietário
- `host_name` - nome do anfitrião
- `neighbourhood_group` - coluna sem valores válidos
- `neighbourhood` - nome do bairro
- `latitude` - coordenada de latitude da propriedade
- `longitude` - coordenada de longitude da propriedade
- `room_type` - tipo de acomodação oferecida
- `price` - valor do aluguel
- `minimum_nights` - menor quantidade de noites para alugar
- `number_of_reviews` - número de reviews
- `last_review` - data do último review
- `reviews_per_month` - quantidade de reviews em um mês
- `calculated_host_listings_count` - quantidade de imóveis do mesmo anfitrião
- `availability_365` - número de dias de disponibilidade dentro de 365 dias

## Aba 1 +

Aba 1 +

Qual a média dos preços de aluguel?

Qual a média dos preços de aluguel?

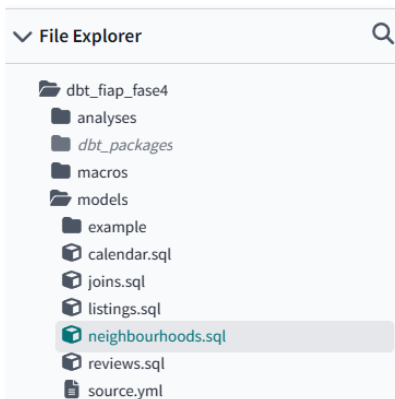


País	Preço médio (R\$)
Portugal	84,29
Brasil	70,17
Estados Unidos	89,80
Reino Unido	91,50
Países Baixos	90,00

## Camada de Transformação DBT CLOUD [↗](#)

Criamos a conexão entre o Bigquery e o DBT Cloud

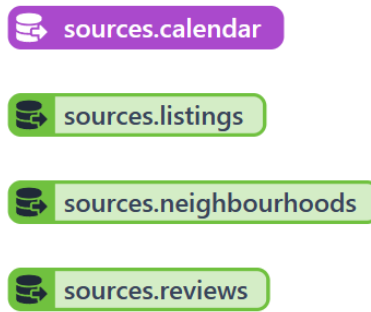
configuramos o arquivo source.yml para acessar as bases de dados do bigquery



Detalhe da configuração do source:

```
models > source.yml
1  version: 2
2
3  sources:
4    - name: sources
5      database: dw-fiap-fase4
6      schema: fiap_fase4
7      tables:
8        Generate model
9        - name: calendar
10       Generate model
11       - name: listings
12       Generate model
13       - name: neighbourhoods
14       Generate model
15       - name: reviews
```

Lineage:



Create the view in DBT cloud:

Criando a view `calc_calendar`

```
models > calendar.sql

1 with calc_calendar as (
2   select
3     *
4     from {{source('sources','calendar')}}
5 )
6 select * from calc_calendar
7
```

Criando a view `calc_listings`

```
models > listings.sql

1 with calc_listings as (
2   select
3     *
4     from {{source('sources','listings')}}
5 )
6 select * from calc_listings
```

Criando a view `calc_neighbourhoods`

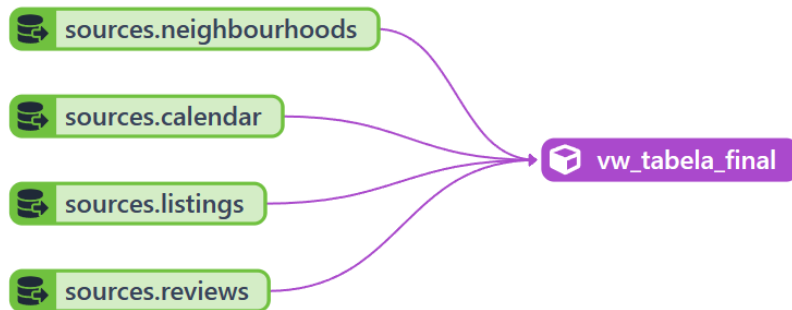
```
models > neighbourhoods.sql

1 with calc_neighbourhoods as (
2   select *
3   from {{source('sources','neighbourhoods')}}
4 )
5 select * from calc_neighbourhoods
```

Criando a view `calc_reviews`

```
models > reviews.sql
1 with calc_reviews as (
2   select *
3   from {{source('sources','reviews')}}
4 )
5 select * from calc_reviews
```

Criando a view final com a união de todas as views



```
models > joins.sql
1 with calc_table_final as (
2   select
3     *
4   from {{source('sources','calendar')}} "calendar"
5   left join {{source('sources','listings')}} "listings" on ("calendar".listing_id = "listings".id)
6   left join {{source('sources','neighbourhoods')}} "neighbourhoods" on ("neighbourhoods".neighbourhood_group = "listings".neighbourho
7   left join {{source('sources','reviews')}} "reviews" on ("reviews".listing_id = "listings".id)
8 )
9 select * from calc_table_final
```

Gravando os dados da view do DBT no Bigquery

The screenshot shows the BigQuery interface. On the left, the Explorer pane lists various tables, including `vw_tabela_final`. The main pane shows a query execution result with the following data:

