

Abstract geometric lines in the top left corner, consisting of several overlapping, irregular polygons and lines in a light beige color.

PROJETO INDIVIDUAL - BC17 -

Stéphanie R. R. Pirajá

VISÃO GERAL

- Escopo
- Ferramentas Utilizadas
- Processos
- Análises

INFRA

O dataset deve estar em Cloud Storage; o arquivo original e o tratado devem ser salvos em MongoDB Atlas; os dataframes devem ser salvos em Cloud Storage

PANDAS

Traduzir os dados para Português-BR; realizar a limpeza dos dados inconsistentes; excluir colunas se necessário

PYSPARK

Montar o StructType do dataframe; realizar as normalizações; alterar o nome de ao menos 2 colunas; criar ao menos 2 colunas; realizar filtros e ordenações; utilizar 2 Window Functions; utilizar SparkSQL

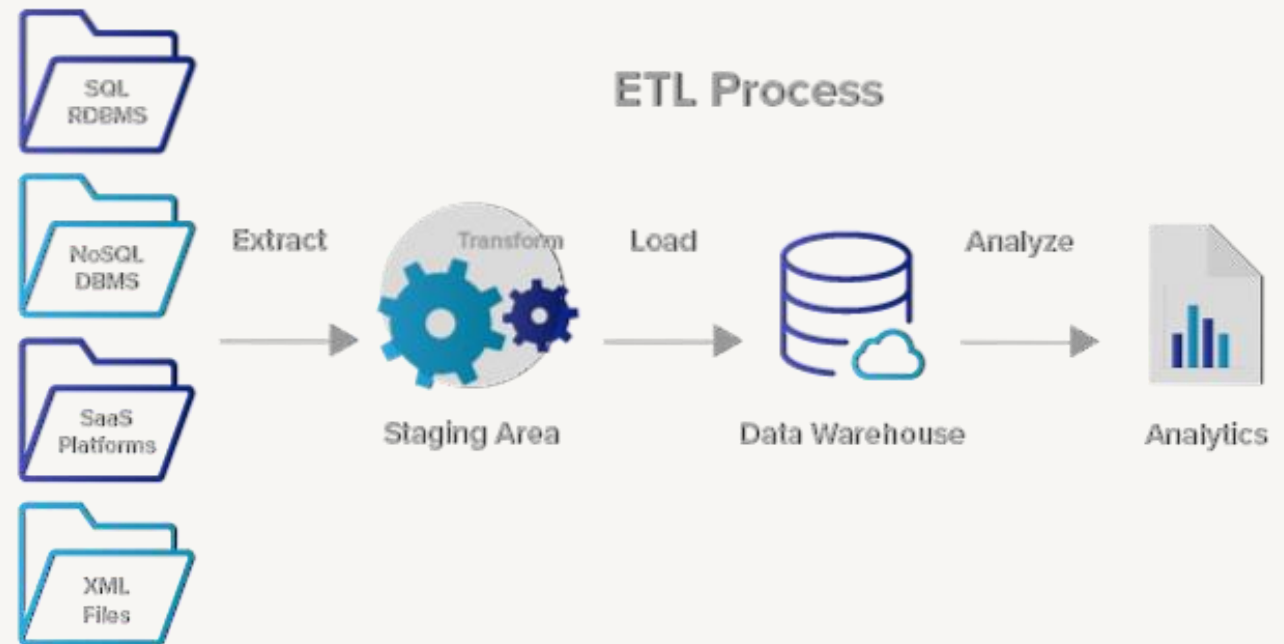
DATASTUDIO

Construir um dashboard para apresentação dos insights

ESCOPO

ETL

SOLUÇÃO





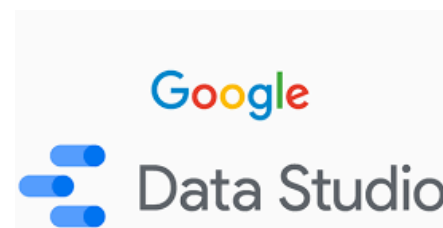
Google Cloud



mongoDB
Atlas



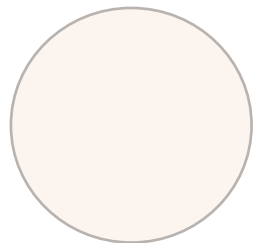
Google Colaboratory



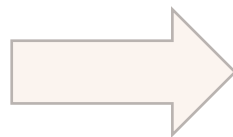
FERRAMENTAS
UTILIZADAS



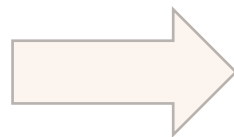
PROCESSOS



Dataset no
classroom



Inserção do
dataset no
Bucket GCP



Configurações
de acesso ao
bucket GCP e
key do service
account



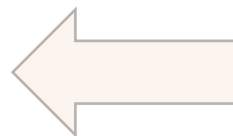
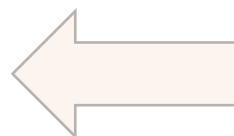
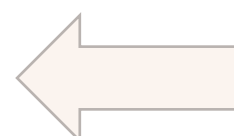
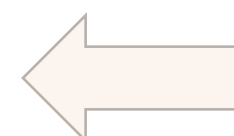
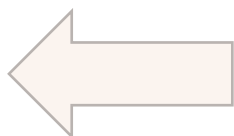
Configurações
de conexão
no MongoDB
Atlas



Criação do
Google
Colab



Conexão
via Python
com GCP e
MongoDB
Atlas



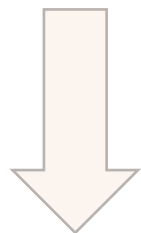
Tratamento
do dataframe
com pandas

Pré análise do
dataframe em
PySpark para
analisar
inconsistências

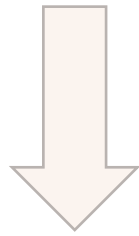
Transformação
do df em
dicionário e
inserção do
dataset bruto
no MongoDB

Transformação
do dataset .csv
em dataframe
pandas

