

**Business Blueprint**

**São Paulo, 15 de novembro de 2020**

|  |  |  |  |
| --- | --- | --- | --- |
| CONTROLE DE VERSÃO | | | |
| Autor | **Versão** | **Data** | Descrição |
| João Franklin | 1.0 | 10/11/2020 | Criação do documento. |
| Nicolas Lira Souza | 1.1 | 18/01/2021 | Revisão do documento |

1. **Introdução**

O objetivo do documento é apresentar em detalhes as regras e requisitos do “Projeto integrado – Gastos parlamentares” seguindo as normas e orientações exigido pela empresa Blueshift, sendo assim adiquirir conhecimento com o treinamento programado pelo Blueshift Academy.

# Solicitação

A principal linha de desenvolvimento foi, apresentar as ocorrencias de consumo e gastos dos parlamentares, para isso foi solicitado a ingestão de dados “Via download” no site de transparencia publica, disponibilizado pela camara dos deputados.

A entrega foi dividida em duas partes. A primeira aplica-se no consumo dos dados via blob storage e repositório, disponíveis nos portais do governo, através de um script em python para otimização do processo. Dados estes que foram transferidos para o compartimento do Datafactory, consumindo na ingestão do banco de dados SQL Sever.

A segunda consiste na modelagem, desenvolvimento do Datawarehouse e construção do dashboard, utilizando PowerBI.

As principais regras de negócio apresentadas no relatório de visualização, são:

* Total de Parlamentares
* Total de Partido
* Total de Gastos com parlamentares
* Total de Gastos por Mês
* Os Parlamentares que mais Gastam
* Total de Gastos por Partido
* Os Estados que tem os Maiores Gastos
* Top 5 Partidos com Maiores Despesas
* Total de Gastos por Fornecedor
* Média de Gastos por Partido
* Média de Gasto dos Ultimos 5 anos

# Premissas da solução

## 3.1.Origem e especificação dos dados

Os dados foram retirados de repositório público no formato CSV, conforme abaixo:

* Cota Parlamentar : *https://brasil.io/dataset/gastos- deputados/cota parlamentar/*

## 3.2. Ambiente de desenvolvimento

## A empresa Blueshift disponibilizou todos os acessos, e ferramnetas necessárias para o desenvolvimento da arquitetura definida na apresentação do Blueprint;

## As etapas de desenvolvimento utilizado para realização desse projeto foram:

## Azure Databrikcs Microsoft;

## Azure Blob Storage ( Containers);

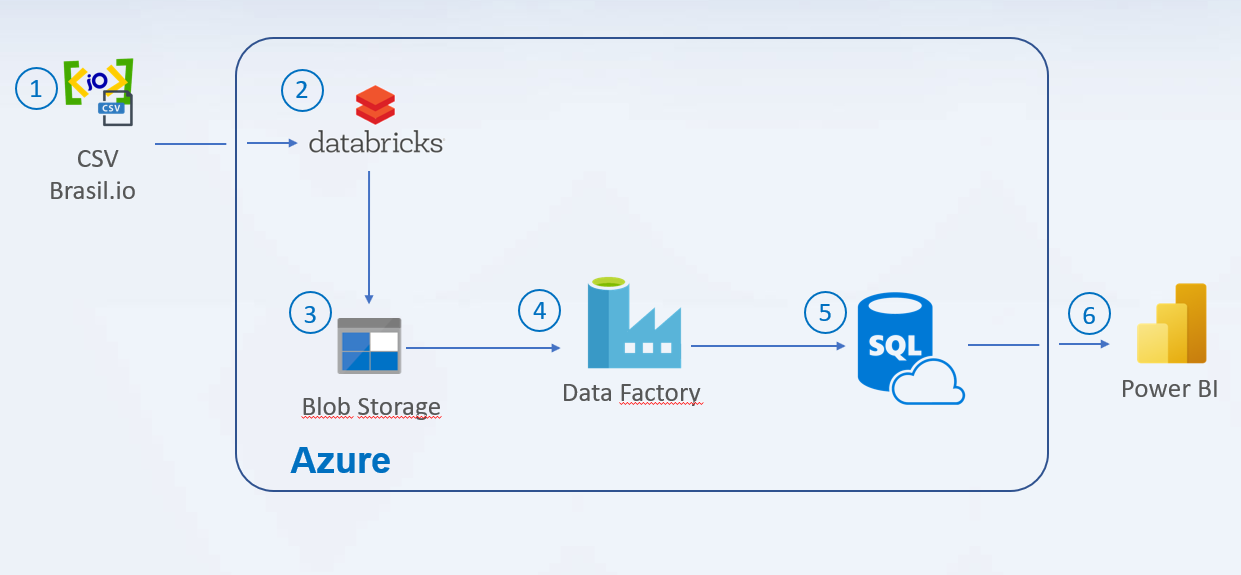
## Azure Data Factory;

