Aula 6 – Exercício de ETL

Prof. Jeovane Honório Alves

1 de abril de 2020

1 Exercício

Visto o impacto do novo coronavírus no mundo como uma pandemia, diversos dados relevantes foram extraídos e disponibilizados na Internet. Para o exercício de ETL, iremos trabalhar com dados disponibilizados pela universidade *Johns Hopkins*. Em específico, iremos trabalhar com os dados disponíveis no seguinte link: <https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports>.

1.1 A base de dados

O link acima estão disponibilizados arquivos no formato .csv relatando dados diários em diversos lugares do mundo. Para cada dia, o relato dos casos, mortes e recuperações é acumulativo, isto é, **não** se trata dos casos do dia mas sim da situação atual naquela região.

Podemos encontrar os seguintes dados nesses arquivos:

- Cidade;
- Estado;
- País;
- Última atualização (dia);
- Latitude e longitude;
- Total de casos (incluso mortes e recuperados);
- Mortes;
- Recuperações;
- Ativos.

Nem sempre teremos todos os dados completos para cada local, sendo necessário tomar decisões em como tratar esses dados. As próximas subseções tratam das atividades a serem desenvolvidas.

1.2 Desenvolvimento do *script* de transformação

Analise os dados disponíveis nos arquivos .csv e, caso necessário realize transformações neles e em suas estruturas. Pode ser feito em qualquer linguagem, contanto que o código, passando a pasta como parâmetro, faça a transformação de todos os arquivos.

Os arquivos .csv estão estruturados de uma forma que dados como a cidade se repetem (e podem não estar normalizados). Ao carregarmos os dados no modelo, dados do tipo precisam estar normalizados (não presentes em colunas sobre os casos, mas sim em uma tabela separada).

Como os dados são atualizados diariamente, teremos arquivos até a data de entrega (e posteriormente). É necessário que até essa data o código desenvolvido funcione.

1.3 Modelagem dimensional

Com base nesses dados, desenvolva o modelo dimensional. Defina qual é a tabela Fato e a granularidade do modelo. É necessário definir a dimensão do tempo e utilizar os dados disponíveis para montar as hierarquias do SnowFlake.

1.4 Desenvolvimento do *script* de carregamento

Agora será necessário realizar o carregamento dos dados no nosso modelo dimensional. Desenvolva o *script* SQL que monte o modelo dimensional no banco e carregue os dados tratados anteriormente no modelo.

1.5 Desenvolvimento de perguntas

Desenvolver pelo menos 10 perguntas que possam ser extraídas desses dados. Para cada pergunta, responder:

- Qual é a pergunta em si?
- Qual é a motivação por trás dessa pergunta?
- Qual foi o *script* desenvolvido para encontrar a resposta dessa pergunta?
- Que respostas foram encontradas?

Evitem trabalhar com perguntas muito parecidas!!! Por exemplo, "qual é o estado na China com mais casos?"e "qual é o estado no EUA com mais casos?".

1.6 Avaliação

Dupla

Método de entrega: Parciais pelo AVA (todos os arquivos desenvolvidos compactados)

Prazo (em andamento): 23h55 do dia 01/04/2020

Prazo (completo): 23h55 do dia 14/04/2020