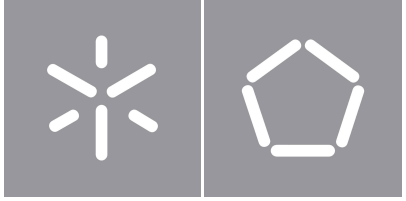
Universidade do Minho
Escola de Engenharia

Joana Rita Araújo Mota

Desenvolvimento de um Sistema de Data Warehousing para Consolidação de Contas

março 2024



Universidade do Minho
Escola de Engenharia

Joana Rita Araújo Mota

Desenvolvimento de um Sistema de Data Warehousing para Consolidação de Contas

Dissertação de Mestrado
Mestrado em Engenharia de Sistemas

Trabalho efetuado sob a orientação de
Professor Doutor Orlando Manuel Oliveira Belo

março 2024

Direitos de Autor e Condições de Utilização do Trabalho por Terceiros

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

Licença concedida aos utilizadores deste trabalho:

CC BY-NC

https://creativecommons.org/licenses/by-nc/4.0/

"Tenho em mim todos os sonhos do mundo."

Fernando Pessoa (Álvaro de Campos) *in* "Tabacaria".

Agradecimentos

Esta dissertação é muito mais do que um projeto individual. É o reflexo de um percurso repleto de aprendizagens, desafios e colaborações enriquecedoras. Durante esta jornada académica, várias pessoas e entidades foram fundamentais para o meu crescimento académico e pessoal. Por isso, gostaria de expressar um agradecimento sincero a todos que contribuíram de alguma forma para a realização desta dissertação:

Ao prof. Doutor Orlando Belo, um obrigada muito especial, pela disponibilidade, apoio e confiança, pelos ensinamentos e conselhos ao longo desta jornada. A sua orientação foi essencial neste último passo do meu percurso académico, e não poderia ter escolhido uma pessoa melhor para me guiar.

À Universidade do Minho, docentes e funcionários, por todos os recursos e materiais disponibilizados que foram imprescindíveis ao longo destes anos.

Aos meus amigos, e pessoas que se cruzaram ao longo do meu percurso académico, pelos bons momentos compartilhados, pelo apoio e aprendizagem mútua. E por fazerem destes anos os melhores da minha vida.

À minha família, em especial, à minha mãe e irmão pelo constante apoio e incentivo em todo o meu percurso. Por acreditarem em mim, por me darem a mão e por serem o meu porto seguro em todos os momentos da minha vida.

Ao Sérgio, pelo suporte e apoio incondicional, pelo companheirismo, pelos conselhos e constante ajuda durante todo o meu percurso. Por ser especial e por tudo aquilo que representa na minha vida. Obrigada por seres a minha âncora.

Declaração de Integridade

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Assinado por: **Joana Rita Araújo Mota**
Num. de Identificação: 30295201
Data: 2024.03.01 14:22:58+00'00'

Universidade do Minho, Braga, março 2024

Resumo

Desenvolvimento de um Sistema de Data Warehousing para Consolidação de Contas

Nas últimas décadas, os dados têm vindo a mudar o mundo a grande velocidade, especialmente, no que concerne ao mercado empresarial. O aumento exponencial da quantidade de dados disponíveis nas empresas implicou que estas procurassem soluções e estratégias para ganhar algumas vantagens competitivas face à sua concorrência. Deste modo, as empresas focaram-se essencialmente na análise aos seus dados, tanto depressa quanto possível, de uma forma eficiente para que pudessem retirar *insights* para suportar os seus processos de tomadas de decisão.

Nesta dissertação, apresenta-se e descreve-se o desenvolvimento de um sistema de *data warehousing* para suporte à consolidação de contas, cujo processo visou otimizar e racionalizar o processo e facilitar a análise de informação de uma empresa. Após um levantamento de requisitos, realizado em conjunto com os agentes de decisão, procedeu-se à modelação do sistema a desenvolver. Concebeu-se um sistema de ETL específico para o *data warehouse* para a consolidação de contas e fez-se o devido armazenamento dos dados coletados e preparados. Posteriormente ao povoamento, os dados foram encaminhados para uma ferramenta de visualização e análise de dados, de forma a alimentar um conjunto de *dashboards* especificamente desenvolvidos para retirar *insights* valiosos que pudessem auxiliar a gestão do negócio da empresa. No final dos trabalhos desta dissertação, obteu-se um sistema para suporte à decisão, automatizado, com capacidade para realizar todas as operações necessárias e de suporte a qualquer processo de análise de dados, desde a fase de extração dos dados, das diversas fontes de informação disponíveis até à sua colocação numa plataforma de análise de dados, concebida especialmente para suporte à consolidação de contas de uma empresa.

Palavras-chave Análise de Dados, *Business Intelligence*, *Dashboards*, *Data Warehouse*, *Data Warehousing*, ETL, Visualização de Dados

Abstract

Development of a Data Warehousing System for Consolidation Accounting

In recent decades, data has been changing the world at a great speed, especially in the business market. The exponential increase in the amount of data available in companies has meant that they seek solutions and strategies to gain some competitive advantages over their competition. In this way, companies have focused essentially on analyzing their data, as quickly and efficiently as possible, so that they can extract insights to support their decision-making processes.

In this dissertation, the development of a data warehousing system to support account consolidation is presented and described, a process aimed at optimizing and rationalizing the process and facilitating the analysis of information in a company. After a requirement survey, carried out in conjunction with decision-makers, the modeling of the system to be developed was carried out. A specific ETL system was designed for the data warehouse for account consolidation, and the collected and prepared data was properly stored. After the population, the data was forwarded to a data visualization and analysis tool, in order to feed a set of dashboards specifically developed to extract valuable insights that could assist in the company's business management. At the end of the work of this dissertation, an automated decision support system was obtained, capable of performing all necessary operations and supporting any data analysis process, from the data extraction phase, from various information sources available, to its placement on a data analysis platform, specially designed to support the consolidation accounting.

Keywords Business Intelligence, Dashboards, Data Analysis, Data Visualization, Data Warehouse, Data Warehousing, ETL

Conteúdo

Lista de Figuras

Lista de Tabelas

Acrónimos

AIC Akaike Information Criterion.

ARIMA Autoregressive Integrated Moving Average.

BIC Bayesian Information Criterion.

BPMN Business Process Model and Notation.

CSV Comma Separated Values.

EBITDA Earnings Before Interest, Taxes, Depreciation and Amortization.

ETL Extract, Transform, Load.

OLAP Online Analytical Processing.

OLTP Online Transaction Processing.

SARIMA Sazonal Autoregressive Integrated Moving Average.

SGBD Sistema de Gestão de Bases de Dados.

SNC Sistema de Normalização Contabilística.

SQL Structured Query Language.

1 Introdução

1.1 Enquadramento e Motivação

Atualmente, a área de *big data* tem-se tornado progressivamente mais presente no quotidiano, tendo, de certa forma, vindo a auxiliar as empresas a tomar melhores decisões. As organizações começam a perceber que podem utilizar os seus dados para melhorar as suas operações e tomar decisões mais conscientes, bem como definir estratégias de atuação especialmente orientadas para a melhoria do negócio. As empresas que adquiriram a capacidade de transformar esses dados em informação e conhecimento têm vindo a se destacar, através da utilização desse conhecimento nos seus processos de tomada de decisões. Consequentemente, obtiveram vantagens competitivas (Vercellis, 2008).

A *business intelligence* representa um conjunto de ferramentas muito vasto e diverso que ajudam as empresas a recolher, organizar, analisar e a interpretar dados, que visam aumentar as suas capacidades de tomada de decisão, transformando os dados em conhecimento (Negash e Gray, 2003). Neste sentido, o processo de implementação de um sistema de *business intelligence* passa por identificar e extrair dados relevantes das fontes de dados, de seguida, transformar esses dados e carregá-los para um sistema de *data warehousing*, que trata de armazená-los para depois serem produzidos *dashboards* e relatórios que possam fornecer informação e conhecimento útil para os gestores. No entanto, a implementação de sistemas de *business intelligence* apresenta grandes desafios, tais como a necessidade de lidar com grandes volumes de dados, muitas vezes, heterogéneos, e desenvolver habilidades técnicas necessárias para o tratamento de dados e garantir a ética, a segurança e a privacidade dos mesmos (Ramos et al., 2019). Para além disto, a falta de informação sistemática, de qualidade e confiável conduz a erros e à perda de oportunidades de negócio (Costa, 2012). Por isso, quanto melhor for a qualidade dos dados, melhor será a informação produzida.

Atualmente, as empresas que investem em ferramentas de *big data* para suportar os seus processos de negócio tendem a ter ganhos de produtividade e lucros acima da sua concorrência (Côrte-Real, 2022). Através destas ferramentas, utilizam-se os dados disponíveis nas organizações para apresentar informação relevante para os processos de tomada de decisão, dado que a informação poderá fornecer

aos gestores um conjunto de indicadores sobre o negócio, que poderão revelar o que aconteceu no passado e delinear possíveis situações para o futuro (Santos e Ramos, 2006). Assim, acredita-se que estes sistemas são uma ferramenta estratégica para o crescimento económico, que permitem a competitividade e o desenvolvimento inovador de muitas organizações, uma vez que estes podem contribuir para acelerar o processo de tomada de decisão (Olszak, 2022).

O grupo empresarial multinacional que deu suporte a este trabalho atua em diversos setores, nomeadamente, serviços financeiros, tecnologia, imobiliário, retalho alimentar e moda, saúde e bem-estar. O volume das ações de negócio que desenvolve diariamente é enorme. Por conseguinte, a quantidade de informação que gera e que possui é enorme, o que torna maior a complexidade das suas operações e leva a grandes desafios no que toca a manter o seu negócio. Assim, uma das necessidades desta organização é manter-se, constantemente, a par da situação financeira, de cada uma das suas empresas, com o objetivo de analisar o desempenho atual e futuro das mesmas.

