
# Laboratório 03 Azure – Projeto Pocco Pamonhas

**São Paulo, 21 de julho de 2022**

| CONTROLE DE VERSÃO | | | |
|---|---|---|---|
| **Autor** | **Versão** | **Data** | **Descrição** |
| Renan Da Silva Ramos | 1.0 | 21/07/2022 | Criação do documento |

## 1. INTRODUÇÃO

Este documento visa detalhar tecnicamente as etapas utilizadas no cumprimento do projeto do cliente Pocco Pamonhas, uma das grandes multinacionais no ramo de pamonhas.

## 2. SOLICITAÇÃO

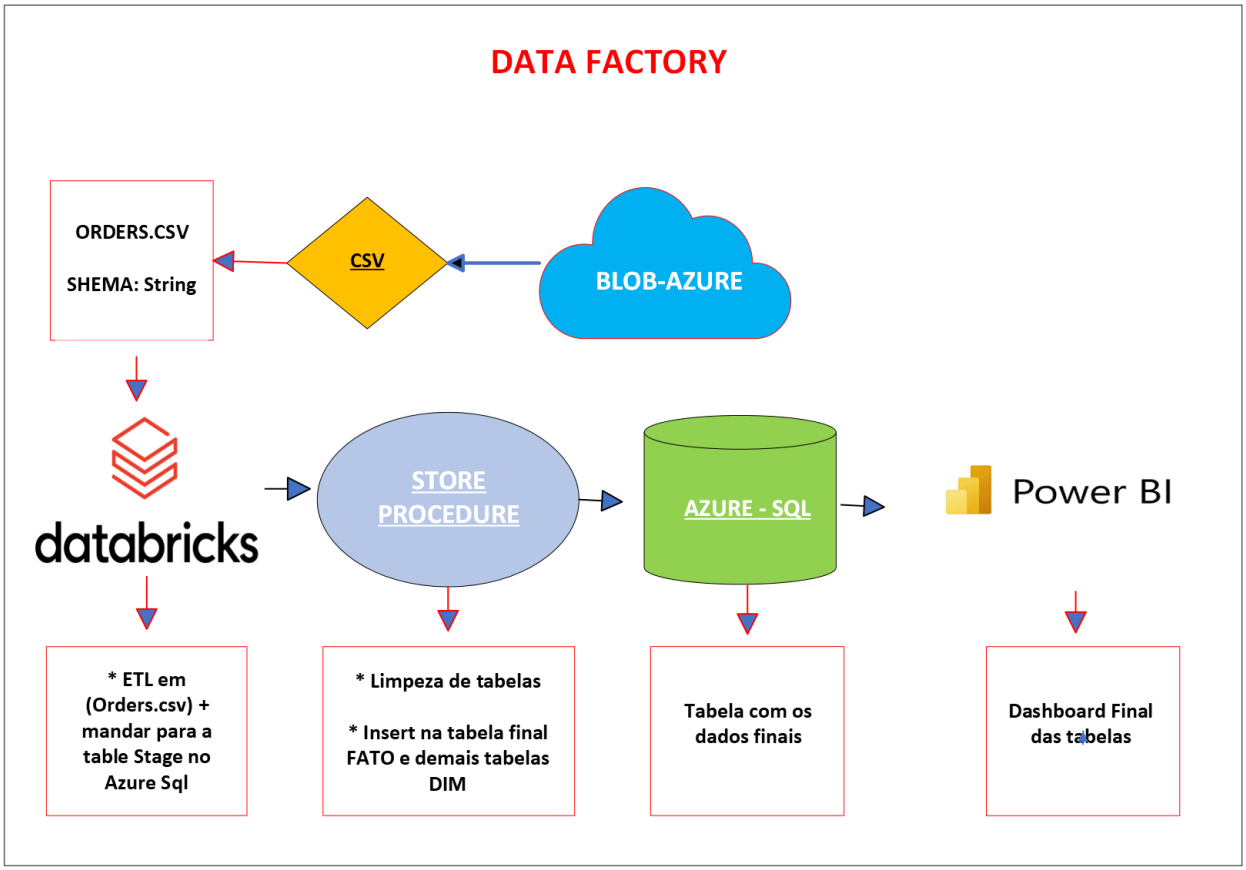
O cliente deseja realizar um upgrade em sua arquitetura de dados que se encontra atualmente no (Apache Nifi). Em negociações amigáveis e produtivas com a (Blueshift Brasil), ambos decidiram que a melhor solução seria utilizar a plataforma do (Databricks) em conjunto com as ferramentas (Azure Sql + Azure Data Factory+ Azure Blob Storage). Além disto também será usado o (Microsoft Power BI) para gerar relatórios e dashboards de algumas tabelas em específico da empresa.

