**Engenharia de Dados**

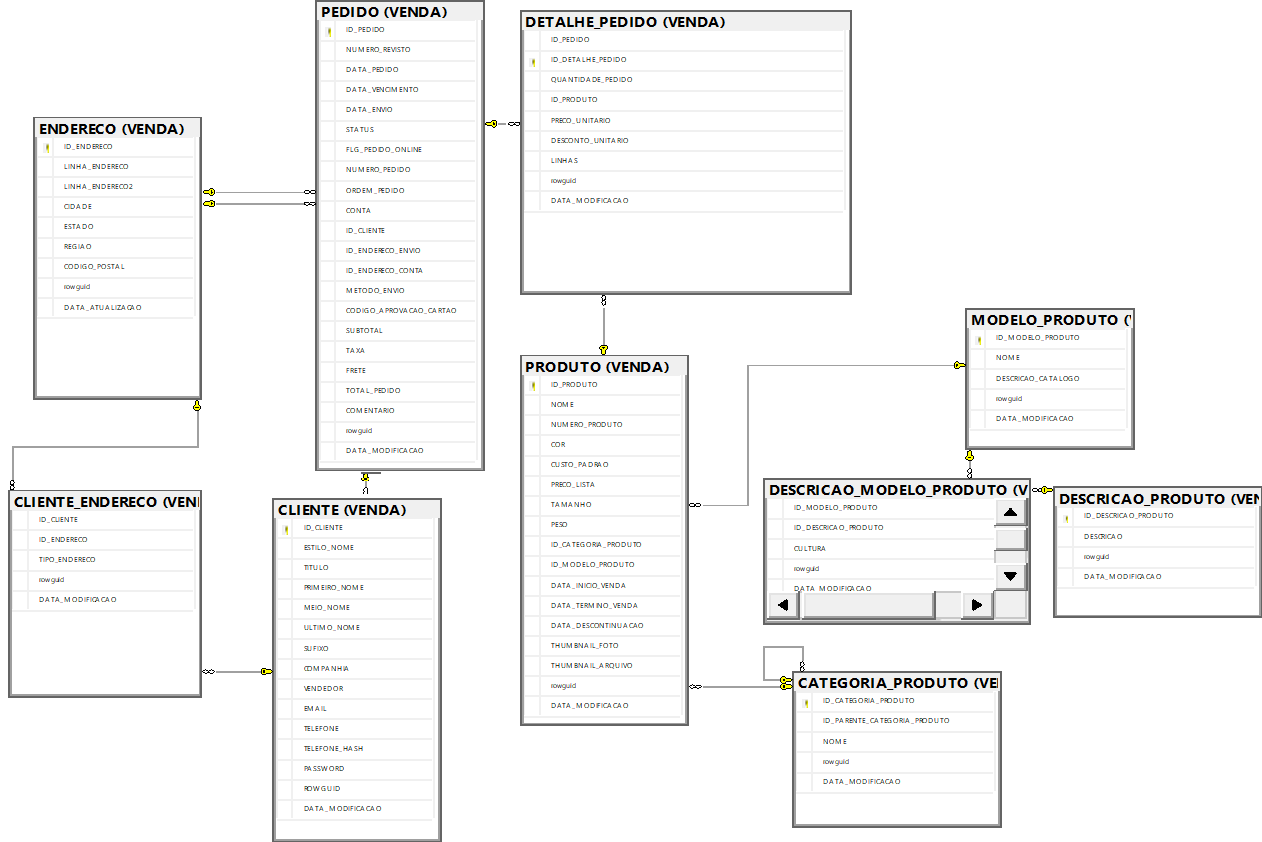
**Exercício Spark**

Atividade em grupo de no máximo 5 alunos para a AC5

**Ambiente de desenvolvimento**

Colab

**Modelo lógico banco de dados**



**Conexão com banco de dados (opcional):**

servidor:sql20211.cp1ri5jswsct.us-east-1.rds.amazonaws.com

login: aluno

senha: aluno123

banco: [AdventureWorksLT2019\_PT]