A consolidação de contas é uma área importante para as empresas, dado que trata de um conjunto de técnicas que visa elaborar as contas consolidadas de um grupo empresarial como se de uma única empresa se tratasse (Adriana e Silva, 2018). Através de uma análise financeira, é possível recolher e estudar a informação de uma empresa para avaliar se os seus objetivos estão a ser cumpridos, particularmente a maximização do seu lucro. Desta forma, as empresas precisam de se adaptar às mudanças e à evolução das tecnologias para que consigam trabalhar com novos mecanismos e ferramentas que possibilitem extrair e aceder aos seus dados de forma eficaz e analisá-los com a finalidade de tomar decisões eficientes que possam gerar valor para a empresa. Nos últimos anos, apesar de algum desenvolvimento na área da consolidação ao nível dos sistemas de informação, ainda se verificou que esta área envolve um grande manuseamento e pouca otimização nos seus sistemas, o que acaba por dificultar as tarefas diárias e a perceção dos rendimentos e gastos dos seus negócios.

Assim, a principal motivação desta dissertação é garantir que o departamento da consolidação de contas, da empresa em questão, consiga otimizar e racionalizar o processo de consolidação de contas, ao possibilitar a recolha, a integração e a análise de dados provenientes de várias fontes, através de ferramentas de *business intelligence*. O processo de implementar um sistema de *business intelligence* pode ser bastante complexo. Porém, se houver um planeamento adequado e um levantamento de requisitos bem realizado é possível desenvolver um sistema de análise de dados robusto e eficaz que seja capaz de ajudar a empresa a tomar decisões mais sustentadas e estratégicas. Desta forma, a organização conseguirá retirar *insights* sobre cada uma das suas empresas, o que lhe permitirá fornecer estratégias para a sua evolução e, consequentemente, facilitar nos seus processos de tomada de decisão.

1.2 Objetivos

O objetivo principal desta dissertação é a aplicação de ferramentas de *business intelligence* na área da consolidação de contas, de forma a facilitar os processos de análise de informação e, consequentemente, extrair conhecimento importante para suportar a execução e o desenvolvimento dos processos que usualmente ocorrem nessa área de trabalho. Uma vez que, a consolidação de contas é uma área com grandes volumes de dados envolvendo frequentemente contas de uma empresa-mãe e das suas subsidiárias, é necessário arranjar formas automatizadas para tornar o processo de análise de informação mais eficiente e eficaz.

Neste projeto, implementou-se um sistema capaz de extrair os dados necessários para os processos de consolidação de uma dada empresa e, seguidamente, fazer todas as transformações de dados essenciais, para que, no final, estes possam ser carregados num sistema de *data warehousing*. Posteriormente, estes dados são encaminhados para o ambiente de uma ferramenta de visualização de dados com o objetivo de alimentar um conjunto de *dashboards* contendo indicadores e métricas intuitivas, de forma a proporcionar uma fácil compreensão da informação tratada e orientada para os processos de tomada de decisão da empresa.

1.3 Metodologia de Investigação

De maneira a atingir os objetivos propostos, adotou-se algumas linhas de orientação de alguns métodos que se consideram adequados para realizar este trabalho de dissertação. Assim, inicialmente, foi necessário estudar o domínio de trabalho para analisar e selecionar ferramentas de desenvolvimento, métodos e estratégias de pesquisa apropriados para desenvolver o sistema de suporte à decisão. A realização dos trabalhos desta dissertação orientou-se através da utilização de duas metodologias: a metodologia *Design Science Research* (DSR) (Hevner et al., 2004) enquanto metodologia de investigação e de trabalho, e a metodologia de *Kimball* (Kimball e Ross, 2002) enquanto metodologia de desenvolvimento do sistema pretendido.

A metodologia de investigação foi considerada eficaz para a realização dos trabalhos e adequada para a investigação de casos de aplicação prática (Hevner et al., 2004). Esta metodologia inclui algumas etapas que devem ser seguidas no desenvolvimento do trabalho, tais como (Geerts, 2011):

- **Identificação do problema e motivação** - Identificação dos problemas de pesquisa a serem resolvidos e justificação adequada da solução.
- **Definição dos objetivos da solução** - Identificação dos objetivos da solução baseando-se nos

problemas de investigação encontrados.

- **Desenho e desenvolvimento da solução** - Criação de um mecanismo que resolva o problema - desenho de arquiteturas, modelos, métodos para o desenvolvimento da solução que visa resolver o problema.
- **Demonstração** - Provar que a solução funciona através da resolução de um ou mais casos do problema.
- **Avaliação** - Avaliação da solução e comparação dos resultados alcançados.
- **Comunicação** - Comunicação do problema e a sua solução desenvolvida para um público.

A metodologia adotada no desenvolvimento do sistema foi a proposta por *Ralph Kimball* para o desenvolvimento de sistemas de *data warehousing*. Esta metodologia apresenta um conjunto bem definido de etapas de trabalho (Figura 1) - ciclo de vida dimensional do negócio - a realizar em qualquer processo de implementação de sistemas de *data warehousing* (Kimball e Ross, 2002).

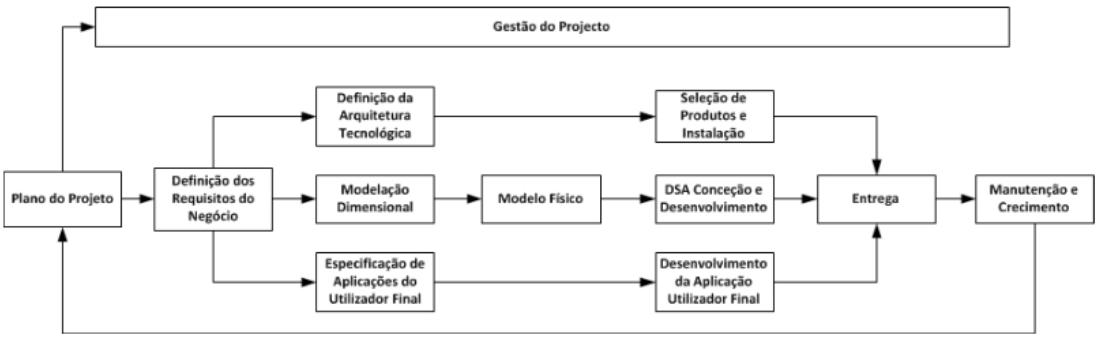


Figura 1: Ciclo de vida dimensional de negócio por Kimball e Ross (2002)

Uma das etapas mais importantes do processo de construção de um sistema de *data warehousing* é a modelação dimensional, uma vez que é aqui que se define o propósito do sistema de *data warehousing* e quais os tipos de análises a realizar. *Kimball* e *Ross* apresentaram o método dos **"4 passos"** (Kimball e Ross, 2002) de forma a ajudar a cumprir esta fase. Os 4 passos são (Kimball e Ross, 2002):

- Selecionar a área de suporte à decisão a analisar, em que é realizado um levantamento de requisitos com o objetivo de compreender as necessidades do negócio. Este é um passo fundamental para o sucesso do sistema de *data warehousing*, uma vez que este deve ser capaz de responder às necessidades com a finalidade de apoiar a tomada de decisão.

- Definir o grão das tabelas de facto que permite analisar o nível de detalhe da informação. Quanto maior for a sua granularidade, maior é o detalhe dos dados, possibilitando assim análises mais profundas.

- Identificar as dimensões de análise que permitem fornecer contexto ao grão definido na etapa anterior. As dimensões fornecem informação detalhada relevante e útil para realizar análises em torno do grão da tabela de factos.

- Identificar as medidas a integrar na estrutura de cada facto, que, basicamente significa, definir medidas numéricas que quantificam um facto e que representam o desempenho, as métricas ou os valores a serem analisados. As medidas que são integradas numa tabela de factos são essenciais para a análise e são acompanhadas por várias dimensões que fornecem o seu contexto.

1.4 Trabalho Realizado

O desenvolvimento dos trabalhos desta dissertação assentou essencialmente em quatro fases. A saber, o conhecimento do problema e das suas necessidades, o estudo do tema, a modelação e o desenvolvimento do sistema de *data warehousing* a implementar e, por fim, a análise dos dados.

A primeira fase consistiu na compreensão do contexto do problema e das necessidades dos agentes de decisão envolvidos, que necessitavam de uma ferramenta capaz de armazenar grandes volumes de dados e que os disponibilizasse de forma interativa e dinâmica, de forma a possibilitar a realização de análises de dados para processos de suporte à decisão.

De seguida, a segunda fase do projeto caracterizou-se pelo estudo e exploração cuidada de todos os conceitos envolvidos no tema em questão, com o propósito de conhecer e compreender os diversos elementos integrantes de um sistema de *data warehousing*, bem como entender o funcionamento e a importância da área da consolidação de contas dentro de uma organização. Esta fase de estudo foi fulcral para absorver o conhecimento necessário para o desenvolvimento do sistema de *data warehousing*.

Na terceira fase começou-se por definir os principais requisitos que o sistema deveria responder, realizando-se uma matriz de decisão para delinear o modelo multidimensional final. Seguidamente, foram conhecidas as fontes de informação essenciais para realizar o mapeamento dos dados para o *data warehouse*. Criou-se o *data warehouse* conforme os requisitos e o modelo multidimensional e realizou-se o processo de **ETL** preparado para carregar os dados no *data warehouse* sempre que houver novas atualizações.

Por fim, na última fase os dados foram encaminhados para uma ferramenta de visualização de dados de forma a realizar *dashboards* dinâmicos que ajudem os utilizadores na elaboração das suas análises e

no suporte à tomada de decisão.

1.5 Estrutura da Dissertação

Para além do presente capítulo, esta dissertação contém mais quatro capítulos, nomeadamente:

- **Capítulo 2** - Sistemas para Análise de Dados -, que é dedicado ao enquadramento teórico dos temas abordados ao longo do projeto, de forma a permitir dar suporte ao desenvolvimento do sistema e fundamentar as decisões executadas durante a realização do mesmo.

