Trabalho de Engenharia de Dados

Dados do Aluno

- **Nome**: Denilson Santos
- **Disciplina**: Engenharia de Dados
- **Pós-Graduação**: Ciência de dados e Analytics

Dados Utilizados

O dataset utilizado foram os dados dos deputados Estatuais da Bahia – Contendo as seguintes informações:

- **Nome do Parlamentar**
- **Sexo**
- **Data de Nascimento**
- **Escolaridade**
- **Profissão**.
- **Filiação**
- **Naturalidade**
- **Mandato Eletivo**
- **URL**
- **ID do Parlamentar**

Origem dos dados

Link:

https://www.al.ba.gov.br/deputados/todos-deputados
https://www.al.ba.gov.br/deputados/todos-deputados?&page=1&size=60
https://www.al.ba.gov.br/deputados/todos-deputados?inicio=1&quantidade=60&page=2&size=60
https://www.al.ba.gov.br/deputados/todos-deputados?inicio=2&quantidade=60&page=3&size=60
https://www.al.ba.gov.br/deputados/todos-deputados?inicio=3&quantidade=60&page=4&size=60
https://www.al.ba.gov.br/deputados/todos-deputados?inicio=4&quantidade=60&page=5&size=60
https://www.al.ba.gov.br/deputados/todos-deputados?inicio=5&quantidade=60&page=6&size=60
https://www.al.ba.gov.br/deputados/todos-deputados?inicio=6&quantidade=60&page=7&size=60
https://www.al.ba.gov.br/deputados/todos-deputados?inicio=7&quantidade=60&page=8&size=60
https://www.al.ba.gov.br/deputados/todos-deputados?inicio=8&quantidade=60&page=9&size=60
https://www.al.ba.gov.br/deputados/todos-deputados?inicio=9&quantidade=60&page=10&size=60
https://www.al.ba.gov.br/deputados/todos-deputados?inicio=10&quantidade=60&page=11&size=60
https://www.al.ba.gov.br/deputados/todos-deputados?inicio=11&quantidade=60&page=12&size=60

Descrição:

Páginas contendo os dados dos parlamentares tanto os atuantes quanto os ex-parlamentares

Definição do Problema

- O Objetivo é buscar a relação dos parlamentares que já exerceram ou exercem mandatos de Deputado estadual no Estado da Bahia.
- De posse com as informações poder disponibilizar em uma tabela os dados para que seja possível criar indicadores ou até mesmo alimentar algum outro processo de BI ou pesquisas e responder alguns questionamentos tais como:
    - Identificação da lista dos 10 parlamentares mais jovens em exercício;
    - Quantidade de Parlamentares do Sexo Feminino;
    - Quantidade de Parlamentares do Sexo Masculino
    - Quantidade de Parlamentar com Nível Superior
    - Identificar o Perfil de Parlamentares tais como como profissão
    - Identificar a quantidade de Parlamentares ativos por faixa de idade
    - Identificar a quantidade de parlamentares por sexo
- O dataset contém informações sobre vários deputados, mas algumas informações estão agrupadas e precisam ser manipuladas, tais como Data de Nascimento e Naturalidade

Para resolver esse problema, iremos realizar um Web Scraping dos dados e gerar uma Flat Table, para encontrar as informações listadas a cima.

Preparação de Dados

FERRAMENTA UTILIZADA PARA EXTRAÇÃO

Para extração dos dados estarei utilizando Python para realizar o Web Scraping dos dados.

BANCO DE DADOS PARA ARMAZENAMENTO E CONSULTA

Para o Armazenamento dos dados foi utilizado o BigQuery ao qual foi criado o Projeto DADOS_DEPUTADOS e a tabela ODS_PARLAMENTAR.

OBSERVAÇÔES

Por medida de segurança a conexão realizado no BIGquery optei por baixar a credencial no ambiente local e configurar o python para conectar

Alguns campos foram mantidos juntos e no formato String para que no SQL seja utilizado as funções de conversão.

EXTRAÇÃO DOS DADOS (ETL)

Para a extração dos dados utilizamos de forma separada dois processos um para buscar os Deputados atuantes (considerados os deputados da legislatura atual) e outro para buscar os ex-deputados.

Foi utilizado essa técnica para possibilitar maior clareza no código e ter a opção de ser gerado os dados apenas para um respectivo grupo.

