Data Mart Implementation (P01)

DECISION SUPPORT SYSTEMS, 2022-23

**João Apresentação (21152), Gonçalo Cunha (21145), Pedro Simões (21140)**

Índice

[1. Introdução 4](#_Toc132052260)

[1.1 Visão 4](#_Toc132052261)

[1.2 Objetivos 4](#_Toc132052262)

[1.3 Descrição de Processos de Negócio 5](#_Toc132052263)

[2. Recurso de dados 6](#_Toc132052264)

[2.1 Data Profiling 7](#_Toc132052265)

[3. Modelação Dimensional 10](#_Toc132052266)

[3.1 Design do modelo de dados dimensional 11](#_Toc132052267)

[4. Implementação da DataMart (ETL) 15](#_Toc132052268)

[4.1 Jobs 15](#_Toc132052269)

[4.2 Transformations 17](#_Toc132052270)

[5. Conclusão 20](#_Toc132052271)

**Índice de imagens**

[Figura 1 - WWI Modelo BaseDados 6](file:///C:\Universidade\3º%20Ano%20-%20Semestre%202\SAD\ETL\Relatório\DSS_G07_P01.docx#_Toc132043312)

[Figura 2 – Job\_dims\_ft 12](file:///C:\Universidade\3º%20Ano%20-%20Semestre%202\SAD\ETL\Relatório\DSS_G07_P01.docx#_Toc132043313)

[Figura 3 - Job\_dims 12](file:///C:\Universidade\3º%20Ano%20-%20Semestre%202\SAD\ETL\Relatório\DSS_G07_P01.docx#_Toc132043314)

[Figura 4 - Transformation Dim\_Costumers 13](file:///C:\Universidade\3º%20Ano%20-%20Semestre%202\SAD\ETL\Relatório\DSS_G07_P01.docx#_Toc132043315)

[Figura 5 -Transformation Dim\_CostumerTransactions 14](file:///C:\Universidade\3º%20Ano%20-%20Semestre%202\SAD\ETL\Relatório\DSS_G07_P01.docx#_Toc132043316)

[Figura 6 - Transformation Dim\_StockItems 15](file:///C:\Universidade\3º%20Ano%20-%20Semestre%202\SAD\ETL\Relatório\DSS_G07_P01.docx#_Toc132043317)

[Figura 7 - Transformation Ft\_invoiceLines 16](file:///C:\Universidade\3º%20Ano%20-%20Semestre%202\SAD\ETL\Relatório\DSS_G07_P01.docx#_Toc132043318)

[Figura 8 - ft\_invoiceLines get\_max 16](file:///C:\Universidade\3º%20Ano%20-%20Semestre%202\SAD\ETL\Relatório\DSS_G07_P01.docx#_Toc132043319)

[Figura 9 - ft\_invoiceLines input 16](file:///C:\Universidade\3º%20Ano%20-%20Semestre%202\SAD\ETL\Relatório\DSS_G07_P01.docx#_Toc132043320)

[Figura 10 - ft\_invoicelines CostumerLookup 17](file:///C:\Universidade\3º%20Ano%20-%20Semestre%202\SAD\ETL\Relatório\DSS_G07_P01.docx#_Toc132043321)

[Figura 11 - ft\_invoicelines StockItemsLookup 17](file:///C:\Universidade\3º%20Ano%20-%20Semestre%202\SAD\ETL\Relatório\DSS_G07_P01.docx#_Toc132043322)

[Figura 12 - ft\_invoicelines SelectValues 18](file:///C:\Universidade\3º%20Ano%20-%20Semestre%202\SAD\ETL\Relatório\DSS_G07_P01.docx#_Toc132043323)

[Figura 13 - ft\_invoicelines output 18](file:///C:\Universidade\3º%20Ano%20-%20Semestre%202\SAD\ETL\Relatório\DSS_G07_P01.docx#_Toc132043324)

[Figura 14 - ft\_orderlines get\_max 19](file:///C:\Universidade\3º%20Ano%20-%20Semestre%202\SAD\ETL\Relatório\DSS_G07_P01.docx#_Toc132043325)

[Figura 15 - ft\_orderlines input 19](file:///C:\Universidade\3º%20Ano%20-%20Semestre%202\SAD\ETL\Relatório\DSS_G07_P01.docx#_Toc132043326)