## Microsoft Azure SQL Server;

## Power BI;

# Modelo da arquitetura sugerida

A figura abaixo apresenta a arquitetura da solução proposta com base no levantamento de requisitos e entendimento do negócio.



1. Arquivo CSV (Portal de transparencia, Brasil.io)
2. Extração dos arquivos via **Databricks**, utilizando um script em Python;
3. Armazenamento dos arquivos no **Azure Blob Storage** em formato CSV;
4. Descompactação dos arquivos com o **Data Factory**, no formato. CSV;
5. Manipulação dos dados, utulizando **SQL Server** Datawarehouse;
6. Desenvolvimento do relatório utilizando **Power BI** Dashboard;

# Extração via Azure Databricks

Foi desenvolvido um script para extração dos dados de forma automatizada via Python, utilizando o Databricks disponível em:

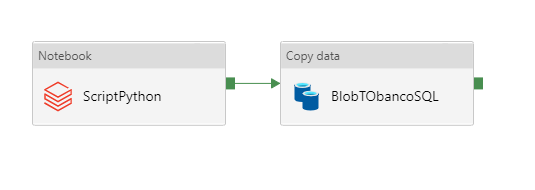
https://github.com/fernandofontebasso/projetointegrado

O código desenvolvido iniciava uma sessão com o blobstorage do Azure e fazia uso da API disponibilizada pelo site Brasil.io para baixar o arquivo cotasparlamentar.csv do site.

Após isso ele descompacta o arquivo e lê usando pyspark e após isso fazemos um tratamento nos dados no databricks limpando campos nulos e inserindo no blobstorage do Azure já com as alterações.

# Tratamento de dados via Datafactory

O tratamento dos dados foi realizado via **Datafactory** utilizando o processo abaixo:

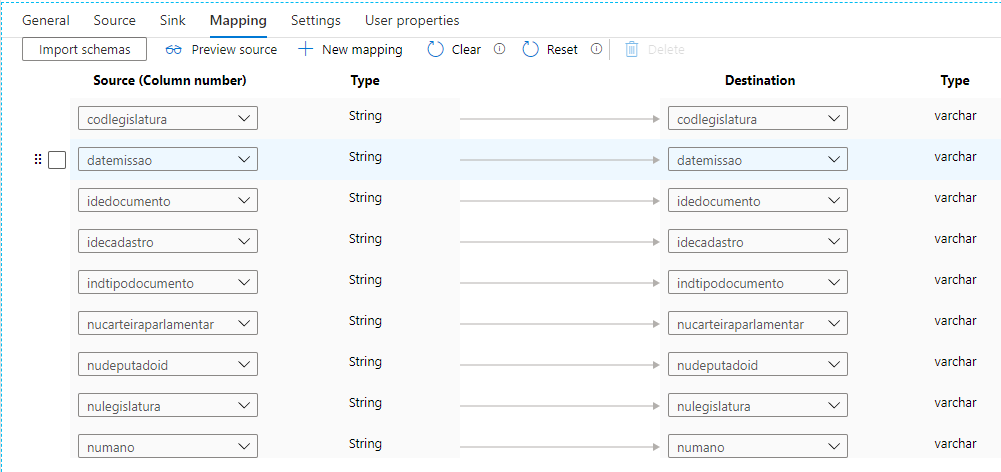


A pipeline foi desenvolvida para suportar o script em Pyspark no Databricks e se conectar com um Copy Data.

O script em python faz o download do arquivo e a limpeza dos dados e manda para o blob storage, após isso o Copy Data pega o arquivo do blob storage e leva ao banco de dados para popular as tabelas fato e dimensão.