Dificuldades encontradas na extração dos dados

1ª – Foi identificado que as páginas não existiam um padrão de layout

Como exemplo tínhamos páginas com informação de Profissão e outras não , porém na páginas que não tinha a informação de profissão não era mostrado o campo profissão em branco.

Solução:

Para que não existisse erro na gravação dos dados foi utilizado a verificação da existência de cada informação que seria levado para ODS e buscava o seu limitador, exemplo:

Para a Tag Datas de Nascimento o próximo campo limitador seria a profissão, se essa tag profissão não existisse iriamos buscando as outras e assim por diante e caso não encontrasse nenhum utilizamos o /n (ENTER) como limitador.

2ª – Foi identificado que a Data de Nascimento e a Naturalidade encontravam-se na mesma TAG

Como exemplo "**Nascimento: 02/06/1909, Fortaleza-CE "**

Solução:

Foi utilizado a técnica de transformação ao qual foi separado o campo em duas partes sendo atribuído a primeira a Data de Nascimento e a segunda a Naturalidade.

CONSTRUÇÃO DA TABELA

Optei pela criação da tabela utilizando o próprio Python com os comandos abaixo no lugar decriar dentro do BIGQuery com o Create Table

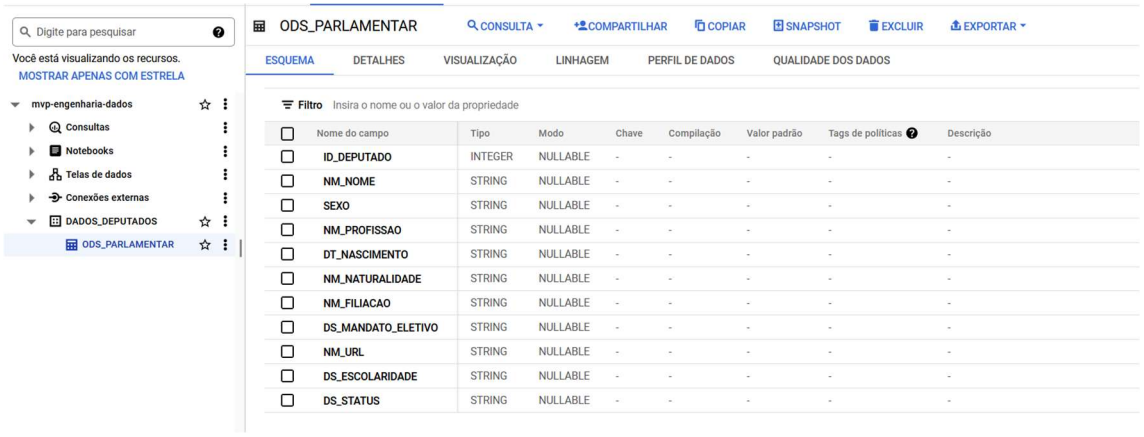
CRIA TABELA NO BANCO

```python
dataset_id = "DADOS_DEPUTADOS"
table_id = "ODS_PARLAMENTAR"
schema = [
    bigquery.SchemaField("ID_DEPUTADO", "INTEGER"),
    bigquery.SchemaField("NM_NOME", "STRING"),
    bigquery.SchemaField("SEXO", "STRING"),
    bigquery.SchemaField("NM_PROFISSAO", "STRING"),
    bigquery.SchemaField("DT_NASCIMENTO", "STRING"),
    bigquery.SchemaField("NM_NATURALIDADE", "STRING"),
    bigquery.SchemaField("NM_FILIACAO", "STRING"),
    bigquery.SchemaField("DS_MANDATO_ELETIVO", "STRING"),
    bigquery.SchemaField("NM_URL", "STRING"),
    bigquery.SchemaField("DS_ESCOLARIDADE", "STRING"),
    bigquery.SchemaField("DS_STATUS", "STRING")

]
table_ref = client.dataset(dataset_id).table(table_id)
table = bigquery.Table(table_ref, schema=schema)
table = client.create_table(table)  # Cria a tabela
```