[Figura 16 - ft\_orderlines CostumersLookup 20](file:///C:\Universidade\3º%20Ano%20-%20Semestre%202\SAD\ETL\Relatório\DSS_G07_P01.docx#_Toc132043327)

[Figura 17 - ft\_orderlines StockLookup 20](file:///C:\Universidade\3º%20Ano%20-%20Semestre%202\SAD\ETL\Relatório\DSS_G07_P01.docx#_Toc132043328)

[Figura 19 - ft\_orderlines SelectValues 21](file:///C:\Universidade\3º%20Ano%20-%20Semestre%202\SAD\ETL\Relatório\DSS_G07_P01.docx#_Toc132043329)

# Introdução

Este projeto de uma forma resumida terá a implementação de um Data Mart com recurso a uma base de dados que armazena atividades de uma empresa de importação e distribuição de produtos, no caso em estudo, Wide World Importers (WWI).

A empresa WWI é uma empresa de importação e distribuição de produtos, que atende principalmente a empresas que revendem para consumidores individuais (público geral). Os clientes da WWI incluem lojas de especialidades, supermercados, lojas de informática, lojas de atrações turísticas e alguns indivíduos.

O projeto passará pela análise de requisitos de negócios, montagem de dimensões e tabelas de factos e implementação de processos ETL.

Nesta introdução, será falada a visão geral do projeto, os objetivos e uma descrição dos processos de negócios envolvidos.

## Visão

O objetivo principal deste projeto é criar um Data Mart eficiente e preciso que possa ser usado para análises de negócios na empresa WWI.

O Data Mart será projetado para fornecer as respostas necessárias e relevantes para os utilizadores, permitindo que estes tomem decisões com base nessas informações.

## Objetivos

Os objetivos deste projeto passam por:

* Realizar a análise de requisitos de negócios para entender as necessidades da empresa WWI;
* Realizar a análise de perfil de dados para entender a qualidade e a integridade dos dados disponíveis;
* Projetar e implementar processos de extração, transformação e carregamento dos dados para a Data Mart;
* Projetar e implementar um modelo dimensional que atenda aos requisitos de negócios da WWI;
* Executar os processos ETL para garantir a integridade dos dados no Data Mart.

## Descrição de Processos de Negócio

Os processos de negócios da WWI inicia-se pelo pedido de produtos, da parte do cliente. Se a WWI não tiver stock suficiente, eles encomendam o stock adicional dos fornecedores. Se os clientes não quiserem esperar pelos produtos que não estão disponíveis, os produtos são enviados posteriormente em um envio separado.

A WWI rentabiliza através dos produtos em stock, convertendo o pedido em fatura. Quando os clientes fazem pedidos de produtos que não estão em stock, esses produtos são colocados em backorder. A WWI entrega os produtos em stock aos clientes, seja por meio de seus próprios veículos de entrega ou por meio de outros correios ou métodos de frete.

Os clientes pagam as faturas à WWI. As faturas e pagamentos são registados na tabela de transações de clientes.

# Uma imagem com diagrama Descrição gerada automaticamenteRecurso de dados

Figura 1 - WWI Modelo BaseDados

Table 1: Summary of important WWI database contents

|  |  |  |
| --- | --- | --- |
| **Event / object** | **Table** | **Nr. Records** |
| Clientes | *Customers* | 663 |
| Encomendas feitas | *Orders* | 73595 |
| Registo de Faturas | *Invoices* | 70510 |
| Transações de dinheiro entre WWI e Costumers | *Customertransactions* | 97147 |
| Inventário de produtos | *Stockitems* | 227 |
| Lista de descontos | *Specialdeals* | 2 |

## Data Profiling

De seguida será feita a análise às tabelas principais da base de dados WWI, fornecida pelo docente.

Para isto foi utilizado o software Open Source Data Quality and Profiling, que tem funcionalidades para analisar uma base de dados e as suas tabelas.

