

ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS

## Projeto Final - ETL

Crislaine Reis

Adriana Goés

Ceará

2022

# Resumo

Este trabalho tem como objetivo demonstrar o resultado da realização de um processo de extração, transformação e carga utilizando três base dados. A linguagem de programação Python foi utilizada como facilitadora desse processo, no jupyter foi possível realizar os joins das bases, bem como os tratamentos necessários para limpeza e padronização dos dados, e por fim a base foi carregada no ambiente da cloud da google, big query.

# 2 - Identificação e Importação das Bases

## Identificação

No primeiro momento, foi realizada uma pesquisa das bases que seriam necessárias para a realização deste trabalho, a base escolhida foi está disponível no portal do governo que pode ser acessada por meio [desse link](https://dados.gov.br/dataset/dados-abertos-de-contratos-administrativos/resource/6a9526c4-bd86-46fb-b7aa-539df37f849a), tendo a estrutura definida foi criado a arquitetura do projeto no [miro](https://miro.com/welcomeonboard/akkwYWpXVXVlb21ta280eHBPWm00eUFpMGhqVXBYMXJoZXRlTUxUcjg0MGU0ZXlxdzlzMkNxRkZHWVlHbkNIRnwzNDU4NzY0NTMyNTA5NjE2NjUy?share_link_id=45995058641), a mesma pode ser vista acessando o referido site.

## Importação dos dados

Após a importação das bibliotecas necessárias para a realização do tratamendo dos dados, as bases extraídas tabela\_contratos.csv, tabelas\_empresas.csv e tabelas\_datas.csv, foram carregadas no jupyter para criação dos seus respectivos data frames.

# 3 - Tratamento dos dados

## 3.1 Realizando o tratamento dos dados

Tendo os dataframes criados, foi realizado um merge da tabela empresas com a tabela contratos e datas, para que essas informações estivessem presente na tabela contrato. Após o merge realizado, as colunas desnecessárias foram apagadas e foi criado o dataframe contratos\_empresas\_final, para melhor leitura e organização, as colunas das datas de início e término dos contratos foram renomeadas.

Após os tratamentos supracitados, foi realizado uma checagem da quantidade de linha do dataframe para cada coluna, para verificar a consistência da mesma. foram verificados os tipos dos dados existentes, após essa verificação foi necessário realizar a alteração do dado do tipo date isso gerou um erro, pois havia uma data na tabela que estava fora do padrão, e precisou ser corrigida.

Tendo a correção realizada, foi realizada uma função para exibir no dataframe o tempo de contrato em dias.

Na última checagem, foi visto que existiam prazos de contratos menor ou igual a 0 dias, esse erro foi corrigido.

# 4 - Carga dos dados

## 4.1 Carregando os dados no Big Query

Com o data frame devidamente criado e tratado, foi possível realizar a carga do mesmo no google big query no ambiente do projeto final - ETL , para realizar essa carga, foi necessário criar uma conta de serviço e gerar a chave de autenticação, para que a biblioteca do google cloud fosse utilizada sem intercorrências.

Todos os arquivos do projeto supracitado está disponível no [github](https://github.com/crislaine27).