Análise de Dados com Base em Processamento Massivo em Paralelo

Lista de Exercícios: Arquitetura de Data Warehousing

Profa. Dra. Cristina Dutra de Aguiar

Observação:

Esta lista contém exercícios classificados como essenciais e complementares. A indicação da classificação de cada exercício é feita junto de sua definição. A resposta de cada exercício encontra-se destacada na cor azul. Recomenda-se fortemente que a lista de exercícios seja respondida antes de se consultar as respostas dos exercícios.

1. (Essencial) Compare os processos de ETL e de ELT, descrevendo quais são as igualdades e diferenças existentes entre esses processos.

   Ambos os processos possuem as etapas de extração (E), transformação (T) e carregamento (L). Contudo, cada processo possui um momento específico para a execução de cada uma das etapas.

   ETL (*Extract*, *Transform*, *Load*): Os dados de interesse das fontes de dados são extraídos, transformados e carregados no *data warehouse* ou nos *data marts*. Ou seja, no processo de ETL, os dados são transformados antes de ocorrer a carga no *data warehouse* ou nos *data marts*.

   ELT (*Extract*, *Load*, *Transform*): Os dados de interesse das fontes de dados são extraídos e armazenados no *data lake* sem sofrer nenhuma transformação, ou sofrendo transformações mínimas apenas. Na sequência, os dados são extraídos do *data lake*, transformados e carregados no *data warehouse* ou nos *data marts*. Ou seja, as transformações acontecem depois que os dados são extraídos do *data lake* e antes de serem armazenados no *data warehouse* ou nos *data marts*.

2. (Essencial) Porque o *data warehouse* é considerado o principal componente do *data warehousing*?

   O *data warehouse* pode ser considerado o principal componente do *data warehousing* porque ele armazena os dados de interesse dos usuários de suporte à decisão (SSD), dados esses que podem oferecer suporte para análises importantes dentro do contexto do negócio. Esses dados são caracterizados por serem estruturados e por estarem organizados multidimensionalmente, de acordo com as diferentes perspectivas de análise dos usuários. Além disso, utilizando o *data warehouse* como base, as consultas analíticas (consultas OLAP) podem ser respondidas rapidamente.

3. (Essencial) Compare os conceitos de *Data Staging Area* e de *Data Lake*, descrevendo quais são as igualdades e diferenças existentes entre esses conceitos.

   Ambos conceitos se encontram na camada de pré-processamento dos dados da arquitetura de *data warehousing*. Além disso, ambos são locais de armazenamento de dados. Contudo, cada conceito possui características particulares, conforme detalhado a seguir.

   Data Staging Area: Contém dados extraídos das fontes de dados, ou de outra *data staging area*, que vão passando por sucessivas modificações até que estejam prontos e que possam ser carregados no *data warehouse*. Logo, os dados armazenados na *data staging area* vão sendo paulatinamente manipulados e processados, utilizando-se de um processo ETL. Após passar por todos os processamentos, os dados são carregados no *data warehouse*.

   Data Lake: Contém um grande volume de dados extraídos das fontes de dados em seu formato nativo (*raw data*). Logo, os dados armazenados nesse componente não sofreram nenhuma transformação, ou sofreram alterações mínimas, decorrentes do processo ELT. O *data lake* pode atuar como uma *data staging area* para carregar os dados no *data warehouse*. Além disso, o *data lake* pode se conectar diretamente com componentes da camada de ferramentas de análise e consulta. Isso significa que as consultas e análises podem ser realizadas diretamente sobre o *data lake*, sem a necessidade de se usar os dados do *data warehouse*.

4. (Essencial) Considere a seguinte “chuva de expressões”:

> “dados consolidados, organizados e estruturados”, “alta latência de disponibilidade”, “maior custo de análise”, “armazena arquivos TXT e JSON”, “armazena apenas dados tabulares”, “dados pré-processados antes de serem carregados”, “esquema em formato nativo (diferentes formatos)”, “baixa latência de disponibilidade”, “esquema estruturado (formato bem definido)”, “maior custo de geração dos dados”, “menor custo de geração dos dados”, “dados estruturados, semiestruturados e não estruturados”, “consultas OLAP”, “dados extraídos e carregados, sem sofrer transformações”, “ELT”, “menor custo de análise”, “ETL”, “tipos de consulta variados”

Preencha a tabela a seguir utilizando as expressões supracitadas:

Data Warehouse	**Data Lake**
. . .	. . .

