



Mercado de Trabalho

EQUIPE



ANDERSON
MELO
OLEMFA



DIEGO
AGUIAR

DIEGOAGUIARDSA



ASKA
PEREIRA
ASKAVICTORIA



JESSICA
STAUDT

JÉSSICALUANASTAUDT

ROSANA
SANTOS

ROSANA-DADOS

PEDRO
BARRIONOVO

PEDROBARRIONOVO



SUMÁRIO

- **Objetivo**
- **Sobre os Dados**
- **Metodologia/Workflow**
- **Perguntas de Negócio**
- **Power BI**
- **Matriz Swot**
- **BigQuery**
- **ETL**
- **Agradecimentos**

OBJETIVO



- **Panorama Mercado de Trabalho Brasil (2012-2022)**
- **Formal vs. Informal**
- **Características Demográficas**
- **Comportamento 10 Anos**
- **Impacto da COVID-19**
- **Pós-Pandemia**
- **Desigualdades e Inclusão Social**

SOBRE OS DADOS



- **Cadastro Geral de Empregados e Desempregados (CAGED)**
- **Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD-C)**
- **Instituto de Pesquisa Econômico Aplicado (IPEA)**
- **Censo 2010 (IBGE)**

METODOLOGIA



KDD: DESCOBERTA DE CONHECIMENTO EM BANCOS DE DADOS



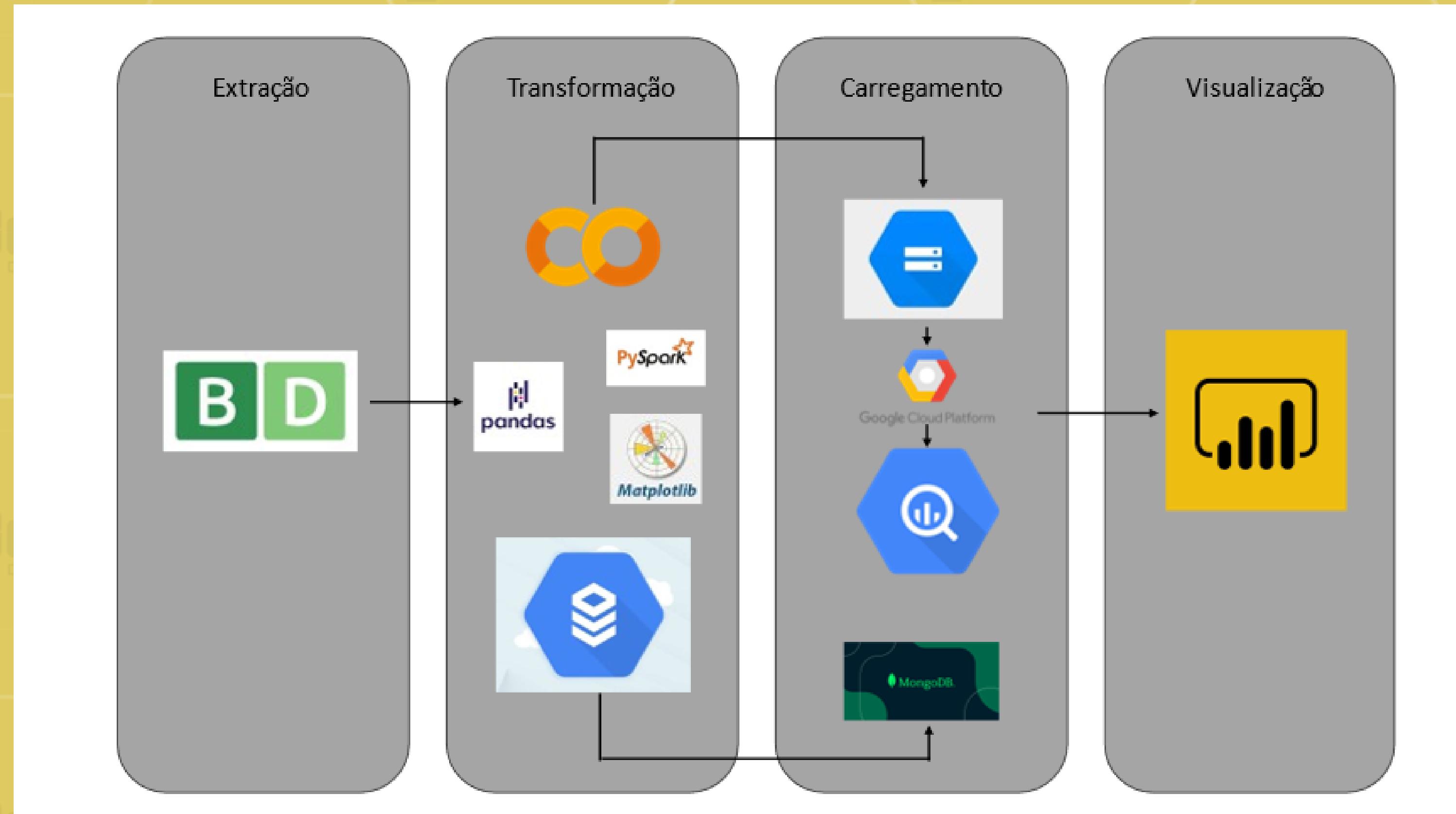
Fonte: tradução adaptada de (FAYYAD *et al.* 1996a)