Linha	listing_id	date	name	host_id	host_name	neighbourhood_group
1	17878	2014-11-10	Condo in Rio de Janeiro - • 4.70 - 2 bedrooms - 2 beds - 1 bath	68997	Matthias	null
2	17878	2020-06-19	Condo in Rio de Janeiro - • 4.70 - 2 bedrooms - 2 beds - 1 bath	68997	Matthias	null
3	17878	2013-06-17	Condo in Rio de Janeiro -	68997	Matthias	null

Uma solução para governança e catálogo de dados:

[view](#)

- Description  
Columns  
Depends On  
Code

## TAGS

untagged

TYP

## PACKAGE

LANGUAGE  
sal

dy-fiap-fase

protected

## CONTRACT

Description

Column8

11

COLUMN	TYPE	DESCRIPTION	CONSTRAINTS	YES	MORE ?
listing_id	INT64				

11

COLUMN	TYPE	CONSTRAINTS	TESTS	MORE
listing_id	INT64			
date	DATE			
available	BOOL			
price	INT64			
adjusted_price	INT64			
minimum_nights	INT64			
maximum_nights	INT64			

- Now

- [calendar](#)

- Compiled
- copy to clipboard

selected



Link Inside Airbnb



### Entregável 3

#### Detalhes da instalação e configuração realizada para implementação da arquitetura sugerida para solução

- Usamos WSL para simular um servidor Ubuntu com Airflow configurado;
- Instalamos Docker para simular o seridor com Airbyte;
- Baixamos imagem Airflow via Terminal (Airflow >> Evidência) ;
- Baixamos os arquivos para pasta local e depois subimos para o Google Drive via script Python;
- clonamos repositório AirByte para acessá-lo via Docker (AirByte >> Evidência);
- Criamos as tabelas usando os conectores no Airbyte (GCS e BigQuery) (Evidências BigQuery);
- Acessamos link da página: [Metabase | Business Intelligence, Dashboards, and Data Visualization](#)  
Em Documentation >> Installation and operation overview >> Running the Metabase JAR file >> Download
- Iniciamos procedimento para instalação no Ubuntu WSL (ver arquivo instalacao\_metabase.txt) (Evidências Metabase)

#### Script(s)/comandos utilizados na transformação de dados, modelo(s) de Machine Learning e qualquer outra etapa adicional/opcional à arquitetura

```
#!/usr/bin/env python3
# download_files.py
import os
import sys
import requests
import shutil

# URL do arquivo a ser baixado
url = "https://storage.googleapis.com/airbyte-repo/airbyte-repo-2023-07-15.tar.gz"

# Nome do arquivo local
local_file = "airbyte-repo-2023-07-15.tar.gz"

# Verificar se o arquivo já existe
if os.path.exists(local_file):
    print(f"Arquivo {local_file} já existe.")
else:
    # Baixar o arquivo
    response = requests.get(url)
    if response.status_code == 200:
        # Salvar o arquivo localmente
        with open(local_file, 'wb') as f:
            f.write(response.content)
        print(f"Arquivo {local_file} baixado com sucesso.")
    else:
        print(f"Erro ao baixar o arquivo. Status code: {response.status_code}")

# Descompactar o arquivo
if os.path.exists(local_file):
    print("Descompactando o arquivo...")
    os.system(f"tar -xzf {local_file}")
    print("Arquivo descompactado com sucesso.")
else:
    print(f"Arquivo {local_file} não encontrado.")
```

**download\_files.py**  
15 jul. 2023, 10:56 PM

```
#!/usr/bin/env bash
# script_config_airflow.txt
set -e

# Definir variáveis de ambiente
AIRFLOW_HOME=/opt/airflow
AIRFLOW_UID=1000
AIRFLOW_GID=1000

# Instalar dependências
apt-get update
apt-get install -y python3-pip python3-venv

# Criar o ambiente virtual do Airflow
python3 -m venv $AIRFLOW_HOME/venv
source $AIRFLOW_HOME/venv/bin/activate

# Instalar o Airflow
pip install apache-airflow

# Criar o usuário airflow
useradd -ms /bin/bash airflow

# Criar o diretório de logs do Airflow
mkdir -p $AIRFLOW_HOME/logs
```

**script\_config\_airflow.txt**  
20 ago. 2023, 09:54 PM

```
#!/usr/bin/env bash
# instalacao_metabase.txt
set -e

# Definir variáveis de ambiente
METABASE_HOME=/opt/metabase
METABASE_UID=1000
METABASE_GID=1000

# Instalar o Metabase
curl -L https://github.com/metabase/metabase/releases/download/v0.44.0/metabase-0.44.0-linux-amd64.tar.gz | tar -xzf -

# Criar o usuário metabase
useradd -ms /bin/bash metabase

# Criar o diretório de logs do Metabase
mkdir -p $METABASE_HOME/logs
```

**instalacao\_metabase.txt**  
02 set. 2023, 06:34 PM

#### Definição detalhada da modelagem de dados