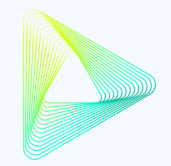


mercafácil

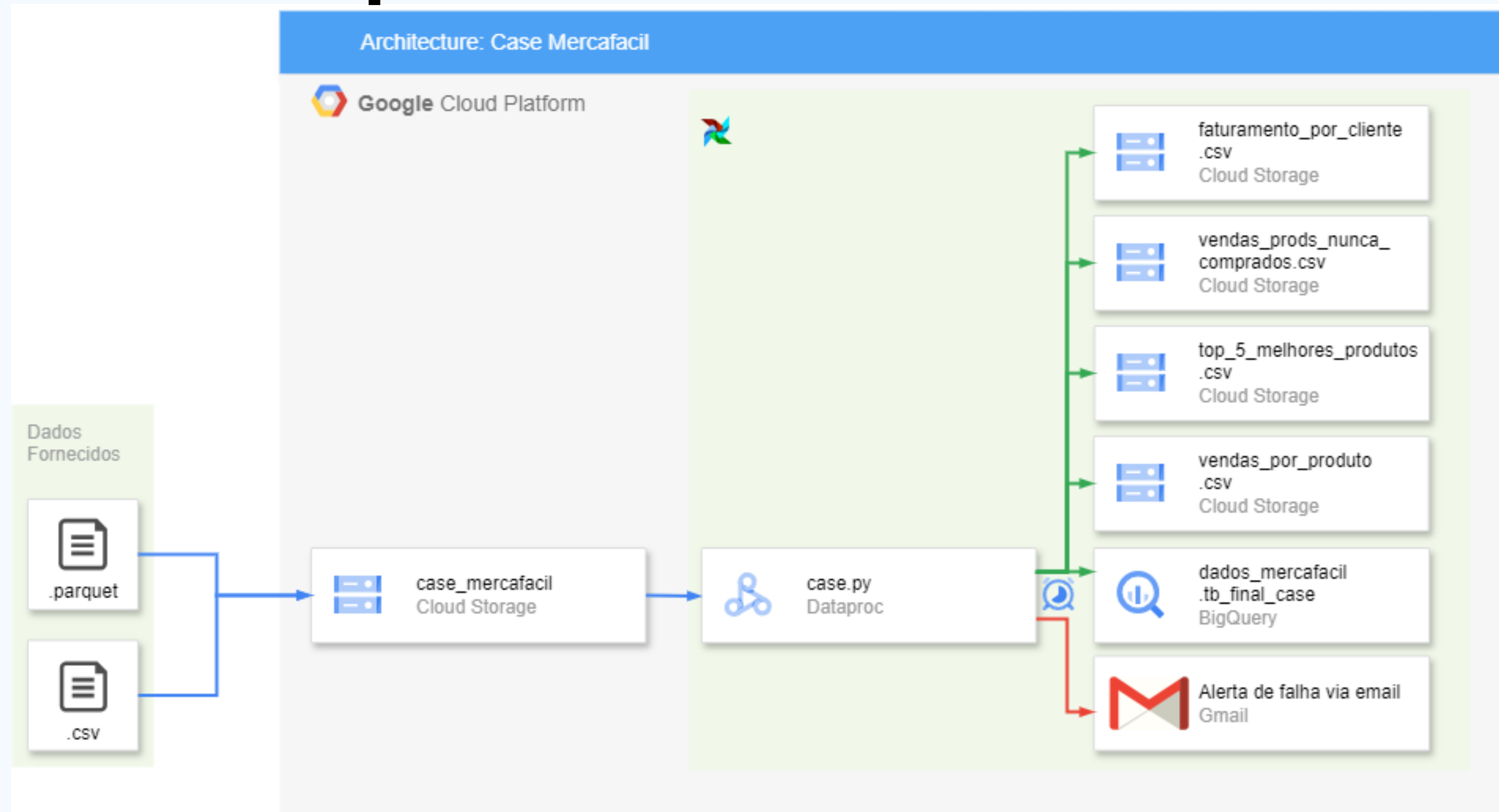
Case Data Eng

Amanda Louise Costa Nascimento

Arquitetura Proposta



mercafácil

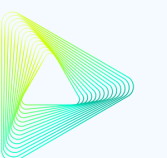


Motivação

- Foi utilizada a versão trial da GCP com o intuito de tornar a solução a mais próxima possível da realidade da empresa. Para isso, todos os dados recebidos foram inseridos em um bucket do Google Cloud Storage

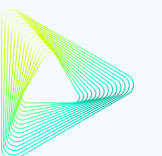
Desenvolvimentos

- Script pyspark armazenado no bucket do Cloud Storage
 - Conversão de tipos de dados adequados.
 - Deduplicação de dados
 - Abertura de arrays
- DAG dag_case_mercafacil.py criada, referenciando o script pyspark case_mercafacil.py
 - Exceptions em cada função
- Tabela dados_mercafacil.tb_final_case criada no BigQuery
 - Joins dos 4 datasets tratados



Desenvolvimentos

- Arquivos com agregações criados no Cloud Storage
 - Vendas por produto ()
 - Total de transações e status de validação
- Monitoramento
 - Exceptions na main do script pyspark para melhor visualização do erro em logs
 - 'email_on_failure' configurado na DAG para envio de email em caso de falha



Desenvolvimentos

- Tabela formada pela união dos datasets armazenada no BigQuery

Google Cloud

My First Project

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

Pesquisa

Seus projetos do BigQuery vão ter novos recursos a partir de 14 de fevereiro de 2024. Os serviços e papéis serão ativados de modo automático para ajudar nessas mudanças.