Inicialmente os dados da empresa contratante estão com o nome de (Orders.csv) e se encontram em um “armazenamento em nuvem” no (Container Blob Azure). Esse arquivo será extraído e manipulado pelo (Databricks), que fará a ingestão juntamente com o (Azure Sql) em uma tabela inicial chamada (Stage). Após a ingestão na primeira tabela, será utilizada uma (Story-Procedure) para desdobramento e inserção destes mesmos dados em outras tabelas que serão utilizadas no (Power Bi) para a construção do (Dashboard) que o cliente deseja consultar sobre seu faturamento de vendas e demais relatórios de sua compania.

Todo este processo será orquestrado com a ferramenta (Azure Data Factory) para se ter um melhor controle otimizado sobre todas as etapas.

## 3. MODELO DA ARQUITETURA

A figura abaixo representa a arquitetura e os principais processos para atender as demandas do cliente



## 4. DADOS INICIAIS - ORIGEM

O arquivo utilizado como fonte dos dados originais se chama (Orders.csv), ele se encontra no (Azure blob storage) e retorna uma tabela com 14 colunas. Estas colunas detalham alguns países e regiões do globo terrestre e também possuem valores monetários e unitários no custo, quantidade, produção e lucro de vendas de pamonhas.

Com a ajuda do (Databricks) criamos um dataframe em CSV e fizemos as manipulações necessárias utilizando (Pyspark + Pandas) e ao final das modificações, enviamos para a tabela inicial chamada de (Stage). Fizemos isso em conjunto com o (Azure Data Factor + Azure Sql Server) através de (links de conexões) que a plataforma (Azure) oferece.

**Abaixo, uma breve ilustração dos dados originais – Parte 1:**

| | Region | Country | Item Type | Sales Channel | Order Priority |
|---|---|---|---|---|---|
| 1 | Sub-Saharan Africa | Chad | Office Supplies | Online | L |
| 2 | Europe | Latvia | Beverages | Online | C |
| 3 | Middle East and North Africa | Pakistan | Vegetables | Offline | C |
| 4 | Sub-Saharan Africa | Democratic Republic of the Congo | Household | Online | C |
| 5 | Europe | Czech Republic | Beverages | Online | C |

**Abaixo, uma breve ilustração dos dados originais – Parte 2:**

| Order Date | Order ID | Ship Date | Units Sold | Unit Price | Unit Cost | Total Revenue | Total Cost | Total Profit |
|---|---|---|---|---|---|---|---|---|
| 27/01/2011 | 292494523 | 12/02/2011 | 4484 | 651,21 | 524,96 | 2920025,64 | 2353920,64 | 566105 |
| 28/12/2015 | 361825549 | 23/01/2016 | 1075 | 47,45 | 31,79 | 51008,75 | 34174,25 | 16834,5 |
| 13/01/2011 | 141515767 | 01/02/2011 | 6515 | 154,06 | 90,93 | 1003700,9 | 592408,95 | 411291,95 |
| 11/09/2012 | 500364005 | 06/10/2012 | 7683 | 668,27 | 502,54 | 5134318,41 | 3861014,82 | 1273303,59 |
| 27/10/2015 | 127481591 | 05/12/2015 | 3491 | 47,45 | 31,79 | 165647,95 | 110978,89 | 54669,06 |

Observe que os dados estão todos em (String) e alguns valores numéricos estão com (,) ao invés de (.). Para resolver este problema, fizemos toda manipulação conforme as exigências do cliente e mandamos para tabela inicial (Stage).

