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. 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.
2. (Essencial) Porque o *data warehouse* é considerado o principal componente do *data warehousing*?
3. (Essencial) Compare os conceitos de *Data Staging Area* e de *Data Lake*, descrevendo quais são as igualdades e diferenças existentes entre esses conceitos.

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
. . .	. . .

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.

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.

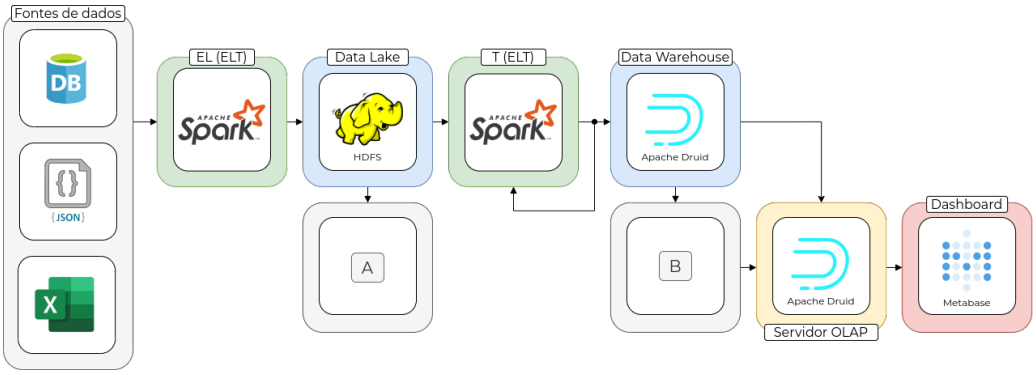

Figura 1: *Pipeline* de processamento de *big data* em lotes.

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*.

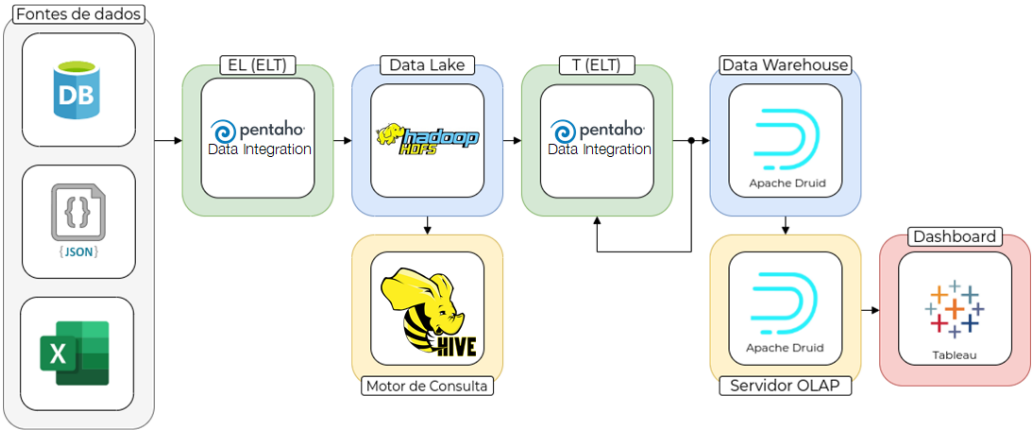

Figura 2: *Pipeline* de processamento de *big data*.