- **Capítulo 3** - Um Sistema para Consolidação de Contas -, que apresenta um levantamento de requisitos a que o sistema deve responder para o seu sucesso e, seguidamente, descreve a modelação multidimensional, na qual caracteriza todos os elementos de dados do *data mart* a ser desenvolvido.

- **Capítulo 4** - Um *Data Warehouse* para Consolidação de Contas -, que detalha a implementação do sistema desenvolvido, descrevendo as fontes de dados e o mapeamento dos dados para o sistema, explica o funcionamento do processo de **ETL** implementado e, por fim, apresenta algumas análises e uma previsão aos dados, demonstrando alguns elementos gráficos relevantes.

- **Capítulo 5** - Conclusões e Trabalho Futuro -, que resume tudo o que foi discutido e realizado ao longo do processo de realização desta dissertação, bem como apresenta possíveis ideias do que poderá ser feito, futuramente, para melhorar o sistema desenvolvido.

2 Sistemas para Análise de Dados

2.1 Big Data

Antes do aparecimento dos sistemas e soluções *big data*, as organizações não tinham capacidade de armazenar todos os seus ficheiros por longos períodos, nem gerir com eficácia enormes volumes de dados, visto que as tecnologias tradicionais possuem armazenamento limitado e ferramentas caras com pouca flexibilidade (Oussous et al., 2018). Atualmente, as empresas começam a perceber a importância de usar cada vez mais os seus dados para apoiar as suas decisões estratégicas para o negócio. Na realidade, quanto maior for o volume de dados mais eficientes serão os seus processos de tomada de decisão, devido às possíveis correlações entre os diversos elementos de dados que seriam difíceis de encontrar se estes fossem analisados em pequenos volumes separados (Ularu et al., 2012). Desta forma, com o passar do tempo, os dados foram crescendo em grande escala, devido ao aumento de fontes de dados e de ferramentas de recolha de dados, que foram produzindo grandes volumes de dados (Magalhães, 2019). Consequentemente, as tecnologias tradicionais passaram a ser menos eficientes, uma vez que começaram a revelar um desempenho lento, incapacidade de escalar e menor precisão (Oussous et al., 2018). Assim, surgiu a necessidade de apostar em tecnologias e soluções para *big data* (Magalhães, 2019).

A definição exata do termo *big data* é quase impossível de encontrar, uma vez que não há um conceito único, devidamente formalizado (Kubina et al., 2015). Há uma variedade de definições diferentes que foram propostas por vários autores ao longo do tempo. Mas, na generalidade, *big data* refere-se a enormes quantidades de dados estruturados, semiestruturados e não estruturados produzidos pelas organizações, indivíduos e máquinas. Esses dados têm uma natureza tão complexa que requerem tecnologias poderosas e avançadas (Oussous et al., 2018).

Kubina et al. (2015) definem *big data* como um conjunto de dados cujo volume está além da capacidade das ferramentas tradicionais de base de dados para o recolher, armazenar, gerir e analisar. Da mesma forma, Davenport (2014) refere que *big data* trata de recolher e interpretar dados de grande dimensão, fornecidos pelo poder da computação que permite aceder e monitorizar informação de várias

fontes. Já Chen et al. (2014) abordam a questão da velocidade dos dados mencionando que o termo *big data*, geralmente, inclui quantidades enormes de dados não estruturados que precisam de ser analisados de forma rápida e em tempo real.

O volume de dados é uma das principais razões pelas quais o *big data* é tão importante, porque cada vez mais há dados e informação disponível, sendo notório a sua dimensão. No entanto, surgiram outras características, tais como: a variedade e a velocidade (Gandomi e Haider, 2015). Estas três características ficaram conhecidas pelos "Três V's" (volume, variedade e velocidade) para caracterizar o significado do conceito de *big data*.

O volume diz respeito à capacidade física que está a ser utilizada, ao tamanho do número de registos, de transações ou de tabelas que vão crescendo à medida que se vão recolhendo cada vez mais dados (Kubina et al., 2015). Com o volume de dados a aumentar, as empresas passam a ter o desafio de armazenar, processar e analisar de forma eficiente toda essa informação, o que se revela uma grande oportunidade para obter *insights* acerca dos seus negócios. Além da explosão de volumes de dados, existe uma grande variedade de dados. A variedade refere-se a um conjunto de dados heterogéneos na sua estrutura, uma vez que, hoje em dia, existem várias formas de criar dados através de pessoas para pessoas, de pessoas para máquinas e de máquinas para máquinas (Côrte-Real, 2022). Com os avanços tecnológicos é possível que as empresas utilizem vários tipos de dados, de várias fontes, estruturados, semiestruturados e não estruturados (Kubina et al., 2015). Para além disto, a velocidade com que os dados são gerados e transmitidos aumentou drasticamente, tornando-se, assim, importante para as empresas terem a capacidade de processar os dados de forma rápida e ágil para obter informações úteis e relevantes em tempo real (Oussous et al., 2018).

Mais tarde, foram identificadas e incluídas mais duas características, nomeadamente, a veracidade e o valor (Figura 2). A veracidade está relacionada com a qualidade e, consequentemente, confiança que se pode obter dos dados (Côrte-Real, 2022), pois se uma fonte apresenta muito ruído, as correlações e as interpretações dos dados serão mal realizadas, originando informação pouco verídica, o que pode conduzir a más decisões. Por último, o valor é o resultado do processamento de grandes volumes, uma vez que os dados que são extraídos, inicialmente, apresentam um valor reduzido em relação ao seu volume. Assim, após a junção de vários dados e, ao seu tratamento, é possível obter dados com maior valor, o que permitirá a analisar e retirar mais *insights* (Gandomi e Haider, 2015).

Posto isto, as soluções de *big data* trazem diversas vantagens competitivas para as empresas, uma vez que permitem obter uma maior transparência da informação e do conhecimento dentro da organização, criar novos modelos de negócio ou melhorar já os existentes, como por exemplo, criar segmentos mais detalhados dos clientes com objetivo de lhes direcionar os seus serviços e, sobretudo, apoiar a tomada

de decisão, identificando correlações úteis que possam estar omitidas (Kubina et al., 2015). De facto, as empresas que usufruam das tecnologias de *big data* estão mais perto de se tornarem líderes de mercado do que aquelas que utilizam as ferramentas tradicionais, acabando estas mesmas por ficar atrás dos seus concorrentes.

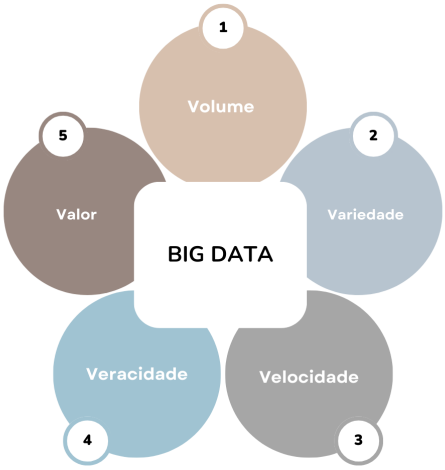


Figura 2: As cinco principais características do *big data*

2.2 Business Intelligence

A *business intelligence* é um conjunto de ferramentas que permitem às organizações transformar dados em informações valiosas para apoiar a tomada de decisão.

Segundo Negash e Gray (2003), os sistemas de *business intelligence* combinam processos de recolha de dados, de armazenamento de dados e de gestão do conhecimento com ferramentas analíticas para apresentar informação interna e competitiva aos agentes de decisão. Para Sezões e Oliveira (2006), o conceito de *business intelligence* é um amplo conjunto de tecnologias para apoio à tomada de decisão que permitem um acesso rápido, partilhado e interativo da informação, bem como a sua consequente análise e manipulação. Através destas ferramentas, os utilizadores podem descobrir tendências e transformar grandes quantidades de dados em conhecimento útil. Também Ranjan (2009) refere que as ferramentas de *business intelligence* são vistas como tecnologias que permitem a eficiência da execução do negócio, agregando valor à informação da organização e, consequentemente, à forma como essas informações são usufruídas.

O objetivo dos sistemas de *business intelligence* passa por aumentar a eficiência na tomada de decisão em todas as áreas de gestão, melhorar os processos e, as relações com os *stakeholders* e alcançar o sucesso nas empresas (Olszak, 2022). A importância do *business intelligence* é cada vez mais

notória nas organizações, uma vez que apresenta sistemas com capacidade de evidenciar informação útil praticamente em tempo real, sugerir mudanças fulcrais para o negócio, que podem resultar em vantagens competitivas. Desta forma, as organizações que têm ao dispor tecnologias de *business intelligence*, conseguem transformar os seus dados em conhecimento, com vista a uma tomada de decisão mais fundamentada e consciente.

Contudo, essas ferramentas precisam de ter a capacidade de extrair dados de diferentes fontes, de transformar esses mesmos dados de acordo com o objetivo da solução e de carregá-los num repositório de dados específico, usualmente um *data warehouse* (Pereira e Costa, 2016). Para tal, são desenvolvidas arquiteturas de sistemas de *business intelligence* que permitem traçar a conceção do sistema.

2.3 Sistemas de Data Warehousing

Os sistemas de *data warehousing* são infraestruturas de dados altamente sofisticadas, integrando um conjunto de componentes, tecnologias e processos que são envolvidos na criação e manutenção de um repositório de dados especializado num ou mais domínios do conhecimento: um *data warehouse*. Através de uma arquitetura de um ambiente de *data warehousing* é possível identificar os componentes que a integram, o relacionamento que existe entre eles e as funcionalidades de cada um deles (Dutra e Ciferri, 2002).