## 5. MANIPULAÇÃO DO AZURE SQL

Nos requisitos do projeto foi necessário a criação de algumas tabelas; “Table stage” , “Table_Fato” *e “Tabelas_Dim” no (Azure Sql).* A tabela “Stage”, contém os dados (brutos) da extração do arquivo inicial (Orders.csv). Já a tabela “Fato/Dw”, terá os dados finais manipulados conforme as especificações que foram dadas no objetivo do projeto. Também foi necessária a criação de uma procedure para a transformação dos dados e inserção nas tabelas de (Dimensões = Dim) utilizadas no (Power BI).

**Schema original do arquivo (Orders.csv):**

```
# VERIFICANDO O SCHEMA
df.printSchema()
```

```
ɔt
-- Region: string (nullable = true)
-- Country: string (nullable = true)
-- Item Type: string (nullable = true)
-- Sales Channel: string (nullable = true)
-- Order Priority: string (nullable = true)
-- Order Date: string (nullable = true)
-- Order ID: string (nullable = true)
-- Ship Date: string (nullable = true)
-- Units Sold: string (nullable = true)
-- Unit Price: string (nullable = true)
-- Unit Cost: string (nullable = true)
-- Total Revenue: string (nullable = true)
-- Total Cost: string (nullable = true)
-- Total Profit: string (nullable = true)
```

**Schema da tabela STAGE_renan_silva.stage:**

| **STAGE_renan_silva.stage** | |
|---|---|
| Order_Date | DATE |
| Unit_Price | FLOAT |
| Total_Revenue | FLOAT |

**Schema da tabela DW_final :**

| **DW_renan_silva.fato** | |
|---|---|
| Order_Date | DATE |
| Unit_Price | FLOAT |
| Total_Revenue | FLOAT |

**Schema da tabela DW_renan_silva.region:**

| DW_renan_silva.region | |
|---|---|
| region | VARCHAR |
| id_region | INT |

**Schema da tabela DW_renan_silva.sales_channel:**

| DW_renan_silva.sales_channel | |
|---|---|
| sales_channel | VARCHAR |
| id_sales_channel | INT |

**Schema da tabela DW_renan_silva.item_type:**

| DW_renan_silva.item_type | |
|---|---|
| item_type | VARCHAR |
| id_item_type | INT |

**Schema da tabela DW_renan_silva.item_type:**

| DW_renan_silva.country | |
|---|---|
| country | VARCHAR |
| id_country | INT |

## Estrutura da Procedure (Codificação) :

```
## CRIANDO PROCEDURE

CREATE PROCEDURE sp_lab3_renan_silva AS BEGIN

TRUNCATE TABLE DW_renan_silva.region
TRUNCATE TABLE DW_renan_silva.sales_channel
TRUNCATE TABLE DW_renan_silva.item_type
TRUNCATE TABLE DW_renan_silva.country
TRUNCATE TABLE DW_renan_silva.fato


INSERT INTO DW_renan_silva.region(region)
SELECT Region FROM STAGE_renan_silva.stage
GROUP BY Region;

INSERT INTO DW_renan_silva.sales_channel(sales_channel)
SELECT Sales_Channel FROM STAGE_renan_silva.stage
GROUP BY Sales_Channel;

INSERT INTO DW_renan_silva.item_type(item_type)
SELECT Item_Type FROM STAGE_renan_silva.stage
GROUP BY Item_Type;

INSERT INTO DW_renan_silva.country(country)
SELECT Country FROM STAGE_renan_silva.stage
GROUP BY Country;

INSERT INTO DW_renan_silva.fato
(Region,Country,Item_Type,Sales_Channel,Order_Priority,Order_Date,Order_ID,Ship_Date,Units_Sold,Unit_Price,Unit_Cost,Total_Revenue,Total_Cost,Total_Profit)

SELECT Region,Country,Item_Type,Sales_Channel,Order_Priority,Order_Date,Order_ID,Ship_Date,Units_Sold,Unit_Price,Unit_Cost,Total_Revenue,Total_Cost,Total_Profit
FROM STAGE_renan_silva.stage
TRUNCATE TABLE STAGE_renan_silva.stage
END;
```

## 6. CONSTRUÇÃO DO DATA FACTORY (PROCESSAMENTO)

O Azure Data Factory é um administrador de todos os processos da pipeline do nosso projeto. Basicamente nossa pipeline é dividida em 3 processos; O primeiro processo é a requisição do arquivo (Orders.csv) que se encontra no (Azure blob storage) através do (Databricks).

