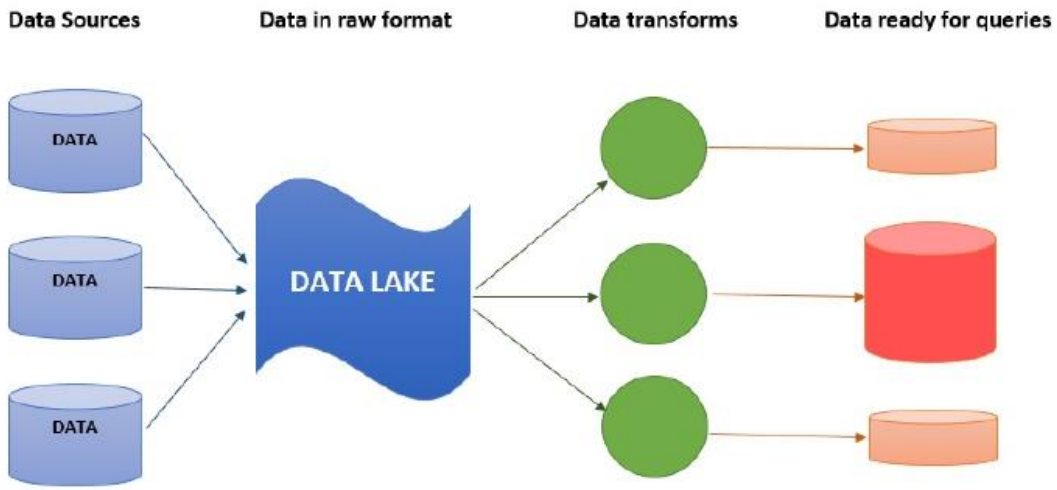
Questão 01



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)

A. Os dados seriam extraídos para o formato .csv, utilizando bibliotecas do Pyspark para a manipulação de arquivos na web em diversos formatos que inclui (JSON, Tuplas, RDF, XML, OWL, .txt, HTML. Será importante fazer uso de modelos incrementais, anexando os dados conforme a periodicidade (mensal, trimestral). Pode-se criar uma base de

repositório de arquivos brutos extraídos de diversas fontes, e colocando todos no formato .csv para disponibilização a camada de staging, criando um modelo de dados brutos. Será importante manter uma base de dados comum com alta disponibilidade (com poder de processamento e paralelismo) e com acesso controlado via políticas de segurança de domínios em rede. É importante seguir os modelo de DW e dataSmart, conforme o interesse da aplicação. É necessário uma solução altamente flexível, que acomode facilmente qualquer novo tipo de dado (não-estruturado e semi-estruturado) e que não seja corrompida por mudanças na estrutura de conteúdo. Por isso é necessário a camada de transformação.

B. Os dados na camada de transformação seriam persistidos em base de dados no formato de dataframe, ou em bases de dados intermediárias com dados tratados, pré-configurados no Spark, carregados em memória e persistidos por meio Tasks no modelo de RDDs do Spark. Já o QlikSense e QlikView utilizam o conceito de bases intermediárias chamadas de .QVD. A depender do modelo de tratamento e persistência dos dados transformados, é importante o mecanismos de tratamento conforme a aplicação de BI que deseja consumir. Sendo necessário tratar informações de datas, valores inconsistentes, valores esparsos, informações faltantes. Deve separar as informações de dimensões e medidas. Bem como as minhas fatos. Que são informações derivadas e calculadas sobre as dimensões de dados. Como existe modelos de demanda de informações d-1 (atém ntem) e em tempo real, pode-se padronizar o resuso de dados e políticas de treinamento da equipe para reutilizar os modelos de dados já transformados e que podem ser utilizados pelos usuários primários (BI, ML BI e API). Quando uma API de serviço web (web service) é escrita de forma aderente as definições REST (*Representational State Transfer*), torna-se conhecida como uma API RESTFull. Por uma API pode-se compartilhar dados em XML/RDF/OWL/JSON. Sendo que o modelo RESTFull mantem a API desacoplada dos detalhes internos da aplicação. Isto resulta em facilidade de escalabilidade e mantém as coisas simples. A interface uniforme garante que cada solicitação seja documentada. Na camada de transformação ajuda a tratar a camada de banco de dados, os dados relacionais são originalmente a escolha popular. Seu uso é cada vez mais problemático porque eles são uma tecnologia centralizada, cuja escalalidade é vertical ou invés de horizontal. Isso não os torna adequado para aplicações que requerem escalabilidade fácil e dinâmica. Por isso a necessidade de uma camada de transformação.

C. Na camada de consumo os dados seriam persistidos em bases de dados tratadas e transformadas. Pode-se utilizar algum serviço em núem, de acrodo com o nível de escalabilidade demandada. Os dados são estruturados em um modelo de dados em formato estrela de modo, que se forme uma tabela que podemos chamar de link, que interliga todas as chaves das tabelas intermediários e fatos, de modo que se consiga extrair informações convergentes e com alta disponibilidade, com painéis e dashboards. Pode criar uma base de dados no modelo tratado e fazer usos de ferramementas de BI como PowerBI, Cognos, Pentaho, QlikSense (Mais moderno) ou QlikView, Knime. Por exemplo o QlikSense e QlikView armazenam os dados da aplicação em uma base interna com o uso do Postgres. Pode-se fazer uso de uma base de dados em Núvem fazendo uso de base de machine leaning baseadas na AWS (Seage Maker) , Azure (Databriks) e Google Analisticas como seus modelo de persistência.

D.