



Instituto Superior de
Engenharia do Porto

Relatório ARMDD

André Conceição | 1200807

Afonso Cruz | 1240434

Índice

1.Introdução	1
2.Objetivos	2
3.Modelo Dimensional.....	3
3.1 Identificação do processo de negócio	3
3.2 Nível de Granularidade	3
3.3 Seleção das Dimensões Relevantes	3
3.3.1 Dimensão Produto (DimProduct).....	4
3.3.2 Dimensão Cliente (DimCustomer).....	5
3.3.3 Dimensão Funcionário (DimEmployee).....	5
3.3.4 Dimensão Data (DimDate).....	6
3.3.5 Dimensão Transportador (DimShipper).....	6
3.3.6 Dimensão Enviado Para (DimShipTo)	7
3.3.7 Dimensão Local da Empresa (DimCompanyLocal)	7
3.3.8 Dimensão Moeda (DimCurrency)	7
3.3.9 Dimensão Fornecedor (DimSupplier).....	8
3.4 Design da Tabela de Factos	8
3.5 Modelo dimensional completo	9
3.6 Estratégias de <i>Slowly Changing Dimensions</i> (SCD)	10
3.6.1 DimCustomer.....	11
4. Staging Area	17
4.1 Descrição do processo etl	17
4.2 SUBtópicos do Processo etl.....	17
4.2.1 Extração.....	17
4.2.2 Transformações	18
4.2.3 Carregamento.....	22
5. Análises Multidimensionais	22
6. Trabalho realizado	23
7. Conclusão	24

Índice de Figuras

Figura 1 - DimProduct.....	4
----------------------------	---

Figura 2 -DimCustomer.....	5
Figura 3 - DimEmployee.....	5
Figura 4 – DimDate.....	6
Figura 5 – DimShipper	6
Figura 6 - DimShipTo	7
Figura 7 - DimCompanyLocal	7
Figura 8 – DimCurrency	8
Figura 9 – DimSupplier	8
Figura 10 - FactOrders	9
Figura 11 - Modelo Dimensional Completo	10
Figura 12 - Slowly Changing Dimension (SCD) DimCustomer	11
Figura 13 - Atributos SCD DimCustomer	12
Figura 14 - Slowly Changing Dimension (SCD) DimEmployee	12
Figura 15 - Atributos SCD DimEmployee	13
Figura 16 - Slowly Changing Dimension (SCD) DimShipTo.....	13
Figura 17 - Atributos SCD DimShipTo	14
Figura 18 - Slowly Changing Dimension (SCD) DimProduct	14
Figura 19 - Atributos SCD DimProduct	15
Figura 20 - Slowly Changing Dimension (SCD) DimSupplier	15
Figura 21 - Atributos SCD DimSupplier	16
Figura 22 - Processo de extração via Excel da base de dados da Sede	18
Figura 23 - Extração de uma tabela (Order) para outra (ShipTo).....	18
Figura 24 - Junção do atributo Name	19
Figura 25 - Tabela Lookup Country	20
Figura 26 - Transformação do atributo Discontinued Delegação to INT	20
Figura 27 - Transformação do atributo Discontinued Sede to INT	20
Figura 28 - Envio para tabela CustomerDQP	21
Figura 29 - Tabela CustomerDQP	21
Figura 30 - Tratamento de duplicados Customer.....	21
Figura 31 - Carregamento tabelas de Dims e de Factos	22

1. Introdução

Este documento abrange a realização e a implementação de um armazém de dados para uma empresa de retalho que se dedica à comercialização de bens alimentares. O projeto aborda sobre os desafios significativos decorrentes da utilização de sistemas operacionais distintos para gestão de encomendas — um implementado na sede nos Estados Unidos e outro desenvolvido especificamente para a subsidiária no Reino Unido. Embora ambos os sistemas tenham o mesmo propósito funcional, a organização divergente dos dados em cada aplicação dificulta a integração, análise abrangente e interpretação consistente das informações, comprometendo a eficiência operacional e a tomada de decisão estratégica.

Este relatório inclui diversos elementos importantes, como a construção de um modelo dimensional sólido, projetado de acordo com a abordagem de *Kimball*, e a implementação de processos eficientes de dados, incluindo extração, transformação, limpeza, integração e carregamento (ETL). Um esforço especial foi dedicado à resolução de questões específicas, como problemas de qualidade dos dados, conversão de moedas e a harmonização de formatos e estruturas de dados entre os diferentes sistemas.

Neste documento, é apresentada uma análise detalhada e uma proposta de solução que visa facilitar a análise de encomendas dos clientes num nível mais complexo. O propósito deste projeto é superar os desafios existentes e fornecer um sistema abrangente, portátil e escalável, capaz de melhorar as capacidades analíticas de decisão e avançar a visão estratégica geral do negócio.

2. Objetivos

Os objetivos da primeira iteração do projeto são os seguintes:

- **Definição da Arquitetura do Armazém de Dados:** Projetar a estrutura geral do armazém de dados para atender às necessidades descritas, garantindo integração eficaz entre as diferentes fontes de dados e suporte às análises pretendidas.
- **Desenvolvimento do Modelo Dimensional Subjacente:** Criar um modelo dimensional que inclua as dimensões e a tabela de fatos necessárias, com a especificação detalhada de atributos, tipos de dados, granularidade e relações entre tabelas.
- **Conceção das Estruturas de Dados na *Staging Area*:** Definir e documentar as tabelas e/ou ficheiros a serem criados na área de *staging* para suporte à extração, transformação e carregamento (ETL) dos dados entre os sistemas fonte e o armazém de dados.
- **Mapeamento de Dados entre Sistemas Fonte, *Staging Area* e Armazém de Dados:** Elaborar o mapeamento detalhado dos dados, especificando a origem de cada atributo, as transformações necessárias (incluindo validação e limpeza de dados) e a estratégia de carregamento no armazém de dados.
- **Estratégia para *Slowly Changing Dimensions* (SCD):** Determinar e documentar a abordagem de gestão de dimensões historicamente variantes para cada dimensão do modelo dimensional.