2.3.1 Arquitetura

Uma arquitetura típica de um sistema de *data warehousing* para dar suporte à tomada de decisão dentro de uma organização tem uma configuração como a que está apresentada na Figura 3. Como se pode ver, esta arquitetura está organizada em cinco camadas, cada uma delas possuindo um conjunto de componentes, processos e tecnologias específicos, que permitem, essencialmente, realizar tarefas de recolha, transformação, armazenamento e análise de dados.

O sistema começa com uma primeira camada relacionada com as fontes de dados, que suporta basicamente a operacionalidade do sistema. Normalmente, os dados que se pretendem analisar provêm de várias bases de dados operacionais dentro de uma organização, mas, em alguns casos, também de fontes de informação externas (Chaudhuri et al., 2011). Contudo, usualmente essas fontes de informação contêm dados com pouca qualidade e veracidade, o que implica a necessidade de se ter técnicas para tratamento de dados. Na camada seguinte, a transformação de dados, realizam-se os processos de **ETL** sobre os dados extraídos das fontes, utilizando ferramentas apropriadas, que possibilitam a integração, a limpeza, a transformação e o carregamento dos dados para as estruturas de armazenamento do sistema.

É, então, que se avança para a próxima camada que integra o *data warehouse* e os diversos *data marts* que armazenam os dados transformados da organização (Chaudhuri et al., 2011). A seguinte camada diz respeito a servidores intermédios que permitem aceder ao *data warehouse* com recurso a várias técnicas, como **OLAP**, que expressa eficientemente a visão multidimensional dos dados para operações de filtragem, de agregação e de detalhe das informações (Chaudhuri et al., 2011). Por fim, a última camada suporta, na fase final do sistema, a análise dos dados armazenados. Para tal, existem várias ferramentas de *front-end* que permitem aos utilizadores o acesso, a manipulação e a visualização da informação, de maneira que possam descobrir *insights* e acompanhar o desempenho do negócio.

É importante ressaltar que, a arquitetura de um sistema de *data warehousing* depende muito das necessidades de cada organização, bem como do volume de dados e dos recursos que esta tem disponíveis. Portanto, é necessário alinhar os objetivos do negócio com a conceção do sistema a implementar. Contudo, o ponto central de qualquer arquitetura de um sistema de *data warehousing* é o *data warehouse*, o qual é considerado como o coração de todo o sistema (Inmon, 2005), uma vez que é este que deve garantir que os dados apropriados estejam disponíveis para o utilizador final no momento certo.

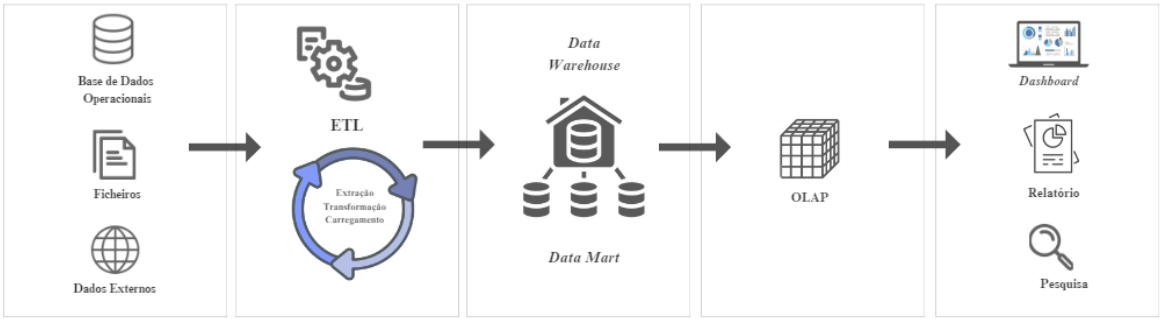


Figura 3: Arquitetura típica de um sistema de *data warehousing* - Figura - adaptada de Chaudhuri et al. (2011)

2.3.2 Sistemas de Base de Dados Operacionais

Para se poder implementar um sistema de *data warehousing* numa entidade é necessário que esta possua dados operacionais armazenados, visto que estes são a matéria prima para a elaboração do desenvolvimento do mesmo. As fontes de dados podem possuir uma variedade de formatos desde **SGBD** relacionais, orientados a objetos até documentos **CSV**, *excel*, *txt*, entre outros (Dutra e Ciferri, 2002).

Deste modo, numa qualquer organização os sistemas **OLTP** têm como finalidade registar dados diariamente em tempo real, o que implica que estes sistemas não estejam otimizados para realizar uma análise integrada dos dados. Esses registos diários são estruturados, repetitivos e consistem em transações isoladas (Chaudhuri e Dayal, 1997). Assim, é fácil de entender o porquê da necessidade de

desenvolver um sistema de *data warehousing* para conceber análises para suporte à tomada de decisão.

2.3.3 Data Warehouses

Segundo Dutra e Ciferri (2002), os *data warehouses* surgiram para colmatar a necessidade de separar ambientes operacionais e dos ambientes analíticos. No passado, a obtenção de informação importante e estratégica para o negócio era efetuada diretamente em bases de dados operacionais, o que implicava processos de extração de dados de diferentes várias fontes de dados heterogéneas. Situações como esta faziam com que acontecesse perda de credibilidade, inconsistências, ou falta de coerência nos dados e, ainda, por vezes, excessiva informação para a geração de relatórios e análises. Devido a estes problemas, os *data warehouses* emergiram como base para o desenrolar das atividades de suporte à tomada de decisão, permitindo extrair dados de várias fontes, realizar o seu tratamento e carregá-los para o seu ambiente, para fins de exploração e de análise de dados.

De forma genérica, um *data warehouse* é um repositório de dados, usualmente centralizado, que armazena dados operacionais devidamente trabalhados, provenientes de várias fontes, e que os torna disponíveis para os utilizadores finais, os agentes de decisão (Yessad e Labiod, 2016). Segundo *Kimball*, um *data warehouse* consiste numa "cópia" de dados transacionais estruturados para consultas e análises com o propósito de fornecer informações para apoiar a tomar decisão numa empresa (Yessad e Labiod, 2016). Por sua vez, Inmon (2005) define *data warehouse* como um conjunto de dados orientados ao assunto, integrados, temporais e não voláteis para suportar o processo de tomada de decisão. Considerando a sua perspetiva, um *data warehouse* caracteriza-se por ser (Inmon, 2005):

- **Orientado ao assunto** - os dados são organizados e repartidos por assuntos específicos de acordo com as necessidades dos utilizadores finais, o que permite uma análise de dados mais focada e direcionada para um determinado tema, uma vez que são excluídos dados irrelevantes para o processo de tomada de decisão.

- **Integrado** - os dados obtidos de várias fontes são integrados num único sistema, o que permite apresentar aos utilizadores uma visão única e abrangente do negócio.

- **Temporal** - os dados são relativos a períodos específicos, de forma a manter o histórico de todas as transações. O *data warehouse* deve assegurar isso através de uma dimensão temporal. O objetivo principal é fornecer informação histórica e atual dos dados aos utilizadores.

- **Não volátil** - os dados não devem ser modificados ou removidos, o que os torna estáveis dentro do sistema de *data warehouse*.

O desenvolvimento de um *data warehouse* justifica-se pela necessidade de integrar dados que estão distribuídos por diferentes fontes de dados (Sezões e Oliveira, 2006), bem como pela necessidade de *reporting* e de análise da informação para a tomada de decisão. De acordo com Dolk (2000), o principal objetivo de um *data warehouse* concentra-se na análise de dados para responder às necessidades da organização, ao contrário de uma base de dados tradicional que regista apenas os dados operacionais. Consequentemente, os *data warehouses* apresentam-se separados das bases de dados operacionais da organização para garantir a autonomia devida das suas funcionalidades. Enquanto uma base de dados operacional suporta sistemas **OLTP**, cujo propósito se centra em registar diariamente transações de novos registos de dados, um *data warehouse* sustenta sistemas analíticos - **Online Analytical Processing (OLAP)** - direcionados para o suporte à tomada de decisão (Chaudhuri e Dayal, 1997).

Estes sistemas analíticos desenvolvem análises focadas em áreas específicas de negócio da organização. Essas áreas de negócio, normalmente, são representadas por subconjuntos organizados dentro do *data warehouse*, designados de *data marts*. Os *data marts* são repositórios de dados multidimensionais, mais pequenos que os *data warehouses*, que juntam um conjunto de tabelas dimensionais de suporte a um determinado processo de negócio Kimball e Ross (2002). Trata-se, assim, de repositórios do *data warehouse* orientados a uma determinada unidade de negócio específica dentro de uma organização. Como se pode observar através da Figura 4, um *data warehouse* pode apresentar vários *data marts* que contêm um determinado assunto dentro do contexto do próprio *data warehouse*.

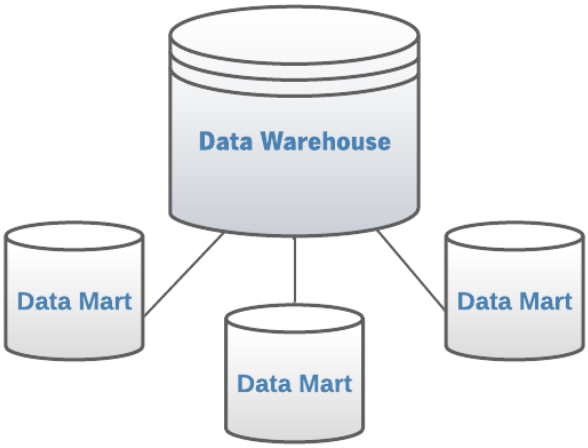


Figura 4: Uma ilustração de um *data warehouse* e seus *data marts*

2.3.4 Modelação Multidimensional

A construção de um *data warehouse* requer a adoção de técnicas de *design* e implementação diferentes daquelas subjacentes aos sistemas de base de dados operacionais (Golfarelli et al., 1998). Uma vez que

o propósito de um *data warehouse* passa pela execução de consultas e carregamento de dados, este necessita de métodos de acesso e técnicas de processamento que possibilitam a realização de consultas altamente eficientes, para facilitar posteriormente os processos de análise de informação (Chaudhuri e Dayal, 1997). Para isso, recorre-se à modelação multidimensional para organizar e estruturar os dados do *data warehouse*, assim como assegurar que todas as necessidades do negócio são respondidas pelo sistema de *data warehousing*. Este tipo de modelação traduz-se num modelo de dados de fácil compreensão e de utilização, que possibilita ao *data warehouse* dar respostas rápidas a consultas mais complexas e aprofundadas por parte dos utilizadores (Vercellis, 2008). O modelo multidimensional consiste num conjunto de tabelas de dimensão e de factos, que são dispostos num dos seguintes esquemas:

- Esquema em estrela.
- Esquema em floco de neve e esquema em constelação.