KNOWLEDGE DISCOVERY IN DATABASES

WORKFLOW



WORKFLOW



PERGUNTAS DE NEGÓCIO



- **EMPREGOS POR SETOR: DISTRIBUIÇÃO E MUDANÇAS DE EMPREGOS POR SETOR?**
- **TAXAS DE DESEMPREGO: VARIAÇÃO TEMPORAL DAS TAXAS DE DESEMPREGO?**
- **PERFIL E EMPREGO: IMPACTO DE IDADE, GÊNERO, RAÇA, EDUCAÇÃO NAS OPORTUNIDADES?**
- **RENDA POR SETOR: DIFERENÇAS DE RENDA ENTRE SETORES?**
- **DIVERSIDADE ÉTNICA: AUTODECLARAÇÃO ÉTNICA E PROPORÇÕES?**
- **NÍVEIS DE EDUCAÇÃO: DISTRIBUIÇÃO DE NÍVEIS EDUCACIONAIS?**

- **FORMAL VS. INFORMAL: PROPORÇÃO FORMAL VS. INFORMAL AO LONGO DO TEMPO?**
- **EDUCAÇÃO E FORMALIDADE: LIGAÇÃO ENTRE EDUCAÇÃO E FORMALIDADE?**
- **EVOLUÇÃO DA FORMALIDADE: MUDANÇAS NA FORMALIDADE AO LONGO DO TEMPO?**
- **GÊNERO E FORMALIDADE: GÊNERO E ESCOLHA ENTRE EMPREGOS FORMAIS E INFORMAIS?**
- **DESEMPREGO PANDEMIA: VARIAÇÃO DO DESEMPREGO DURANTE A PANDEMIA?**
- **INFORMALIDADE PANDEMIA: IMPACTO DA PANDEMIA NA INFORMALIDADE?**

POWER BI



MATRIZ SWOT



MATRIZ SWOT

F Forças	O Oportunidades	F Fraquezas	A Ameaças
<ul style="list-style-type: none">• Diversidade Demográfica• Reforma Trabalhista	<ul style="list-style-type: none">• Capacitação e Educação• Análise Comparativa	<ul style="list-style-type: none">• Desigualdades Sociais• Educação	<ul style="list-style-type: none">• Riscos de Saúde Pública• Insegurança Econômica

ETL



- SCHEMA
- ESCOLHA DAS COLUNAS
- CONSULTAS PARA VERIFICAÇÃO DE INCONSISTÊNCIAS
- TRATAMENTO DA TABELA
- TRADUÇÃO DA TABELA

