Análise de Dados com Base em Processamento Massivo em Paralelo

Lista de Exercícios: Modelagem Conceitual de ETL/ELT

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) Descreva qual a importância de se modelar conceitualmente um *workflow* de ETL/ELT antes de implementá-lo.

   A anuência de que o processo de ETL/ELT é a etapa mais custosa de todo o projeto de *data warehousing* já é algo consolidado tanto na literatura quanto no mercado. Sendo assim, é importante modelá-lo conceitualmente a fim de contribuir para diminuir o esforço dos projetistas e desenvolvedores durante a implementação do *workflow*. Além disso, o esquema conceitual é um recurso precioso para a documentação das decisões tomadas na construção do processo de ETL/ELT, para a análise de impacto das alterações necessárias para atendimento de demandas que ocorrem no ciclo de vida do *data warehouse* (tais como alterações nas fontes de dados, evolução dos requisitos ou das regras de negócio, necessidade de melhoria no desempenho das consultas, correção de erros cometidos durante a fase de projeto, entre outros) e para facilitar a exploração de cenários alternativos para a solução desejada.

2. (Essencial) Por que é interessante o uso de um modelo específico para projetar o processo de ETL/ELT (como o Modelo Intuitive) e não o uso de um modelo de processos genérico (como o Modelo BPMN - *Business Process Model and Notation*)?

   O uso de um modelo específico para projetar o processo de ETL/ELT justifica-se pelo fato desse modelo prover clareza na representação do processo. O modelo é baseado em operadores que representam graficamente as operações usualmente presentes em processos de ETL/ELT. Como resultado, o projeto final provê melhor compreensão por parte do usuário final e não evidencia detalhes de implementação. Modelos genéricos também podem ser utilizados. Entretanto, esses modelos definem operadores mais genéricos, os quais não descrevem de forma gráfica e visual as operações usualmente presentes em processos de ETL/ELT.

3. (Essencial) Considere as seguintes categorias de operadores:

   (a) Operadores de armazenamento;

   (b) Operadores de manipulação de dados;

   (c) Operadores de inicialização;

   (d) Operadores de agregação;

   (e) Operadores de fluxo;

   (f) Operadores especiais.

   Descreva, de forma sucinta, o objetivo de cada uma das categorias supracitadas.

   Operadores de armazenamento: Os operadores de armazenamento podem ser usados para representar áreas de armazenamento de dados, tais como repositórios, arquivos ou bases de dados.

   Operadores de manipulação de dados: Os operadores de manipulação de dados são usados para representar as tarefas de transformação e de limpeza que são aplicadas aos dados extraídos das diversas fontes para torná-los compatíveis com a estrutura proposta para o *data warehouse*.

   Operadores de inicialização: Os operadores de inicialização de dados servem para representar a atribuição de valores específicos para atributos de um conjunto de dados.

   Operadores de agregação: Os operadores de agregação podem ser usados para representar funções que, quando aplicadas a um conjunto de dados, processam os valores de um atributo específico e retornam um único valor como resultado. Se forem definidos atributos de agrupamento, as funções processam os valores de um atributo e retornam um único valor para cada atributo de agrupamento.

   Operadores de fluxo: Os operadores de fluxo de dados permitem representar uma alteração no fluxo dos dados no *workflow* de ETL, sem impactar esses dados.

   Operadores especiais: Os operadores especiais representam operações que envolvem especificidades, complementando as funcionalidades dos demais operadores.

4. (Essencial) Considere o *workflow* de ETL modelado na Figura 1, o qual ilustra a extração de funcionários de duas bases de origem: (i) funcionarioRelacional, a qual representa um sistema gerenciador de banco de dados relacional; e (ii) colaboradorJSON, a qual representa uma coleção de documentos JSON.