3. Modelo Dimensional

A metodologia de *Kimball*, amplamente reconhecida e adotada no desenvolvimento de armazéns de dados, foi a base metodológica deste projeto devido à sua abordagem orientada ao negócio e foco em entregar valor direto às análises empresariais. O projeto seguiu os passos principais da metodologia, garantindo um modelo dimensional sólido e eficiente para atender às necessidades identificadas. A seguir, detalhamos os passos adotados.

3.1 IDENTIFICAÇÃO DO PROCESSO DE NEGÓCIO

O primeiro passo no desenvolvimento do armazém de dados foi identificar o processo de negócio central que ele deve suportar: **a gestão de encomendas dos clientes**. Este processo foi escolhido devido à sua importância estratégica para a empresa, pois oferece informações fundamentais para a tomada de decisões em várias áreas.

3.2 NÍVEL DE GRANULARIDADE

O armazém de dados será projetado para trabalhar com a granularidade mais detalhada possível, focada em cada linha individual de encomenda. Isso significa que cada entrada na tabela de factos corresponderá a detalhes específicos de uma encomenda.

Essa escolha de granularidade proporciona a máxima flexibilidade analítica, permitindo consultas detalhadas e abrangentes, como:

- **Análise das encomendas:** Permitir consultas por produto, cliente, fornecedor, categoria ou região, facilitando a identificação de padrões de consumo e desempenho comercial.
- **Monitoramento logístico:** Avaliar o desempenho de transportadores, custos de frete, e a eficiência das entregas, possibilitando a otimização das operações.
- **Tendências temporais:** Oferecer uma visão flexível e abrangente de tendências ao longo do tempo, com a capacidade de realizar análises diárias ou agregações em níveis maiores, como meses, trimestres, semestres ou anos.

3.3 SELEÇÃO DAS DIMENSÕES RELEVANTES

Com base no processo de negócio identificado, foram selecionadas as dimensões que fornecem o contexto necessário para a análise das métricas associadas às encomendas. As dimensões escolhidas são:

3.3.1 Dimensão Produto (DimProduct)

Fornecer informações detalhadas sobre os produtos vendidos e os seus respectivos fornecedores, essenciais para análises como identificação dos itens mais vendidos, avaliação de margens de lucro e tendências de consumo.



DimProduct
ProductKey(PK)
ProductIdUK
ProductIdUSA
ProductName
CategoryName
CategoryDescription
QuantityPerUnit
UnitPriceUK
UnitPriceUSA
UnitsInStockUK
UnitsInStockUSA
UnitsOnOrderUK
UnitsOnOrderUSA
ReorderLevel
Discontinued
EffectiveDate
ExpiredDate
IsCurrent

Figura 1 - DimProduct

3.3.2 Dimensão Cliente (DimCustomer)

Identifica os clientes, permitindo a analisar quais os principais clientes e comportamentos de compra.

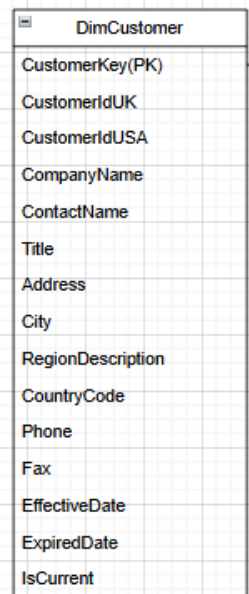


Figura 2 -DimCustomer

3.3.3 Dimensão Funcionário (DimEmployee)

Permite identificar os funcionários responsáveis pelo processamento de pedidos, sendo útil para análises de desempenho interno e produtividade.

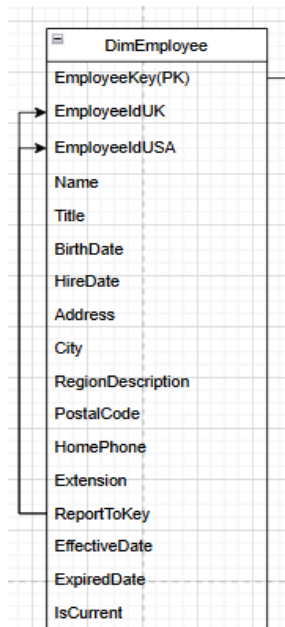
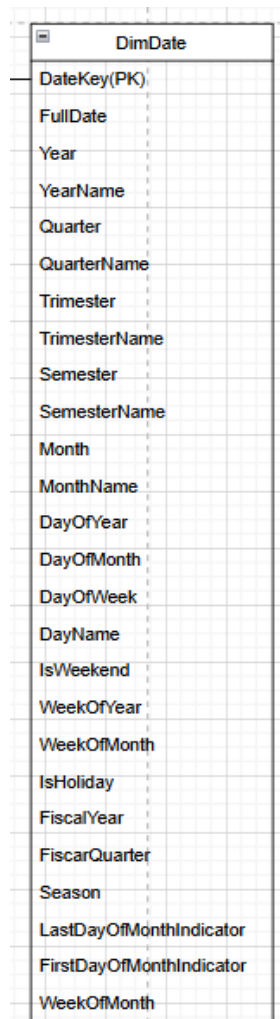


Figura 3 - DimEmployee

3.3.4 Dimensão Data (DimDate)

A dimensão data permite realizar análises temporais detalhadas, como tendências de vendas ao longo do tempo, sazonalidade e comparações entre períodos.



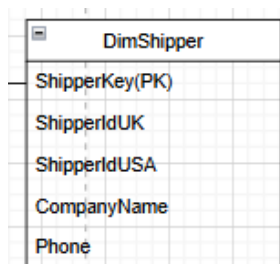
The diagram shows a table named 'DimDate' with a primary key 'DateKey(PK)'. The table contains various date-related attributes organized into hierarchical levels: FullDate, Year, YearName, Quarter, QuarterName, Trimester, TrimesterName, Semester, SemesterName, Month, MonthName, DayOfYear, DayOfMonth, DayOfWeek, DayName, IsWeekend, WeekOfYear, WeekOfMonth, IsHoliday, FiscalYear, FiscalQuarter, Season, LastDayOfMonthIndicator, FirstDayOfMonthIndicator, and WeekOfMonths.