BIGQUERY

TRATAMENTO CAGED 2012-2019

```
WITH Tabela AS (
  SELECT
  SAFE_CAST(ano AS INT64) Ano,
  SAFE_CAST(sigla_uf AS STRING) Estado,
  SAFE_CAST(admitidos_desligados AS STRING) Movimentacao,
  SAFE_CAST(salario_mensal AS FLOAT64) Salario_Mensal,
  SAFE_CAST(indicador_aprendiz AS STRING) Aprendiz,
  CASE
    WHEN SAFE_CAST(indicador_trabalho_intermitente AS STRING) = '1' THEN '1'
    WHEN SAFE_CAST(indicador_trabalho_intermitente AS STRING) = '0' THEN '0'
    ELSE '3'
  END AS Trabalho_Intermitente,
  CASE
    WHEN SAFE_CAST(indicador_trabalho_parcial AS STRING) = '1' THEN '1'
    WHEN SAFE_CAST(indicador_trabalho_parcial AS STRING) = '0' THEN '0'
    ELSE '3'
  END AS Trabalho_Parcial,
  SAFE_CAST(indicador_portador_deficiencia AS STRING) PCD,
  CASE
    WHEN SAFE_CAST(grau_instrucao AS STRING) = '01' OR SAFE_CAST(grau_instrucao AS STRING) = '02' OR SAFE_CAST(grau_instrucao AS STRING) = '03' OR
    SAFE_CAST(grau_instrucao AS STRING) = '04' THEN '1'
    WHEN SAFE_CAST(grau_instrucao AS STRING) = '05' OR SAFE_CAST(grau_instrucao AS STRING) = '06' THEN '2'
    WHEN SAFE_CAST(grau_instrucao AS STRING) = '07' OR SAFE_CAST(grau_instrucao AS STRING) = '08' THEN '3'
    WHEN SAFE_CAST(grau_instrucao AS STRING) = '09' OR SAFE_CAST(grau_instrucao AS STRING) = '10' OR SAFE_CAST(grau_instrucao AS STRING) = '11' THEN '4'
    ELSE '99'
  END AS Escolaridade,
  SAFE_CAST(idade AS INT64) Idade,
  SAFE_CAST(sexo AS STRING) Sexo,
```

```
CASE
  WHEN SAFE_CAST(raca_cor AS STRING) = '01' THEN '01'
  WHEN SAFE_CAST(raca_cor AS STRING) = '02' THEN '02'
  WHEN SAFE_CAST(raca_cor AS STRING) = '04' THEN '04'
  WHEN SAFE_CAST(raca_cor AS STRING) = '06' THEN '06'
  WHEN SAFE_CAST(raca_cor AS STRING) = '08' THEN '08'
  ELSE '99'
END AS Raca_Cor,
SAFE_CAST(REPLACE(subsetor_ibge, ' ', '') AS INT64) AS subsetor
from `basedosdados-staging.br_me_caged_staging.microdados_antigos`
WHERE Ano IN ('2019', '2018', '2017', '2016', '2015', '2014', '2013', '2012')
  AND idade NOT IN ('000', '0000000', '0000010', '0000011', '0000012', '0000013',
'010', '011', '012', '013', '11', '13')
)
SELECT Ano, Estado, Movimentacao, Salario_Mensal, Aprendiz, Trabalho_Intermitente,
Trabalho_Parcial, PCD,
Escolaridade, Idade, Sexo, Raca_Cor,
CASE
  WHEN subsetor <= 13 THEN '01'
  WHEN subsetor = 14 OR subsetor >= 18 OR subsetor <= 23 THEN '02'
  WHEN subsetor = 15 THEN '03'
  WHEN subsetor = 16 OR subsetor = 17 THEN '04'
  WHEN subsetor = 24 THEN '05'
  WHEN subsetor = 25 THEN '06'
  ELSE '99'
END AS Subsetor_Economico
FROM Tabela
WHERE (salario_mensal <= 121500 AND salario_mensal >= 243 AND idade >= 14);
```