Estas designações são atribuídas aos esquemas para caracterizar a forma como as tabelas de factos e as dimensões estão organizadas. O esquema em estrela é o mais comum nas configurações dos elementos de dados de *data warehouses*.

As tabelas de dimensão acolhem as "variáveis" de análise que fornecem contexto aos dados de uma tabela de factos (Kimball e Ross, 2002). Estas tabelas possuem atributos que são definidos de acordo com as necessidades apresentadas pelos utilizadores e, como tal, pretendem categorizar uma tabela de factos através de elementos de dados relacionados com clientes, produtos, localização, período de tempo, entre outros. Por norma, estes atributos apresentam-se em formato textual e possuem um número inferior de registos, quando comparados com a tabela de factos.

Em relação à tabela de factos, esta trata de armazenar medidas de desempenho relacionadas com a unidade de negócio que se pretende que seja analisada (Kimball e Ross, 2002). Nesta tabela encontram-se as relações com as tabelas de dimensão, assim como os atributos relativos aos indicadores de análise que poderão corresponder a quantidade vendida, preço unitário, receita, etc.

A configuração de um esquema em estrela pressupõe a criação de tabelas de dimensão, que ficam relacionadas entre si através de uma tabela de factos com a utilização de chaves estrangeiras (Han et al., 2022). A tabela de factos armazena todos os valores de medição para cada combinação possível (Bellatreche, 2003). Geralmente, este modelo de dados multidimensional é o mais utilizado para modelar um *data warehouse* (Chaudhuri e Dayal, 1997). Na Figura 5 está apresentado um exemplo de um esquema em estrela.

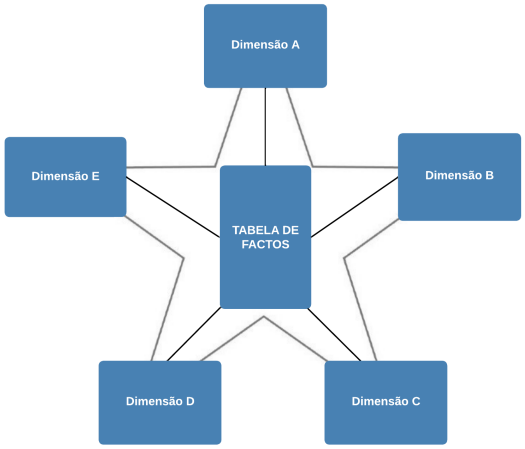


Figura 5: Ilustração de um esquema em estrela

Um esquema em floco de neve (Figura 6) trata-se de um esquema mais complexo do que um esquema em estrela. Usualmente, resulta da normalização das dimensões, agrupando os dados hierarquicamente em várias tabelas, evitando assim redundâncias (Han et al., 2022). Por outro lado, não surgem anomalias e o esquema é mais simples de crescer (Bellatreche, 2003). O esquema é representado por uma tabela de factos interligada a várias tabelas de dimensão que, por sua vez, interligam-se a outras tabelas de dimensão (Abreu, 2021), tal como se pode observar no esquema da Figura 6. No entanto, as tabelas de subdimensões podem levar à perda de desempenho no processamento de consultas.

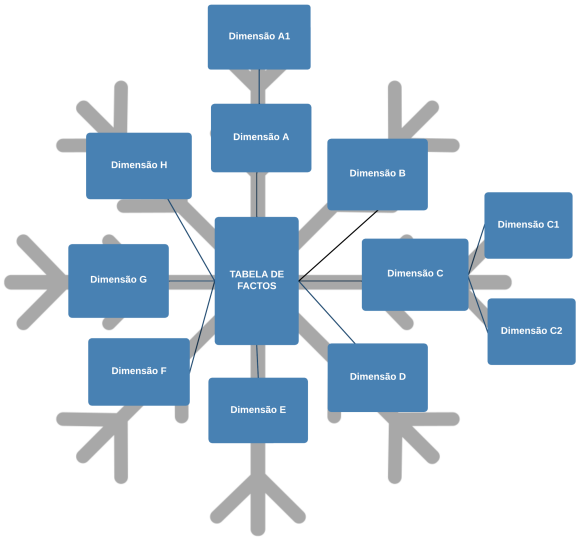


Figura 6: Ilustração de um esquema em floco de neve

Um esquema em constelação é uma estrutura ainda mais complexa do que os esquemas anteriores. Este é representado por várias tabelas de factos, que partilham algumas ou todas as tabelas de dimensão

(Han et al., 2022) (Figura 7).

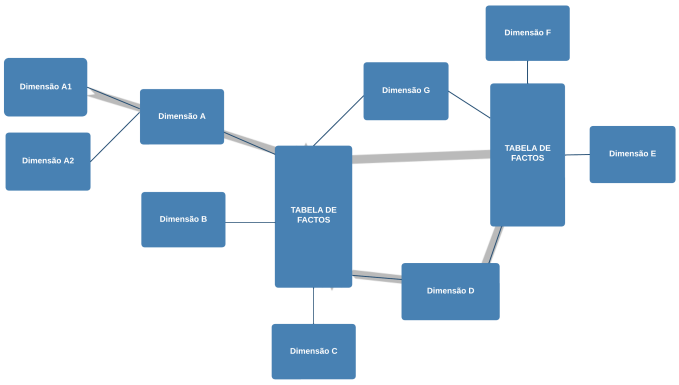


Figura 7: Ilustração de um esquema em constelação

2.3.5 Os Processos de ETL

Um processo de **ETL** considera a realização de diversas operações, incluindo extração, transformação e carregamento de dados (Figura 8). Os processos de **ETL** são considerados a "espinha dorsal" na arquitetura de um sistema de *data warehousing* (Al-Rahman et al., 2023), uma vez que é uma parte fundamental na gestão de informação nas organizações, especialmente, quando se lida com grandes volumes de dados provenientes de várias fontes de informação. Através dos processos de ETL, as empresas são capazes de recolher dados que se encontram em diferentes locais e integrá-los num sistema, adquirindo uma visão única dos dados (Sreemathy et al., 2021).

A primeira fase do processo de **ETL** é relativa à extração de dados, com foco principal na seleção e extração dos dados necessários e pertinentes para o *data warehouse* (Sreemathy et al., 2021). Normalmente, esses dados são extraídos de várias fontes de dados diferentes, que tanto podem ser internas, como externas à própria organização. Geralmente, os dados são brutos, heterogéneos e, frequentemente, apresentam várias irregularidades, o que motiva a transição para a próxima fase do processo. A fase seguinte de um processo de **ETL** diz respeito à transformação de dados. Nesta fase realizam-se procedimentos de limpeza, integração e transformação de dados, sendo considerada a mais crítica e demorada fase na construção de um *data warehouse* (Ferreira et al., 2010). Os dados previamente extraídos são temporariamente armazenados numa área de retenção, na qual são feitas as operações de limpeza e de transformação necessárias (Han e Kamber, 2011). A finalidade desta atividade é aprimorar a qualidade e a consistência dos dados, tornando-os homogéneos através da correção de anomalias, como inconsistências, valores duplicados, valores em falta, valores com erros, entre outros (Vercellis, 2008). A fase

final do processo envolve o carregamento de dados no *data warehouse* (Sreemathy et al., 2021).

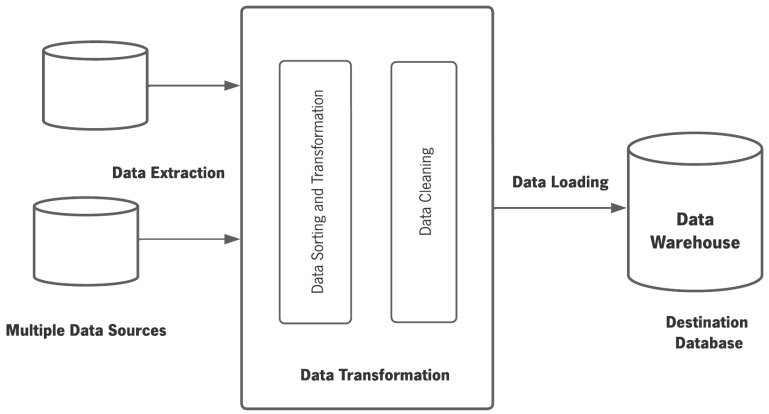


Figura 8: Ilustração de um processo de ETL - Figura - adaptada de Wang e Liu (2020)

2.3.6 Visualização de Dados

Atualmente, as empresas cada vez mais recolhem enormes quantidades de dados que precisam de ser analisados constantemente de maneira a obter *insights* para as suas atividades, para o seu negócio. Uma forma de o fazer com eficácia e com rapidez é compreender e explicar os dados que têm em posse de uma maneira que faça sentido. Portanto, a visualização de dados é uma forma poderosa de apresentar *storytelling* acerca da informação que as empresas possuem (Qin et al., 2020), uma vez que coloca os dados visualmente interativos. Esta é importante na ajuda ao olho humano na observação de ocorrências que são difíceis ou impossíveis de detetar em grandes conjuntos de dados ou em informação abstrata em formato textual (Rodríguez et al., 2015). A visualização de dados permite fornecer visualizações interativas, nas quais o utilizador pode interagir e, consequentemente, detetar padrões, tendências e correlações interessantes (Rodríguez et al., 2015). Alguns dos elementos de visualização mais utilizados num processo de visualização de dados são:

- ***CountPlot*** - um gráfico de barras que mostra a contagem dos diferentes valores que surgem numa coluna.
- ***BarPlot*** - um gráfico de barras com eixo xx e eixo yy, que permite analisar a relação entre duas variáveis.
- ***DistPlot*** - um gráfico de barras para quantidades quantitativas, divididas em intervalos. Trata-se de um histograma para entender a distribuição dos dados.