DimDate	
DateKey(PK)	
FullDate	
Year	
YearName	
Quarter	
QuarterName	
Trimester	
TrimesterName	
Semester	
SemesterName	
Month	
MonthName	
DayOfYear	
DayOfMonth	
DayOfWeek	
DayName	
IsWeekend	
WeekOfYear	
WeekOfMonth	
IsHoliday	
FiscalYear	
FiscalQuarter	
Season	
LastDayOfMonthIndicator	
FirstDayOfMonthIndicator	
WeekOfMonths	

Figura 4 – DimDate

3.3.5 Dimensão Transportador (DimShipper)

Contém informações sobre os transportadores responsáveis pelo envio das encomendas, permitindo avaliar os custos e desempenho logístico.



The diagram shows a table named 'DimShipper' with a primary key 'ShipperKey(PK)'. The table contains attributes: ShipperIdUK, ShipperIdUSA, CompanyName, and Phone.

DimShipper	
ShipperKey(PK)	
ShipperIdUK	
ShipperIdUSA	
CompanyName	
Phone	

Figura 5 – DimShipper

3.3.6 Dimensão Enviado Para (DimShipTo)

Representa as informações detalhadas sobre o destino final das encomendas, permitindo análises por local de expedição, cidade ou país.



Figura 6 - DimShipTo

3.3.7 Dimensão Local da Empresa (DimCompanyLocal)

Permite diferenciar os locais de operação da empresa (sede e delegações), fornecendo contexto geográfico e operacional para análises. É importante para diferenciar as encomendas feitas à sede e à delegação.

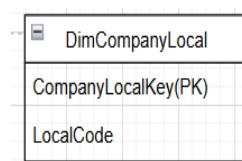


Figura 7 - DimCompanyLocal

3.3.8 Dimensão Moeda (DimCurrency)

Essa dimensão possibilita análises financeiras, uma vez que estamos a trabalhar com valor em dólares (encomendas da sede) e em libras (encomendas da delegação). É essencial para padronizar valores e realizar conversões monetárias consistentes.

DimCurrency
CurrencyKey (PK)
CurrencyCode
CurrencyName

Figura 8 – DimCurrency

3.3.9 Dimensão Fornecedor (DimSupplier)

Representa informações detalhadas sobre os fornecedores, permitindo análises abrangentes baseadas em sua localização, como país, região e cidade, além de outros atributos relevantes relacionados à empresa fornecedora.

DimSupplier
SupplierKey
SupplierIdUK
SupplierIdUSA
CompanyName
ContactName
ContactTitle
Address
City
Region
PostalCode
Country
Phone
Fax
EffectiveDate
ExpiredDate
IsCurrent

Figura 9 – DimSupplier

3.4 DESIGN DA TABELA DE FACTOS

A tabela de factos principal do modelo foi definida como **FactOrders**, contendo as métricas essenciais para análise. Cada linha da tabela representa uma linha de pedido, ou seja, um produto específico encomendado em uma encomenda. A tabela armazena factos quantitativos, como:

- Valor unitário;
- Quantidade de produtos solicitada;
- Valor do frete;
- Valor do desconto;
- Valor total com desconto

Além disso, a tabela inclui chaves estrangeiras para cada dimensão seleccionada, conectando os fatos às dimensões contextuais. Isso permite realizar análises detalhadas e agregações a partir de diferentes pontos de vista, como vendas por transportador, categorias de produtos mais vendidas ou total de vendas por empregado.

<i>FactOrders</i>	
OrderDate(FK)(PK)	◀
RequiredDate(FK)	◀
ShippedDate(FK)	◀
ProductKey(FK)(PK)	◀
▶ CustomerKey(FK)(PK)	
▶ EmployeeKey(FK)(PK)	
ShipperKey(FK)(PK)	◀
▶ ShipToKey(FK)(PK)	
SupplierKey(FK)(PK)	◀
CurrencyKey(FK)	◀
▶ CompanyLocalKey(FK)	
OrderId	
UniPriceUK	
UniPriceUSA	
Quantity	
TotalLineUK	
TotalLineUSA	
DiscountCostUK	
DiscountCostUSA	
FreightUK	
FreightUSA	

Figura 10 - FactOrders

3.5 MODELO DIMENSIONAL COMPLETO

Com base nas tabelas de dimensões e a tabela factos que referimos anteriormente, o modelo dimensional completo foi estruturado conforme a seguinte imagem:

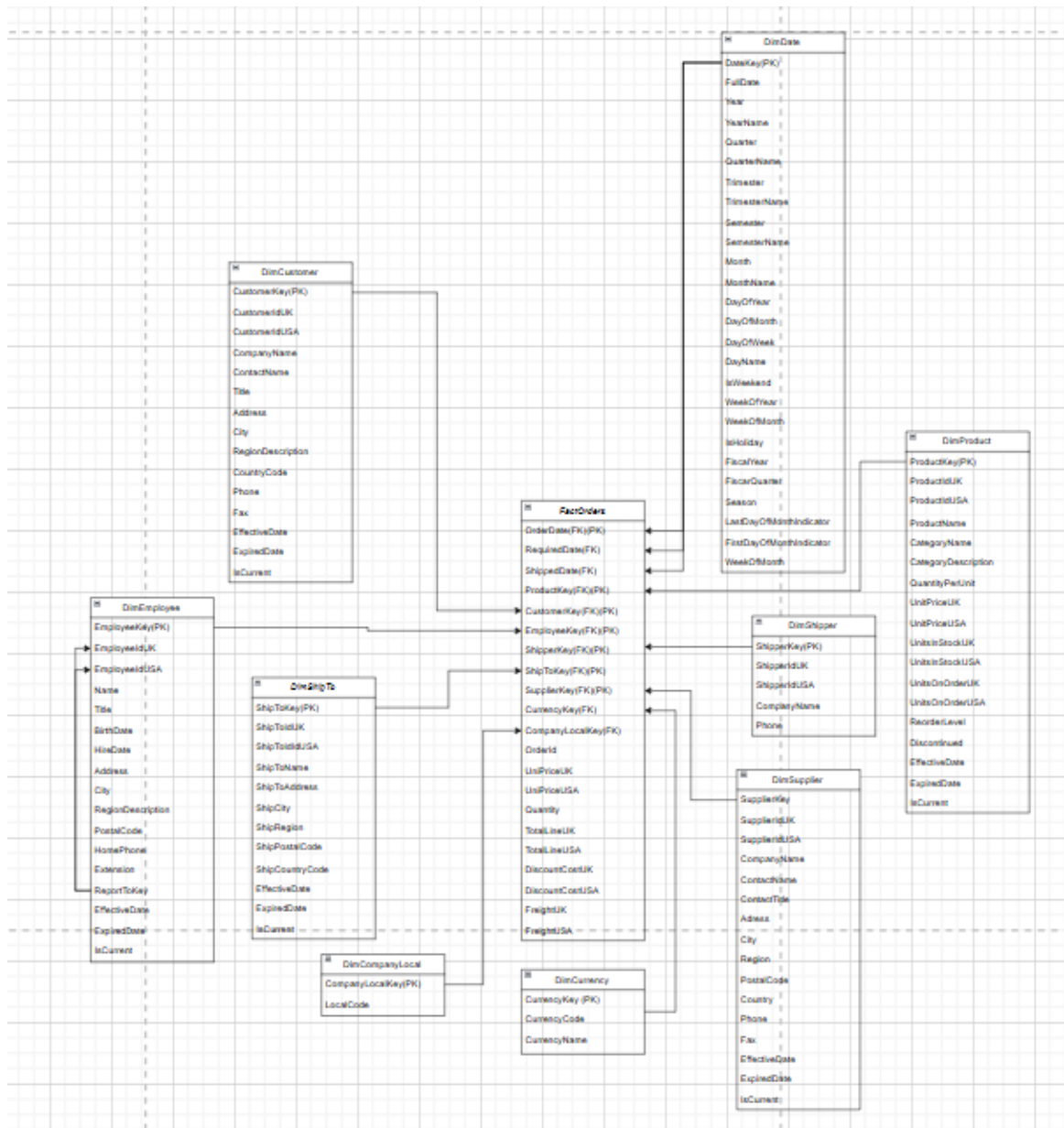


Figura 11 - Modelo Dimensional Completo

3.6 ESTRATÉGIAS DE *SLOWLY CHANGING DIMENSIONS* (SCD)

Para garantir a consistência histórica e a precisão nas análises temporais, adotamos a estratégia de ***Slowly Changing Dimensions* (SCD)** do tipo 2, que adiciona uma nova linha às tabelas de dimensão sempre que ocorre uma alteração em atributos relevantes para análise. Essa abordagem, amplamente explorada nas aulas de PL, permite capturar e preservar o histórico das mudanças de forma detalhada, assegurando a rastreabilidade das informações.

Após uma análise criteriosa, identificamos as dimensões para as quais o histórico seria essencial para análises futuras: **DimShipTo**, **DimEmployee**, **DimCustomer**, **DimProduct** e **DimSupplier**. Para implementar a estratégia SCD Tipo 2, foram incluídas colunas específicas nas dimensões:

- **EffectiveDate**: Data de início de validade do registo.
- **ExpiredDate**: Data de expiração do registo (quando aplicável).
- **IsCurrent**: Indicador que identifica se o registo é o mais recente.

No processo de atualização de um atributo utilizando a estratégia SCD Tipo 2, as etapas seguem este fluxo:

- **Detetar a mudança**: Durante o processo ETL, compara-se o valor do atributo no sistema fonte com o valor no registo atual da dimensão para identificar alterações.
- **Marcar o registo antigo**: O registo atual é atualizado, preenchendo a coluna **ExpiredDate** com a data da mudança e configurando **IsCurrent** como 0 (falso), indicando que está desatualizado.
- **Inserir um novo registo**: Um novo registo é adicionado à dimensão, contendo os valores atualizados do atributo, uma nova **chave substituta**, a data da mudança em **EffectiveDate**, e **IsCurrent** como 1 (verdadeiro).

Esta estrutura garante o registo preciso das alterações nas dimensões, permitindo análises temporais robustas e consultas baseadas no contexto histórico, fundamentais para decisões de negócio informadas.

3.6.1 DimCustomer

A Slowly Changing Dimension que realizamos para a tabela DimEmployee é possível ser observado na Figura 12:

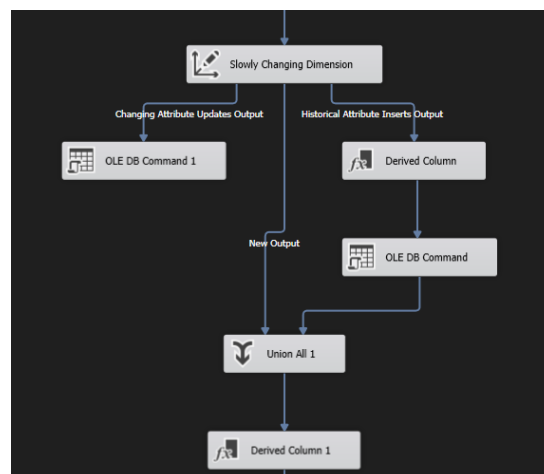


Figura 12 - Slowly Changing Dimension (SCD) DimCustomer

Para a SCD, da tabela DimCustomer, definimos os atributos, como Tipo 1 (Changing Attribute) e Tipo 2 (Historical Attribute), da seguinte maneira, como é possível observar na Figura 13:

Dimension Columns	Change Type
Address	Changing attribute
CityName	Historical attribute
CompanyName	Changing attribute
ContactName	Changing attribute
Country	Historical attribute
CountryCode	Historical attribute
Fax	Changing attribute
Phone	Changing attribute
PostalCode	Changing attribute
RegionDescription	Historical attribute
Title	Historical attribute

Figura 13 - Atributos SCD DimCustomer

Definimos como Tipo 2 os atributos: *CityName*, o *Country*, o *CountryCode*, o *RegionDescription* e o *Title*.

