

Teste Boticário – Bruno Flammarion C. Boscolo

1 - Realizar a importação dos dados dos 3 arquivos em uma tabela criada por você no banco de dados de sua escolha;


Existem várias formas de desenvolver essa ingestão, priorizei um desenvolvimento de um pipeline python com dataflow.

Outras formas de fazer essa ingestão:

1. bq load
2. dataflow com fluxos de template
3. datafusion
4. via interface do Bigquery
5. Outras ferramentas Nifi, Streamsets e outras de ETL.

Para o projeto criei um ambiente na GCP, todo o teste foi em cima desse projeto.

Selecione um projeto

 **NOVO PROJETO**

 Pesquisar projetos e pastas

RECENTE

TODOS

Nome	ID
 testeboticario 	testeboticario-294416

Para iniciar o processo de ingestão, precisei configurar o ambiente com gcloud para enviar os arquivos, levando em consideração que um sistema ou fluxo ETL entregaria os arquivos no cloud storage.

```
gcloud auth login
gcloud config set project testeboticario-294416
gsutil mb gs://testeboticario
gsutil cp Base*.csv gs://testeboticario
```

Alguns pré-requisitos para rodar dataflow:

1. Habilitar o dataflow API
2. Um account service com as devidas permissões (fiz a criação do dataflow@testeboticario-294416.iam.gserviceaccount.com)
3. Instalar as dependências
4. Ter o dataset criado, criei um dataset com o nome (teste_boticatio)

Instalação de dependências: `python3.7 -m pip install -r /home/testeboticariobruno/pipeline/requirements.txt`

Para rodar o dataflow, deve executar os seguintes scripts:

Base_2017_1.csv:

```
python3.7 /home/testeboticariobruno/pipeline/df_csv_bq.py --project
testeboticario-294416 --autoscaling_algorithm THROUGHPUT_BASED --
service_account_email dataflow@testeboticario-
294416.iam.gserviceaccount.com --job_name df-boticario-csv-bq --
runner DataflowRunner --input gs://testeboticatio/Base_2017_1.csv
--staging_location gs://testeboticatio/process --temp_location
gs://testeboticatio/temp --requirements_file
/home/testeboticariobruno/pipeline/requirements.txt --region us-
central1
```

Base_2018_2.csv:

```
python3.7 /home/testeboticariobruno/pipeline/df_csv_bq.py --project
testeboticario-294416 --autoscaling_algorithm THROUGHPUT_BASED --
service_account_email dataflow@testeboticario-
294416.iam.gserviceaccount.com --job_name df-boticario-csv-bq --
runner DataflowRunner --input gs://testeboticatio/Base_2018_2.csv
--staging_location gs://testeboticatio/process --temp_location
gs://testeboticatio/temp --requirements_file
/home/testeboticariobruno/pipeline/requirements.txt --region us-
central1
```

Base_2019_3.csv:

```
python3.7 /home/testeboticariobruno/pipeline/df_csv_bq.py --project
testeboticario-294416 --autoscaling_algorithm THROUGHPUT_BASED --
service_account_email dataflow@testeboticario-
294416.iam.gserviceaccount.com --job_name df-boticario-csv-bq --
runner DataflowRunner --input gs://testeboticatio/Base_2019_3.csv
--staging_location gs://testeboticatio/process --temp_location
gs://testeboticatio/temp --requirements_file
/home/testeboticariobruno/pipeline/requirements.txt --region us-
central1
```

Print de uma das execuções:

The screenshot displays the Google Cloud Platform Dataflow console. The main view shows a job named 'df-boticario-csv-bq' with a status of 'Finalizado' (Completed). The job details on the right include:

- Nome do job: df-boticario-csv-bq
- ID do job: 2020-11-02_11_11_23-8982206869640596659
- Tipo de job: Lote
- Status do job: Finalizado
- Versão do SDK: Apache Beam Python 3.7 SDK 2.15.0
- Região do job: us-central1
- Local do worker: us-central1-b
- Quantidade atual de workers: 0
- Status mais recente do worker: Worker pool stopped.
- Horário de início: 2 de novembro de 2020 16:11:26 GMT-3
- Tempo decorrido: 5 min 41 s
- Tipo de criptografia: Chave gerenciada pelo Google

The job metrics on the right show 1 vCPU atual (active) and 0.00 vCPUs reserved. The main view shows a DAG with three steps: 'Read from a File', 'Format', and 'Write to BigQuery', all marked as 'Finalizado'.

