Projeto: CIENTISTA DE DADOS - FOCO EM ENGENHARIA DE DADOS
OBSERVATÓRIO DA INDÚSTRIA

Intro:
Foi solicitado à equipe de AI+Analytics do Observatório da Indústria/FIEC para que definissem o projeto arquitetural, assim como a realização da adaptação de alguns códigos do processo ETL de algumas bases, para nosso novo *data lake*.

O projeto será realizado em parceria com a equipe de cientistas de dados, tendo em mente suas necessidades de disponibilidade, bem como as necessidades de disponibilização dos dados para clientes que possuem equipe de analistas própria e utilize a ferramenta Microsoft Power BI ou que consumam dados usando APIs REST.

O que queremos receber:

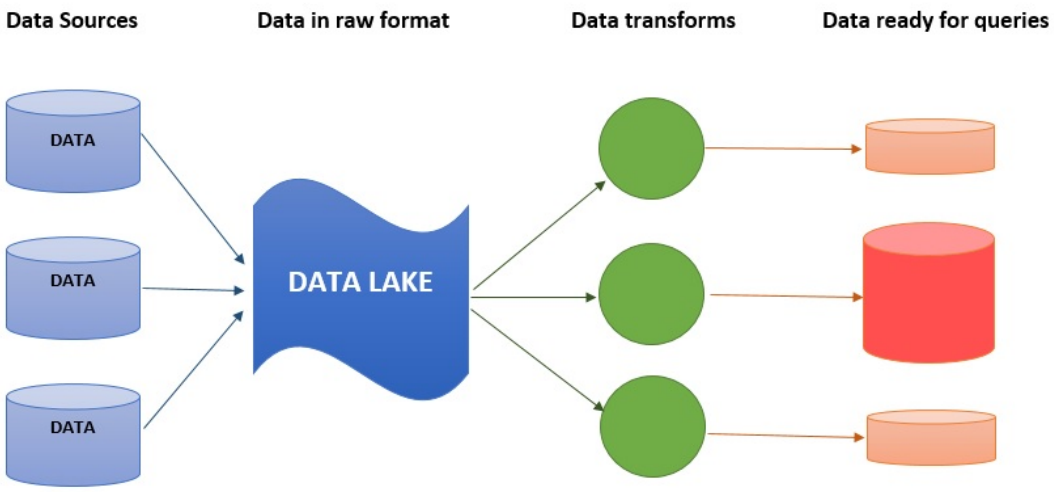
Prova:

- Arquivo .txt com resposta da questão 1.
- Um repositório no Github com scripts:
    1. Código da questão 2;
    2. Código da questão 3.

Além do repositório, queremos também receber os scripts finais 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

Questão 1 (3 pontos):
O nosso *data lake* é atualmente estruturado em um cluster em nuvem no Microsoft Azure. A sua estruturação base segue o modelo bastante comum representado na imagem a seguir:



Nosso engenheiro de dados será responsável por definir a forma que será realizada o processo ETL e uma parte importante deste processo inclui definir como os dados serão armazenados em cada parte do *pipeline*, levando em consideração as fontes e a forma de consumo que será feito destes dados.
Na tabela abaixo estão descritas 3 das nossas fontes de interesse e algumas de suas particularidades:

Base	**Extração**	**Quantidade de Tabelas**	**Periodicidade de Atualização**	Tabela com Dados Históricos	**Atualiza registros passados**	**Usos primários**
PIMPF	API	1	Mensal	Sim	Sim	BI
Receita Federal	Web crawling	2	Trimestral	Não	Sim	Machine learning, BI
ANTAQ	Web crawling	5	Mensal	Sim	Não	Disponibilização via API

Tendo em mente as diversas formas de armazenamento de dados como: bases de dados relacionais (SQL), não relacionais (NoSQL) e sistemas de arquivos; assim como as características particulares de cada conjunto de dados:

A. Como você armazenaria os dados na camada Raw após coletá-los de suas fontes? Justifique. (1 ponto)
B. Como você armazenaria os dados na camada de staging, onde estes serão transformados para a camada de consumo? Justifique. (1 ponto)
C. Como você armazenaria os dados na camada de consumo, levando em conta os usos primários de cada base? Justifique. (1 ponto)

Questão 2 (5 pontos):
Das bases apresentadas, nossa equipe de economistas atualmente possui um interesse especial pela ANTAQ. Eles solicitaram informações sobre as atracações e cargas contidas

nas atracações dos últimos 3 anos (2018 a 2020) para o estado do Ceará. No entanto, existe a necessidade de que algumas tabelas sejam unidas para que as informações de interesse sejam geradas.
Sendo assim, desenvolva scripts Python usando PySpark que extraia os dados do anuário e os transforme em duas tabelas fato: atracacao_fato e carga_fato.

Os dados se encontram no link: http://web.antaq.gov.br/Anuario/ (é possível que o acesso seja problemático fora do Internet Explorer).

Cada tabela deve conter as seguintes colunas:

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

carga_fato:

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 e Carga conteinerizada = Peso sem contêiner)

Atente-se para o tipo de carga conteinerizada, pois cada contêiner pode ter mais de uma mercadoria.

Soluções com download automático e transformações otimizadas contam como pontuação extra (1 ponto).

Questão 3 (2 pontos):
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.