O segundo processo será utilizar (Pandas + Pyspark no Databricks) na tipagem de dados e também substituição do padrão de (,) por (.) nos valores numéricos.

Já o terceiro processo, a (Store-Procedure) entra em ação e faz a extração dos dados da tabela inicial, para a tabela final (Dw_Fato) e também para as tabelas (Dimensões finais) que serão utilizadas no (Power BI).

**Ilustração do 1° processo – obtendo (Orders.csv) no (Databricks) :**

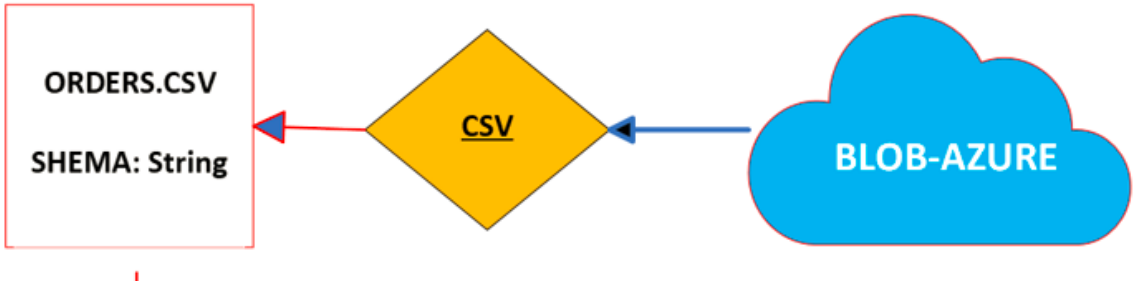


**Ilustração do 2° processo – ETL (DATABRICKS PARTE - 1) :**

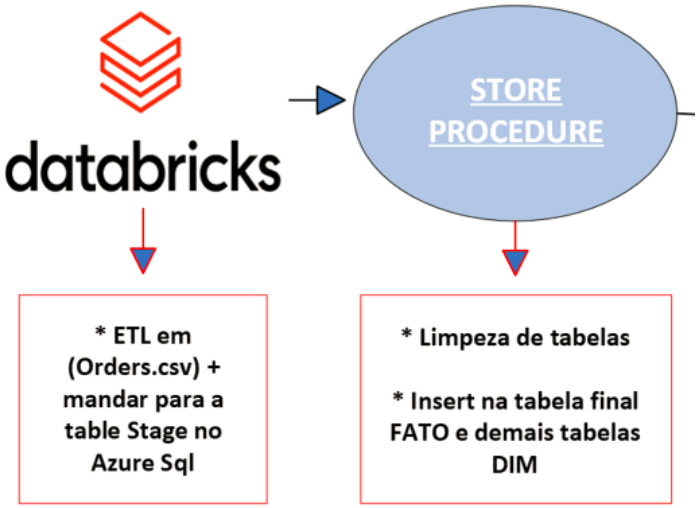
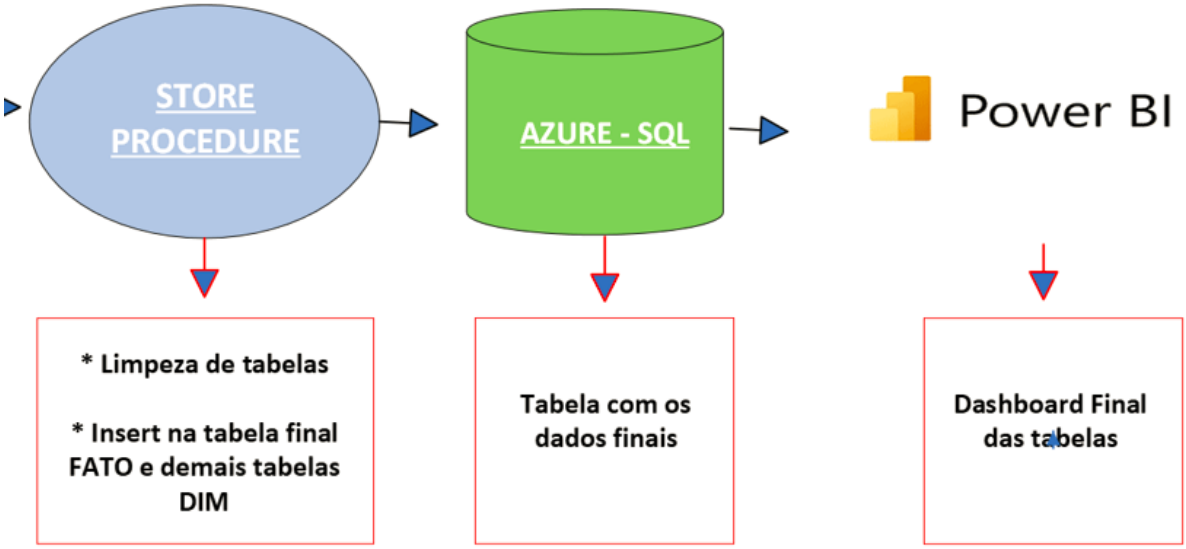