Saiba mais

DISPENSAR

Explorer

+ ADICIONAR

<

Digite para pesquisar

Você está visualizando os recursos.

MOSTRAR APENAS COM ESTRELA

eco-palisade-416013

Consultas

Notebooks

Conexões externas

dados_mercafacil

tb_final_case

tb_vendas

RESUMO

tb_final_case

eco-palisade-416013.dados_mercafacil

Última modificação

4 de mar. de 2024, 17:18:36 UTC-3

Local dos dados

US

tb_final_case

CONSULTA

COMPARTILHAR

COPIAR

SNAPSHOT

EXCLUIR

EXPORTAR

ATUALIZAR

ESQUEMA

DETALHES

VISUALIZAÇÃO

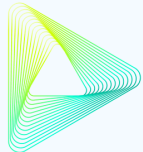
LINHAGEM

PERFIL DE DADOS

QUALIDADE DOS DADOS

Linha	COD_ID_CATEG	COD_ID_CLIENT	DES_SEXO_C	DES_TIPO_C	COD_ID_PRODU	COD_ID_LOJA	NUM_ANOMES	VAL_VALOR_SE	VAL_VALOR_DE	VAL_VALOR_CO	VAL_QUANTIDA
1	9306	1047684	F	F	97228	65	20220109	17.88	0.0	17.88	0.448
2	9306	1071971	M	F	97228	65	20220123	17.32	0.0	17.32	0.434
3	9306	1010730	F	F	97228	65	20220115	13.41	0.0	13.41	0.336
4	9307	298327	M	F	68828	30	20220122	15.29	0.0	15.29	0.668
5	9307	1351635	F	F	79160	55	20220107	14.73	0.0	14.73	0.328
6	9307	630099	M	F	79160	55	20220129	30.3	0.0	30.3	0.675
7	9307	612002	M	F	79160	55	20220114	13.29	0.0	13.29	0.296
8	9307	721838	M	F	79160	55	20220121	11.76	0.0	11.76	0.262
9	9307	1268699	M	F	96899	23	20220102	42.93	0.0	42.93	1.724
10	9307	928155	M	F	96899	95	20220130	16.75	0.83	15.92	0.842
11	9307	691631	M	F	96899	75	20220103	19.67	0.0	19.67	0.79
12	9307	1365446	M	F	96899	22	20220129	16.99	0.0	16.99	0.776
13	9307	795250	F	F	96899	85	20220107	18.79	0.0	18.79	0.858
14	9307	1080	M	F	96899	95	20220105	19.44	0.0	19.44	0.93

Resultados por página: 50 1 – 50 de 15975704







Pontos de atenção

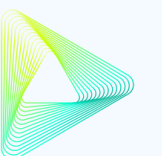
- Memória do ambiente da GCP
 - Cluster do Dataproc no Compute Engine limitado
- Arquivos que não foram imputados no Cloud Storage
 - Evitando assim o erro de *'Unable to infer schema for Parquet'*
 - Apenas a extensão *'zstd.parquet'* foi considerada

Intervalos > case_mercafacil > clientes

FAZER UPLOAD DE ARQUIVOS CARREGAR PASTA CRIAR PASTA TRANSFERIR

Filtrar apenas pelo prefixo do nome ▼ Filtro Filtrar objetos e pastas

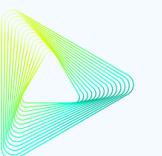
<input type="checkbox"/>	Nome	Tamanho	Tipo
<input type="checkbox"/>	 ._SUCCESS.crc	8 B	application/octet-stream
<input type="checkbox"/>	 .part-00000-2bbe8d8e-6fe1-4006-...	156,9 KB	application/octet-stream
<input type="checkbox"/>	 ._SUCCESS	0 B	application/octet-stream
<input type="checkbox"/>	 part-00000-2bbe8d8e-6fe1-4006-...	19,6 MB	application/octet-stream



Links para os arquivos:

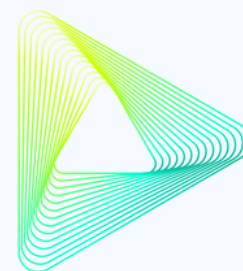
- Jupyter notebook conectado ao Dataflow
- Script Python
- Airflow
- Arquivos csv

[https://drive.google.com/drive/folders/1CgOFFWrBXUrhQA1agJOmyMKf-3f2RVOh?
usp=sharing](https://drive.google.com/drive/folders/1CgOFFWrBXUrhQA1agJOmyMKf-3f2RVOh?usp=sharing)



OBRIGADA

Amanda Louise Costa Nascimento



mercafacil