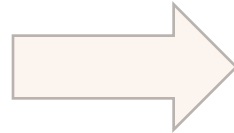
Extração
da base
bruta do
Bucket GCP



Transformação
em df PySpark



Normalização
do StructType



Funções,
Window
Functions e
SparkSQL



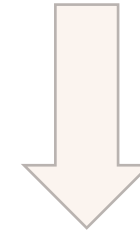
Transformação
do df para
Pandas,
conversão para
.csv e inserção
do dataset
tratado no
Bucket



Extração do
.csv tratado
do Bucket
GCP para o
Data Studio





Confecção
dos
dashboards
no Google
Data Studio



Finalização
e entrega
do projeto

COLETA DE DADOS



Buckets > projetos_ste > Brutos			
UPLOAD FILES UPLOAD FOLDER CREATE FOLDER MANAGE HOLDS D			
Filter by name prefix only ▼ ≡ Filter Filter objects and folders			
<input type="checkbox"/>	Name	Size	Type
<input type="checkbox"/>	 dfPandasBruto	227 KB	application/octet-stream
<input type="checkbox"/>	 marketing_campaign.csv	215 KB	text/csv

Os dados foram fornecidos pelos professores através do classroom e colocados em nuvem no Google Cloud Storage através de armazenamento em Bucket.

EXTRAÇÃO DOS DADOS DA GOOGLE CLOUD

```
1 #CÓDIGO QUE ACESSA A BUCKET GCP CRIADA E FAZ O DOWNLOAD DOS ARQUIVOS VIA PANDAS
2 client = storage.Client()
3 #CRIAR UMA VARIÁVEL CHAMADA BUCKET QUE VAI RECEBER O NOME DA BUCKET DO CLOUD STORAGE
4 bucket = client.get_bucket('projetos_ste')
5 #USAR O MÉTODO BLOB PARA RETORNAR O NOME DO ARQUIVO (JSON, CSV, PARQUET)
6 bucket.blob('marketing_campaign.csv')
7 #CRIAR UMA VARIÁVEL PATH PARA COLOCAR O CAMINHO DO CSV
8 #path = 'gs://datasets_pyspark/arquivo_geral.csv'
9 path = 'gs://projetos_ste/Brutos/marketing_campaign.csv'
```

```
[ ] 1 #Puxando o dataset do Bucket e transformando em dataframe (com Pandas)
    2 df_pandas = pd.read_csv(path, sep=';')
```

```
[ ] 1 df_pandas
```

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines
0	5524	1957	Graduation	Single	58138.0	0	0	2012-09-04	58	635
1	2174	1954	Graduation	Single	46344.0	1	1	2014-03-08	38	11
2	4141	1965	Graduation	Together	71613.0	0	0	2013-08-21	26	426