Os atributos *CityName*, o *Country*, o *CountryCode*, o *RegionDescription* e o *Title* garantem a identificação de alterações históricas que possam impactar análises de clientes, como mudanças na localização geográfica ou no título. Esta identificação é essencial para compreender a evolução do perfil dos clientes ao longo do tempo e avaliar métricas relacionadas a regiões ou categorias específicas. Os restantes serão do Tipo 1, porque caso algum dos campos se altere, podemos sobrepor a informação dos mesmos pela nova, sem ser necessário guardar histórico.

3.6.2 DimEmployee

A Slowly Changing Dimension que realizamos para a tabela DimEmployee foi o da Figura 14:

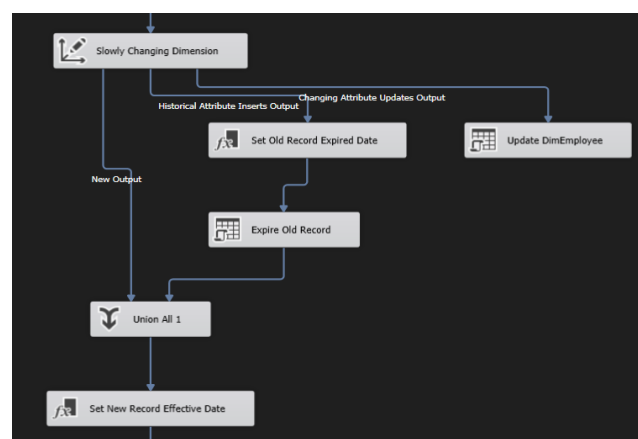


Figura 14 - Slowly Changing Dimension (SCD) DimEmployee

Para a SCD, da tabela DimEmployee, definimos os atributos, como Tipo 1 (Changing Attribute) e Tipo 2 (Historical Attribute), da seguinte maneira, como é possível observar na Figura 15:

Address	Changing attribute
BirthDate	Changing attribute
CityName	Historical attribute
Extension	Changing attribute
HireDate	Historical attribute
HomePhone	Changing attribute
Name	Changing attribute
PostalCode	Changing attribute
RegionDescription	Historical attribute
ReportsToKey	Changing attribute
Title	Historical attribute

Figura 15 - Atributos SCD DimEmployee

Definimos como Tipo 2 os atributos: *CityName*, o *HireDate*, o *RegionDescription* e o *Title*.

Os atributos *CityName*, *HireDate* e *Title* devem ser identificados historicamente, pois refletem mudanças organizacionais relevantes, enquanto *RegionDescription* deve manter histórico caso esteja associada a métricas. Os restantes serão do Tipo 1, porque caso algum dos campos se altere, podemos sobrepor a informação dos mesmos pela nova, sem ser necessário guardar histórico.

3.6.3 DimShipTo

A Slowly Changing Dimension que realizamos para a tabela DimShipTo foi o da Figura 16:

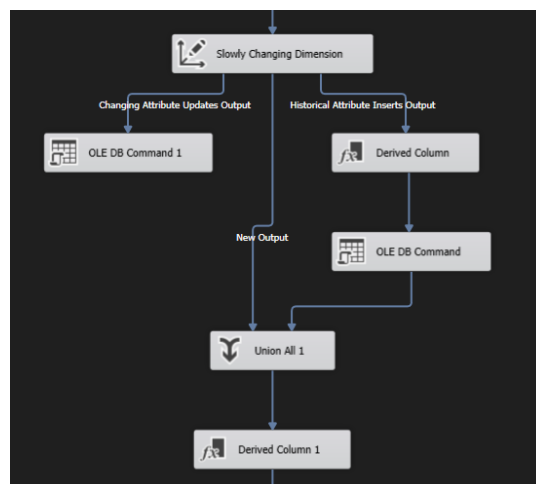


Figura 16 - Slowly Changing Dimension (SCD) DimShipTo

Para a SCD, da tabela DimShipTo, definimos os atributos, como Tipo 1 (Changing Attribute) e Tipo 2 (Historical Attribute), da seguinte maneira, como é possível observar na Figura 17:

Dimension Columns	Change Type
ShipCity	Historical attribute
ShipCountry	Historical attribute
ShipCountryCode	Historical attribute
ShipPostalCode	Changing attribute
ShipRegion	Changing attribute
ShipToAddress	Changing attribute
ShipToName	Changing attribute

Figura 17 - Atributos SCD DimShipTo

Definimos como Tipo 2 os atributos: *ShipCity*, o *ShipCountry* e o *ShipCountryCode*.

Os atributos *ShipCity*, o *ShipCountry* e o *ShipCountryCode* permitem a identificação de mudanças históricas relacionadas às localizações de envio. Esta identificação é importante para análises temporais, como identificar alterações nos padrões de envio e compreender o impacto dessas mudanças em métricas logísticas e operacionais. Os restantes serão do Tipo 1, porque caso algum dos campos se altere, podemos sobrepor a informação dos mesmos pela nova, sem ser necessário guardar histórico.