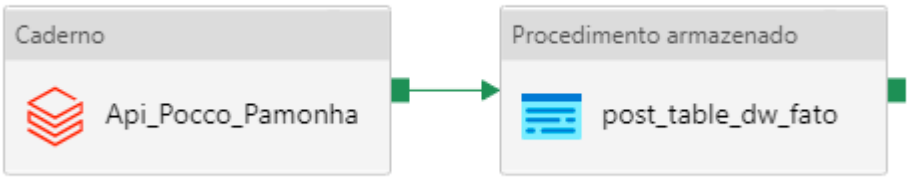

**Ilustração do 2° processo – ETL (DATABRICKS - PARTE 2) :**

```
# TRANSFORMAÇÃO DOS DADOS (ETL) => NOME DOS CAMPOS / TIPO DE DADOS / E (,) POR (.)

pronto = df.withColumnRenamed("Item Type","Item_Type")\
.withColumnRenamed("Sales Channel","Sales_Channel")\
.withColumnRenamed("Order Priority","Order_Priority")\
.withColumnRenamed("Order Date","Order_Date")\
.withColumnRenamed("Order ID","Order_ID")\
.withColumnRenamed("Ship Date","Ship_Date")\
.withColumnRenamed("Units Sold","Units_Sold")\
.withColumnRenamed("Unit Price","Unit_Price")\
.withColumnRenamed("Unit Cost","Unit_Cost")\
.withColumnRenamed("Total Revenue","Total_Revenue")\
.withColumnRenamed("Total Cost","Total_Cost")\
.withColumnRenamed("Total Profit","Total_Profit")\
.withColumn("Order_Date",to_date("Order_Date","dd/MM/yyyy"))\
.withColumn("Ship_Date",to_date("Ship_Date","dd/MM/yyyy"))\
.withColumn("Order_ID",col("Order_ID").cast("int"))\
.withColumn("Units_Sold",col("Units_Sold").cast("int"))\
.withColumn("Unit_Price",regexp_replace("Unit_Price",",",".").cast("float"))\
.withColumn("Unit_Cost",regexp_replace("Unit_Cost",",",".").cast("float"))\
.withColumn("Total_Revenue",regexp_replace("Total_Revenue",",",".").cast("float"))\
.withColumn("Total_Cost",regexp_replace("Total_Cost",",",".").cast("float"))\
.withColumn("Total_Profit",regexp_replace("Total_Profit",",",".").cast("float"))

```

**Ilustração do 3° processo – (Dados finais) :**




**Resultado final do pipeline no (Azure Data factor) :**



## 7. IMPLEMENTAÇÃO DO (POWER BI)

A proposta de implementar um (Dashboard) no (Power BI) foi exigida pelo cliente, pois o mesmo, deseja saber visualmente alguns resultados em que sua empresa vem desempenhando, abaixo segue a relação das pequenas amostras do (Dashboard).