Foi desenvolvido dessa forma pensando em criar uma cloud function para iniciar o processo batch assim que arquivo chegar no cloud Storage;

Segue print da tabela criada, até desenvolvi o pipeline para trabalhar com tabela particionada, mas quando fui fazer a agenda do Bigquery no Airflow encontrei alguns problemas quando as consultas no BigQueryOperator. Acabei voltando meu código e tabela para trabalhar sem o particionamento.

The screenshot shows the Google Cloud Platform interface with the BigQuery console open. The left sidebar contains navigation options like 'Histórico de consultas', 'Consultas salvas', 'Histórico de jobs', 'Transferências', 'Consultas programadas', 'Reservas', 'BI Engine', and 'Recursos'. The main area displays the 'Editor de consultas' for the table 'base_boticario'. The 'Detalhes' tab is active, showing the table's description, schema, and a table of information.

Informações da tabela	
Código da tabela	testeboticario-294416:base_boticario
Tamanho da tabela	161,08 KB
Número de linhas	3.000
Criada em	2 de nov. de 2020 16:16:10
Expiração da tabela	Nunca
Última modificação	2 de nov. de 2020 16:36:38
Local dos dados	US

2 - Com os dados importados, modelar 4 novas tabelas e implementar processos que façam as transformações necessárias e insiram as seguintes visões nas tabelas:

Para executar o Airflow foi necessário habilitar o Composer na GCP, fiz a escrita de 4 Dags para gerar as visões.

The screenshot shows the Airflow web interface. The top navigation bar includes 'Airflow', 'DAGs', 'Data Profiling', 'Browse', 'Admin', 'Docs', and 'About'. The main content area is titled 'DAGs' and features a search bar and a table of DAGs.

DAG	Schedule	Owner	Recent Tasks	Last Run	DAG Runs	Links
airflow_monitoring	None	airflow		2020-11-02 19:48		
tabela1	0 0 * * *	airflow		2020-11-01 00:00		
tabela2	0 0 * * *	airflow		2020-11-01 00:00		
tabela3	0 0 * * *	airflow		2020-11-01 00:00		
tabela4	0 0 * * *	airflow		2020-11-01 00:00		

The screenshot shows the Google Cloud Platform Storage console. The left sidebar contains navigation options like 'Storage', 'Navegador', 'Monitoramento', and 'Configurações'. The main area displays the 'Detalhes do bucket' for 'us-central1-testeboticario-9e578c0c-bucket'. The 'OBJETOS' tab is active, showing a list of objects.

Nome	Tamanho	Tipo	Horário da criação
airflow_monitoring	729 B	text/x-python	2 de nov. de 2020 1...
tabela1.py	2,2 KB	text/plain	2 de nov. de 2020 1...
tabela2.py	2,2 KB	text/plain	2 de nov. de 2020 1...
tabela3.py	2,4 KB	text/plain	2 de nov. de 2020 1...
tabela4.py	2,4 KB	text/plain	2 de nov. de 2020 1...

Tabela1: Consolidado de vendas por ano e mês;

← → ↻ console.cloud.google.com/bigquery?hl=pt-br&organizationId=0&project=testebotuario-294416&p=testebotuario-294416&d=teste_boticario&t=tabela1&page=table

Status do período de teste gratuito: R\$ 1.387,88 de crédito e 91 dias restantes. Com uma conta completa, você tem acesso ilimitado a todos os recursos do Google Cloud Platform.

DISPENSAR ATIVAR

Google Cloud Platform testebotuario Pesquisar produtos e recursos

BigQuery RECURSOS E INFORMAÇÕES ATALHO

Histórico de consultas

Consultas salvas

Histórico de jobs

Transferências

Consultas programadas

Reservas

BI Engine

Recursos

testebotuario-294416

teste_boticario

tabela1

tabela2

tabela3

tabela4

Editor de consultas

+ ESCREVER NOVA CONSULTA

OCULTAR EDITOR

TELA CHEIA

1

Executar Salvar consulta Salvar visualização Programar consulta Mais

tabela1

CONSULTAR TABELA

COMPARTILHAR TABELA

COPIAR TABELA

EXCLUIR TABELA

EXPORTAR

Esquema Detalhes Visualização

Linha	ano	mes	total
1	2018	1	2773
2	2018	2	2211
3	2018	3	2532
4	2018	4	2193
5	2018	5	2717
6	2018	6	2363
7	2018	7	2801
8	2018	8	2510
9	2018	9	2500
10	2018	10	3125