3.6.4 DimProduct

A Slowly Changing Dimension que realizamos para a tabela DimProduct foi o da Figura 18:

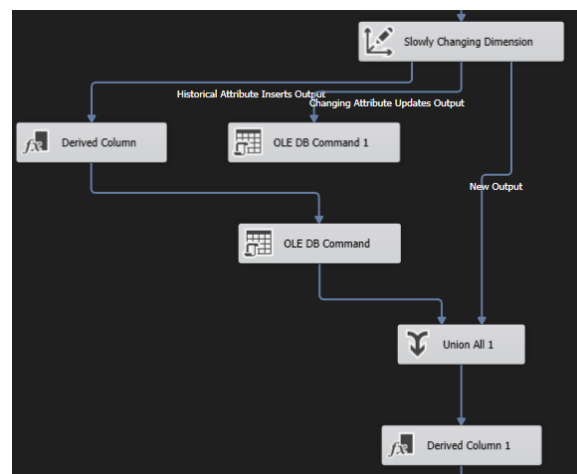


Figura 18 - Slowly Changing Dimension (SCD) DimProduct

Para a SCD, da tabela DimProduct, definimos os atributos, como Tipo 1 (Changing Attribute) e Tipo 2 (Historical Attribute), da seguinte maneira, como é possível observar na Figura 19:

Dimension Columns	Change Type
CategoryName	Historical attribute
Discontinued	Changing attribute
ProductName	Changing attribute
QuantityPerUnit	Changing attribute
ReorderLevel	Changing attribute
UnitPriceUK	Historical attribute
UnitPriceUSA	Historical attribute
UnitsInStockUK	Changing attribute
UnitsInStockUSA	Changing attribute
UnitsOnOrderUK	Changing attribute
UnitsOnOrderUSA	Changing attribute

Figura 19 - Atributos SCD DimProduct

Definimos como Tipo 2 os atributos: *CategoryName*, o *UnitPriceUK* e o *UnitPriceUSA*.

Os atributos *CategoryName*, o *UnitPriceUK* e o *UnitPriceUSA* representam informações que impactam diretamente análises históricas e temporais. A categoria de um produto é essencial para permitem mudanças organizacionais e de mercado ao longo do tempo, enquanto o preço é fundamental para análises de evolução de valores e tendências comerciais, garantindo a precisão de relatórios e a rastreabilidade das alterações. Os restantes serão do Tipo 1, porque caso algum dos campos se altere, podemos sobrepor a informação dos mesmos pela nova, sem ser necessário guardar histórico.

3.6.5 DimSupplier

A Slowly Changing Dimension que realizamos para a tabela DimSupplier foi o da Figura 20:

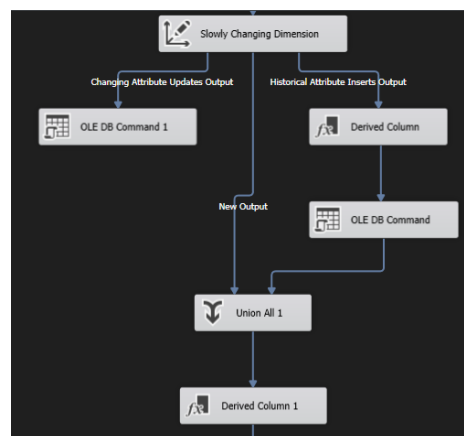


Figura 20 - Slowly Changing Dimension (SCD) DimSupplier

Para a SCD, da tabela DimSupplier, definimos os atributos, como Tipo 1 (Changing Attribute) e Tipo 2 (Historical Attribute), da seguinte maneira, como é possível observar na Figura 21:

Dimension Columns	Change Type
Address	Changing attribute
City	Historical attribute
CompanyName	Historical attribute
ContactName	Changing attribute
ContactTitle	Historical attribute
Country	Historical attribute
CountryCode	Historical attribute
Fax	Changing attribute
Phone	Changing attribute
PostalCode	Changing attribute
Region	Historical attribute

Figura 21 - Atributos SCD DimSupplier

Definimos como Tipo 2 os atributos: *City*, o *CompanyName*, o *ContactTitle*, o *Country*, o *CountryCode* e o *Region*.

Os atributos *City*, o *CompanyName*, o *ContactTitle*, o *Country*, o *CountryCode* e o *Region* representam informações importantes para análises históricas e temporais relacionadas aos fornecedores. Estas alterações devem ser identificadas para garantir a identificação de mudanças organizacionais, geográficas ou de contacto, permitindo uma análise mais detalhada do impacto destas alterações ao longo do tempo. Os restantes serão do Tipo 1, porque caso algum dos campos se altere, podemos sobrepor a informação dos mesmos pela nova, sem ser necessário guardar histórico.

4. Staging Area

4.1 DESCRIÇÃO DO PROCESSO ETL

O processo de ETL (*Extract, Transform e Load*) é essencial na construção de um armazém de dados, garantindo a movimentação e a conversão dos dados desde os sistemas fonte até ao seu destino final. Normalmente está dividido em três fases:

- Extração (*Extract*) – Responsável maioritariamente pela procura dos sistemas fonte e pela sua transferência para a *Staging Area*. Essa extração pode ser feita a partir do carregamento de ficheiros CSV, apesar de uma das tabelas ser carregada apenas depois de carregar todos os ficheiros Excel, mais concretamente a tabela *ShipTo*, para a base de dados da USA (sede) ou então o processo de extração pode ser feito através do carregamento de uma base de dados da UK (delegação).
- Transformação (*Transform*) – Nesta fase, os dados, anteriormente extraídos, sofrem alguns processos de limpeza e padronização. As transformações abrangem alterações nos formatos, cálculos, entre outros processos de transformação.
- Carregamento (*Load*) – Por fim, nesta etapa, os dados já limpos e transformados são inseridos nas tabelas, de dimensões e de factos, da DW (*Data Warehouse*). Incluindo também a incorporação de novos dados e a utilização de SCD para facilitar o registo de histórico.

4.2 SUBTÓPICOS DO PROCESSO ETL

4.2.1 Extração

A informação será obtida através da leitura de ficheiros excel, convertendo os dados para um formato compatível. Na subsidiária será utilizada os dados já anteriormente extraídos de um sistema relacional. Será feita uma validação de forma a saber se os dados esperados foram realmente extraídos.

Para o processo de extração da base de dados da sede, aparecia-nos um erro ao extrair os dados do ficheiro CSV, pelo método que o professor ensinou nas aulas, optando por utilizar um método diferente onde foi necessário alterar o formato do ficheiro para *.xlsx*. Para além disto, optámos pela criação da tabela *ShipTo*, que extrai a informação única, ou seja, sem duplicados, dos conteúdos *ShipTo* contidos na tabela *Order*.



