



Instituto Superior de
Engenharia do Porto

Relatório ARMDD

André Conceição | 12008

Afonso Cruz | 1240434

Índice

2.	Objetivos	3
3.	Modelo Dimensional.....	4
3.1	Identificação do processo de negócio	4
3.2	Nível de Granularidade	4
3.3	Seleção das Dimensões Relevantes.....	4
3.3.1	<i>Dimensão Produto (DimProduct)</i>	5
3.3.2	<i>Dimensão Cliente (DimCustomer)</i>	6
3.3.3	<i>Dimensão Funcionário (DimEmployee)</i>	6
3.3.4	<i>Dimensão Data (DimDate)</i>	7
3.3.5	<i>Dimensão Transportador (DimShipper)</i>	7
3.3.6	<i>Dimensão Enviado Para (DimShipTo)</i>	8
3.3.7	<i>Dimensão Local da Empresa (DimCompanyLocal)</i>	8
3.3.8	<i>Dimensão Moeda (DimCurrency)</i>	8
3.4	Design da Tabela de Factos	9
3.5	Modelo dimensional completo	10
3.6	Estratégias de <i>Slowly Changing Dimensions (SCD)</i>	10
4.	Operações de Transformação e Limpeza de Dados	12
4.1	Transformações	12
4.1.1	<i>Conversão de Formatos de Data</i>	12
4.1.2	<i>Separação de Atributos Compostos</i>	12
4.1.3	<i>Normalização da Representação do País</i>	12
4.2	Limpeza de dados	13
4.2.1	<i>Gestão de Valores Nulos</i>	13
4.2.2	<i>Remoção de duplicados</i>	13
5.	Staging Area	14
5.1	Descrição do processo etl	14
5.2	SUBtópicos do Processo etl.....	14
5.2.1	<i>Extração</i>	14
5.2.2	<i>Transformação</i>	14
5.2.3	<i>Carga</i>	14
6.	Conclusão	15

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 de vendas:** 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)
ProductID
ProductName
CategoryName
CategoryDescription
QuantityPerUnit
UnitPriceUK
UnitPriceUSA
UnitsInStockUK
UnitsInStockUSA
UnitsOnOrderUK
UnitsOnOrderUSA
Discontinued
SupplierID
SupplierCompanyName
SupplierContactName
SupplierContactTitle
SupplierAddress
SupplierCity
SupplierRegion
SupplierPostalCode
SupplierCountry
SupplierPhone
SupplierFax
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.

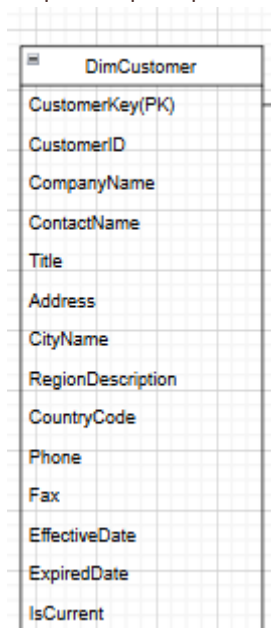


Figura 2 -DimCustomer

3.3.3 Dimensão Funcionário (DimEmployee)

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

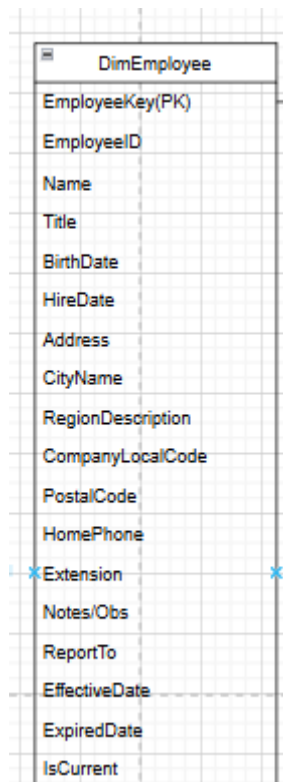
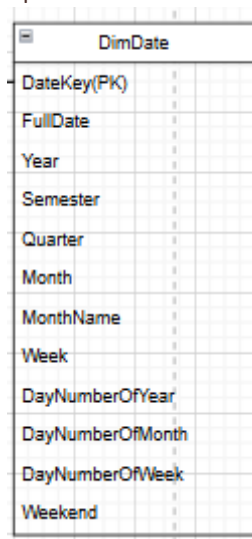


Figura 3 - DimEmployee

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.

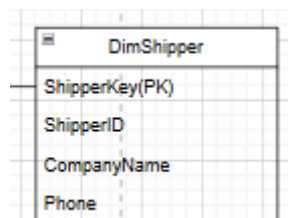


DimDate	
DateKey(PK)	
FullDate	
Year	
Semester	
Quarter	
Month	
MonthName	
Week	
DayNumberOfYear	
DayNumberOfMonth	
DayNumberOfWeek	
Weekend	

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.