# Ingestão no SQL Server

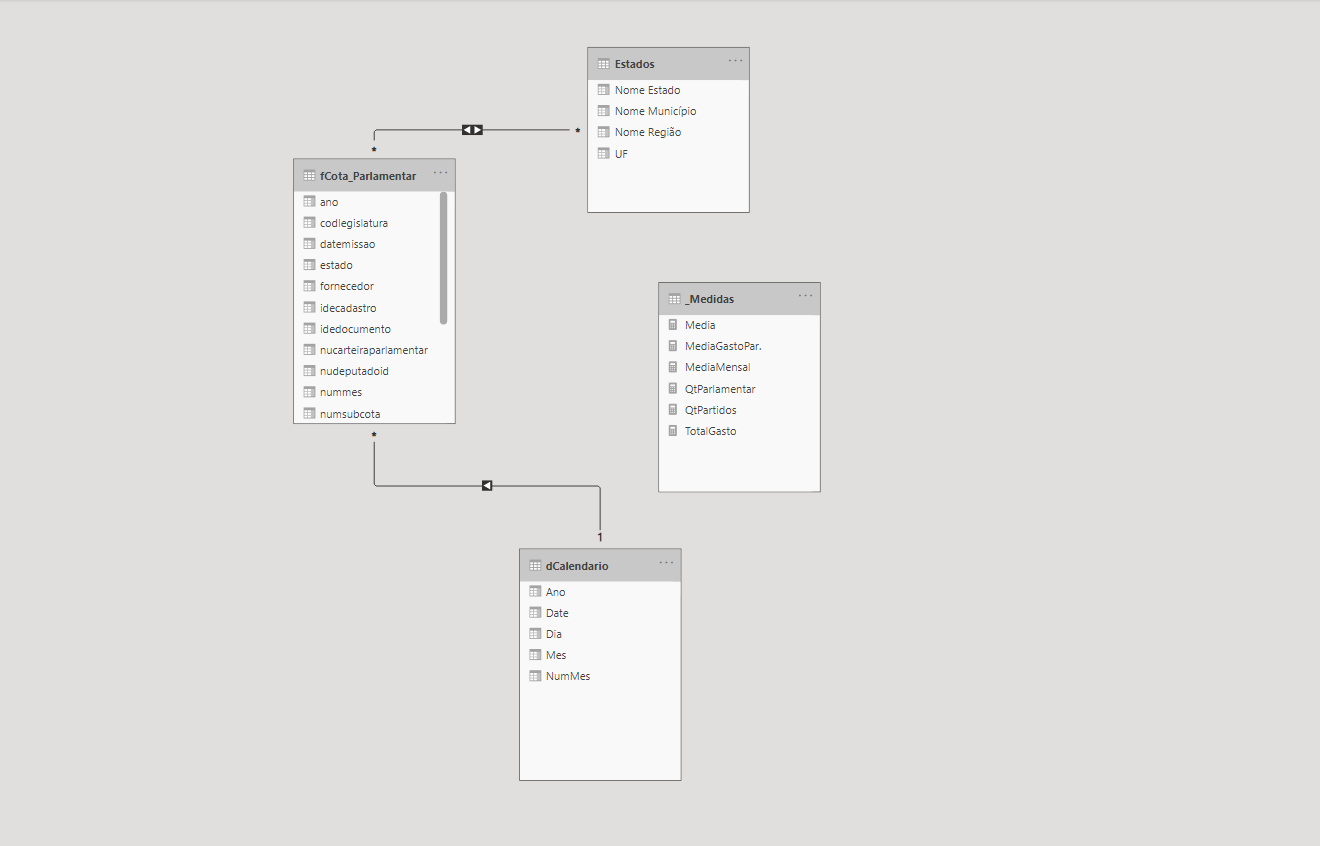
Seguindo as orientação propostas da empresa, foi realizado a criação de um modelo dimencional.



Na ingestão do SQL Server é usado o Data Factory para que as tabelas sejam populadas como String e no banco de dados é feita a alteração para float,varchar,etc, assim o banco é populado e as tabelas formatadas .

# Relacionamento

Segue abaixo o desenvolvimento do relacionamento de entre tabelas via **Power BI**.



# Dashboards

Será realizado o desenvolvimento de dois *dashboards / views* no Power BI que consumirão os dados disponibilizados nas tabelas de fato e dimensão SQL Server.