BIGQUERY

TRATAMENTO CAGED 2020-2023

```
WITH Tabela AS (
  SELECT
  SAFE_CAST(ano AS INT64) Ano,
  SAFE_CAST(sigla_uf AS STRING) Estado,
  CASE
    WHEN SAFE_CAST(tipo_movimentacao AS STRING) = '10' OR
    SAFE_CAST(tipo_movimentacao AS STRING)= '20' OR SAFE_CAST(tipo_movimentacao AS STRING) = '25'
      OR SAFE_CAST(tipo_movimentacao AS STRING) = '35' OR
    SAFE_CAST(tipo_movimentacao AS STRING) = '70' OR SAFE_CAST(tipo_movimentacao AS STRING) = '97' THEN '01'
    WHEN SAFE_CAST(tipo_movimentacao AS STRING) = '31' OR
    SAFE_CAST(tipo_movimentacao AS STRING) = '32' OR SAFE_CAST(tipo_movimentacao AS STRING) = '33'
      OR SAFE_CAST(tipo_movimentacao AS STRING) = '40' OR
    SAFE_CAST(tipo_movimentacao AS STRING) = '43' OR SAFE_CAST(tipo_movimentacao AS STRING) = '45'
      OR SAFE_CAST(tipo_movimentacao AS STRING) = '50' OR
    SAFE_CAST(tipo_movimentacao AS STRING) = '60' OR SAFE_CAST(tipo_movimentacao AS STRING)= '80'
      OR SAFE_CAST(tipo_movimentacao AS STRING) = '90' OR
    SAFE_CAST(tipo_movimentacao AS STRING) = '98' THEN '02'
      ELSE '99'
  END AS Movimentacao,
  SAFE_CAST(salario_mensal AS FLOAT64) Salario_Mensal,
  SAFE_CAST(indicador_aprendiz AS STRING) Aprendiz,
  SAFE_CAST(indicador_trabalho_intermitente AS STRING) AS Trabalho_Intermitente,
  SAFE_CAST(indicador_trabalho_parcial AS STRING) AS Trabalho_Parcial,
  CASE
```

```
WHEN SAFE_CAST(tipo_deficiencia AS STRING) = '0' THEN '0'
  ELSE '1'
END AS Pcd,
CASE
  WHEN SAFE_CAST(grau_instrucao AS STRING) = '1' OR SAFE_CAST(grau_instrucao AS STRING) = '2'OR SAFE_CAST(grau_instrucao AS STRING)= '3'
    OR SAFE_CAST(grau_instrucao AS STRING) = '4' THEN '1'
    WHEN SAFE_CAST(grau_instrucao AS STRING) = '5' OR SAFE_CAST(grau_instrucao AS STRING) = '6' THEN '2'
      WHEN SAFE_CAST(grau_instrucao AS STRING) = '7' OR SAFE_CAST(grau_instrucao AS STRING) = '8' THEN '3'
        WHEN SAFE_CAST(grau_instrucao AS STRING) = '9' OR SAFE_CAST(grau_instrucao AS STRING) = '10'OR SAFE_CAST(grau_instrucao AS STRING) = '11'
          OR SAFE_CAST(grau_instrucao AS STRING) = '80' THEN '4'
          ELSE '99'
  END AS Escolaridade,
  SAFE_CAST(idade AS INT64) Idade,
  CASE
    WHEN SAFE_CAST(sexo AS STRING) = '1' THEN '01'
    ELSE '02'
  END AS Sexo,
  CASE
    WHEN SAFE_CAST(raca_cor AS STRING) = '5' THEN '01' #Indigena
    WHEN SAFE_CAST(raca_cor AS STRING) = '1' THEN '02' #Branco
    WHEN SAFE_CAST(raca_cor AS STRING) = '2' THEN '04' #Preto
    WHEN SAFE_CAST(raca_cor AS STRING) = '4' THEN '06' #Amarelo
    WHEN SAFE_CAST(raca_cor AS STRING) = '3' THEN '08' #Pardo
    ELSE '99' #Não Informado
  END AS Raca_Cor,
  CASE
```

BIGQUERY

TRATAMENTO CAGED 2020-2023

```
WHEN SAFE_CAST(cnae_2_secao AS STRING) = 'B' OR cnae_2_secao = 'C'  
THEN '01' #Indústria  
WHEN SAFE_CAST(cnae_2_secao AS STRING) = 'F' THEN '03' #Construção  
Civil  
WHEN SAFE_CAST(cnae_2_secao AS STRING)= 'G' THEN '04' #Comércio  
WHEN SAFE_CAST(cnae_2_secao AS STRING) = 'O' THEN '05' #Adm Pública  
WHEN SAFE_CAST(cnae_2_secao AS STRING) = 'A' THEN '06' #Agropecuária  
ELSE '02' #Serviço  
END AS Subsetor_Econômico,  
FROM `basedosdados.br_me_caged.microdados_movimentacao`  
)  
  
SELECT Ano, Estado, Movimentacao, Salario_Mensal, Aprendiz,  
Trabalho_Intermitente, Trabalho_Parcial, PCD,  
Escolaridade, Idade, Sexo, Raca_Cor,Subsetor_Econômico,  
FROM Tabela  
  
WHERE (Salario_Mensal <= 172950 AND Salario_Mensal >= 350 AND Idade >= 14 AND  
Trabalho_Intermitente != '9' AND Trabalho_Parcial != '9');
```

CONSULTA

```
SELECT  
    tipo_movimentacao,  
    Count(*) AS qtd  
FROM  
    `basedosdados.br_me_caged.microdados_movimentacao`  
GROUP BY  
    tipo_movimentacao  
ORDER BY  
    qtd DESC;  
Select *  
FROM  
    `basedosdados.br_me_caged.microdados_movimentacao`
```