**Atividades**

**Parte 1 - Integração de dados (EXTRACT AND LOAD)**

1. Instalar e importar as bibliotecas do spark e pyspark
2. Criar um spark context com o app name = “AC5 DataEng”
3. Criar uma conexão com o banco de dados AdventureWorksLT2019\_PT
4. Ler todas as tabelas com o pandas, transformar em um dataframe do spark e gravar como parquet em uma pasta chamada “adventure” no colab. Lembrando que as tabelas estão no schema VENDA no banco de dados (VENDA.PEDIDO,VENDA.CLIENTE, etc)

**Parte 2 - Transformação de dados (TRANSFORM) - Utilize os dados gravados no colab.**

1. Crie um dataframe juntando as tabelas CLIENTE, PEDIDO e DETALHES\_PEDIDO. Agrupando pelo id do cliente, some a quantidade pedido, valor unitário e desconto. Depois crie uma coluna chamada lucro (quantidade pedido \* valor unitario) - desconto. Grave esse dataframe no formato json com o nome de “TRANSFORMACAO/PEDIDO\_CLIENTE”.
2. Crie um dataframe juntando os dados de CATEGORIA\_PRODUTO, PRODUTO E MODELO\_PRODUTO. Selecione as seguintes colunas,

* Nome da categoria do produto.
* Nome, número, cor, custo padrão, preço lista, tamanho e peso do produto
* Nome do modelo do produto.

**Aplicando as regras abaixo:**

* O início e o fim do nome do produto não pode ter espaços e tem que ter todas as letras maiúsculas.
* Concatenar um hífen e o nome da cor no final do nome do produto
* Se o tamanho vier como “M” deverá ser 49, se vier “P” deverá ser 32, se vier “L” deverá ser 52, se vier XL deverá ser 70 e ser for null deverá ser 0. Nos outros casos, mantenha o valor do tamanho.
* Criar uma coluna chamada Codigo\_Produto. Essa coluna deve conter os dois primeiros dígitos do número do produto.
* Substituir o hífen no nome do modelo do produto por um espaço.

Gravar esse dataframe como parquet com o nome de “TRANSFORMACAO/PRODUTOS”, particionado por categoria\_produto

1. Crie um dataframe trazendo as seguintes colunas das tabelas (PRODUTO, DETALHE\_PEDIDO)

Obs: Filtrar apenas os produtos que não possuem data\_termino\_venda.

● nome produto,

● soma da quantidade\_pedido

● contagem distinta de pedido

● data da última compra

● data da primeira compra

Agrupado por nome\_produto e grave como parquet com o nome de “TRANSFORMACAO/COMPRA\_PRODUTO”

1. Crie um dataframe trazendo as seguintes colunas das tabelas (ENDERECO,CLIENTE\_ENDERECO):

● endereco (fica na coluna linha endereço 2, caso essa coluna estiver

vazia, traga no lugar os dados da coluna linha endereço).

● Número\_endereço (pegar apenas os digitos, aplique a função do spark.sql.functions “regexp\_replace(column\_name, '[^0-9]', ''))”.

● Cidade (letras minúsculas).

● Estado (letras maiúsculas).

● Duas primeiras letras da Região.

● tipo\_endereço (se o valor for “Main Office”, traga como “principal” se for “shipping” traga como “entrega”). Gravar como parquet com o nome de “TRANSFORMACAO/ENDERECO\_CLIENTE”, particionado por estado,cidade.

1. Crie um dataframe utilizando a tabela DESCRICAO\_PRODUTO.

Crie uma coluna chamada DESCRICAO\_LISTA transformando a coluna DESCRICAO em uma lista. Use a função spark.sql.functions.split(DESCRICAO,’ ‘).

Crie uma coluna chamada KEY\_DESCRICAO do tipo chave valor, contando como chave a coluna ID\_DESCRICAO\_PRODUTO e como valor a coluna DESCRICAO\_LISTA

Use a função spark.sql.functions.struct(KEY\_DESCRICAO ,DESCRICAO\_LISTA ).

Grave como parquet com o nome de “TRANSFORMACAO/DESCRICAO\_CHAVE\_VALOR”

**Entregável**

Apenas o notebook com o nome dos integrantes