Projeto: Engenheiro de dados
OBSERVATÓRIO DA INDÚSTRIA

O que queremos receber:

Um repositório no Github com texto da resposta 1a) e scripts das demais com read documentando sua solução.

Além do repositório, queremos também receber por e-mail.

Auto avaliação:

Auto avalie suas habilidades nos requisitos de acordo com os níveis especificados usando o link abaixo:
https://forms.gle/dqbhRYKjENThgmWk7

1) Foi solicitado à equipe de AI+Analytics do Observatório da Indústria/FIEC, um projeto envolvendo os dados do Anuário Estatísticos da ANTAQ (Agência Nacional de Transportes Aquáticos).
O projeto consiste em uma análise pela equipe de cientistas de dados, bem como a disponibilização dos dados para o cliente que possui uma equipe de analistas própria que utiliza a ferramenta de BI (*business intelligence*) da Microsoft.

Para isto, o nosso *cientista de dados* tem que entender a forma de apresentação dos dados pela ANTAQ e assim, fazer o ETL dos dados e os disponibilizar no nosso data lake para ser consumido pelo time de cientistas de dados, como também, elaborar uma forma de entregar os dados tratados ao time de analistas do cliente da melhor forma possível.

Informações Importantes:
Link dos dados e dicionários (abrir no Internet Explore):
http://anuario.antaq.gov.br
Tutorial de acesso à fonte de dados: bit.ly/intrucaoantaq
Banco SQL da FIEC: SQL Server
Banco NoSQL da FIEC: Mongo DB
Ferramenta dos analistas do cliente: Power BI

Supondo que você seja nosso Engenheiro de dados:

a) Olhando para todos os dados disponíveis na fonte citada acima, em qual estrutura de banco de dados você orienta guardá-los no nosso Data Lake? SQL ou NoSQL? Discorra sobre sua orientação. (**1 pts**)

b) Nosso cliente estipulou que necessita de informações apenas sobre as atracações e cargas contidas nessas atracações dos últimos 3 anos (2019-2021). Logo, o time de cientistas de dados, em conjunto com você, analisaram e decidiram que duas tabelas, uma para atracação e outra para carga, seriam suficientes tanto para o trabalho do Observatório como para o trabalho do time externo.
Assim, desenvolva *scripts* em *python* que extraia os dados do anuário, e transforme-os em duas tabelas fato (atracacao_fato e carga_fato), com as respectivas colunas abaixo.
Como os dados têm um volume considerável e periodicidade mensal, os *scripts* automatizados e em pyspark ganham pontos extras. (
Scripts de extração em python: 1 pt
Scripts de extração em python com solução automatizada: 2,5 pts
Scripts de transformação em python: 1pt
Scripts de transformação em pyspark: 2,5pt)

Colunas da tabela atracacao_fato:

IDAtracacao	Tipo de Navegação da Atracação
CDTUP	Nacionalidade do Armador
IDBerco	FlagMCOperacaoAtracacao
Berço	Terminal
Porto Atracação	Município
Apelido Instalação Portuária	UF
Complexo Portuário	SGUF
Tipo da Autoridade Portuária	Região Geográfica
Data Atracação	Nº da Capitania

Data Chegada	Nº do IMO
Data Desatracação	TEsperaAtracacao
Data Início Operação	TEsperaInicioOp
Data Término Operação	TOperacao
Ano da data de início da operação	TEsperaDesatracacao
Mês da data de início da operação	TAtracado
Tipo de Operação	TEstadia

Colunas da tabela carga_fato: (atente-se que para o tipo de carga conteinerizada, pois cada contêiner pode ter mais de uma mercadoria)

IDCarga	FlagTransporteViaInterioir
IDAtracacao	Percurso Transporte em vias Interiores
Origem	Percurso Transporte Interiores
Destino	STNaturezaCarga
CDMercadoria (Para carga conteinerizada informar código das mercadorias dentro do contêiner.)	STSH2
Tipo Operação da Carga	STSH4
Carga Geral Acondicionamento	Natureza da Carga
ConteinerEstado	Sentido
Tipo Navegação	TEU
FlagAutorizacao	QTCarga
FlagCabotagem	VLPesoCargaBruta
FlagCabotagemMovimentacao	Ano da data de início da operação da atracação
FlagConteinerTamanho	Mês da data de início da operação da atracação

FlagLongoCurso	Porto Atracação
FlagMCOperacaoCarga	SGUF
FlagOffshore	Peso líquido da carga (Carga não conteinerizada = Peso bruto; Carga conteinerizada = Peso sem contêiner)

c) Essas duas tabelas ficaram guardadas no nosso Banco SQL SERVER. Nossos economistas gostaram tanto dos dados novos que querem escrever uma publicação sobre eles. Mais especificamente sobre o tempo de espera dos navios para atracar. Mas eles não sabem consultar o nosso banco e apenas usam o Excel. Nesse caso, pediram a você para criar uma consulta (query) otimizada em sql em que eles vão rodar no excel e por isso precisa ter o menor número de linhas possível para não travar o programa.
Eles querem uma tabela com dados do Ceará, Nordeste e Brasil contendo número de atracações, para cada localidade, bem como tempo de espera para atracar e tempo atracado por meses nos anos de 2020 e 20121. Segundo tabela abaixo: (2pts +1pt para a coluna bônus)

Localidade	Número de Atracações	Variação do número de atracação em relação ao mesmo mês do ano anterior - Bônus	Tempo de espera médio	Tempo atracado médio	Mês	Ano

Questão Bônus!

Finalmente, este processo deverá ser automatizado usando a ferramenta de orquestração de *workflow* Apache Airflow. Escreva uma DAG para a base ANTAQ levando em conta as características de uso da base. Esta também deve conter operadores para enviar avisos por email quando necessário (e.g.: caso os dados não sejam encontrados, quando o processo for finalizado, etc).

Todos os passos do processo ETL devem ser listados como tasks e orquestrados de forma otimizada, porém não é necessário implementar o código chamado em cada uma das *tasks*. Foque em mostrar o fluxo de *tasks* e as estruturas básicas de uma DAG. (3 pontos)