[13] Python

Resutado de como ficou no BIGQuery



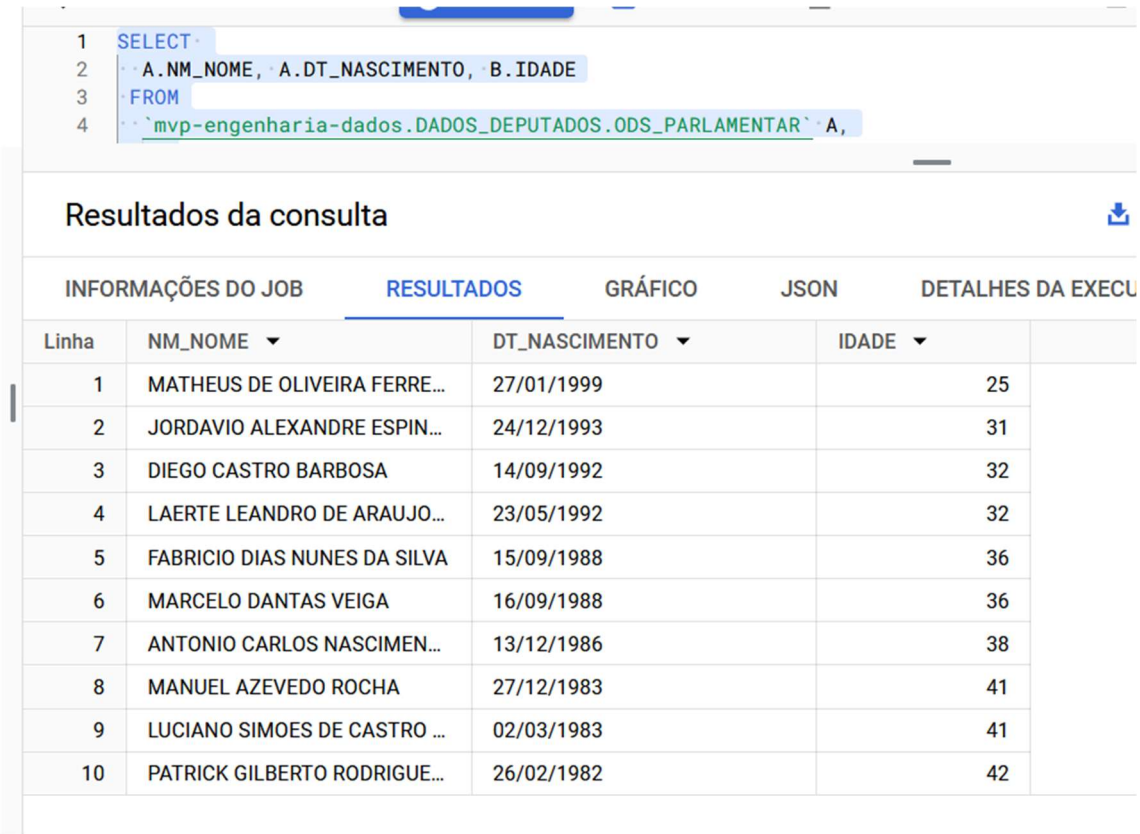
Análise dos dados

1ª Identificação da lista dos 10 parlamentares mais jovens em exercício

SQL

```
SELECT
  A.NM_NOME, A.DT_NASCIMENTO, B.IDADE
 FROM
  `mvp-engenharia-dados.DADOS_DEPUTADOS.ODS_PARLAMENTAR` A,
   (
   SELECT
    ID_DEPUTADO,  NM_NOME,   DT_NASCIMENTO,
    EXTRACT(YEAR FROM (CURRENT_DATE()))  - EXTRACT(YEAR
FROM(PARSE_DATE('%d/%m/%Y',  DT_NASCIMENTO))) IDADE,
   FROM
    `mvp-engenharia-dados.DADOS_DEPUTADOS.ODS_PARLAMENTAR`
   WHERE
    DS_STATUS = 'DEPUTADO'
   ) B
WHERE
  A.ID_DEPUTADO = B.ID_DEPUTADO
  AND A.DS_STATUS = 'DEPUTADO'
ORDER BY  B.IDADE ASC
LIMIT 10
```

Resultado



```
1  SELECT
2    A.NM_NOME, A.DT_NASCIMENTO, B.IDADE
3  FROM
4    `mvp-engenharia-dados.DADOS_DEPUTADOS.ODS_PARLAMENTAR` A,
```

Resultados da consulta

INFORMAÇÕES DO JOB | RESULTADOS | GRÁFICO | JSON | DETALHES DA EXECU

Linha	NM_NOME	DT_NASCIMENTO	IDADE
1	MATHEUS DE OLIVEIRA FERRE…	27/01/1999	25
2	JORDAVIO ALEXANDRE ESPIN…	24/12/1993	31
3	DIEGO CASTRO BARBOSA	14/09/1992	32
4	LAERTE LEANDRO DE ARAUJO…	23/05/1992	32
5	FABRICIO DIAS NUNES DA SILVA	15/09/1988	36
6	MARCELO DANTAS VEIGA	16/09/1988	36
7	ANTONIO CARLOS NASCIMEN…	13/12/1986	38
8	MANUEL AZEVEDO ROCHA	27/12/1983	41
9	LUCIANO SIMOES DE CASTRO …	02/03/1983	41
10	PATRICK GILBERTO RODRIGUE…	26/02/1982	42

