Relatório Processo ETL

Nesse projeto foi elaborado um processo de ETL. Nosso grupo utilizou duas fontes de dados, uma do Instituto Nacional de Meteorologia (INMET), que contém dados climáticos sobre o território brasileiro, e outra da Agência Nacional de Energia Elétrica (ANEEL), que contém informações sobre empreendimentos de geração de energia no Brasil.

As bases de dados foram extraídas manualmente dos seus respectivos sites.

Para os **dados climáticos** foi acessada a base de dados da Agência Nacional de Energia Elétrica, o grupo seguiu o seguinte processo:

* Acessar o site do INMET pelo link: <https://portal.inmet.gov.br/dadoshistoricos> ;
* Ir até a aba de "Dados metereológicos";
* Selecionar "Histórico de dados metereológicos";
* Selecionar os dados do ano de 2022;
* Após isso um download de um arquivo rar vai se iniciar, basta extraí-lo para utilizar a base de dados.

Para os **dados de empreendimento** foi acessada a base de dados do Instituto Nacional de Meteorologia, o grupo seguiu o seguinte processo:

* Acessar o site dos dados abertos da Agência Nacional de Energia Elétrica pelo link: [https://dadosabertos.aneel.gov.br](https://dadosabertos.aneel.gov.br/) ;
* Procuramos por diversas tabelas com termos gerais como “Geração de energia”;
* Encontramos a tabela pelo nome de “[SIGA - Sistema de Informações de Geração da ANEEL](https://dadosabertos.aneel.gov.br/dataset/siga-sistema-de-informacoes-de-geracao-da-aneel)”;
* Fizemos o download do arquivo csv pelo nome de “siga-empreendimentos-geracao.csv” ;

Para cada uma dessas fontes de dados, a equipe utilizou o Jupyter Notebook e a biblioteca Pandas para fazer a leitura dos dados, realizar a transformação necessária e, em seguida, salvar os dados em arquivos no formato CSV.

Para os dados de empreendimento:

* Selecionar as colunas relevantes para a análise (Data de Geração, Nome do Empreendimento, Sigla da UF, Tipo de Geração, Fonte de Combustível, Potência Fiscalizada e Potência Outorgada).
* Ler o arquivo CSV e armazenar os dados filtrados em um dataframe genérico que utilizaremos para as tabelas fato e dimensão;
* Criamos um indíce através do axis do dataframe;
* Para a tabela fato utilizamos os campos escolhidos para a mesma (Data de Geração, Potência Fiscalizada e Potência Outorgada);
* Para a tabela dimensão utilizamos os campos escolhidos para a mesma (Nome empreendimento, Unidade Federativa, Abreviação do tipo de geração, Nome da fonte de combustível);
* Em seguida traduzimos ambos os dataframes em seus respectivos arquvios csv.

No caso dos dados de empreendimentos de geração de energia, a equipe selecionou apenas as colunas relevantes para a análise que estavam presentes no arquivo original e salvou os dados filtrados em um novo arquivo CSV. Já que a mesma estava formatada em uma tabela simples, não foi necessário um trabalho maior além da retirada de colunas não importantes.

Para os dados meteorológicos:

* Criamos um dataframe a partir das primeiras sete linhas de cada arquivo CSV. Essas linhas continham informações sobre a estação de medição, como a região geográfica, nome da estação, latitude, longitude, altitude, entre outras. Sendo essa nossa tabela dimensão;
* Transformamos o dataframe transpondo a tabela e resetando o index, para que cada estação de medição ficasse em uma única linha;
* Para deixar o dataframe mais organizado, removemos algumas colunas que não eram relevantes para a análise;
* Criamos outro dataframe a partir das linhas do arquivo CSV que continham informações meteorológicas. Para isso, a equipe pulou as primeiras oito linhas do arquivo, que contêm informações de cabeçalho;
* Removemos as linhas que continham dados faltantes em algumas das colunas relevantes;
* Selecionamos apenas as colunas que eram relevantes para a análise, como a data e hora do registro, precipitação, temperatura, umidade relativa do ar, entre outras.
* Adicionamos o dataframe resultante à lista de dados meteorológicos;
* Em ambas tabelas adicionamos um id;
* Em ambas as tabelas renomeamos os campos para termos colunas mais legíveis.

Já no caso dos dados meteorológicos, a equipe fez o processamento em duas etapas. Primeiro, foi necessário ler vários arquivos CSV que continham informações meteorológicas de diferentes estações de medição do INMET. Em seguida, a equipe selecionou apenas as colunas relevantes e salvou os dados em novos arquivos CSV. Esses dados precisaram de um pouco mais de trabalho, visto que as tabelas estavam separadas para cada estação metereológica em arquivos CSV diferentes. Além disso os arquivos efetivamente continham tanto a tabela dimensão quanto a tabela fato ao mesmo tempo.

Podemos resumir o processo feito nas duas tabelas da seguinte forma:

* A equipe coletou manualmente os dados através de base de dados espalhadas pela internet;
* Fizemos o tratamento dos dados de acordo com suas necessidades para apoiar o escopo da nossa pesquisa;
* Carregamos esses dados tratados em arquivos CSV filtrados e prontos para serem utilizados para pesquisa.