Resposta

Data Warehouse	**Data Lake**
dados consolidados, organizados e estruturados	dados estruturados, semiestruturados e não estruturados
alta latência de disponibilidade	baixa latência de disponibilidade
menor custo de análise	maior custo de análise
dados pré-processados antes de serem carregados	dados extraídos e carregados, sem sofrer transformações
armazena apenas dados tabulares	armazena arquivos TXT e JSON
esquema estruturado (formato bem definido)	esquema em formato nativo (diferentes formatos)
maior custo de geração dos dados	menor custo de geração dos dados
consultas OLAP	tipos de consulta variados
ETL	ELT

5. (Essencial) Uma empresa líder de mercado deseja começar a realizar análises de *big data*. Segundo os gestores, esse tipo de análise permite identificar uma série de padrões a respeito dos clientes da empresa. Porém, para que as análises sejam fidedignas, os gestores especificaram que querem trabalhar com *petabytes* de dados coletados em pequenos intervalos de tempo. Além disso, o conjunto de dados a ser coletado deve englobar cliques dos clientes nas páginas da empresa, fotos e vídeos compartilhados nas redes sociais com a *hashtag* da empresa e textos nos *tweets* realizados com a *hashtag* da empresa, dentre outros. Por fim, os gestores desejam que esses dados sejam exibidos de forma clara e interativa para facilitar o processo de tomada de decisão estratégica.

   Considerando o contexto descrito, o gestor deve escolher para auxílio na tomada de decisão um *data warehouse* ou um *data lake*? Justifique a sua resposta usando como base os conceitos relacionados aos 7Vs.

   Devido ao gigantesco *volume* de dado gerado pelos clientes, pela *variedade* dos dados (como dados de cliques dos clientes e *tweets*) e pela alta *velocidade* exigida para que os dados sejam exibidos de forma interativa, um *data lake* é a melhor solução, pois os dados não precisam estar estruturados e possuem diferentes formatos.

6. (Essencial) Considere uma empresa que utiliza uma aplicação de *data warehousing* baseada no *pipeline* ilustrado na Figura 1. O volume de dados está crescendo e o *pipeline* deve se adequar a essa mudança. Para tanto, é necessário adicionar: (i) um motor de consulta para melhor explorar os dados armazenados no *data lake*; e (ii) um *data mart* para acelerar as consultas de um conjunto de dados do *data warehouse*. Em qual das opções abaixo os elementos (i) e (ii) devem ser encaixados para atender à demanda?

   (a) O motor de consulta deve ser adicionado em A e o *data mart* deve ser adicionado em B.

   (b) O motor de consulta deve ser adicionado em B e o *data mart* deve ser adicionado em A.

   (c) O motor de consulta e o *data mart* devem ser adicionados em A.

   (d) O motor de consulta e o *data mart* devem ser adicionados em B.