BIGQUERY

UNINDO TABELAS

```
CREATE TABLE `sympnone-project.projeto_final.caged-total` AS  
SELECT * FROM `sympnone-project.projeto_final.caged-2012-2019`  
UNION ALL  
SELECT * FROM `sympnone-project.projeto_final.caged-2020-2023`
```

VISUALIZANDO OS DADOS DA NOVA TABELA

```
SELECT * FROM `sympnone-project.projeto_final.caged-total`
```

BIGQUERY

TRADUÇÃO DA TABELA CAGED TOTAL

```
SELECT  
SAFE_CAST(Ano AS INT64) Ano,  
SAFE_CAST(Estado AS STRING) Estado,  
CASE  
WHEN SAFE_CAST(Movimentacao AS STRING) = '1' THEN 'Admissão'  
WHEN SAFE_CAST(Movimentacao AS STRING) = '2' THEN 'Desligamento'  
ELSE 'Não Informado'  
END AS Movimentacao,  
SAFE_CAST(Salario_Mensal AS FLOAT64) Salario_Mensal,  
CASE  
WHEN SAFE_CAST(Aprendiz AS STRING) = '1' THEN 'Sim'  
WHEN SAFE_CAST(Aprendiz AS STRING) = '0' THEN 'Não'  
ELSE 'Não Informado'  
END AS Aprendiz,  
CASE  
WHEN SAFE_CAST(Trabalho_Intermitente AS STRING) = '1' THEN 'Sim'  
WHEN SAFE_CAST(Trabalho_Intermitente AS STRING) = '0' THEN 'Não'  
ELSE 'Não se Aplica'  
END AS Trabalho_Intermitente,
```

```
CASE  
WHEN SAFE_CAST(Trabalho_Parcial AS STRING) = '1' THEN 'Sim'  
WHEN SAFE_CAST(Trabalho_Parcial AS STRING) = '0' THEN 'Não'  
ELSE 'Não se Aplica'  
END AS Trabalho_Parcial,  
CASE  
WHEN SAFE_CAST(PcD AS STRING) = '1' THEN 'Sim'  
WHEN SAFE_CAST(PcD AS STRING) = '0' THEN 'Não'  
ELSE 'Não Informado'  
END AS PcD,  
CASE  
WHEN SAFE_CAST(Escolaridade AS STRING) = '1' THEN 'Fundamental Incompleto'  
WHEN SAFE_CAST(Escolaridade AS STRING) = '2' THEN 'Fundamental Completo'  
WHEN SAFE_CAST(Escolaridade AS STRING) = '3' THEN 'Médio Completo'  
WHEN SAFE_CAST(Escolaridade AS STRING) = '4' THEN 'Superior Completo'  
ELSE 'Não Informado'  
END AS Escolaridade,  
SAFE_CAST(Idade AS INT64) Idade,  
CASE  
WHEN SAFE_CAST(Sexo AS STRING) = '1' THEN 'Masculino'  
ELSE 'Feminino'  
END AS Sexo,
```

BIGQUERY

TRADUÇÃO DA TABELA CAGED TOTAL

```
CASE
    WHEN SAFE_CAST(Raca_Cor AS STRING) = '2' THEN 'Branco'
    WHEN SAFE_CAST(Raca_Cor AS STRING) = '4' THEN 'Preto'
    WHEN SAFE_CAST(Raca_Cor AS STRING) = '6' THEN 'Amarelo'
    WHEN SAFE_CAST(Raca_Cor AS STRING) = '8' THEN 'Pardo'
    WHEN SAFE_CAST(Raca_Cor AS STRING) = '1' THEN 'Indígena'
    ELSE 'Não Informado'
END AS Raca_Cor,
CASE
    WHEN SAFE_CAST(Subsetor_Economico AS STRING) = '1' THEN 'Indústria'
    WHEN SAFE_CAST(Subsetor_Economico AS STRING) = '2' THEN 'Serviço'
    WHEN SAFE_CAST(Subsetor_Economico AS STRING) = '3' THEN 'Construção Civil'
    WHEN SAFE_CAST(Subsetor_Economico AS STRING) = '4' THEN 'Comércio'
    WHEN SAFE_CAST(Subsetor_Economico AS STRING) = '5' THEN 'Adm. Pública'
    WHEN SAFE_CAST(Subsetor_Economico AS STRING) = '6' THEN 'Agropecuária'
    ELSE 'Não Informado'
END AS Subsetor_Economico
FROM `symphone-project.projeto_final.caged-total`;
```