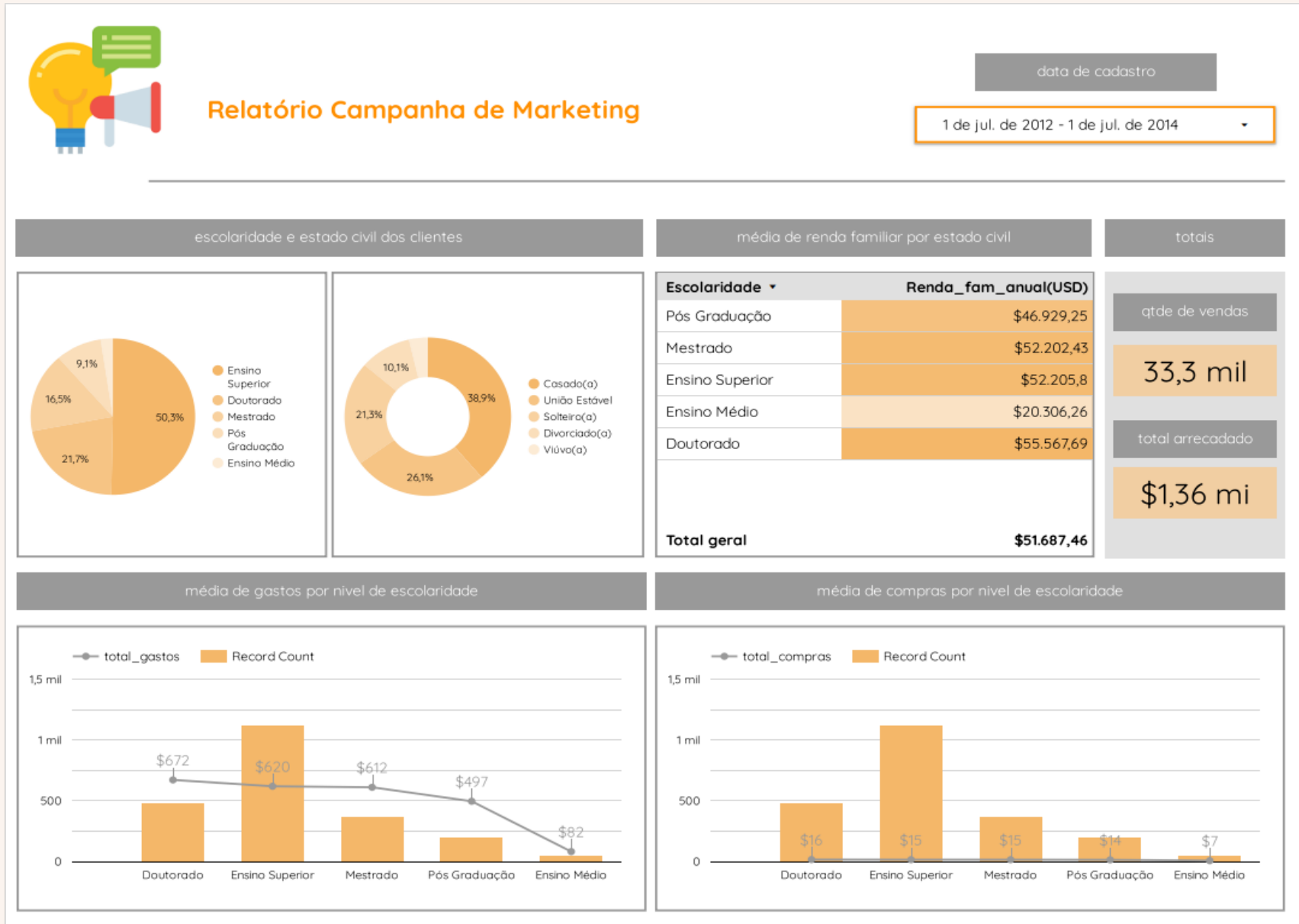
Foi feita a conexão com o Bucket GCP e o arquivo .csv foi transformado em dataframe Pandas



TRATAMENTO DOS DADOS

- Pré análise das inconsistências utilizando o Pyspark
- Tratamento utilizando biblioteca Pandas
- Tradução para o Português - BR
- Normalização dos valores nulos
- Exclusão de colunas irrelevantes para a análise
- Substituição dos valores booleanos por 'Sim' e 'Não'
- Transformação do dataframe pandas em dataframe Pyspark
- Normalização dos tipos de dados dos atributos
- Criação de 2 colunas com a soma dos Gastos e das Compras
- Insights realizados com funções, Window Function e SparkSQL

DATA STUDIO



“O SUCESSO É O ACÚMULO
DE PEQUENOS ESFORÇOS
REPETIDOS DIA APÓS DIA.”

Robert Coller

Two thin, intersecting orange lines in the top-left corner of the slide.

AGRADECIMENTOS

SoulCode

Professores Adriano, Bismark, Felipe e Igor

Colegas da turma

A series of thin, light brown lines forming an abstract geometric pattern in the top left corner of the slide. The lines intersect to create various triangular and polygonal shapes.

MUITO OBRIGADA!

Stéphanie Pirajá

(11) 9 4038-7444

stepirajadev@gmail.com

<https://www.linkedin.com/in/stepiraja>