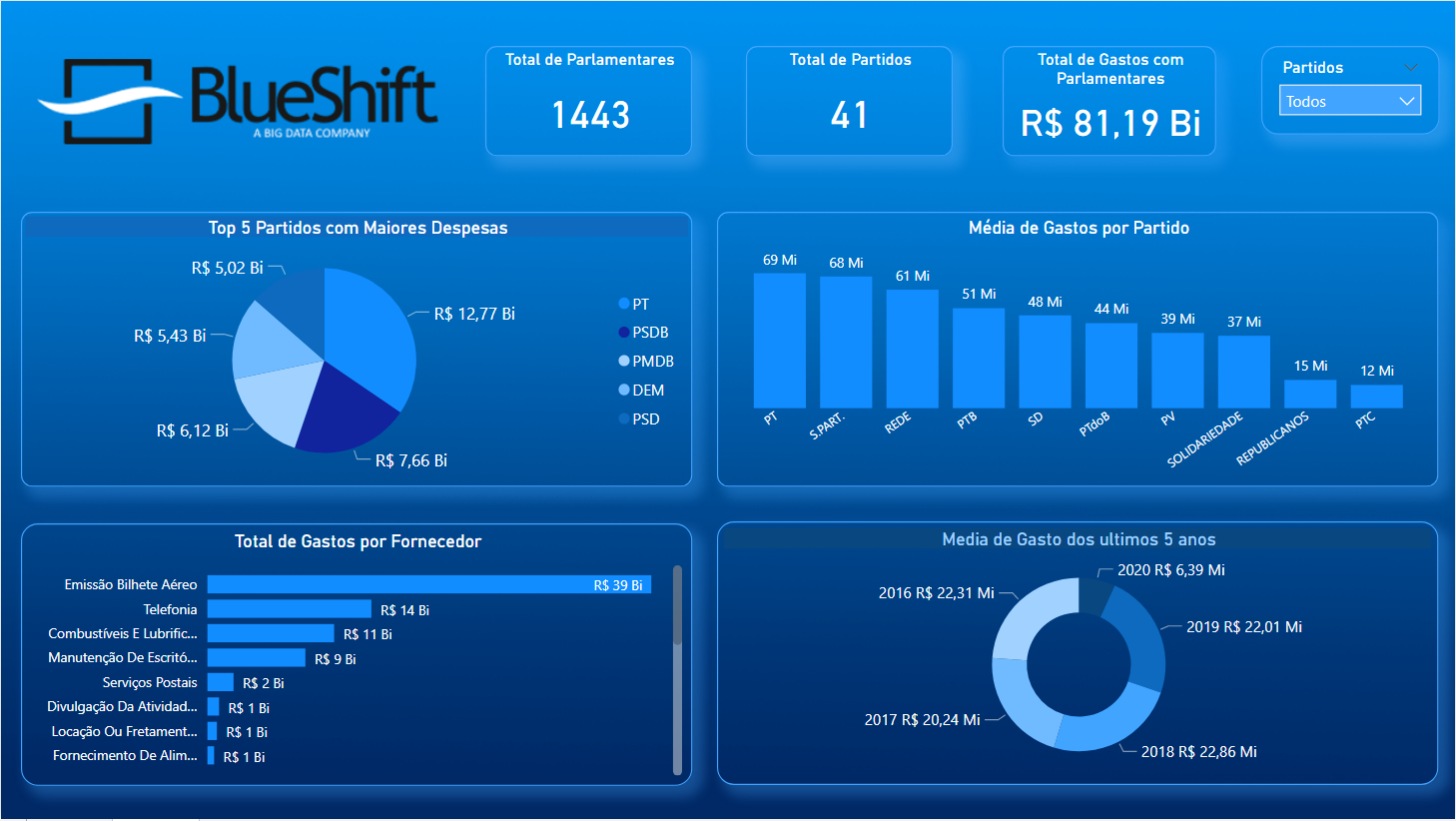
## Tabela Fato

A visualização do painel consolidado contém uma planilha tabela retratando as colunas a seguir:

* + COD. LEGISLUTURA
  + DATA EMISSÃO
  + ID DOCUMENTO
  + ID CADASTRO
  + N° CARTEIRA PARLAMENTAR
  + N° DEPUTADO
  + ANO
  + MÊS
  + N° SUBCOTA
  + PARTIDO
  + ESTADO
  + PARLAMENTAR
  + DESCRIÇÃO
  + FORNECEDOR
  + NUMERO
  + PASSAGEIRO
  + TRECHO
  + VALOR DOCUMENTO