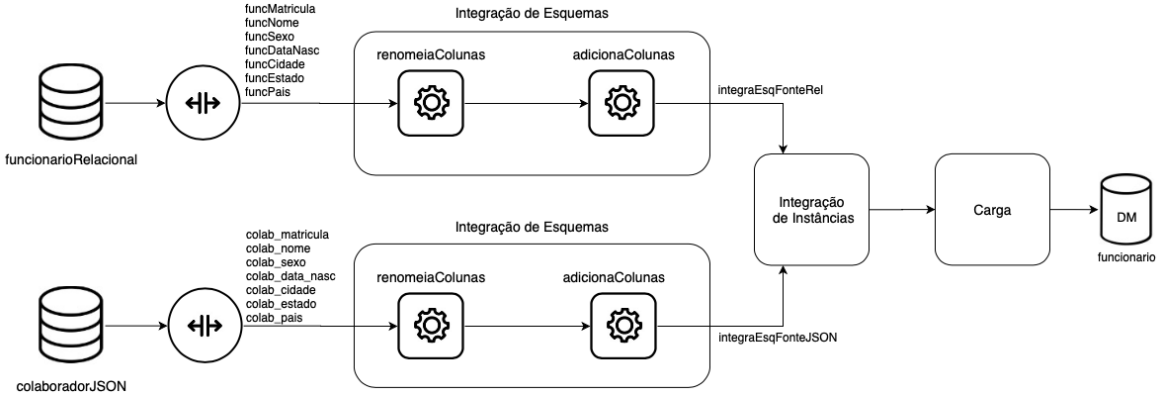

Figura 1: Visão geral do processo de ETL da **BI Solutions**.

Considere que a empresa **BI Solutions**, responsável pela manutenção do *workflow* ilustrado na Figura 1, necessita incluir mais uma fonte de dados no processo de ETL. Essa nova fonte de dados, denominada empregadoPlanilha, representa uma planilha Excel que contém os seguintes dados de funcionários: "Matrícula do Empregado", "Nome do Empregado", "Sexo do Empregado", "Data de Nascimento", "Cidade de Residência", "Estado de Residência".

Estenda o *workflow* de ETL para incluir essa nova fonte de dados. Modele apenas as etapas anteriores ao subfluxo de integração de instâncias.

Resposta:

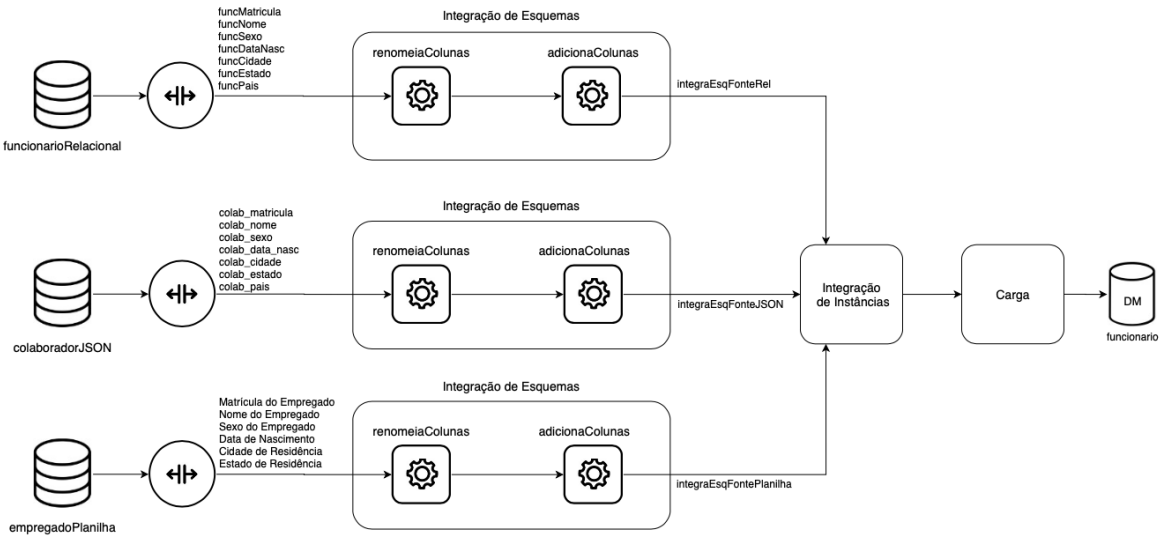

Figura 2: Resposta da questão 4.

