DOCUMENTO:

Projeto: Engenheiro de dados SFIEC– Jammesson Cabral
OBSERVATÓRIO DA INDÚSTIRA

Respostas:

1. Após análise dos arquivos que estão disponível na base de dados da ANATAQ arquivos esses em formato .txt com dados pré-definidos ou seja arquivos estruturado, podemos guardar eles no mesmo formato coletado, na hora da equipe de cientista de dados forem fazer suas analises poderíamos tanto jogar em um data warehoue como também poderíamos mudar o formato do arquivo para um “parquet” ou “avro” e utilizar ferramentas de processamento paralelo para ganho de desempenho exemplo disso seria o spark, também podemos observar que esse o formato da extração desses dados seria em batch com tempo de 1mês, então não deve ter um grande nível complexidade para escalar pois podemos ter uma estimativa de quanto de dados estamos falando, então a resposta na minha visão seria, depende da situação, depende da equipe e depende da quantidade de dados que a equipe vai utilizar para fazer analise, logo eu deixaria os dados no seu formato bruto no datalake, faria as transformações necessárias e conversaria com a equipe e definiria o formato de como esses dados seriam apresentados.

2. Foi exportado conforme solicitado pela equipe os dados dos últimos 3 anos bem como apenas as duas tabelas “carga_fato” e “atracação_fato”.

3. Para a letra C foi solicitado criar uma (query) com informações de tempo médio bem como informações baseados a estados e regiões do brasil, existia duas colunas no arquivo baixado sem transformações contendo essas informações para se obter uma query otimizada seria interessante acrescentar essas duas colunas ‘**SGUF**’ ‘**Região Geográfica**’ no banco de dados SQL SERVER, porém foi feito a query tanto com as duas colunas no banco como sem as colunas, apenas para nível de teste de performance.