Para este caso irá ser feita a análise das tabelas que formarão as tabelas de facto:

* Invoicelines e invoice
* Orderlines e orders

Uma imagem com mesa

Descrição gerada automaticamente

Figura 3 – Tabela de Metadados da invoicelines

Uma imagem com mesa

Descrição gerada automaticamente

Figura 4 - Tabela de Metadados da invoices

Analisando primeiramente, verifica-se em invoices que todos os campos do tipo varchar, são nullable dado a entender que estes são apenas atributos com o objetivo de suporte a quem interage com a base de dados, sendo assim, estes atributos não serão relevantes para a data mart.

De seguida, é notável que ambas as tabelas trabalham com bastantes ids, tornando a dimensão formada por ambas, numa forte concorrente a ser uma tabela de factos da data mart.

Uma imagem com mesa

Descrição gerada automaticamente

Uma imagem com mesa

Descrição gerada automaticamenteFigura 5 - Tabela de Metadados da orderlines

Figura 6 - Tabela de Metadados da orders

Primeiramente, é possível observar que estas tabelas são semelhantes às tabelas invoices e invoicelines, trabalhando também com vários Ids e mostrando-se uma concorrente a ser uma tabela de factos.

Pelas informações dadas pela database, sabe-se que é possível fazer encomendas nos quais é preciso fazer um pedido de itens em falta de stock, tendo em conta que nem todos as encomendas necessitarão de fazê-lo, o campo backorderorderid é nullable. Os outros campos nullable são semelhantes aos da tabela invoices, tendo assim, as mesmas razões de esses atributos serem possivelmente nulos.

Uma imagem com mesa

Descrição gerada automaticamente

Figura 7 - Tabela invoicelines

Uma imagem com mesa

Descrição gerada automaticamente

Figura 8 - Tabela orderlines

Analisando agora as possíveis tabelas de facto, tendo em conta que orders está em várias orderlines e invoices está em várias invoicelines, foi feita uma observação das tabelas orderlines e invoicelines.

Como é possível verificar, temos bastantes dados inseridos nestas tabelas o que faz com que estas tenham um grau de importância maior.

Com o grande número de chaves estrangeiras e registos nestas tabelas, podem ser determinadas como tabelas de facto da data mart a ser construída

# Modelação Dimensional

Table 2: Data Warehouse Matrix

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **DIMENSIONS**  **BUSINESS PROCESSES** | Dim\_costumers | Dim\_stockitems | Dim\_costumertransactions | Ft\_orderlines | Ft\_invoicelines |
| Top de países que compraram mais o item X | X | X |  | X |  |
| Top de items mais comprados no WWI |  | X |  | X |  |
| Top pessoas que mais fizeram encomendas | X |  |  | X |  |
| Mês que tem mais vendas em quantidade |  |  |  |  | X |
| Top faturas com mais lucrativas |  |  |  |  | X |
| Média da quantidade de vendas de sempre |  |  |  |  | X |
| Top de países que mais dinheiro gastaram com o WWI | X |  |  |  | X |
| Top marcas mais vendidas |  | X |  |  | X |
| Qual a categoria de clientes que mais gera receita à WWI | X |  |  |  | X |
| Qual é o período médio entre a receção do pedido e a entrega do produto? |  |  |  | X | X |
| Top de países que demora mais a receber produtos? | X |  |  | X | X |
| Média de lucro mensal |  |  |  |  | X |
| Percentagem de pedidos que cumpriram o prazo expectável de entrega |  |  |  | X | X |

## Design do modelo de dados dimensional

Abaixo está representado, em tabelas, informações sobre os atributos mais importantes de cada tabela

