PROJETO DE CONCLUSÃO

ALUNOS: Henrique Moura de Mesquita, matrícula 201803355085
& Ana Carolina de Oliveira Farah, matrícula 201803355093

Rio de Janeiro, 24 de novembro de 2022

CONTROLE DE VERSÃO			
Autor	**Versão**	**Data**	**Descrição**
Henrique Moura de Mesquita	1.0	09/10/2022	Criação do documento
Ana Carolina de Oliveira Farah	1.1	24/11/2022	Inserção dos dados e revisão

Sumário

Lista de Figuras

1 Introdução

Esta documentação tem por propósito realizar o detalhamento do projeto “Pocco Pamonhas” sob o ponto de vista técnico, considerando - quando aplicável - possíveis soluções, premissas e atividades executadas.

2 Solicitação

Realizar a migração de cargas de trabalho atualmente realizadas na plataforma Apache NiFi para a plataforma Databricks, estruturar um Data Warehouse seguindo especificações da arquitetura proposta e desenvolver dashboards que gerem insights para a área de negócios fazer uso do mesmo para consulta e tomadas de decisão.

3 Premissas da solução

Nesta seção discutiremos as premissas da solução.

Origem e especificação dos dados

- O dataset em formato CSV será extraído de um conteiner provisionado em uma conta de armazenamento do Azure Blob Storage através de script no Databricks utilizando linguagem python e a interface de alto nível PySpark.Após, será carregado no SQL Server e tratado em tabelas separadas por fatos e dimensões, seguindo esquema em estrela.
- As tabelas serão carregadas no Power BI através de integração da plataforma em questão com o SQL Server para o desenvolvimento.

Ambiente de desenvolvimento

- O cliente deverá disponibilizar acessos aos ambientes de desenvolvimentom todas as ferramentas especificadas na arquitetura proposta neste documento.

4 Modelo da arquitetura sugerida

A Figura abaixo apresenta a arquitetura da solução proposta levando em consideração o levantamento de requisitos e entendimento do negócio.

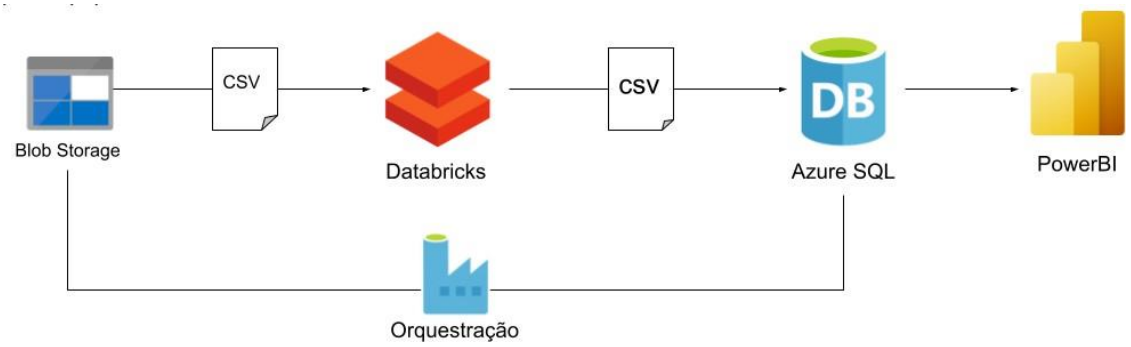

Figura 1: Arquitetura do projeto.

- Data Factory: Orquestração da pipeline
- Azure Blob Storage: Conta de armazenamento onde estão os dados
- Databricks: Normalização, tratamento e processamento de dados
- Azure SQL: Estruturação do Data Warehouse
- Power BI: Dashboard com visuais e indicadores

5 Dicionário de dados

De acordo com a seção 5

- Tabela Stage [STAGE_ANA_FARAH_lab3].[POCCO_ORDERS_TABLE]

Campo	Tipo	Tamanho	Restrição	Descrição
region	varchar	40	null	Região geográfica
country	varchar	40	null	País
item_type	varchar	20	null	Tipo de mercadoria
sales_channel	varchar	07	null	Canal de vendas
order_priority	varchar	01	null	Prioridade da venda
order_date	date	n/a	null	Data da venda
order_id	int PK	n/a	not null	ID da venda
ship_date	date	n/a	null	Data do envio
units_sold	int	n/a	null	Unidades vendidas
unit_price	decimal	10,2	null	Preço unitário
unit_cost	decimal	10,2	null	Custo unitário
total_revenue	decimal	10,2	null	Receita total
total_cost	decimal	10,2	null	Custo total
total_profit	decimal	10,2	null	Lucro total