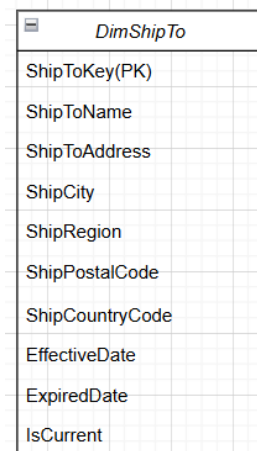


DimShipper	
ShipperKey(PK)	
ShipperID	
CompanyName	
Phone	

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.

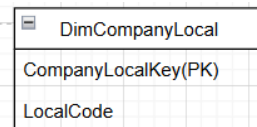


DimShipTo	
ShipToKey(PK)	
ShipToName	
ShipToAddress	
ShipCity	
ShipRegion	
ShipPostalCode	
ShipCountryCode	
EffectiveDate	
ExpiredDate	
IsCurrent	

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.

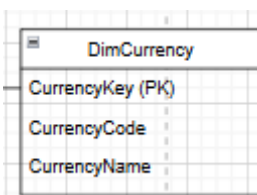


DimCompanyLocal	
CompanyLocalKey(PK)	
LocalCode	

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

3.4 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 total sem desconto;
- 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.

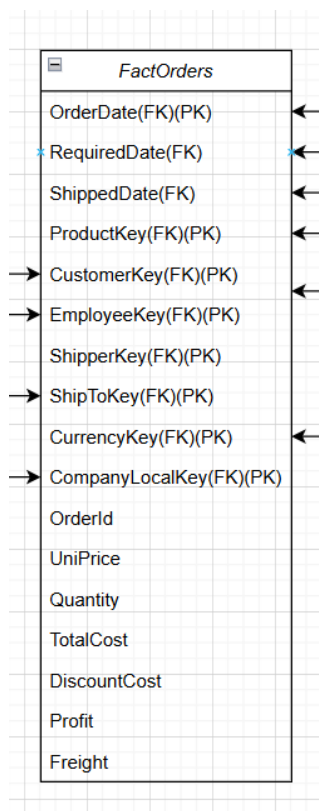


Figura 9 - FactOrders

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:

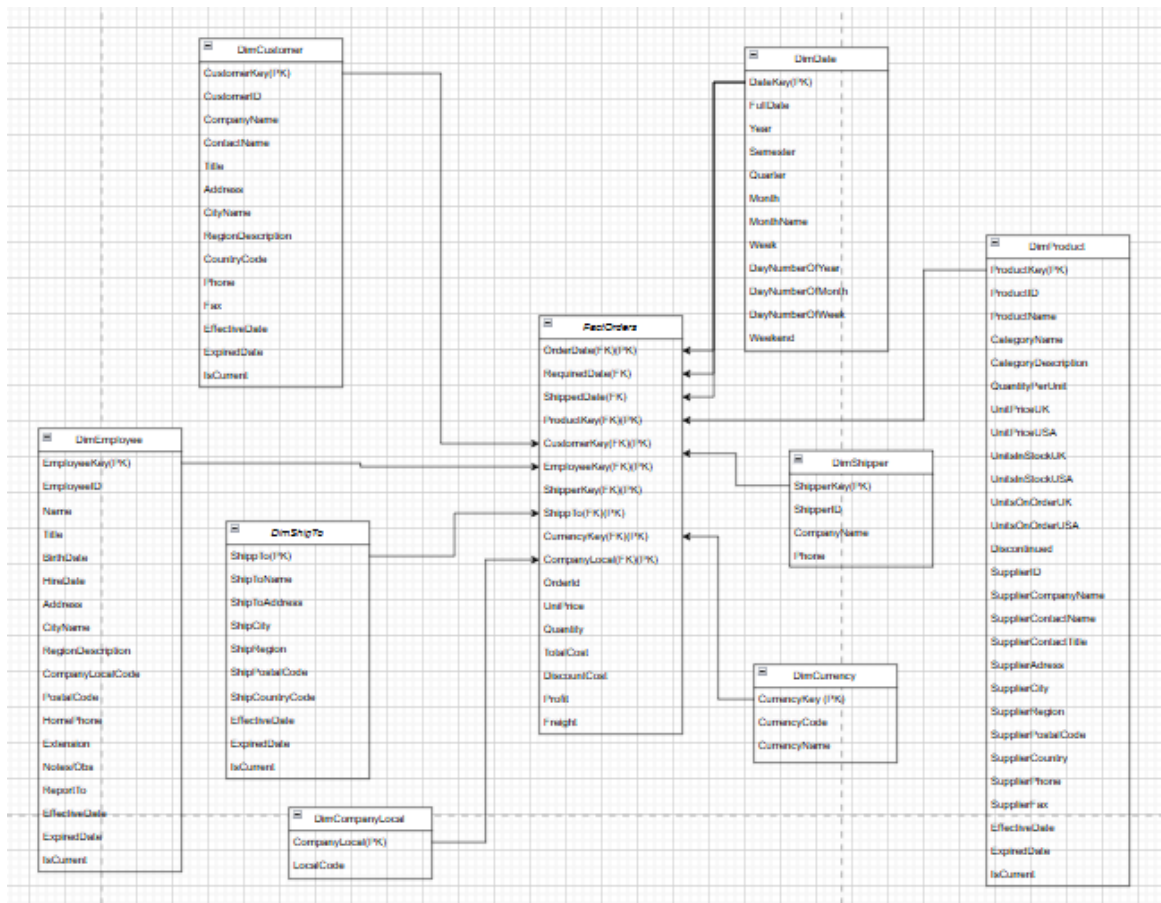


Figura 10 - Modelo Dimensional Completo

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** e **DimProduct**. Para implementar a estratégia SCD Tipo 2, foram incluídas colunas específicas nas dimensões:

- **EffectiveDate:** Data de início de validade do registro.
- **ExpiredDate:** Data de expiração do registro (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.

4. Operações de Transformação e Limpeza de Dados

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.

4.1 TRANSFORMAÇÕES

4.1.1 Conversão de Formatos de Data

Os dados adotados e utilizados pelos sistemas da sede (EUA) usam o formato de data *MM-DD-AAAA*, enquanto os dados da subsidiária (UK) utilizam o formato *AAAA-MM-DD*. Desta maneira, para unificar a base de dados, foi decidido que todas as datas serão representadas no formato padrão ISO *AAAA-MM-DD*, pois é amplamente preferido em sistemas globais. Esta medida garante análises temporais consistentes, eliminando assim possíveis erros provenientes de variações de representação.

4.1.2 Separação de Atributos Compostos

Na estrutura da subsidiária os nomes dos funcionários são armazenados em três campos *TitleOfCourtesy*, *FirstName* e *LastName*. Contudo na estrutura da sede o nome apenas existe um campo, que é o campo *Name* que irá conter estes três atributos, com o objetivo de manter a conformidade.

4.1.3 Normalização da Representação do País

Na sede, os países são registados com uma abreviatura do nome, por exemplo PT para Portugal, contudo na subsidiária estes registos têm o nome completo. Para garantir a uniformidade, todos os registos serão consolidados no formato de código de dois caracteres, conforme é definido pela norma internacional ISO 3166-1. Facilitando assim a realização de análises e assegurando uma representação geográfica uniforme. Existem também dados que no sistema da subsidiária 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*. Desta forma passará apenas a aparecer como *yes/no*, de forma a simplificar e normalizar os dados.

4.2 LIMPEZA DE DADOS

4.2.1 Gestão de Valores Nulos

Serão garantidos processos de validação e consistência em campos de dados importantes, como o *ProductName*, *CustomerID*, entre outros. Registos com valores nulos nestes campos serão desviados pelas tabelas específicas de erros na *Staging Area*, juntamente com logs para investigação manual e possíveis correções.

4.2.2 Remoção de duplicados

Os dados que estejam duplicados serão identificados e eliminados no processo de extração, devido a serem cruciais para evitar o enviesamento ou informações incorretas quando forem realizadas análises.

5. Staging Area

5.1 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, ou então através do carregamento de uma base de dados.
- 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.
- Carga (*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.

5.2 SUBTÓPICOS DO PROCESSO ETL

5.2.1 Extração

A informação será obtida através da leitura dos CSVs indexados, 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.

5.2.2 Transformação

Tal como foi referido no tópico 4 serão realizadas todas essas transformações e de limpeza dos dados.

5.2.3 Carga

Após a preparação dos dados para as tabelas de dimensões, estes serão carregados primeiramente, utilizando chaves vizinhas (FK) e mantendo o histórico através das SCD tipo 2. Mal as tabelas de dimensões estejam carregadas, proceder-se-á ao carregamento da tabela de factos.

6. Trabalho realizado

Para a realização deste trabalho, não foram estabelecidas divisões rígidas de responsabilidades, optando por uma abordagem colaborativa em que ambos os membros da equipa se apoiaram mutuamente ao longo de todo o processo. No entanto, o aluno Afonso Cruz se destacou mais na análise e elaboração do modelo de domínio, enquanto o aluno André Conceição se destacou mais na elaboração do Logical Data Map. 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 planeamento 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.