- ***PieChart*** - um gráfico circular dividido por setores. Permite representar para cada categoria a percentagem de um total.
- ***LineChart*** - um gráfico de linhas que mostra uma série de pontos ligados por segmentos de linha reta. Permite analisar a relação entre duas variáveis.
- ***ScatterPlot*** - um gráfico de dispersão utilizado para mostrar possíveis correlações entre duas variáveis.
- ***Table*** - uma tabela que serve para exibir dados textuais. Cada coluna corresponde a uma variável.

Posto isto, a etapa final no desenvolvimento de um sistema de *business intelligence* consiste na utilização de ferramentas visuais que permitem o desenvolvimento de gráficos em *dashboards* que ajudam os utilizadores a explorar e interpretar a informação, extraindo conhecimento para dar suporte às suas tomadas de decisão (Freitas, 2021).

Os *dashboards* são ferramentas de trabalho bastante poderosas especialmente orientados para a apresentação visual de dados. Usualmente permitem exibir, num único ecrã, a informação mais crucial para um ou mais utilizadores, de forma que estes sejam capazes de identificar, explorar e comunicar as áreas problemáticas que exigem ações corretivas (Velcu-Laitinen e Yigitbasioglu, 2012). Como se pode observar na Figura 9, um *dashboard* pode possuir diferentes elementos visuais para a consulta de dados. A seleção apropriada destes elementos desempenha um papel fundamental na interação de visualizações intuitivas para os utilizadores (Freitas, 2021).

Figura 9: Exemplo de um *dashboard*

2.4 Previsão de Dados

Uma previsão consiste na tentativa de antecipar o comportamento futuro com base nas condições atuais, envolvendo dados do passado. A previsão de dados deve ser uma parte integrante nas atividades de suporte à tomada de decisão, visto que pode desempenhar um papel fulcral em diversas áreas da organização.

Segundo Petropoulos et al. (2022), a previsão baseia-se na premissa de que o conhecimento atual e passado pode ser usado para realizar previsões sobre o futuro. Embora seja impossível prever o futuro assertivamente, os modelos de previsão procuram ao máximo aproximar-se dessa realidade. Os métodos de previsão podem ser simples ou altamente complexos, dependendo dos dados disponíveis, incluindo dados históricos e o conhecimento de eventuais eventos futuros que possam afetar as previsões (Ferreira, 2017). Particularmente em séries temporais, existe a crença de que é possível identificar padrões nos valores históricos e implementá-los com sucesso no processo de previsão de dados futuros.

2.4.1 Séries Temporais

Uma série temporal pode ser definida como um conjunto de observações relacionadas a um fenómeno específico, que variam ao longo de um período de tempo e que estão estatisticamente relacionadas (Pais, 2017). É possível que qualquer observação que seja registada sequencialmente ao longo do tempo seja considerada uma série temporal. A análise de dados de séries temporais são úteis quando se pretende prever tendências ou comportamentos que estão em mudança ao longo do tempo, como por exemplo, o volume de vendas, o valor das ações, a inflação, etc (Ferreira, 2017).

Uma série temporal pode ser decomposta em quatro elementos:

1. **Tendência** - Movimentos regulares e consistentes durante períodos longos. Esta componente descreve o comportamento padrão ao longo do período de tempo, isto é, ascendente ou descendente.

2. **Sazonalidade** - Variações de movimentos que ocorrem com periodicidade curta. Ou seja, são oscilações que acontecem consecutivamente no mesmo período temporal, devido a vários fatores externos, como aumento de vendas em épocas festivas.

3. **Cíclico** - Movimentos cíclicos que ocorrem periodicamente e que constituem flutuações positivas e negativas não obedecendo a nenhum período temporal específico, como expansão ou recessão económica.

4. **Ruído** - Movimentos esporádicos e aleatórios. Representa todos os movimentos que não foram possíveis de identificar.

2.4.2 O Modelo ARIMA

Um dos métodos mais utilizados em análise de séries temporais é apresentado como modelo **ARIMA** (Stellwagen e Tashman, 2013). Este tem como finalidade realizar previsões futuras com base em três componentes essenciais, nomeadamente, a: componente autorregressiva (AR), a diferenciação (I) e a componente média móvel (MA) -, estas componentes serão representadas por p, d e q, respetivamente (Stellwagen e Tashman, 2013).

A componente autorregressiva (AR) é um tipo de modelo de regressão, no qual a variável dependente é baseada em valores passados. A ordem desta componente, representada por p, indica quantos períodos passados são considerados para a futura previsão. Portanto, um modelo AR considera as observações dos últimos períodos p para prever o valor atual da série temporal. Quanto maior for o valor de p, mais dados do passado são considerados, permitindo que o modelo observe mais padrões ao longo do tempo. Em baixo, está representada a fórmula 2.1 da componente autorregressiva AR, na qual t corresponde ao período atual, c corresponde à constante do modelo, p corresponde à ordem da componente autorregressiva, fp corresponde à previsão para ordem a p, Xt corresponde ao valor da série no período t e, por último, et que corresponde ao erro do período t (Borges, 2015).

$$X_t = c + f_1 X_{t-1} + f_2 X_{t-2} + \ldots + f_p X_{t-p} + e_t \tag{2.1}$$

A componente de diferenciação, representada por d, diz respeito ao número de diferenciações que são necessárias para transformar uma série temporal não estacionária em estacionária. Uma série temporal estacionária corresponde aos dados que não apresentam tendências, nem movimentos periódicos e que flutuam em torno de uma média constante, independentemente do tempo. Já uma série não estacionária implica que haja movimentos ao longo do tempo (Pais, 2017). Uma vez que os modelos **ARIMA** apenas são aplicados em séries temporais estacionárias, é necessário que haja uma transformação prévia dos dados caso a série não seja estacionária, eliminando assim a tendência e a sazonalidade (Borges, 2015).

Por fim, a componente média móvel (MA), que permite analisar os erros do modelo em que verifica as diferenças entre os valores observados da série temporal e os valores previstos pelo modelo. Esta componente é representada por q e indica quantos erros observados de médias móveis são considerados. Na fórmula 2.2 está apresentada a equação da componente média móvel (MA), em que q corresponde à ordem da componente média móvel, c corresponde à constante do modelo, t corresponde ao período atual, Qq corresponde à média móvel de ordem a q e, por fim, et corresponde ao erro do período t

(Borges, 2015).

$$X_t = c - q_1 e_{t-1} + -q_2 e_{t-2} - \ldots - q_q e_{t-q} + e_t \tag{2.2}$$

Em suma, o modelo **ARIMA** é representado pelos parâmetros p, d e q. O parâmetro p refere-se ao número de termos autorregressivos, o parâmetro d é relativo ao número de diferenciações que são necessárias para transformar uma série em estacionária e, por fim, o parâmetro q que diz respeito ao número de médias móveis (Pais, 2017). Este modelo pode ser representado pela equação que está apresentada na fórmula 2.3.

$$X_t = c + f_1 X_{t-1} - q_1 e_{t-1} + e_t \tag{2.3}$$

Além do modelo **ARIMA**, existe ainda outro modelo que leva em consideração componentes sazonais, conhecido como **SARIMA** (Wilher, 2022). Este modelo é utilizado para modelar séries temporais que apresentam padrões de sazonalidade, ou seja, padrões que se repetem ao longo de vários períodos de tempo regulares.

O modelo SARIMA(p,d,q)(P,D,Q)m apresenta uma estrutura de parâmetros com o mesmo significado ao do modelo **ARIMA**. A diferença reside nos parâmetros, onde as letras minúsculas são destinadas à parte não sazonal, enquanto as letras maiúsculas são destinadas à parte sazonal (Pais, 2017). Além disso, o parâmetro m refere-se ao período de sazonalidade na série temporal.

2.4.3 Modelação

Quando se pretende utilizar um modelo de previsão, um dos desafios cruciais é determinar qual o método mais apropriado a aplicar. Para isso, é comum utilizar o método de *Box-Jenkins* (Stellwagen e Tashman, 2013), que apresenta um algoritmo matemático para avaliar os diversos modelos existentes. De seguida, a partir de várias observações históricas, calcula diferentes estatísticas com base nessa informação. Isto permite a identificação do modelo mais adequado para cada série temporal (Pais, 2017). Assim, a construção de um modelo passa pelas seguintes etapas: (Costa, 2015):

1. Identificação do modelo.
2. Estimativa dos parâmetros do modelo (AR, I, MA) identificado na primeira etapa.
3. Verificação e avaliação do modelo.
4. Previsão do modelo.

Os parâmetros ótimos que forem encontrados são determinados por meio de critérios de informação, tais como **AIC** (Lin, 2021) e **BIC** (Lin, 2021), que auxiliam na avaliação da qualidade do ajuste entre os vários potenciais modelos para uma série temporal específica (Pais, 2017). O critério de informação **AIC** é uma medida amplamente utilizada para medir a qualidade de modelos estatísticos. Este tem como finalidade selecionar o melhor modelo entre vários, tendo em consideração o ajustamento do modelo aos dados, bem como a sua complexidade. Além disso, o **AIC** penaliza modelos mais complexos, incentivando à escolha de modelos mais simples. Assim como o **AIC**, o critério de informação **BIC** visa encontrar um equilíbrio entre o ajuste do modelo aos dados e a complexidade deste. Este favorece ainda mais que o **AIC** os modelos mais simples, particularmente quando há um grande volume de dados. Quanto menor o valor de **AIC** e **BIC**, melhor é o modelo, pois indica que o modelo se ajusta bem aos dados observados.