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Type of table** | | **Nr. Records** | **Description** |
| dim\_costumers | Dimension | | 664 | Dimensão que apresenta os dados pessoais de um cliente |
| **Target (Data mart)** | | | |  |
| **Column** | **Data type** | **Description** | | |
| Costumerkey | Int | Chave identificadora do cliente | | |
| Costumerid | Int | Identificador do cliente | | |
| Costumername | Varchar | Nome do cliente | | |
| Cityname | Varchar | Nome da cidade em que o cliente vive | | |
| Stateprovincename | Varchar | Nome do estado em que o cliente vive | | |
| Formalname | Varchar | Nome do país em que o cliente vive | | |
| Costumercategoryname | Varchar | Nome do tipo de categoria do cliente | | |
| Isoncredithold | Bit |  | | |
| Phonenumber | Varchar | Número de telemóvel do cliente | | |
| Faxnumber | Varchar | Número de telefone fixo do cliente | | |
| Deliveryadressline1 | Varchar | Morada de entrega de encomendas do cliente | | |
| Deliverypostalcode | Varchar | Código postal de entrega de encomendas do cliente | | |
| Deliverylocation | Varchar | Localização em coordenadas de entrega de encomendas do cliente | | |
| Postaladressline1 | Varchar | Morada do cliente | | |
| Postalpostalcode | Varchar | Código postal do cliente | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Type of table** | **Nr. Records** | **Description** |
| dim\_stockitems | Dimension | 228 | Dimensão que apresenta os dados sobre os items em stock e suas características |
| **Target (Data mart)** | | |  |
| **Column** | **Data type** | **Description** | |
| Stockitemkey | Int | Chave identificadora do item | |
| Stockitemid | Int | Identificador do item | |
| Stockitemname | Varchar | Nome do item | |
| Colorname | Varchar | Cor do item | |
| Unitprice | Numeric | Preço unitário do item | |
| Brand | Varchar | Marca à qual o item pertence | |
| Size | Varchar | Tamanho do item | |
| Ischillerstock | Varchar | Atributo que verifica se o item é fresco ou não | |
| Barcode | Varchar | Número do código de barras do item | |
| Taxrate | Numeric | Taxa aplicada no item para a sua venda | |
| Recommendedretailprice | Numeric | Preço recomendado para revenda do item | |
| Typicalweightperunit | Numeric | Peso por unidade | |
| Tags | Text | Etiquetas que relacionam-se ao item | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Type of table** | **Nr. Records** | **Description** |
| dim\_customertransaction | Dimension | 26638 | Dimensão que apresenta todas as transações que ocorreram no WWI |
| **Target (Data mart)** | | |  |
| **Column** | **Data type** | **Description** | |
| Customertransactionkey | Int | Chave identificador da transação | |
| Customertransactionid | Int | Identificador da transação | |
| Costumerid | Int | Identificador do cliente | |
| Invoiceid | Varchar | Identificador da fatura | |
| Transactiontypename | Int | Nome do tipo de transição efetuada | |
| Paymentmethodname | Varchar | Nome do tipo de pagamento efetuado | |
| Transactiondate | Datetime | Data em que a transação foi efetuada | |
| Taxammount | Numeric | Taxa aplicada na transação | |
| Transactionamount | Numeric | Valor monetário enviado a partir da transação | |
| Finalizationdate | Datetime | Data de finalização da transação | |
| Isfinalized | Bit | Atributo que confirma se a transação está terminada ou não | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Type of table** | **Nr. Records** | **Description** |
| Ft\_orderlines | Fact Table | 1157060 | Tabela de factos representante de cada linha de cada encomenda |
| **Target (Data mart)** | | |  |
| **Column** | **Data type** | **Description** | |
| Orderlineid | Int | Identificador da linha de encomenda | |
| Orderid | Int | Identificador da encomenda | |
| Costumerkey | Int | Chave identificadora do cliente | |
| Stockitemkey | Int | Chave identificadora do item a comprar | |
| Orderdate | Datetime | Data em que a encomenda foi feita | |
| Orderprice | Numeric | Preço total da encomenda | |
| Expecteddeliverydate | Datetime | Data expectável de receção da encomenda por parte do cliente | |
| Costumerpurchaseordernumber | Varchar | Número de pagamento da encomenda do cliente | |
| Isundersupplybackordered | bit | Atributos que regista se a encomenda atual tem stock em falta para fazer uma encomenda desse mesmo material | |
| Pickingcompletedwhen | Datetime | Data em que a obtenção de todos os itens a serem enviados por parte da WWI | |
| Backorderid | Int | Identificador da encomenda de stock em falta | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | | **Type of table** | **Nr. Records** | **Description** |
| Ft\_invoicelines | | Fact Table | 456530 | Tabela de factos representante de cada linha de cada fatura |
| **Target (Data mart)** | | | |  |
| **Column** | **Data type** | | **Description** | |
| Invoicelineid | Int | | Identificador da linha de fatura | |
| Invoiceid | Int | | Identificador da fatura | |
| Orderid | Int | | Identificador da encomenda | |
| Costumerkey | Int | | Chave identificadora do cliente | |
| Stockitemkey | Int | | Chave identificadora do item a comprar | |
| Extendedprice | Numeric | | Resultado do preço final do produto, calculado a partir de unitprice, taxrate e quantity | |
| Lineprofit | Numeric | | Lucro retirado da linha da fatura | |
| Invoicedate | Datetime | | Data na qual a fatura foi tirada | |
| Quantity | Int | | Quantidade de unidades registadas na fatura | |
| Unitprice | Numeric | | Preço unitário registado na fatura | |
| Taxrate | Numeric | | Taxa aplicada ao comprador por linha de fatura | |
| Iscreditnote | Bit | |  | |
| Totaldryitems | Int | | Quantidade total de items secos | |
| Totalchilleritems | Int | | Quantidade total de items frescos | |
| Returneddeliverydata | Datetime | |  | |
| Confirmeddeliverytime | Datetime | | Data de confirmação da receção da encomenda | |