Tabela2: Consolidado de vendas por marca e linha;

← → ↻ console.cloud.google.com/bigquery?hl=pt-br&organizationId=0&project=testebotuario-294416&p=testebotuario-294416&d=teste_boticario&t=tabela2&page=table

Status do período de teste gratuito: R\$ 1.387,88 de crédito e 91 dias restantes. Com uma conta completa, você tem acesso ilimitado a todos os recursos do Google Cloud Platform.

DISPENSAR ATIVAR

Google Cloud Platform testebotuario Pesquisar produtos e recursos

BigQuery RECURSOS E INFORMAÇÕES ATALHO

Histórico de consultas

Consultas salvas

Histórico de jobs

Transferências

Consultas programadas

Reservas

BI Engine

Recursos

testebotuario-294416

teste_boticario

tabela1

tabela2

tabela3

tabela4

Editor de consultas

+ ESCREVER NOVA CONSULTA

OCULTAR EDITOR

TELA CHEIA

1

Executar Salvar consulta Salvar visualização Programar consulta Mais

tabela2

CONSULTAR TABELA

COMPARTILHAR TABELA

COPIAR TABELA

EXCLUIR TABELA

EXPORTAR

Esquema Detalhes Visualização

Linha	marca	linha	total
1	BELEZA NA WEB	CABELOS	1254
2	BELEZA NA WEB	HIDRATANTES	1315
3	BELEZA NA WEB	MAQUIAGEM	1351
4	BELEZA NA WEB	PERFUMARIA	1160
5	BELEZA NA WEB	SOLAR	1351
6	BOTICÁRIO	CABELOS	1645
7	BOTICÁRIO	HIDRATANTES	1162
8	BOTICÁRIO	MAQUIAGEM	1211
9	BOTICÁRIO	PERFUMARIA	1312
10	BOTICÁRIO	SOLAR	1220

Linhas por página: 100 1 - 25 de 25 Primeira página < > Última página

Tabela3: Consolidado de vendas por marca, ano e mês;

The screenshot shows the Google Cloud Platform BigQuery interface. On the left, a sidebar lists project resources under 'testeboticario-294416', including 'base_boticario', 'tabela1', 'tabela2', 'tabela3' (selected), and 'tabela4'. The main area displays 'Tabela3' with a table view. The table has columns: 'Linha', 'marca', 'ano', 'mes', and 'total'. It contains 10 rows of data for the year 2018, with 'marca' consistently being 'BELEZA NA WEB'.

Linha	marca	ano	mes	total
1	BELEZA NA WEB	2018	1	612
2	BELEZA NA WEB	2018	2	446
3	BELEZA NA WEB	2018	3	445
4	BELEZA NA WEB	2018	4	439
5	BELEZA NA WEB	2018	5	618
6	BELEZA NA WEB	2018	6	535
7	BELEZA NA WEB	2018	7	589
8	BELEZA NA WEB	2018	8	530
9	BELEZA NA WEB	2018	9	457
10	BELEZA NA WEB	2018	10	761

Tabela4: Consolidado de vendas por linha, ano e mês;

The screenshot shows the Google Cloud Platform BigQuery interface. On the left, a sidebar lists project resources under 'testeboticario-294416', including 'base_boticario', 'tabela1', 'tabela2', 'tabela3', and 'tabela4' (selected). The main area displays 'Tabela4' with a table view. The table has columns: 'Linha', 'linha', 'ano', 'mes', and 'total'. It contains 10 rows of data for the year 2018, with 'linha' consistently being 'CABELOS'.

Linha	linha	ano	mes	total
1	CABELOS	2018	1	680
2	CABELOS	2018	2	387
3	CABELOS	2018	3	436
4	CABELOS	2018	4	451
5	CABELOS	2018	5	526
6	CABELOS	2018	6	590
7	CABELOS	2018	7	601
8	CABELOS	2018	8	562
9	CABELOS	2018	9	558
10	CABELOS	2018	10	605

7 - Criar um processo de captura de dados através da API do Twitter, que utilize os seguintes parâmetros:

- Palavras a serem pesquisadas: “Boticário” e o nome da linha com mais vendas no mês 12 de 2019 (conforme item 2.d.);
- Recuperar os 50 twitts mais recentes;
- Recuperar apenas twitts que estejam em português.