VISUALIZANDO OS DADOS DA TABELA TRATADA

```
SELECT * FROM `symphone-project.projeto_final.caged-total-final`
```

BIGQUERY

CONSULTAS CENSO 2010

SELECT

```
sigla_uf AS Estado,  
COUNT(*) AS QTD  
FROM `basedosdados.br_ibge_censo_demografico.microdados_pessoa_2010`  
GROUP BY Estado  
ORDER BY Estado ASC
```

```
CASE  
WHEN SAFE_CAST(v0614 AS STRING) = '1' OR SAFE_CAST(v0614 AS STRING) = '2' OR  
SAFE_CAST(v0615 AS STRING) = '1' OR SAFE_CAST(v0615 AS STRING) = '2' OR  
SAFE_CAST(v0616 AS STRING) = '1' OR SAFE_CAST(v0616 AS STRING) = '2' OR  
SAFE_CAST(v0617 AS STRING) = '1' THEN '1'#Sim  
ELSE  
'0' #Não  
END  
AS PCD,  
FROM  
`basedosdados.br_ibge_censo_demografico.microdados_pessoa_2010`  
SELECT  
*  
FROM  
Tabela  
WHERE  
Idade >= 14  
AND Idade <= 121;
```

WITH

```
Tabela AS (  
SELECT  
SAFE_CAST(sigla_uf AS STRING) Estado,  
CASE  
WHEN SAFE_CAST(v0601 AS STRING) = '1' THEN '01'#Masculino  
ELSE  
'02'#Feminino  
END  
AS Sexo,  
CASE  
WHEN SAFE_CAST(v0606 AS STRING) = '1' THEN '02' #'Branco'  
WHEN SAFE_CAST(v0606 AS STRING) = '2' THEN '04' #'Preto'  
WHEN SAFE_CAST(v0606 AS STRING) = '3' THEN '06' #'Amarelo'  
WHEN SAFE_CAST(v0606 AS STRING) = '4' THEN '08' #'Pardo'  
WHEN SAFE_CAST(v0606 AS STRING) = '5' THEN '01' #'Indígena'  
ELSE  
'99' #Não Informado  
END  
AS Raca_Cor,  
SAFE_CAST(v6036 AS INT64) Idade,  
SAFE_CAST(v6400 AS STRING) Escolaridade,  
CASE
```

BIGQUERY

NOVA TABELA CENSO 2010

CREATE TABLE

```
`sympnone-project.projeto_final.censo2010-final` AS
SELECT
  SAFE_CAST(Estado AS STRING) Estado,
  CASE
    WHEN SAFE_CAST(Sexo AS STRING) = '1' THEN 'Masculino'
    ELSE
      'Feminino'
  END
  AS Sexo,
  CASE
    WHEN SAFE_CAST(Raca_Cor AS STRING) = '2' THEN 'Branco'
    WHEN SAFE_CAST(Raca_Cor AS STRING) = '4' THEN 'Preto'
    WHEN SAFE_CAST(Raca_Cor AS STRING) = '6' THEN 'Amarelo'
    WHEN SAFE_CAST(Raca_Cor AS STRING) = '8' THEN 'Pardo'
    WHEN SAFE_CAST(Raca_Cor AS STRING) = '1' THEN 'Indígena'
    ELSE
      'Não Informado'
```

END

```
  AS Raca_Cor,
  SAFE_CAST(Idade AS INT64) Idade,
  CASE
    WHEN SAFE_CAST(Escolaridade AS STRING) = '1' THEN 'Fundamental Incompleto'
    WHEN SAFE_CAST(Escolaridade AS STRING) = '2' THEN 'Fundamental Completo'
    WHEN SAFE_CAST(Escolaridade AS STRING) = '3' THEN 'Médio Completo'
    WHEN SAFE_CAST(Escolaridade AS STRING) = '4' THEN 'Superior Completo'
    ELSE
      'Não Informado'
  END
  AS Escolaridade,
  CASE
    WHEN SAFE_CAST(PcD AS STRING) = '1' THEN 'Sim'
    WHEN SAFE_CAST(PcD AS STRING) = '0' THEN 'Não'
    ELSE
      'Não Informado'
  END
  AS PcD
FROM
  `sympnone-project.projeto_final.censo2010`
```

