

Documentação do Plano de Carga das Dimensões e Tabela Fato (Apache Hop)

Ferramentas Utilizadas

Para a construção do pipeline de carga das dimensões, foi utilizado o Apache Hop, empregando as etapas (steps) *Microsoft Excel Writer*, *Select Values*, e *Insert/Update*. A seguir, detalhamos cada uma dessas etapas e seu papel no processo.

Passo 1: Microsoft Excel Writer

Objetivo: Escrever os dados das dimensões provenientes de fontes externas em arquivos Excel, facilitando a visualização e verificação dos dados antes de carregá-los no banco de dados.

Configuração:

- Configurar o caminho e nome do arquivo Excel.
- Definir as colunas e formatos apropriados.
- Garantir que os dados sejam exportados corretamente e que os arquivos Excel contenham todas as colunas necessárias para as dimensões.

Etapas:

1. **Definir o caminho do arquivo Excel:** Especificar onde os arquivos serão salvos.
2. **Mapear as colunas:** Certificar-se de que todas as colunas necessárias estejam presentes no arquivo.
3. **Executar e verificar:** Executar o writer e revisar os arquivos gerados para garantir a integridade dos dados.

Passo 2: Select Values

Objetivo: Selecionar e, se necessário, transformar as colunas dos dados antes de inseri-los nas tabelas de dimensão.

Configuração:

- Selecionar as colunas necessárias dos arquivos Excel.
- Realizar quaisquer transformações de dados necessárias, como mudanças de formato ou limpeza de dados.

Etapas:

1. **Selecionar as colunas:** Escolher as colunas relevantes dos dados exportados.
2. **Transformar os dados:** Realizar qualquer transformação necessária, como mudança de tipos de dados ou limpeza de valores nulos.
3. **Verificar os dados:** Garantir que os dados estejam prontos para inserção nas tabelas de dimensão.

Passo 3: Insert/Update

Objetivo: Inserir novos registros ou atualizar os registros existentes nas tabelas de dimensão no banco de dados PostgreSQL.

Configuração:

- Conectar ao banco de dados PostgreSQL.
- Configurar as chaves primárias e campos de atualização.
- Definir as condições de inserção e atualização.

Etapas:

1. **Conectar ao banco de dados:** Estabelecer a conexão com o PostgreSQL.
2. **Configurar as chaves:** Definir quais colunas serão usadas como chaves para identificar registros únicos.
3. **Definir operações de inserção/atualização:** Especificar como os novos registros serão inseridos e os registros existentes serão atualizados.
4. **Executar e verificar:** Executar o processo de inserção/atualização e revisar os registros no banco de dados para garantir a precisão e integridade dos dados.

Configuração para Cada Tabela de Dimensão

Tabela dim_rota:

- **Microsoft Excel Writer:** Exportar dados de rotas para um arquivo Excel.
- **Select Values:** Selecionar e transformar colunas como rota.
- **Insert/Update:** Inserir ou atualizar registros na tabela dim_rota com base na coluna rota_id.

Tabela dim_cliente:

- **Microsoft Excel Writer:** Exportar dados de clientes para um arquivo Excel.
- **Select Values:** Selecionar e transformar colunas como codigo, status, nome_fantasia, razao_social, segmento.
- **Insert/Update:** Inserir ou atualizar registros na tabela dim_cliente com base na coluna cliente_id.

Tabela dim_funcionario:

- **Microsoft Excel Writer:** Exportar dados de funcionários para um arquivo Excel.
- **Select Values:** Selecionar e transformar colunas como funcionario.
- **Insert/Update:** Inserir ou atualizar registros na tabela dim_funcionario com base na coluna funcionario_id.

Tabela dim_localidade:

- **Microsoft Excel Writer:** Exportar dados de localidades para um arquivo Excel.
- **Select Values:** Selecionar e transformar colunas como bairro, municipio, estado.

- **Insert/Update:** Inserir ou atualizar registros na tabela dim_localidade com base na coluna localidade_id.

Tabela dim_tempo:

- **Microsoft Excel Writer:** Exportar dados de tempo para um arquivo Excel.
- **Select Values:** Selecionar e transformar colunas como mes e ano.
- **Insert/Update:** Inserir ou atualizar registros na tabela dim_tempo com base na coluna tempo_id.

Tabela fato_faturamento

Passo 1: Microsoft Excel Writer

Objetivo: Exportar os dados da tabela fato para um arquivo Excel a partir de fontes externas.

Configuração:

- Definir o caminho e nome do arquivo Excel.
- Especificar as colunas e formatos necessários para exportação.

Etapas:

1. **Definir o caminho do arquivo Excel:** Especificar onde o arquivo será salvo.
2. **Mapear as colunas:** Garantir que todas as colunas necessárias estejam presentes no arquivo.
3. **Executar o writer:** Exportar os dados da tabela fato para o arquivo Excel.

Passo 2: Database Lookup para Todas as Dimensões

Objetivo: Consultar as tabelas de dimensão para obter as chaves correspondentes aos dados da tabela fato.

Configuração:

- Conectar ao banco de dados PostgreSQL.
- Configurar as consultas SQL para cada dimensão.

Etapas:

1. **Conectar ao banco de dados:** Estabelecer a conexão com o PostgreSQL.
2. **Configurar as consultas SQL:** Escrever consultas SQL para cada dimensão a fim de obter as chaves primárias correspondentes aos dados da tabela fato.
3. **Executar as consultas:** Realizar as consultas SQL para obter as chaves das dimensões.

