Projeto de Engenharia de dados – Vendas por Região

1. Descrição do projeto e das fontes de dados

O projeto de nome Vendas por região tem como fontes de dados três arquivos no formato CSV, contendo informações de produtos vendidos por região em determinada data.

Os arquivos estão armazenados na pasta de nome Dataset no Github.

A seguir uma amostra da estrutura e dos dados do arquivo de PRODUTO.CSV .

CODIGO_PRODUTO	DESCRICAO_PRODUTO	SIGLA_PRODUTO
101	ZINCO	ZI
104	COBRE	CO
107	OURO	OU
110	PRATA	PR
113	FERRO	FE
116	PLATINA	PL
119	NIQUEL	NI

Abaixo uma amostra do arquivo REGIAO.CSV.

CODIGO_REGIAO	DESCRICAO_REGIAO	NOME_PAIS
201	SUL	BRASIL
204	SUDESTE	BRASIL
207	NORTE	BRASIL
210	NORDESTE	BRASIL

Segue dados do arquivo VENDAS.CSV.

CODIGO_PRODUTO	CODIGO_REGIAO	DATA_VENDA	QTDE_VENDA	VALOR_VENDA
101	201	01/01/2021	351	350649
104	204	01/01/2021	384	383616
107	207	01/02/2021	417	416583
110	210	04/02/2021	450	449550
113	201	07/02/2021	483	482517
116	201	10/03/2021	516	515484
119	207	10/03/2021	549	548451
101	207	13/04/2021	219	218781
104	210	16/04/2021	219	218781
107	201	19/05/2021	252	251748
110	201	16/05/2021	285	284715
113	207	19/06/2021	318	317682
116	204	20/06/2021	120	119880
119	207	24/06/2021	153	152847

2. Arquitetura de software utilizada na construção do projeto

Este projeto fez uso do Databricks versão Comunity Edition e da ferramenta Airflow para orquestração da execução dos scripts de carga de dados.

Foram criados e executados scripts na linguagem Pyspark .

Os scripts estão separados pelas camadas: Bronze, Silver e Gold.

A seguir a relação dos scripts criados para carregar os dados dos arquivos PRODUTO.CSV, REGIAO.CSV e VENDAS.CSV em tabelas no formato Delta do Databricks.

- cria_db_bronze_silver_gold.py;

- bronze_di_produto.py;

- bronze_di_regiao.py;

- bronze_fa_venda.py;

- silver_di_produto.py;

- silver_di_regiao.py;

- silver_fa_venda.py;

- gold_di_tempo.py;

- gold_di_produto.py;

- gold_di_regiao.py;

- gold_fa_venda.py.

3. Consumo dos dados disponíveis na camada Gold

Ao final da execução dos scripts foi disponibilizado no ambiente Databricks na camada Gold as seguintes tabelas no formato Delta:

- di_produto;

- di_reguão;

- di_tempo;

- fa_venda.

As tabelas da camada Gold podem ser consumidas por qualquer ferramenta de visualização basta ter disponível um conector para o ambiente Databricks (Spark).

4. Orquestração dos scripts de carga de dados

Foi criado um script no ambiente do Airflow de nome **dag_vendas_regiao** responsável pela orquestração da execução dos scripts de carga de dado.