BIGQUERY

TRATAMENTO DO PNADC

```
WITH Tabela AS (
  SELECT
    SAFE_CAST(ano AS INT64) Ano,
    SAFE_CAST(sigla_uf AS STRING) Estado,
    CASE
      WHEN SAFE_CAST(V2007 AS STRING) = '1' THEN '01'
      ELSE '02'
    END AS Sexo,
    CASE
      WHEN SAFE_CAST(V2010 AS STRING) = '1' THEN '02' #'Branco'
      WHEN SAFE_CAST(V2010 AS STRING) = '2' THEN '04' #'Preto'
      WHEN SAFE_CAST(V2010 AS STRING) = '3' THEN '06' #'Amarelo'
      WHEN SAFE_CAST(V2010 AS STRING) = '4' THEN '08' #'Pardo'
      WHEN SAFE_CAST(V2010 AS STRING) = '5' THEN '01' #'Indigena'
      ELSE '99'                                #'Não Informado'
    END AS Raca_Cor,
    SAFE_CAST(V2009 AS INT64) AS Idade,
    CASE
      WHEN SAFE_CAST(VD3004 AS STRING) = '1' OR SAFE_CAST(VD3004 AS
      STRING) = '2' THEN '1'
      WHEN SAFE_CAST(VD3004 AS STRING) = '3' OR SAFE_CAST(VD3004 AS
      STRING) = '4' THEN '2'
      WHEN SAFE_CAST(VD3004 AS STRING) = '5' OR SAFE_CAST(VD3004 AS
      STRING) = '6' THEN '3'
      WHEN SAFE_CAST(VD3004 AS STRING) = '7' THEN '4'
      ELSE '99'
    END AS Escolaridade
  FROM `basedosdados.br_ibge_pnadc.microdados`)
SELECT *
FROM Tabela
WHERE Idade >= 14 AND Idade <= 121;
```

TRADUÇÃO DA TABELA PNADC

```
CREATE TABLE `symphone-project.projeto_final.pnad-c-ibge-final` AS
SELECT
  SAFE_CAST(Ano AS INT64) Ano,
  SAFE_CAST(Estado AS STRING) Estado,
  CASE
    WHEN SAFE_CAST(Sexo AS STRING) = '1' THEN 'Masculino'
    ELSE 'Feminino'
  END AS Sexo,
  CASE
    WHEN SAFE_CAST(Raca_Cor AS STRING) = '2' THEN 'Branco'
    WHEN SAFE_CAST(Raca_Cor AS STRING) = '4' THEN 'Preto'
    WHEN SAFE_CAST(Raca_Cor AS STRING) = '6' THEN 'Amarelo'
    WHEN SAFE_CAST(Raca_Cor AS STRING) = '8' THEN 'Pardo'
    WHEN SAFE_CAST(Raca_Cor AS STRING) = '1' THEN 'Indigena'
    ELSE 'Não Informado'
  END AS Raca_Cor,
  SAFE_CAST(Idade AS INT64) Idade,
  CASE
    WHEN SAFE_CAST(Escolaridade AS STRING) = '1' THEN 'Fundamental
    Incompleto'
    WHEN SAFE_CAST(Escolaridade AS STRING) = '2' THEN 'Fundamental
    Completo'
    WHEN SAFE_CAST(Escolaridade AS STRING) = '3' THEN 'Médio Completo'
    WHEN SAFE_CAST(Escolaridade AS STRING) = '4' THEN 'Superior
    Completo'
    ELSE 'Não Informado'
  END AS Escolaridade
FROM `symphone-project.projeto_final.pnad-c-ibge-novo` ;
```

CONSULTAS BIGQUERY



ETL COLAB



COLAB

Extração

O próximo passo é fazer a extração dos dados. Essa etapa vai depender de onde estarão localizados os dados. Neste projeto em específico serão mostradas como obter os dados localizados em um Google Drive, mas também pelo serviço de armazenamento da GCP (Google Cloud Platform).

Google Cloud

```
[ ] 1 # CONFIGURANDO DA CHAVE DE SEGURANCA - ACESSO O PROJETO
2 serviceAccount = '/content/symphone-project-2b6e69a15a32.json' # Chave criada no IAM
3 os.environ['GOOGLE_APPLICATION_CREDENTIALS'] = serviceAccount

[ ] 1 # Configurações Google Cloud Storage - ACESSO AO BUCKET
2 client = storage.Client()
3 bucket = client.get_bucket('mercado-de-trabalho-projeto')
4 bucket.blob('ipea.xls')
5 path = 'gs://mercado-de-trabalho-projeto/Base de Dados/IPEA/Bruto/ipea.xls'

▶ 1 # Abertura da base de dados e cópia de segurança (bucket)
2 # Ler o arquivo XLS, utilizando pandas
3 path = "gs://mercado-de-trabalho-projeto/Base de Dados/IPEA/Bruto/ipea.xls" # Substitua pelo caminho real do arquivo XLS
4 df = pd.read_excel(path, sheet_name='TrimestreMovel')
```

