DOCUMENTAÇÃO - PROJETO FINAL BC8 - ENERGIA

Aurélia Covre / João Victor Guimarães / Kely Fernandes / Ricardo Rowedder / Robson Motta

- Teste das ferramentas em ambiente local

- Primeiramente criamos um repositório no GitHub (https://github.com/JoaoVictorGuimaraesUFJF/Projeto_Final) para armazenar os códigos do projeto e um Trello (https://trello.com/b/Y3tsBQuz/projeto-final-energia) para acompanhamento das etapas, em seguida, utilizamos as 4 planilhas abaixo para o projeto.
    - Série histórica de Geração Distribuida - Geração Distribuída - Conjuntos de dados - Portal Brasileiro de Dados Abertos
    - Empreendimento geração distribuída - Relação de empreendimentos de Geração Distribuída - Conjuntos de dados - Portal Brasileiro de Dados Abertos
    - Tarifa fornecimento residência e Tarifa média fornecimento - Tarifas - ANEEL

- Criado o ambiente de execução em nuvem, com:
    - Um cluster com um nó master e dois nós workers, ambos com 4 CPU's e 15GB de memória
    - Uma instância para o MySQL, banco de dados SQL utilizado no projeto
    - Uma VM para o banco de dados Cassandra, o banco de dados NoSQL utilizado no projeto
    - Buckets para o armazenamento dos arquivos brutos e tratados, jobs e outros.

- Ao iniciarmos a leitura dos arquivos utilizando a biblioteca Pandas, nos deparamos com alguns problemas onde foi necessária a utilização do encoding utf-16 e encoding latin-1 por conta dos caracteres especiais

- Na leitura da planilha “Tarifa média de fornecimento” foi identificada uma coluna de texto com vírgula, e devido a vírgula ser um separador de arquivo csv, a vírgula quebrava parte do texto formando uma nova coluna. Fizemos uma concatenação do texto que foi separado em apenas uma coluna.

- Foram realizados tratamentos nas tabelas "Tarifa mensal Fornecimento", “Tarifa Fornecimento Residencial” e “Geração distribuída” - exclusão das linhas com espaços vazios utilizando o pandas (dropna).

- Tratamento realizado na tabela “Empreendimento geração distribuída” preenchimento dos campos faltantes (FIllna). Pois nessa tabela haviam dados relevantes nas linhas que não queríamos perder e por isso optamos pelo preenchimento ao invés do drop.

- Tabela "empreendimentos" coluna "PotenciaInstaladaKW" números com o separador decimal a vírgula , no momento da inserção o algoritmo estava entendendo como uma coluna a mais a ser inserida, o tratamento foi: converter a coluna "PotenciaInstaladaKW" de vírgula para ponto

- Ajustes de tratamento pois detectamos apóstrofo (') entre o texto e isso estava atrapalhando a inserção

- Criação do database MySQL no Google Cloud
    - Triggers
    - Procedures

- Inserção das dados no database MySQL

- Migração dos dados do database MySQL para o database Cassandra

- Salvamos os arquivos em parquet apenas em três das nossas tabelas. Na tabela empreedimentosGD tivemos um problema de limitação de hardware em nossos computadores onde os mesmos não tinham memória disponível. Este problema foi solucionado quando foi feito o mesmo procedimento em nuvem.

- Criamos Keyspace e tabelas no database Cassandra na nuvem

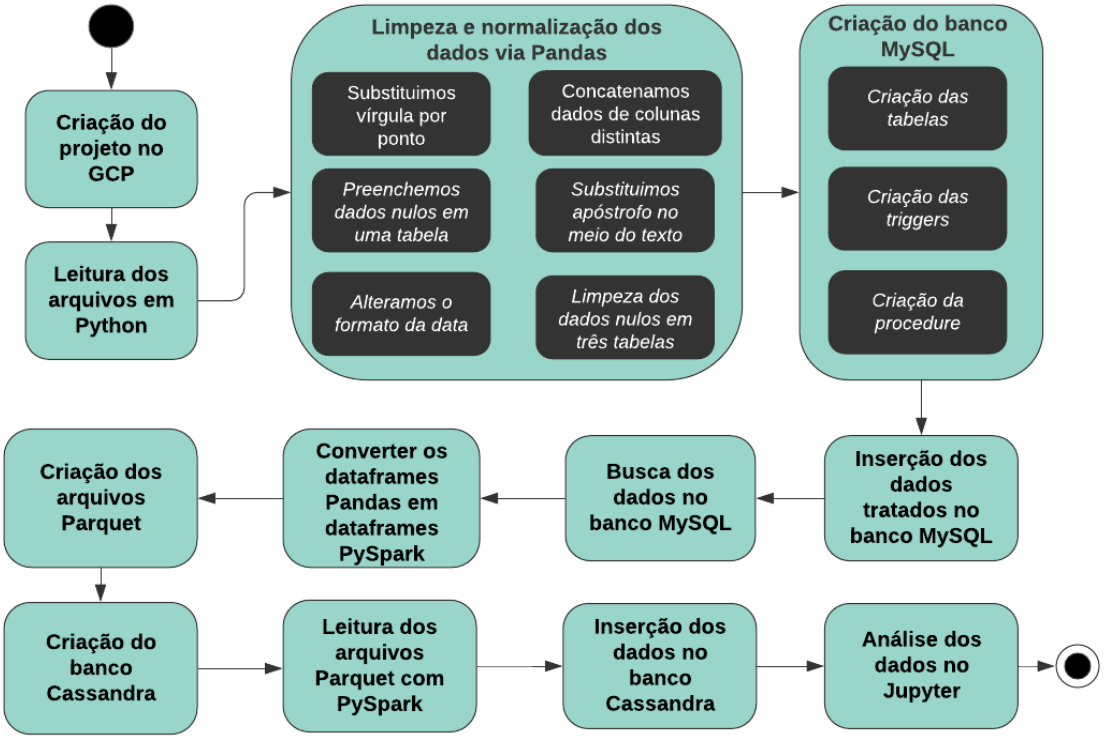
- Leitura e inserção dos arquivos Parquet no database Cassandra

- Início das análises

- Geração dos gráficos

- Criação da Apresentação

- Geração da documentação



FERRAMENTAS UTILIZADAS NA ELABORAÇÃO DO PROJETO

- Utilizamos o Trello, um aplicativo de gestão de projeto que possui ferramentas para gerenciamento de tarefas que facilitam a comunicação com o objetivo de otimizar o tempo e produtividade
- Utilizamos o GitHub para armazenar os códigos e para todos da equipe terem acesso durante a realização do projeto
- No Python utilizamos as bibliotecas Pandas e Numpy para limpeza, normalização e visualização dos dados
- Utilizamos o banco MySQL e o NoSQL Cassandra para a inserção dos dataset
- Utilizamos a interface PySpark para criação de arquivos parquet
- Utilizamos o Google Cloud para a execução do nosso projeto em nuvem
- Utilizamos o Jupyter Notebook para a realização das análises e resultados