Tabela 1: Tabela Stage

- Tabela Dimensão Região [DW_ANA_FARAH_lab3].[DW_DIM_REGION_TABLE]

Campo	Tipo	Tamanho	Restrição	Descrição
region_id	int identity PK	n/a	not null	ID da região
region	varchar	40	null	Região geográfica

Tabela 2: Tabela Dimensão Região

- Tabela Dimensão País [DW_ANA_FARAH_lab3].[DW_DIM_COUNTRY_TABLE]

Campo	Tipo	Tamanho	Restrição	Descrição
country_id	int identity PK	n/a	not null	ID do país
country	varchar	40	null	País
region	varchar	40	null	Região geográfica

Tabela 3: Tabela Dimensão País

- Tabela Dimensão Canais de Venda [DW_ANA_FARAH_lab3].[DW_DIM_CHANNEL_TABLE]

Campo	Tipo	Tamanho	Restrição	Descrição
channel_id	int identity PK	n/a	not null	ID do canal (online / offline)
sales_channel	varchar	07	null	Canal de venda

Tabela 4: Tabela Dimensão Canais de Venda

- Tabela Fato Vendas [DW_ANA_FARAH_lab3].[DW_FACT_SALES_TABLE]

Campo	Tipo	Tamanho	Restrição	Descrição
region_id	int FK	n/a	not null	ID da região
country_id	int FK	n/a	not null	ID do país
channel_id	int FK	n/a	not null	ID do canal de venda
order_id	int FK	n/a	not null	ID da venda
item_type	varchar	20	null	Tipo de mercadoria
order_priority	varchar	01	null	Prioridade da venda
order_date	date	n/a	null	Data da venda
ship_date	date	n/a	null	Data do envio
units_sold	int	n/a	null	Unidades vendidas
unit_price	decimal	10,2	null	Preço unitário
unit_cost	decimal	10,2	null	Custo unitário
total_revenue	decimal	10,2	null	Receita total
total_cost	decimal	10,2	null	Custo total
total_profit	decimal	10,2	null	Lucro total

Tabela 5: Tabela Fato Vendas

6 Processo de desenvolvimento até etapa final

As atividades do pipeline estão enumeradas a seguir.

1. **Extração dos dados:** Os dados foram extraídos de uma conta de armazenamento do Azure Blob Storage onde as cargas já existiam previamente no formato parquet.

2. **Procedimentos no Databricks:**

   - A plataforma foi utilizada para extração e tratamento dos dados, respeitando as boas práticas de ETL, realizando teste lógico para identificação de outliers, implementando etapas da arquitetura medalhão. Foi realizada conversão de tipos de dados para string afim de otimizar a modelagem e gerar cargas na camada bronze;
   - Utiliza-se Spark para ingestão dos datasets do histórico armazenado como dataframes, é realizada nova conversão de tipo dos dados conforme regra de negócios para cargas na camada prata;
   - Novamente é utilizado spark para ingestão dos dados históricos e, seguindo as regras de negócios, é realizada ordenação decrescente das datas de vendas para serem armazenadas no banco de dados onde os dados serão normalizados em etapa seguinte, consumindo a carga desta última etapa na camada ouro.
   - A biblioteca MSAL é utilizada para gerar token e provisionar a pipeline entre o Databricks e o SQL Server na tabela Stage criada para receber as cargas.

3. **Modelagem de dados:** É realizada modelagem dimensional dos dados seguindo parâmetros do conceito Star Schema de Data Warehousing a partir de construção de tabelas Fatos e Dimensões.

4. **Stored Procedure:** A procedure parametriza a ingestão da carga de dados a partir da tabela Stage normalizando os dados nas demais tabelas Fatos e Dimensões afim de melhorar sua granularidade.

7 Dashboards

- Link do Dashboard: Clique para visualizar ou com o botão direito do mouse para copiar o link
- Dashboard de Vendas:

Figura 2: Dashboard Pocco Pamonhas

- Card de lucro total
- Card de receita total
- Card de receita total dos últimos 30 dias
- Card de receita total dos últimos 10 dias
- Card de unidades totais vendidas
- Mapa mundial com amostras do tipo bolhas vinculadas ao volume da receita total por país e região
- Gráficos de colunas empilhadas para receitas totais dos últimos 30 dias e dos últimos 10 dias
- Filtro por canais de vendas
- Filtro por data