Passo 3: Select Values para Selecionar Chaves Primárias e Métricas

Objetivo: Selecionar as chaves primárias das dimensões e as métricas necessárias para inserção/atualização na tabela fato.

Configuração:

- Mapear as colunas das chaves primárias e métricas.
- Definir quais colunas serão utilizadas no processo de carga.

Etapas:

1. **Mapear as colunas das chaves e métricas:** Selecionar e transformar as colunas necessárias para a carga na tabela fato.
2. **Executar o Select Values:** Realizar a seleção das colunas conforme configurado.

Passo 4: Insert/Update para Carga da Tabela Fato

Objetivo: Inserir novos registros ou atualizar os registros existentes na tabela fato com as chaves e métricas selecionadas.

Configuração:

- Conectar ao banco de dados PostgreSQL.
- Configurar as operações de inserção e atualização na tabela fato.

Etapas:

1. **Conectar ao banco de dados:** Estabelecer a conexão com o PostgreSQL.
2. **Definir as operações de Insert/Update:** Especificar como os novos registros serão inseridos e os registros existentes serão atualizados na tabela fato.
3. **Executar a carga:** Inserir ou atualizar os registros na tabela fato com base nas chaves e métricas selecionadas.
- 4.

Configuração da Tabela Fato fato_faturamento

Microsoft Excel Writer

1. **Nome do Arquivo Excel:** fato_faturamento.xlsx
2. **Caminho do Arquivo:** C:\Caminho\Para\Arquivos\fato_faturamento.xlsx
3. **Mapeamento das Colunas:**
 - rota_id
 - cliente_id
 - funcionario_id
 - localidade_id
 - tempo_id
 - fat_mai
 - fat_abr
 - fat_mar

- fat_fev
- fat_jan
- fat_dez

Database Lookup para Todas as Dimensões

1. Conexão ao Banco de Dados PostgreSQL:

- Host: localhost
- Porta: 5432
- Banco de Dados: nome_do_banco
- Usuário: usuario
- Senha: senha

Select Values para Selecionar Chaves Primárias e Métricas

1. Mapeamento das Colunas Seleccionadas:

- Chaves Primárias:
 - rota_id
 - cliente_id
 - funcionario_id
 - localidade_id
 - tempo_id
- Métricas:
 - fat_mai
 - fat_abr
 - fat_mar
 - fat_fev
 - fat_jan
 - fat_dez

Insert/Update para Carga da Tabela Fato

1. Conexão ao Banco de Dados PostgreSQL:

- Mesmas configurações que no Database Lookup

2. Operações de Insert/Update:

- Inserir novos registros se não existirem chaves primárias duplicadas.
- Atualizar os registros existentes com as métricas atualizadas.

3. Chaves para Inserção/Atualização:

- Chaves Primárias:
 - rota_id
 - cliente_id
 - funcionario_id
 - localidade_id
 - tempo_id

Considerações Finais

Ao longo deste processo de construção do pipeline de ETL (Extract, Transform, Load) para as dimensões e tabela fato em um ambiente PostgreSQL utilizando o Apache Hop, foram adotadas várias etapas e configurações para garantir a integridade, consistência e eficiência na manipulação e carga dos dados. Aqui estão algumas considerações finais:

1. **Integração de Fontes de Dados Externas:**
 - Utilização do Microsoft Excel Writer para exportar dados de fontes externas para arquivos Excel, permitindo uma visualização e verificação dos dados antes da carga no banco de dados.
2. **Consulta e Mapeamento de Dimensões:**
 - O uso do Database Lookup para consultar todas as dimensões e obter as chaves correspondentes aos dados da tabela fato, garantindo a consistência e integridade das relações dimensionais.
3. **Seleção de Chaves Primárias e Métricas:**
 - A etapa Select Values foi empregada para selecionar as chaves primárias das dimensões e as métricas relevantes para a carga na tabela fato, facilitando o processo de transformação e seleção dos dados.
4. **Carga na Tabela Fato:**
 - A utilização do Insert/Update na carga da tabela fato assegura que novos registros sejam inseridos adequadamente e registros existentes sejam atualizados com as métricas corretas, mantendo a consistência dos dados.
5. **Configuração do Ambiente PostgreSQL:**
 - Todas as operações de consulta, inserção e atualização foram configuradas de acordo com as especificações do banco de dados PostgreSQL, garantindo a compatibilidade e eficiência no processo de carga.
6. **Validação e Verificação:**
 - Durante todo o processo, foram realizadas validações e verificações para garantir a integridade e precisão dos dados, desde a exportação até a carga final no banco de dados.
7. **Documentação e Rastreabilidade:**
 - A documentação detalhada de todo o plano de carga, incluindo as configurações, consultas SQL, mapeamentos e operações, proporciona uma maior rastreabilidade e entendimento do processo de ETL.
8. **Melhoria Contínua:**
 - Este pipeline de ETL pode ser continuamente aprimorado e otimizado para atender a novas necessidades e exigências do ambiente de dados, garantindo um fluxo de dados eficiente e confiável.

Em suma, a combinação do Apache Hop com suas diversas etapas e ferramentas, juntamente com a configuração adequada das operações no PostgreSQL, resulta em um processo robusto e confiável para a manipulação e carga de dados em um ambiente de Data Warehouse.