5. (Essencial) Considere o subfluxo relacionado à "Integração de Instâncias" da **BI Solutions**, representado tanto no diagrama conceitual da Figura 3 quanto no *workflow* da Figura 4. Esse subfluxo considera como entradas dados oriundos das fontes de dados funcionarioRelacional e colaboradorJSON. Estenda o diagrama conceitual e o workflow de forma que o subfluxo relacionado à Integração de Instâncias também considere como entrada os dados oriundos da fonte de dados empregadoPlanilha.

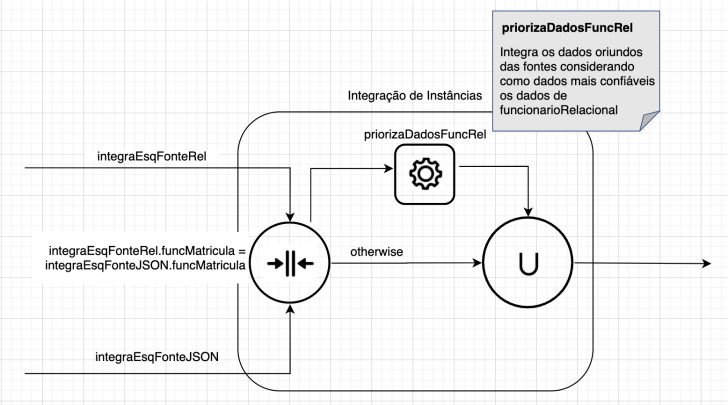

Figura 3: Diagrama conceitual para o subfluxo de "Integração de Instâncias" da **BI Solutions**.

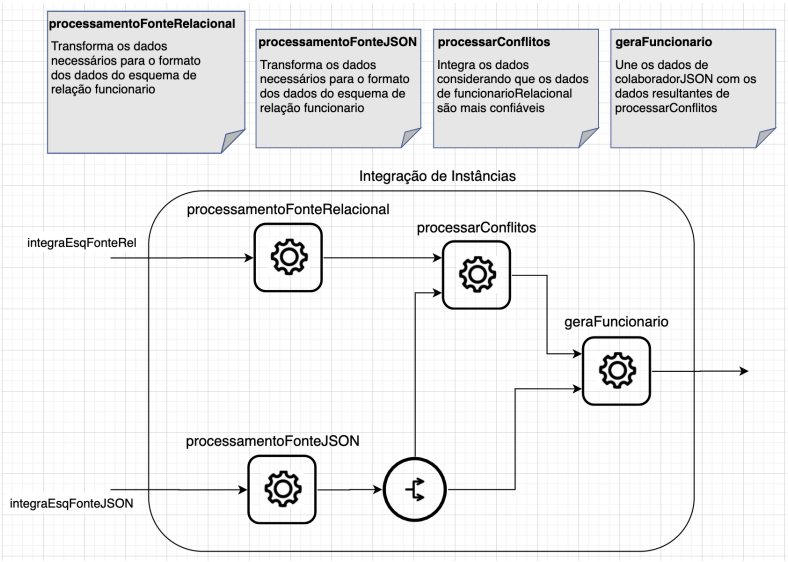

Figura 4: *Workflow* para o subfluxo de "Integração de Instâncias" da **BI Solutions**.

Resposta:

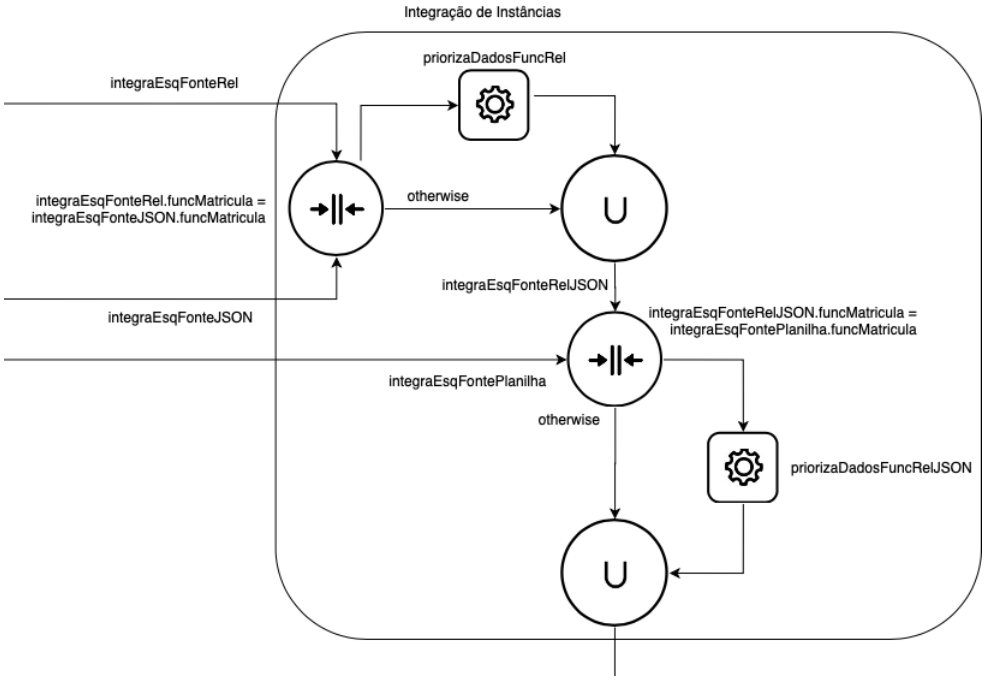

Figura 5: Resposta da primeira parte da questão 5.

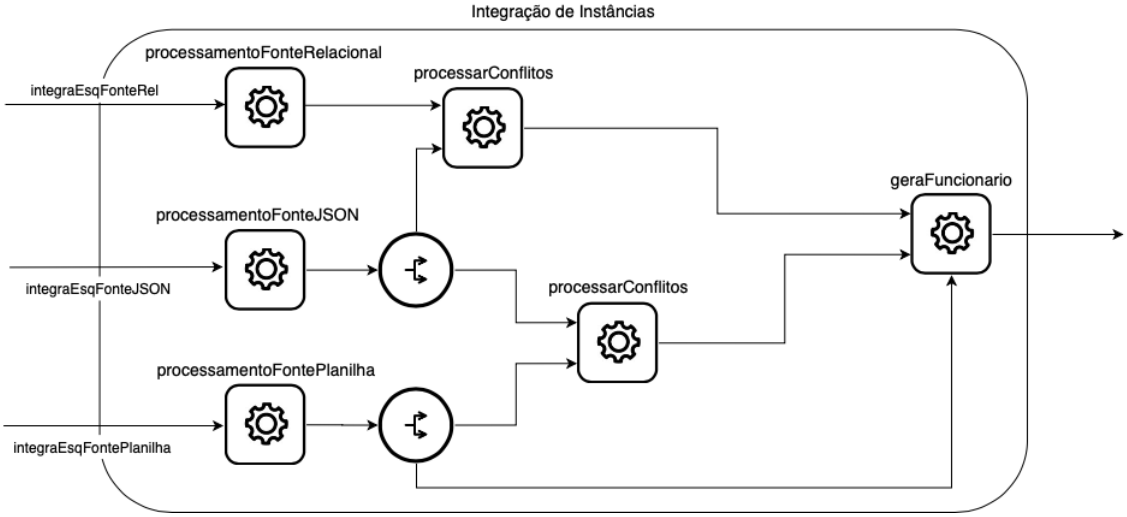

Figura 6: Resposta da segunda parte da questão 5.

6. (Complementar) Considere o exemplo do *data mart* apresentado nas aulas da disciplina, referente à folha de pagamento da empresa **BI Solutions**. Nele, são considerados dados de funcionários, datas, cargos e departamentos. Escolha uma dessas perspectivas e desenvolva um modelo completo para seu *workflow* de ETL (exceto a perspectiva de funcionários, visto que esta já foi modelada nas aulas). Considere diferentes fontes de dados e englobe as etapas de integração de esquemas, integração de instâncias e carga em seu modelo.

   Questão elaborada para gerar discussões nas tutorias. Não há uma única resposta correta.