Para o processo do Twitter, iniciei com uma tentativa com Datafusion, mas tive alguns problemas com Account Service no Dataproc, acabei fazendo em python mesmo, aproveitei um projeto existente:

https://github.com/TDehaene/blogposts/blob/master/got_sentiment/4_streaming_pipeline/streaming_tweet.py

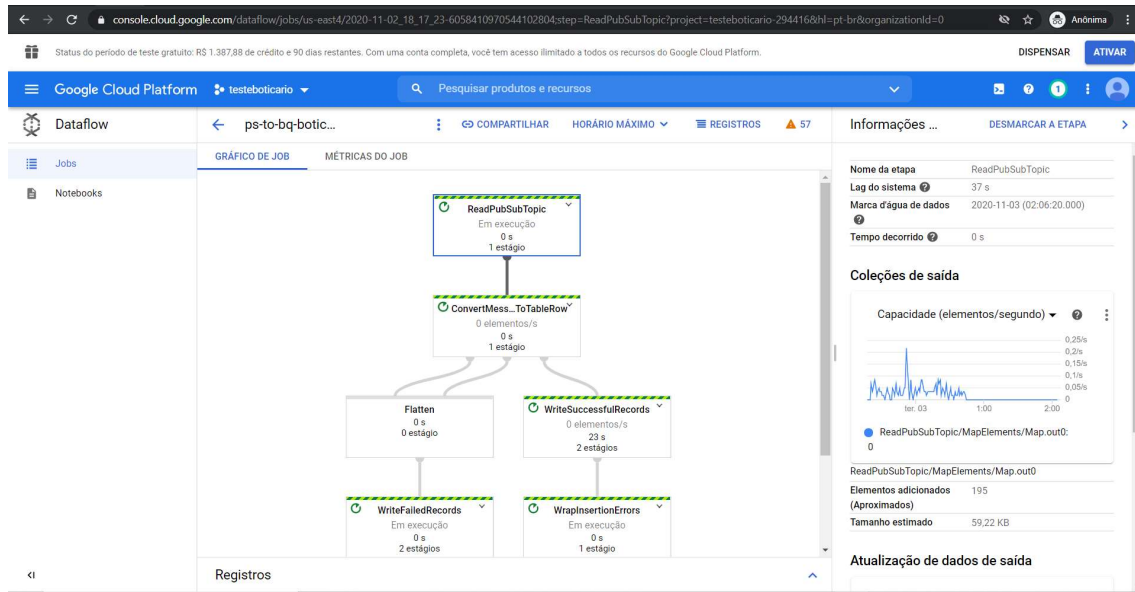
Utilizei o python para fazer as requisições no Twitter e jogar os dados para um Pubsub tópico chamado boticário.

Para enviar os dados para o bigquery utilizei um template pronto:

Equivalência do comando:

```
gcloud dataflow jobs run ps-to-bq-boticario --gcs-location
gs://dataflow-templates-us-east4/latest/PubSub_to_BigQuery --region
us-east4 --staging-location gs://testeboticario/temp --parameters
inputTopic=projects/testeboticario-294416/topics/boticario2,outputTableSpec=testeboticario-294416:teste_boticatio.twitter_boticario
```


Print da execução:



Print da tabela

The screenshot displays the Google Cloud BigQuery console. The query editor shows the following SQL query:

```
1 select * from `teste_boticatio.twitter_boticario` where text like '%boticario%'
2
```

The query results are displayed in a table with the following columns: **Linha**, **text**, **user_id**, **user_name**, **id**, and **posted_at**.

Linha	text	user_id	user_name	id	posted_at
1	hidratante pimenta rosa e framboesa da natura é tudo p mim, pau a pau com morango e leite da o boticario	4828136480	mariazinha	1323458948058501122	2020-11-03 02:56:16
2	ÁCIDO HIALURÔNICO. CREME. #botik #boticario @oboticario 👉👉👉 VENDEDORA @bolosda.enica https://t.co/ztUzFP60m	1263112709790449669	Ysa santana	1323466720556781568	2020-11-03 03:27:09