Figura 22 - Processo de extração via Excel da base de dados da Sede

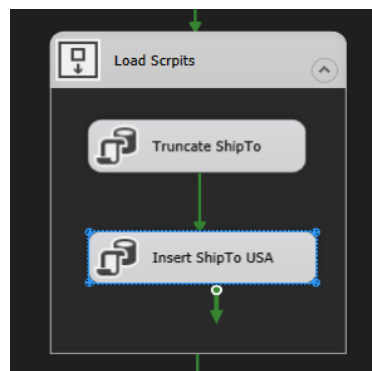


Figura 23 - Extração de uma tabela (Order) para outra (ShipTo)

Relativamente ao processo de extração da base de dados da delegação, apenas foram carregadas as bases de dados anteriormente fornecidas pelo professor.

4.2.2 Transformações

O funcionamento eficaz de um armazém de dados depende, em grande parte da fiabilidade e consciência das informações armazenadas. Por isso, os sistemas fonte, foram analisados cuidadosamente para identificar e planear os processos necessários de transformação, limpeza e integração. Estes processos têm como objetivo resolver problemas relacionados com a estrutura, significado e qualidade de dados, garantindo desta forma que a informação armazenada possa ser utilizada para análise e tomada de decisões.

4.2.2.1 Junção de Atributos Compostos

Na estrutura da delegação, os nomes dos funcionários estão armazenados em três campos distintos: *TitleOfCourtesy*, *FirstName* e *LastName*. No entanto, na estrutura da sede, o nome dos funcionários é representado por um único campo denominado *Name*, que agrega os três atributos

mencionados. Para manter a integridade e a consistência dos dados entre as estruturas, foi necessário realizar a junção desses atributos da delegação em um único campo ficando no formato igual á da sede, uma vez que não achamos que fosse necessário guardar esses atributos separadamente na dimensão.

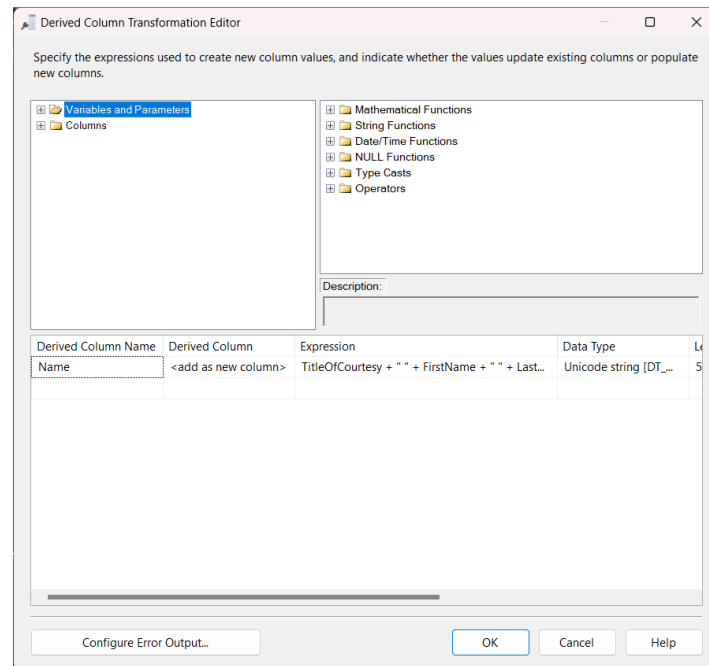


Figura 24 - Junção do atributo *Name*

4.2.2.2 Normalização da Representação do País e do atributo Discontinued

Na sede, os países são registados com uma abreviatura do nome, por exemplo PT para Portugal, contudo na delegação estes registos têm o nome completo. Para garantir a uniformidade, foi criado uma tabela *LookUpCountry* com os nomes dos pais e os respetivos códigos, conforme é definido pela norma internacional ISO 3166-1. Para a sede é feito um *LookUpCode* e para a delegação é feito um *LookUpCountry*, adicionando na dimensão ambos os campos, *Country* e *CountryCode* na dimensão, facilitando assim a realização de análises e assegurando uma representação geográfica uniforme e variada.

	CountryName	CountryCode
1	Argentina	AR
2	Austria	AT
3	Australia	AU
4	Barbados	BB
5	Belgium	BE
6	Brazil	BR
7	Canada	CA
8	Switzerland	CH
9	Costa Rica	CR
10	Czech Republic	CZ
11	Germany	DE
12	Denmark	DK
13	Spain	ES
14	Ethiopia	ET
15	Finland	FI
16	France	FR
17	United Kingdom	GB
18	Gambia	GM
19	Guatemala	GT
20	Guinea-Bissau	GW
21	Honduras	HN
22	Haiti	HT

Figura 25 - Tabela Lookup Country

Existem também dados que no sistema da delegação estão com informação de bit, no caso da tabela de *Products* em que a coluna *Discontinued* está no formato bit, enquanto no sistema da sede está como *TRUE* ou *FALSE*. Para garantir a integridade e a consistência dos dados entre os dois sistemas, ambos os formatos serão unificados, adotando int como padrão, 1/0.

Derived Column Name	Derived Column	Expression	Data Type	
DiscontinuedInt	<add as new column>	Discontinued == "0" ? 0 : 1	four-byte signed inte...	

Figura 26 - Transformação do atributo Discontinued Delegação to INT

Derived Column Name	Derived Column	Expression	Data Type	
DiscontinuedInt	<add as new column>	Discontinued == "False" ? 0 : 1	four-byte signed inte...	

Figura 27 - Transformação do atributo Discontinued Sede to INT

4.2.2.3 Gestão de Valores Nulos

Serão garantidos processos de validação e consistência em campos de dados importantes, como por exemplo na tabela *Customer*, o *CityId*, o *RegionId* e o *CountryCode*. Registos com valores nulos nestes campos, encontrados nos Lookups, serão desviados para tabelas específicas de erros, tal como é possível observar na figura 28, na *Staging Area*, neste caso o *CustomerDQP*, juntamente com logs para investigação manual e possíveis correções, tal como é possível observar na figura 29.