COLAB

Tradução

+ Código

+ Texto

Primeiramente serão traduzidos ou renomeados os atributos do dataframe com o objetivo de melhorar o entendimento sobre os mesmos sem precisar constantemente recorrer ao dicionário de dados.

```
[ ] 1 # Renomeando colunas
2
3 df.rename(columns={
4     'anomesfinaltrimmovel':'Ano',
5     'populacao': 'Populacao' ,
6     'popocup': 'Ocupacao',
7     'popdesocup': 'Desocupacao',
8     'popforadaforca': 'Fora_da_forca',
9     'empregprivcomcart':'Privado_com_cart',
10    'empregprivsemcart': 'Privado_sem_cart', # pode ter formais e informais
11    'domesticocomcart': 'Domestico_com_cart',
12    'domesticosemcart': 'Domestico_sem_cart', # pode ter formais e informais
13    'empregpublicomcart':'Publi_com_cart',
14    'empregpublsemcart': 'Publi_sem_cart',
15    'contapropriacomcnpj':'Conta_prop_com_cnpj',
16    'contapropriasemcnpj': 'Conta_prop_sem_cnpj',
17    'trabfamauxiliar':'Trabalhador_familiar'},inplace=True)
18
19
20 df
```

COLAB - SCHEMA

```
1 # Criação do schema com pyspark
2
3 from pyspark.sql.types import *
4 from pyspark.sql.types import StructType,StructField, StringType, IntegerType, DoubleType
5
6
7 schema = StructType([
8     StructField('Ano', IntegerType(), True), \
9     StructField('Populacao', IntegerType(), True), \
10    StructField('Ocupacao', IntegerType(), True), \
11    StructField('Desocupacao', IntegerType(), True), \
12    StructField('Fora_da_forca', IntegerType(), True), \
13    StructField('Privado_com_cart', IntegerType(), True), \
14    StructField('Privado_sem_cart', IntegerType(), True), \
15    StructField('Domestico_com_cart', IntegerType(), True), \
16    StructField('Domestico_sem_cart', IntegerType(), True), \
17    StructField('Publi_com_cart', IntegerType(), True), \
18    StructField('Publi_sem_cart', IntegerType(), True), \
19    StructField('Conta_prop_com_cnpj', DoubleType(), True), \
20    StructField('Conta_prop_sem_cnpj', DoubleType(), True), \
21    StructField('Trabalhador_familiar', IntegerType(), True), \
22])
23
24 # Criar um novo DataFrame com o novo schema
25
26 df_py = spark.createDataFrame(df, schema=schema)
27
28 # Exibir o schema do novo DataFrame
29
30 df_py.printSchema()
31
```

MONGODB

Tratamento

```
1 # gsutil da base de dados tratada diretamente da bucket
2
3 path = 'gs://mercado-de-trabalho-projeto/Base de Dados/IPEA/Tratado/ipea_tratado.csv'
4 df = pd.read_csv(path,
5                  encoding='ISO-8859-1',
6                  )
```

Carregamento

```
[ ] 1 # Conector MongoDB
2
3 uri = "mongodb+srv://symphone.q44pclg.mongodb.net/?authSource=%24external&authMechanism=MONGODB-X509&retryWrites=true&w=majority"
4
5 # Conexão
6 # apos fazer o download do certificate do mongo, faça o upload dessa chave no google colab, copie o caminho e insira abaixo
7 client = MongoClient(uri, tls=True, tlsCertificateKeyFile='/content/X509-cert-4703031099688872260.pem') # path da chave certificare do mongo

[ ] 1 # Carregamento da base de dados tratada no MongoDB
2 # criando a pasta dos dados tratados no mongo
3 db2 = client['pandasmongo']
4 collection2 = db2['tratados_ipea']
5 collection2.count_documents({})

0

[ ] 1 # Conversão de dados para MongoDB
2 # transformando de tabela para dicionario mongo
3 df_dict = df.to_dict("records")
4 collection2.insert_many(df_dict)

In[5]: pymongo.results.InsertManyResult at 0x7d881fe8b20
```



OBRIGADO!