### Ilustração da proposta do cliente – (Requisição) :

**Especificação para construçãos dos dashboard em Power BI**

**VENDAS = quantidade de itens multiplicados pelo valor do item.**

- O acumulado de vendas do último ano por Região e País. Ele gostaria de ter essa visão através de um Mapa Mundial diretamente no Relatório.
- Quantidade de vendas dos últimos 10 dias através de um gráfico de colunas.
- Quantidade de vendas e a Quantidade acumulada de vendas dos últimos 30 dias.
- Uma visão acumulada das vendas do último ano por Canal e País. De forma que seja possível ver a distribuição das vendas um determinado país por canal.

**Especificação para reestruturação do Data Warehouse**

**Os relatórios desenvolvidos no Power BI deverão apontar para as seguintes tabelas:**

- Dimensão com as regiões.
- Dimensão com país.
- Dimensão com Canais de venda.
- Fato com as vendas

### Ilustração da proposta do cliente – (Gráfico) :

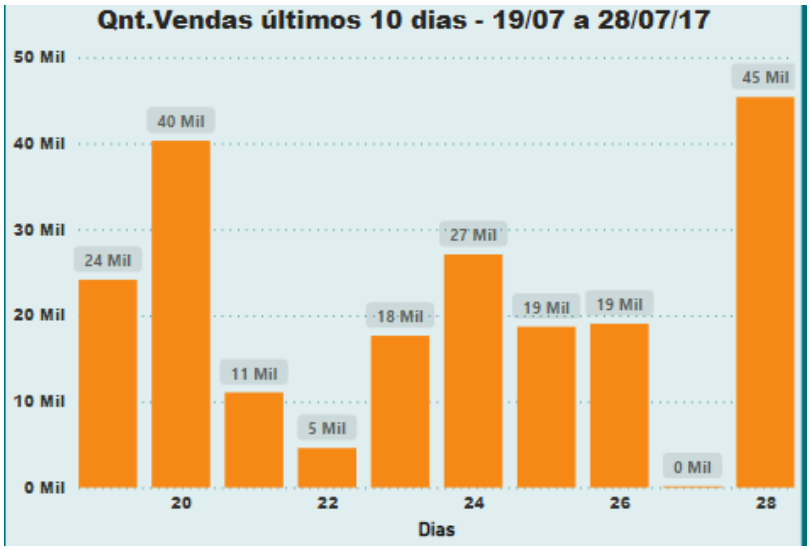


**Ilustração da proposta do cliente – (Gráfico) :**

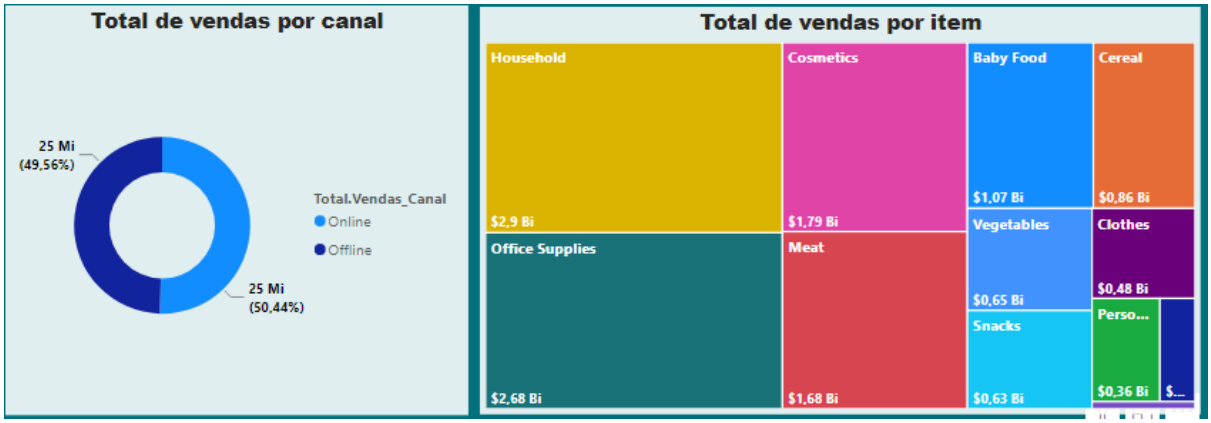



**Ilustração da proposta do cliente – (Geral) :**