4.2.3 Carregamento

Após a preparação dos dados para as tabelas de dimensões, estas são carregadas em primeiro lugar, garantindo a preservação do histórico através da implementação das SCD Tipo 2, quando aplicável, e utilizando chaves estrangeiras (FK) para assegurar a integridade referencial. O carregamento é realizado sequencialmente para cada tabela de dimensão, conforme apresentado na Figura 31, **DimEmployee**, **DimSupplier**, **DimShippers**, **DimShipTo**, **DimCustomer** e **DimProduct**. Após a conclusão do carregamento das dimensões, as restrições de chaves estrangeiras na tabela de factos (**FactOrder**) são temporariamente desativadas para otimizar o processo de inserção de dados. Por fim, os dados da tabela de factos são carregados, e as restrições de chaves estrangeiras são recriadas para garantir a consistência e integridade do modelo de dados.

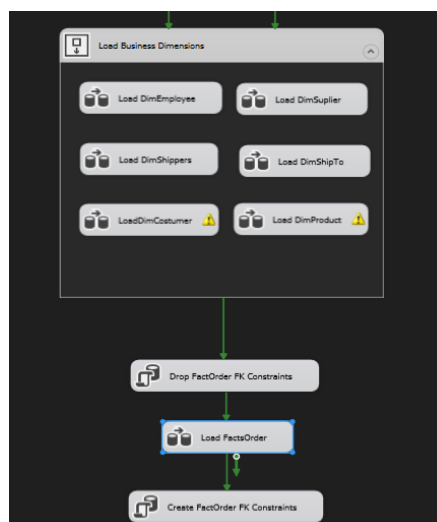


Figura 31 - Carregamento tabelas de Dims e de Factos

5. Análises Multidimensionais

As 10 análises multidimensionais, criadas pelo grupo, e realizadas sobre o armazém de dados criado, utilizando um cubo de dados desenvolvido no **Analysis Services Project do Visual Studio** foram:

11. Valores totais com desconto das encomendas efetuadas no segundo semestre de 2021, com possibilidade de análise detalhada (i.e., drill down) ao nível do trimestre e do mês, detalhados por cidade do cliente e por produto.
12. Quantidades de produtos encomendados no 3º trimestre de 2022, detalhadas por transportador e por país do cliente, com possibilidade de análise detalhada (i.e., drill down) ao nível do funcionário responsável.
13. Valores totais de vendas com e sem desconto durante o inverno de 2021, detalhados por trimestre, país de expedição e fornecedor do produto, com possibilidade de análise agregada (i.e., roll up) ao nível da categoria do produto.

14. Valores totais de fretes (em libras) no 2º semestre de 2022, detalhados por transportador, por mês e por região de expedição, com possibilidade de análise detalhada (i.e., drill down) ao nível país do cliente e da cidade do cliente.
15. Valores totais de descontos aplicados no dia 15 de cada mês do ano de 2021, detalhados por categoria do produto e por fornecedor, com possibilidade de análise detalhada (i.e., drill down) ao nível da cidade do cliente e do produto.
16. Valores totais com desconto e respectivas quantidades de encomendas processadas no verão de 2022, detalhadas por funcionário responsável, por região do cliente e por transportador, com possibilidade de análise agregada (i.e., roll up) ao nível do trimestre.
17. Quantidades de unidades encomendadas no primeiro quadrimestre de 2022, detalhadas por fornecedor do produto, por mês e por país de expedição, com possibilidade de análise detalhada (i.e., drill down) ao nível da cidade.
18. Valores totais sem desconto das encomendas efetuadas durante o 4º trimestre de 2021, detalhadas por país de expedição, por funcionário responsável e pela categoria do produto, com possibilidade de análise detalhada (i.e., drill down) ao nível do produto.
19. Valores totais com e sem desconto de encomendas efetuadas durante os fins de semana de 2022, detalhados por cliente, cidade de expedição e categoria do produto, com possibilidade de análise detalhada (i.e., drill down) ao nível do produto e do fornecedor.
20. Quantidades de unidades encomendadas durante feriados nacionais em 2021, detalhadas por funcionário responsável, por mês e por país de expedição, com possibilidade de análise agregada (i.e., roll up) ao nível do trimestre.

6. Trabalho realizado

Para a primeira parte deste trabalho, adotou-se uma abordagem colaborativa, sem divisões rígidas de responsabilidades, permitindo que ambos os membros da equipa se apoiassem mutuamente ao longo de todo o processo. Apesar disso, algumas contribuições individuais se destacaram: Afonso Cruz teve maior ênfase na análise e elaboração do modelo dimensional, enquanto André Conceição se destacou na construção do *Logical Data Map*.

Na segunda parte do trabalho, a dinâmica de colaboração manteve-se semelhante à da primeira parte, com ambos os membros trabalhando de forma integrada e equilibrada.

Em resumo, acreditamos que o empenho e o esforço de ambos os alunos foram equivalentes e complementares, sendo fundamentais para o sucesso deste trabalho. Essa colaboração mútua foi essencial para superar os desafios e alcançar os objetivos propostos com qualidade.

7. Conclusão

A execução deste armazém de dados constitui uma solução abrangente e completa, que integra e consolida dados de dois sistemas operacionais distintos para análises mais eficazes. Ao adotar o modelo de *Kimball*, juntamente com processos de ETL bem projetados e implementados, garantimos a consistência e qualidade dos dados carregados no armazenamento.

Os processos de transformação e limpeza resolverão questões cruciais, incluindo valores em falta, redundância de dados e informações incorretas nos sistemas fonte. Com esta infraestrutura, a empresa estará preparada para realizar análises multidimensionais detalhadas, melhorar o planejamento estratégico e promover o desenvolvimento organizacional. Este esforço estabelece as bases para a exploração contínua dos dados e o aumento da eficiência operacional, por meio de uma abordagem produtiva à análise de dados.