# Implementação da DataMart (ETL)

Para implementar a nossa datamart foi utilizado o Kettle, ferramenta que permite fazer operações ETL em dados inseridos no mesmo.

Foram criadas transformações neste software de forma a montar as dimensões e tabelas de factos pretendidas para os nossos objetivos, adicionalmente foram implementados alguns jobs de forma a dar apoio ao funcionamento da montagem e atualização da datamart.

## Jobs

Primeiramente é executado o **job\_dims\_ft** que está a fazer inicialmente a execução do job “job\_dims”, que será analisado mais abaixo no relatório, depois é feito o carregamento das transformações das tabelas de factos ft\_invoicelines e ft\_orderlines, no caso de erro de qualquer um destes processo, o processo ETL é cancelado identificado o erro na consola.

Este é o “main job” que faz o trabalho todo de registar as informações na data mart.

Uma imagem com gráfico

Descrição gerada automaticamente

Figura 2 – Job\_dims\_ft

O outro job é **jobs\_dims** que tem como função fazer a execução das transformações referentes às dimensões geradas para a nossa data mart, no caso de erro o processo ETL é parado e o erro é registado na consola.

Uma imagem com diagrama

Descrição gerada automaticamente

Figura 3 - Job\_dims

## Transformations

Nas transformações houve 2 tipos de abordagens: uma para as **dimensões** e outra para as **tabelas de factos**.