## Distribuição dos filtros no painel

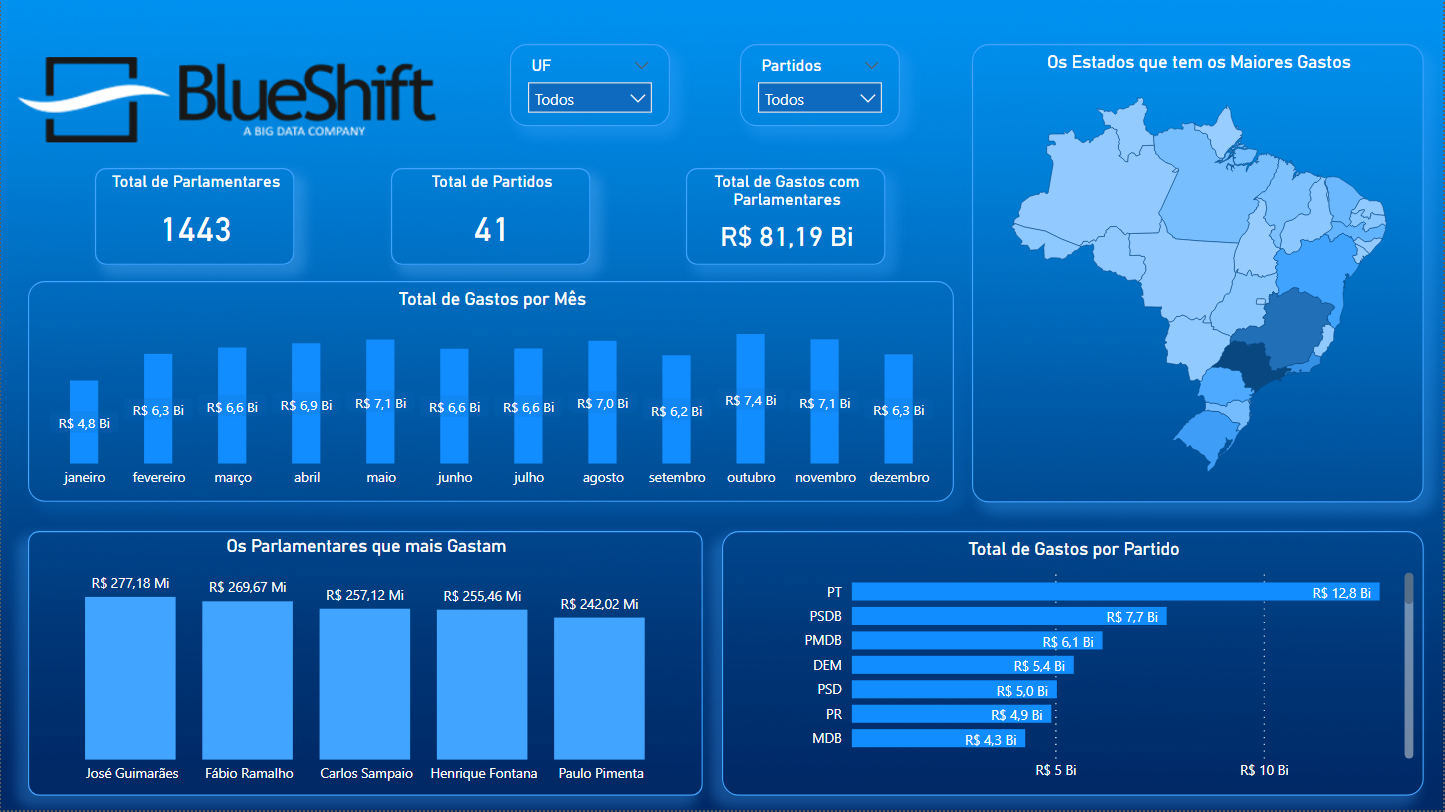
Contará com dois filtros, sendo estes:

* Partidos
* UF

## Distribuição das analises no painel

Nesta visualização temos as regras principais do projeto, sendo assim uma análise do acumulado de vendas e quantidade vendida dos ultimos 10 e 30 dias.

* + - Quantidade de vendas dos ultimos 10 dias - gráfico de barras
    - Acumulado ultimos 30 dias - gráfico de barras
    - Quantidade de vendas dos ultimos 10 dias - gráfico de barras



## Distribuição das analises no painel

Nesta visualização temos as regras como sendo uma análise do acumulado de vendas de 2017 e acumulado de vendas por região e canal

* + - Acumulado por País e Canal - gráfico tornado