3 Um Sistema para Consolidação de Contas

Antes de implementar qualquer sistema de *business intelligence*, o primeiro passo será compreender os objetivos e as necessidades do negócio. Neste capítulo, começou-se por apresentar a área de trabalho base, a consolidação de contas, e logo de seguida o que são demonstrações financeiras e a forma como se realizam. De seguida, tendo sido apresentado a problemática em mãos, passou-se, propriamente, ao processo de desenvolvimento do sistema de *data warehousing* requerido. Assim, numa fase seguinte, definiu-se e descreveu-se os principais requisitos, nomeadamente, de descrição, de exploração e de controlo e acesso, dos vários agentes de decisão envolvidos. Além disso, realizou-se a modelação multidimensional, para projetar os esquemas dos sistemas de dados, através da definição das tabelas de dimensões e das tabelas de factos, orientados para o suporte a processos de tomada de decisão, baseando-se, especialmente, nos requisitos solicitados. Este processo é fundamental para compreender os objetivos do negócio e validar o sistema com os agentes de decisão antes de ser efetivamente implementado. Como se verificou anteriormente, esta é uma das atividades mais relevantes no desenvolvimento e conceção de um sistema de *data warehousing*.

3.1 A Consolidação de Contas

Na última década, com o crescimento económico e do grupo de sociedades mais complexos, aliado à globalização, ao aparecimento de novos mercados e, também, ao desenvolvimento tecnológico, houve a necessidade de proporcionar aos *stakeholders* informação financeira que permitisse uma análise adequada a um conjunto de empresas pertencentes ao mesmo grupo económico (Garcia, 2019). De forma a acompanhar a evolução da economia, verifica-se que as empresas cada vez mais apostam em diversas formas de organização, seja através de aquisições de capital de outras sociedades, de fusões com outras entidades, ou através de outros mecanismos de cooperação, como alianças estratégicas (Ribeiro, 2017). Isto acontece devido à necessidade estratégica das organizações conseguirem dar resposta à elevada competição empresarial proporcionada pela globalização das economias. Assim, as entidades unem-se com a empresa-mãe para conceber uma unidade económica de direção única, formando um grupo de

sociedades.

Um grupo de sociedades pode-se definir como um conjunto de entidades ligadas financeira e economicamente a uma empresa-mãe que controla e gere cada uma delas, sem perder autonomia jurídica (Marques, 2019). A empresa-mãe, bem como cada uma das suas subsidiárias, além de apresentar individualmente a prestação de contas, tem a obrigatoriedade de apresentar em conjunto a prestação de contas (Ribeiro, 2017). Para executar este processo, recorre-se à consolidação de contas.

Basicamente, pode-se ver a consolidação de contas como um processo que visa o fornecimento de informação económica e financeira dos grupos de sociedades (Ribeiro, 2010), que envolve um conjunto de técnicas de natureza contabilística que permitem a elaboração de demonstrações financeiras consolidadas para um grupo empresarial, como se de uma única entidade se tratasse (Couto e Dias, 2023). A contabilidade de um grupo de empresas está rigorosamente ligada à consolidação de contas, dado que o património e a rentabilidade não podem ser refletidos pelas contas individuais de cada uma das empresas do grupo (Couto e Dias, 2023). Portanto, o objetivo da consolidação de contas é agregar toda a informação financeira das empresas que pertencem ao mesmo grupo económico (Garcia, 2019), fornecendo uma imagem de transparência sobre os seus resultados. Esta permite uma visão mais completa da situação financeira e económica do grupo, uma comparação da *performance* do grupo com outras empresas do setor, auxilia na tomada de decisões estratégicas e demonstra a evolução do grupo aos longo dos anos. Trata-se, assim, de um processo com elevada importância para os utilizadores, e pode ser utilizado como um instrumento de (Marques, 2019):

- **Gestão, que** - permite à empresa-mãe avaliar a situação patrimonial, financeira e económica do grupo.
- **Controlo interno, para** - assegurar a qualidade da informação, contribuindo para a harmonização de processos e normalização da informação financeira.
- **Avaliação externa, que** - permite fornecer a realidade financeira do grupo aos seus *stakeholders*.

No entanto, as contas consolidadas possuem algumas limitações aquando da sua análise, nomeadamente, entidades que contenham fraca *performance* dentro de um grupo empresarial serem encobertas por outras empresas que obtenham melhores resultados, dificuldade em realizar comparações de resultados com outras organizações, uma vez que cada entidade empresarial possui as suas próprias características e condições e, ainda, a falta de informação detalhada individualmente sobre cada uma das empresas que compõem o perímetro de consolidação (Garcia, 2019).

O processo de consolidação de contas segue, normalmente, as seguintes etapas (Marques, 2019):

1. Definição do perímetro de consolidação.
2. Escolha dos métodos de consolidação.
3. Levantamento das demonstrações financeiras individuais de cada entidade.
4. Conversão das demonstrações financeiras para a moeda de relato da empresa-mãe.
5. Harmonização das políticas contabilísticas.
6. Agregação das demonstrações financeiras individuais.
7. Elaboração das demonstrações financeiras consolidadas.

Antes da elaboração das contas consolidadas, o primeiro passo é determinar o perímetro de consolidação, que, basicamente, consiste na definição do conjunto das empresas subsidiárias cujas contas vão ser objeto de consolidação (Ribeiro, 2010). Esta delimitação tem por base a percentagem de controlo detida pela empresa-mãe (Marques, 2019). O processo de identificação de perímetro é feito através de uma organograma organizacional que consta a empresa-mãe e suas subsidiárias. O perímetro pode ser alterado com alguma regularidade, o que provoca alguma diferença entre períodos. Assim, anualmente, deve-se redefinir o perímetro, uma vez que pode surgir entradas e saídas de empresas no grupo (Couto e Dias, 2023).

3.2 Demonstrações Financeiras

As demonstrações financeiras são um conjunto de relatórios contabilísticos que refletem a situação económico-financeira de uma empresa, num determinado período, e destacam os seus pontos fortes e fracos (Quintaneiro et al., 2000). Além disso, permitem reunir um conjunto de elementos importantes para evidenciar o panorama geral financeiro de uma empresa e, por conseguinte, elaborar decisões estratégicas. Um conjunto completo de demonstrações financeiras inclui balanço, demonstração do resultado do exercício, demonstração do fluxo de caixa e demonstração das contas de capitais próprios (Quintaneiro et al., 2000). Neste trabalho apenas se abordará a demonstração de resultados do exercício. Esta visa destacar os resultados do exercício, quanto ao lucro ou prejuízo que foram obtidos na atividade da empresa. Existem dois tipos de demonstrações de resultados (Ribeiro e Morais, 2021):

1. Demonstração de resultados por natureza, que consiste na junção dos resultados apurados pela empresa, evidenciando os proveitos e custos.

2. Demonstração de resultados por funções, que consiste em reunir os resultados conforme as funções empresariais (produção, distribuição, administrativa e financeira). Neste trabalho apenas utilizou-se as demonstrações dos resultados por natureza.

Na Figura 10 é possível ver uma demonstração de resultados por natureza. Trata-se de um documento contabilístico que fornece informações resumidas dos resultados das operações financeiras da empresa durante um determinado período específico, no qual se pretende expor os rendimentos e os gastos desse mesmo período de exercício (Quintaneiro et al., 2000). Através deste relatório financeiro é possível realizar uma avaliação do desempenho da empresa relativa a um ano e face ao ano anterior. Várias empresas possuem centenas de transações todos os meses que envolvem muitos gastos e rendimentos. Por isso, seria impensável submeter todas essas transações apenas numa única rubrica. Além disso, também impossibilitava de estabelecer relações entre as várias rubricas. Assim, originou-se um conjunto de contas de receitas e despesas ao longo do mapa financeiro.

O principal objetivo final da demonstração de resultados do exercício é a obtenção do resultado líquido do exercício que se traduz no valor total de vendas do período e, de outros eventuais ganhos, subtraindo todos os custos do mesmo período (Quintaneiro et al., 2000). A demonstração de resultados do exercício é uma ferramenta fundamental para avaliar o desempenho financeiro de uma empresa, a sua rentabilidade e a capacidade de gerar lucros. Esta é frequentemente comparada com os resultados de períodos anteriores o que permite identificar tendências e padrões que podem indicar a saúde financeira da empresa, ajudando, assim, os *stakeholders* a tomar decisões financeiras e estratégicas.

Os rendimentos e gastos são os elementos essenciais da demonstração de resultados do exercício, que permitem identificar principais fontes de receita e de despesas da empresa. Numa demonstração de resultados, os rendimentos dizem respeito à parte considerada agradável de todo o exercício de uma empresa. Trata-se, assim, de um aumento no benefício económico durante o período contabilístico. Ao contrário dos gastos, a parte menos agradável de todo o exercício, que são diminuições nos benefícios económicos da empresa (Quintaneiro et al., 2000). Ou seja, rendimentos são valores recebidos pela empresa em decorrência das suas atividades operacionais, enquanto que gastos são os valores pagos pela empresa. No documento da demonstração do resultado do exercício, tanto os rendimentos e gastos estão divididos por rubricas, sendo que uma rubrica pode conter várias sub-rubricas. Na Figura 10 pode-se observar os rendimentos e gastos que integram o relatório financeiro da demonstração de resultados.