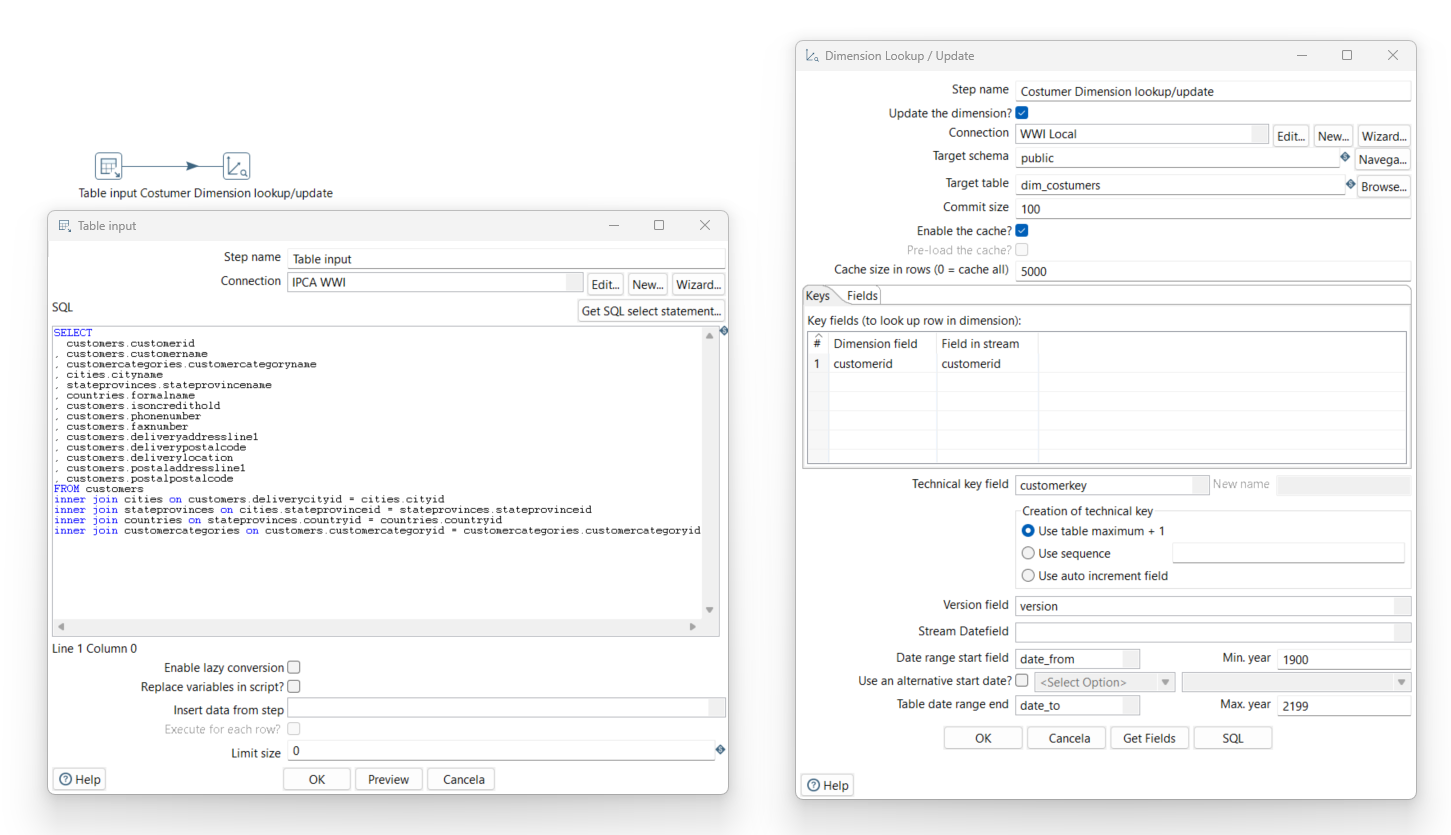
Segue-se abaixo um exemplo de uma dimensão, neste caso, dim\_costumers:

Figura 4 - Transformation Dim\_Costumers

Na construção de dimensões, é preciso apenas estar preocupado em pegar na tabela no formato que já queremos, incluído assim já todas as colunas das tabelas necessárias para construir a dimension e, por fim, aplicar o dimension lookup para fazer a criação da key da nossa dimensão.

Este passo permite também a criação da tabela na nossa data mart e o povoamento a partir dos dados obtidos do table input.

Estes passos foram feitos para as restantes dimensões:

* Dim\_stockitems
* Dim\_costumertransactions

Uma imagem com texto, captura de ecrã, interior, computador portátil

Descrição gerada automaticamente

Figura 5 -Transformation Dim\_CostumerTransactions

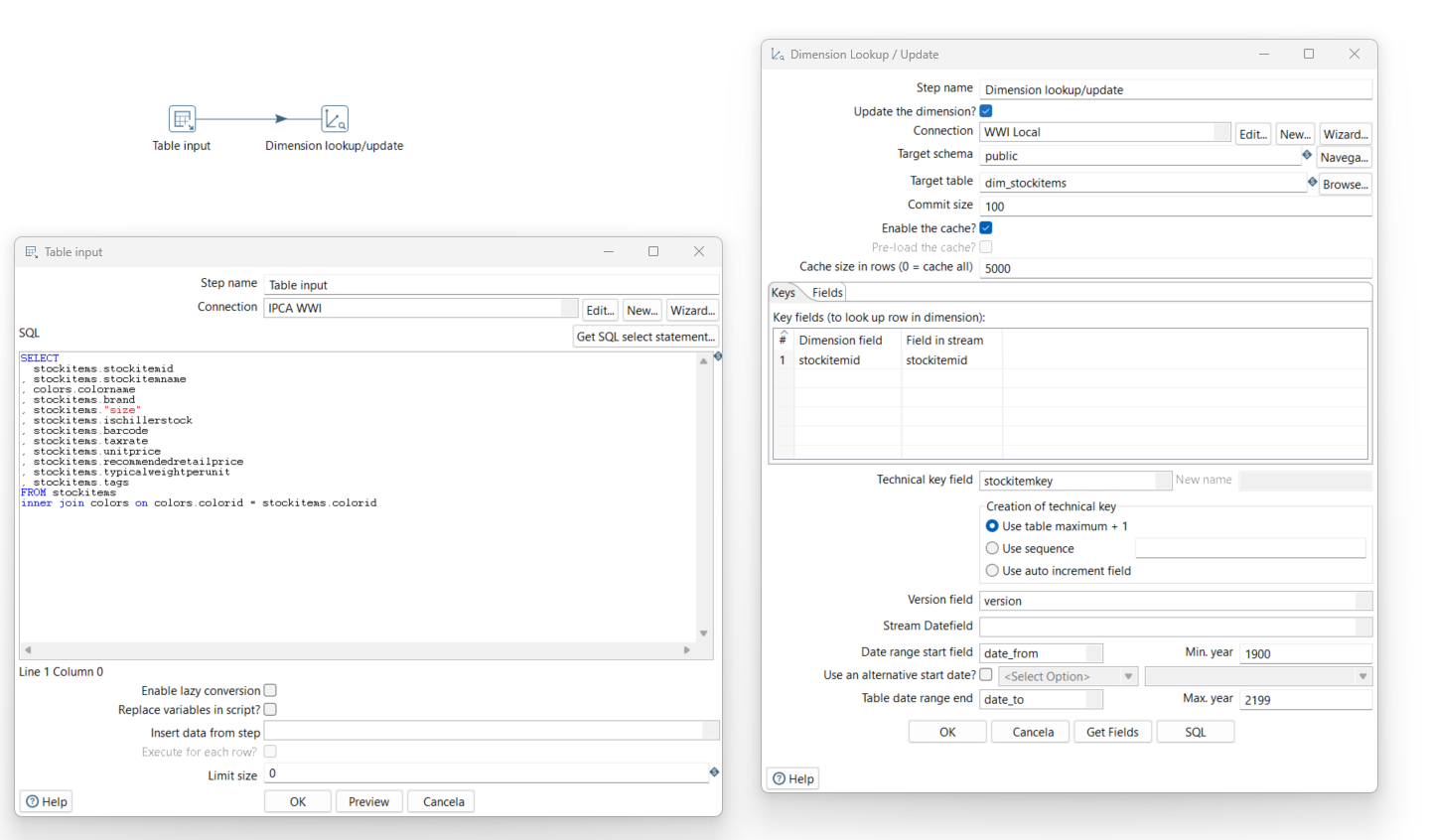


Figura 6 - Transformation Dim\_StockItems

Para o caso das tabelas de factos, a abordagem foi mais exigente, tendo em conta que estamos a trabalhar com uma fact table, foi necessário fazer o lookup para cada atributo que possuía um id de outra dimensão.

Segue-se abaixo as transformações de ft\_invoicelines e ft\_orderlines:

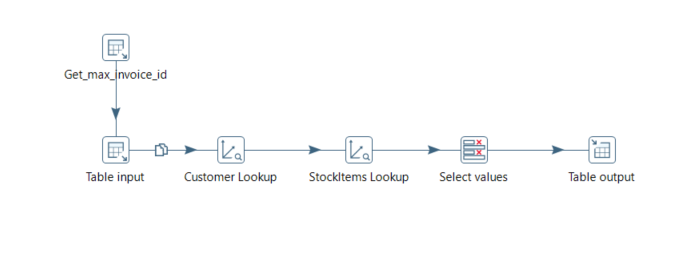


Figura 7 - Transformation ft\_invoicelines

Uma imagem com diagrama

Descrição gerada automaticamente

Figura 8 - ft\_orderlines

Inicialmente são obtidos os dados das tabelas pretendidas para a montagem das tabelas de factos.

De seguida, dependendo das colunas adicionas, terá de ser feito então um lookup a todas as chaves estrangeiras exceto aos ids principais da própria fact table, por exemplo para ft\_invoicelines não foi necessário este passo para invoiceid e invoicelineid.

Por fim são removidas colunas em excesso e é feito o output para a datamart da tabela de factos pronta.

# Conclusão

* Revisão crítica do trabalho feito
* Pontos fortes e fracos
* Se possível, o que se mudaria/adicionaria