Análise de Dados com Base em Processamento Massivo em Paralelo

Arquitetura de Data Warehousing: Lista de Exercícios

Profa. Dra. Cristina Dutra de Aguiar Ciferri
André Perez
Guilherme Muzzi da Rocha
Jadson José Monteiro Oliveira
João Pedro de Carvalho Castro
Leonardo Mauro Pereira Moraes
Piero Lima Capelo

Observação:

Recomenda-se fortemente que a lista de exercícios seja respondida antes de se consultar as respostas dos exercícios.

1. Qual a diferença entre os processos de ETL e de ELT?
2. Por que o *data warehouse* é considerado o "coração" do *data warehousing*?
3. Considere uma empresa que possui os seguintes conjuntos de dados:
   (a) Tabelas de transformação, contendo dados de etapas intermediárias do processo de ETL.
   (b) Fotografias e vídeos de incêndios no território brasileiro, associados às respectivas localizações geográficas e em formato nativo.
   (c) *Dataset* integrado, histórico, não volátil e orientado a um assunto muito específico de um departamento da empresa.
   (d) *Dataset* integrado, histórico, não volátil e orientado a assunto, englobando todo o escopo de atuação da empresa.
   (e) Dados de documentos JSON extraídos de páginas web, sem sofrer transformações.
   (f) Dados estruturados em uma grande quantidade de tabelas normalizadas.

Associe cada conjunto de dados aos componentes da arquitetura de *data warehousing* listados a seguir. É permitido associar mais de um conjunto de dados a um mesmo componente. Também é permitido não associar nenhum conjunto de dados a um determinado componente.

- Repositório de Metadados:
- *Data Lake*:
- *Data Warehouse*:
- Bancos de Dados Operacionais:
- *Data Staging Area*:
- *Data Marts*:

4. Considere a seguinte "chuva de expressões":

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

Preencha a tabela a seguir utilizando as expressões supracitadas:

Data Warehouse	Data Lake

5. Uma empresa líder de mercado deseja começar a realizar análises de *big data*. De acordo com 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, 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, indique quais conceitos do modelo de 7Vs são imprescindíveis para garantir a satisfação dos gestores da empresa em seu processo de análise de *big data*.

6. Considere uma empresa que utiliza uma aplicação de *data warehousing* baseada no *pipeline* na nuvem ilustrado na Figura 1. 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 a tecnologia de DW e servidor OLAP Amazon (AWS) Redshift. Escolha apenas 1 única proposta, mesmo que mais do que uma proposta possa ser utilizada.

   (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*.

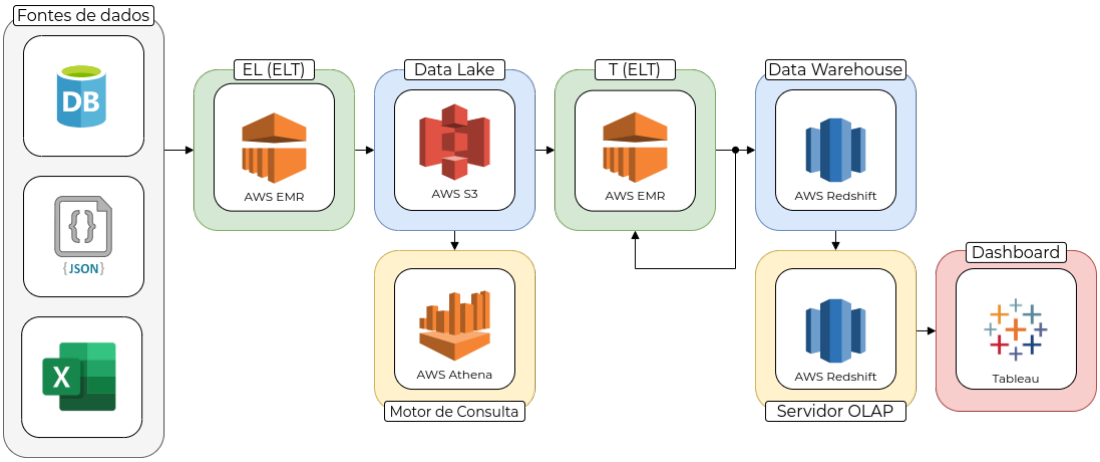

Figura 1: *Pipeline* de processamento de *big data* em lotes na nuvem.

7. Considere uma empresa que utiliza uma aplicação de *data warehousing* baseada no *pipeline* ilustrado na Figura 2. 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 DW. Em qual das opções abaixo as 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.

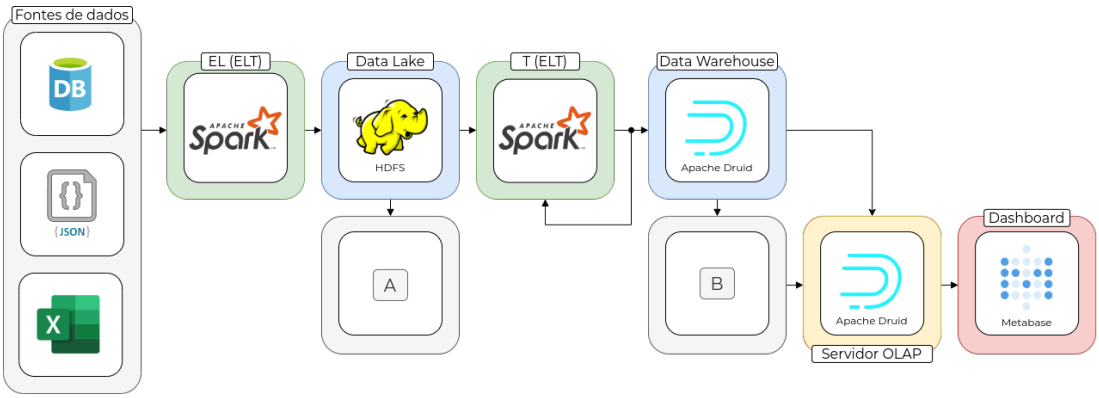

Figura 2: *Pipeline* de processamento de *big data* em lotes.