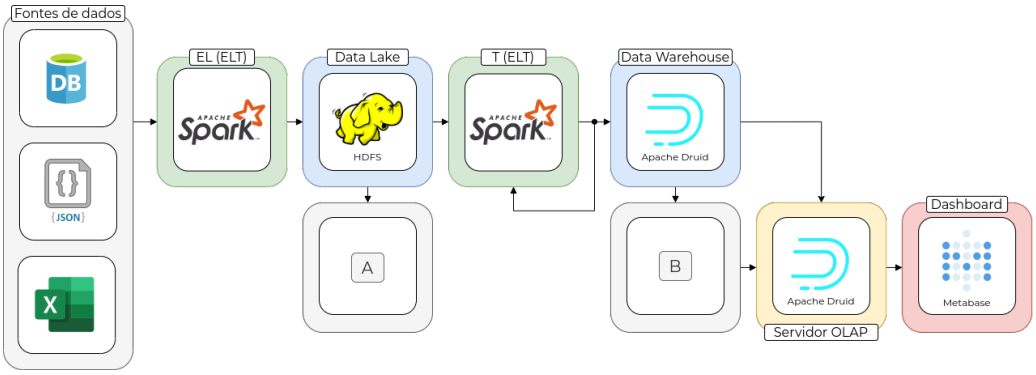

Figura 1: *Pipeline* de processamento de *big data* em lotes.

Opção (a). O motor de consulta deve se conectar ao *data lake* para facilitar a execução de consultas executadas contra os dados armazenados nessa área de armazenamento. O *data mart* deve se conectar ao *data warehouse*, desde que seus dados são um subconjunto dos dados armazenados no *data warehouse*.

7. (Complementar) Considere uma empresa que utiliza uma aplicação de *data warehousing* baseada no *pipeline* na nuvem ilustrado na Figura 2. Para reduzir os custos da arquitetura, decidiu-se substituir a solução proprietária e paga da ferramenta de construção de *dashboards* interativos Tableau por uma versão de *software* livre e gratuita. Faça a substituição solicitada escolhendo uma das propostas de solução sugeridas a seguir. Note que a solução deve ser compatível com as tecnologias ilustradas na Figura 2. Escolha apenas uma única proposta, mesmo que mais do que uma proposta possa ser adequada para a substituição.

   (a) Proposta 1. Substituir Tableau por Metabase. Detalhes sobre Metabase podem ser obtidos em *https://www.metabase.com/docs/latest/faq/setup/which-databases-does-metabase-support.html*.

   (b) Proposta 2. Substituir Tableau por Grafana. Detalhes sobre Grafana podem ser obtidos em *https://grafana.com/docs/grafana/latest/features/datasources/*.

   (c) Proposta 3. Substituir Tableau por Redash. Detalhes sobre Redash podem ser obtidos em *https://redash.io/help/data-sources/querying/supported-data-sources*.

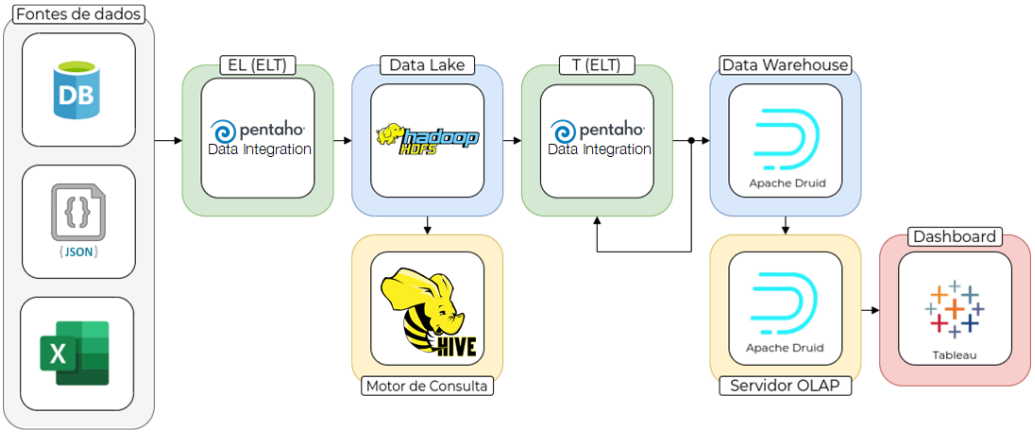

Figura 2: *Pipeline* de processamento de *big data*.

Todas as propostas são ferramentas de construção de *dashboards* interativos de *sofware* livre e gratuitas, sendo compatíveis com a tecnologia do *data warehouse* e do servidor OLAP Apache Druid. Portanto, pode-se escolher qualquer uma dessas opções. Portanto, não existe apenas uma resposta certa. Essa questão é livre para discussão durante as tutorias.