2ª Quantidade de Parlamentares por Sexo

Query

```
SELECT
  SEXO , DS_STATUS,
  COUNT(*) AS QTD_PARLAMENTARES
FROM `mvp-engenharia-dados.DADOS_DEPUTADOS.ODS_PARLAMENTAR`
GROUP BY SEXO , DS_STATUS
ORDER BY SEXO, DS_STATUS
```

Resultado

Linha	SEXO	DS_STATUS	QTD_PARLAMENTAR
1		DEPUTADO	34
2		EX-DEPUTADO	345
3	Feminino	DEPUTADO	6
4	Feminino	EX-DEPUTADO	23
5	Masculino	DEPUTADO	23
6	Masculino	EX-DEPUTADO	297

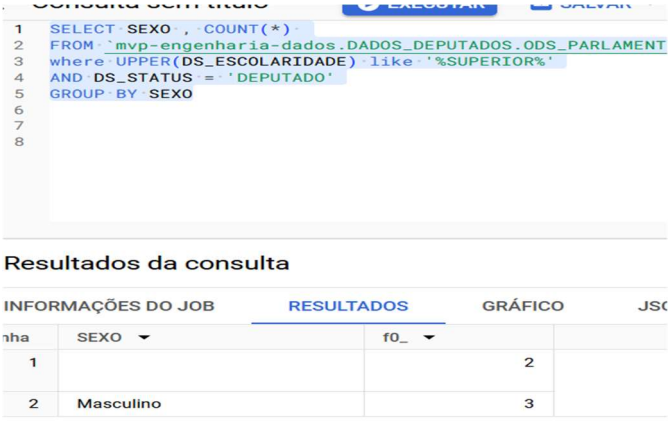
Com essa análise podemos identificar que a origem de dados não estar bem consistente pois existe parlamentar cadastrado sem a informação do Sexo, outra análise que podemos observar é que o número de parlamentares do sexo feminino é muito baixo

3ª Quantidade de Parlamentares por Sexo contendo a escolaridade de Nível superior

Query

```
SELECT SEXO , COUNT(*) FROM `mvp-engenharia-dados.DADOS_DEPUTADOS.ODS_PARLAMENTAR`
where UPPER(DS_ESCOLARIDADE) like '%SUPERIOR%' AND DS_STATUS = 'DEPUTADO'
GROUP BY SEXO
```

Resultado



4ª Identificação do perfil da casa legislativa buscando os Parlamentares por Profissão contendo os parlamentares com a mesma profissão.

Query

```
SELECT A.ID_DEPUTADO , A.NM_NOME , A.NM_PROFISSAO , A.SEXO , A.NM_URL
FROM
  `mvp-engenharia-dados.DADOS_DEPUTADOS.ODS_PARLAMENTAR` A ,
  (SELECT
     NM_PROFISSAO , count(*)
   FROM `mvp-engenharia-dados.DADOS_DEPUTADOS.ODS_PARLAMENTAR`
  GROUP BY NM_PROFISSAO
  HAVING COUNT(*) > 1
  ) B
WHERE A.NM_PROFISSAO = B.NM_PROFISSAO
```

Resultado

Autoavaliação

Ao finalizar este trabalho, foi possível identificar que os objetivos iniciais foram alcançados tais como identificar o parlamentar mais novo, o nível de escolaridade, a quantidade de parlamentares por sexo.

Durante a execução deste trabalho tive como dificuldade a construção do processo de extração e gravação dos dados ao qual foi a primeira vez que utilizo o método de Web Scraping e a utilização de armazenamento em nuvem, durante a análise dos dados foi possível identificar que a ODS pode servir de base para construção de novos indicadores em tabelas fato e novas dimensões tais como Município, Sexo, Profissão, e até mesmo a busca de mais dados não contemplados nesse trabalho.

Apesar das dificuldades o que prevaleceu foi o aprendizado que foi de grande enriquecimento e com isso posso aplicar em trabalhos futuros, tais como mineração de dados e busca de novas informação que possam ajudar nos projetos internos da empresa.