DEMONSTRAÇÃO (INDIVIDUAL/CONSOLIDADA) DOS RESULTADOS POR NATUREZAS
PERÍODO FINDO EM XX DE YYYYYYY DE 20NN

UNIDADE MONETÁRIA (1)

RENDIMENTOS E GASTOS	NOTAS	PERÍODOS	
		N	N-1
Vendas e serviços prestados		+	+
Subsídios à exploração		+	+
Ganhos/perdas imputados de subsidiárias, associadas e empreendimentos conjuntos		+ / -	+ / -
Variação nos inventários da produção		+ / -	+ / -
Trabalhos para a própria entidade		+	+
Custo das mercadorias vendidas e das matérias consumidas		-	-
Fornecimentos e serviços externos		-	-
Gastos com o pessoal		-	-
Imparidade de inventários (perdas/reversões)		- / +	- / +
Imparidade de dívidas a receber (perdas/reversões)		- / +	- / +
Provisões (aumentos/reduções)		- / +	- / +
Imparidade de investimentos não depreciáveis/amortizáveis (perdas/reversões)		- / +	- / +
Aumentos/reduções de justo valor		+ / -	+ / -
Outros rendimentos		+	+
Outros gastos		-	-
Resultado antes de depreciações, gastos de financiamento e impostos		=	=
Gastos/reversões de depreciação e de amortização		- / +	- / +
Imparidade de investimentos depreciáveis/amortizáveis (perdas/reversões)		- / +	- / +
Resultado operacional (antes de gastos de financiamento e impostos)		=	=
Juros e rendimentos similares obtidos		+	+
Juros e gastos similares suportados		-	-
Resultado antes de impostos		=	=
Imposto sobre o rendimento do período		- / +	- / +
Resultado líquido do período		=	=
Resultado das atividades descontinuadas (líquido de impostos) incluído no resultado líquido do período			
Resultado líquido do período atribuível a: (2)			
Detentores do capital da empresa-mãe			
Interesses que não controlam			
		=	=
Resultado por ação básico			

Figura 10: Demonstração de Resultados por natureza - tabela - retirada do SNC (Diário da República - 1.ª série — N.º 173, 2009)

3.3 Levantamento de Requisitos

O levantamento de requisitos é uma etapa crucial que visa identificar e documentar os requisitos que o sistema deve fazer para atender às necessidades dos utilizadores. É essencial que o sistema seja capaz de fornecer a informação necessária para a tomada de decisão. O levantamento de requisitos para o desenvolvimento do *data warehouse* foi realizado com a participação de todos os, futuros, utilizadores do sistema, através de uma entrevista. Todos os requisitos fornecidos foram documentados para garantir que todos os envolventes tivessem uma visão clara sobre o que o sistema deve fazer.

3.3.1 Requisitos de Descrição

Os requisitos de descrição representam os elementos de especificação mais importantes de um sistema, que servem para revelar como o sistema deve funcionar e quais as suas características e funcionalidades. Estes requisitos são de extrema importância, uma vez que fazem parte do processo inicial de desenvolvi-

mento, e ajudam a garantir que o sistema responda de forma adequada às necessidades dos utilizadores. É necessário que estes requisitos sejam elaborados de forma clara e concisa. Além disso, devem ser facilmente compreendidos, sobretudo, pelo responsável de desenvolvimento do sistema.

Tendo em consideração o contexto deste caso de aplicação, elaborou-se os seguintes requisitos:

- Para a análise do valor de cada transação deverá ser considerada a informação acerca da rubrica, do período do exercício, do local, do perímetro e da empresa.
- O período do exercício deve especificar o mês, o número do mês, o trimestre e o ano.
- Em cada rubrica deve constar o seu código do **SNC**, bem como a sua descrição detalhada e geral.
- Uma empresa deve ser caracterizada por um identificador único, pelo seu nome, pelo local onde se situa e pelo perímetro a que pertence.
- Cada empresa do grupo deve possuir um histórico com o número de modificação e a data de atualização dos dados de cada uma delas.
- O local de cada empresa deve possuir o seu identificador único, o continente, o país, a sigla do país e a sede.
- O perímetro do grupo deve conter o nome e o setor que o carateriza.

3.3.2 Requisitos de Exploração

Os requisitos de exploração descrevem as necessidades dos utilizadores para obter *insights*, realizar análises e descobrir informações importantes acerca do negócio. Deste modo, estes requisitos vão definir a forma como o sistema responderá às necessidades dos seus utilizadores em processos de exploração dos seus dados. De seguida, apresentam-se vários requisitos de exploração que o sistema deve ser capaz de satisfazer no final:

- Analisar o período do exercício relativo ao ano 2015 e 2016.
- Analisar a evolução do resultado líquido do grupo por mês.
- Identificar o valor do EBTIDA do grupo por ano.
- Identificar o volume de negócios por ano.
- Identificar a margem de lucro por ano.

- Analisar a evolução dos gastos e receitas por mês.
- Identificar o top 3 de empresas que obtiveram lucro e prejuízo durante os períodos de 2015 e 2016.
- Averiguar o número de empresas por país.
- Analisar as receitas das vendas mensalmente.
- Identificar, especificamente, quais foram as maiores despesas e receitas por empresa.
- Prever o volume de negócios para o ano seguinte, por mês.

3.3.3 Requisitos de Controlo e Acesso

De forma a proteger a integridade e a confidencialidade dos dados no sistema, é necessário definir medidas de permissão e de acesso. Assim, os requisitos de controlo e acesso descrevem quem são os utilizadores que terão permissão para visualizar e manipular dados no sistema. Estes requisitos são fundamentais para garantir a segurança dos dados armazenados. Assim, alguns requisitos de controlo que foram definidos são:

- Todas as equipas do departamento da "Consolidação de Contas" têm permissões de consulta a toda a informação contida no sistema.
- A equipa de "Suporte & Desenvolvimento" desse departamento tem permissões ilimitadas.

Por fim, os requisitos levantados foram analisados minuciosamente e validados com o departamento de modo a garantir que ambas as partes estivessem alinhadas. O sucesso de um projeto está intimamente ligado à qualidade do levantamento de requisitos, uma vez que requisitos mal definidos podem levar a complicações durante o desenvolvimento e implementação do sistema e, consequentemente, levar ao seu fracasso.

3.4 Modelação Multidimensional

3.4.1 Caracterização do Data Mart e do Grão

Para atender aos requisitos solicitados e fazer o planeamento da modelação do *data warehouse* foi definida uma matriz de decisão para auxiliar, no processo de definição e identificação dos dados a armazenar no *data warehouse* (Belo, 2012). A matriz desenvolvida (Figura 1) descreve o *data mart* "financeiro". Este

data mart acolhe a informação para suportar as tomadas de decisão no domínio das demonstrações financeiras, fornecendo dados acerca dos proveitos e custos ao longo de um determinado período do exercício que foram efetuados por cada empresa. A principal motivação para a elaboração deste *data mart* foi garantir a gestão e o controlo de custos e de receitas de forma a avaliar e a potenciar melhorias no desempenho financeiro do grupo.

De modo a estruturar o *data mart* foi necessário definir o nível de detalhe da informação que ele acolherá, ou seja, o grão da sua tabela de factos. Esta é uma das etapas mais delicadas no processo de modelação (Belo, 2012). O grão define o contexto e a amplitude dos dados e, por isso, é o nível de detalhe da informação que se pretende manter na estrutura de dados do *data warehouse*. Quanto mais alto for a sua granularidade, maior é o detalhe dos dados. Isso possibilita a realização de análises mais detalhadas, a identificação de tendências mais específicas e a tomada de decisões baseada em informação detalhada. A sua má definição pode implicar que a exploração de dados de uma tabela de factos revele resultados inconsistentes ou pouco coerentes.

De seguida, identificou-se o grão para integrar na tabela de factos: o resultado de uma transação relativa a uma determinada rubrica, efetuada por uma empresa específica, num certo período do exercício. Esta definição é baseada, sobretudo, nos requisitos anteriormente fornecidos, que se pretende explorar e analisar os valores monetários dos proveitos e custos realizados ao longo do período do exercício. O grão definido corresponde ao nível de informação mais detalhado da estrutura do *data warehouse*. Por conseguinte, a estrutura base do *data mart* "financeiro" é composta por uma única tabela de factos que armazena dados sobre os resultados das transações financeiras das empresas do grupo. Esta tabela relaciona-se com três tabelas de dimensão que sustentam o período, a rubrica e a empresa a que os dados acerca dessas mesmas se referem.

Por fim, o *data mart* foi configurado como um sistema transacional, com atualizações de dados feitas mensalmente, uma vez que a elaboração das demonstrações financeiras ocorrem no final de cada mês para cada empresa. Essas atualizações ficam disponíveis para os utilizadores responsáveis por consolidar os dados financeiros das empresas do grupo, para a apresentação de análises e para relatórios financeiros. O desenvolvimento do *data mart* financeiro é um recurso fulcral, pois fornece dados importantes e confiáveis para a tomada de decisão financeira, permitindo avaliar o desempenho financeiro de cada empresa, identificando aquelas que apresentam mais e menos lucros, com o propósito final de potenciar o crescimento de cada uma delas. Na Tabela 1 pode-se verificar, resumidamente, a caraterização final do *data mart* financeiro.

Caracterização do Data Mart	
Identificação: Financeiro	
Descrição geral: Informação para suporte à tomada de decisão na área financeira de um grupo de empresas com base em dados acerca das receitas e gastos realizados por estas ao longo de um determinado período. Servirá como motivação à gestão e controlo da situação económica e financeira e ao incentivo do crescimento destas.	
Estrutura base	
Tabela de Factos >>>	Transação
<<<Dimensões	
Período	√
Empresa	√
Rubrica	√
Local	√
Perímetro	√
Número de Dimensões	5
Tipo	Transacional
Periocidade	Mensal
Descrição	Transações financeiras de um grupo de empresas
Utilidade Estratégica	Avaliar o desempenho financeiro de cada empresa. Identificar as empresas mais lucrativas. Avaliar e compreender as empresas que apresentam prejuízo e realizar incentivos para o seu crescimento.
Utilizadores	Equipa da Consolidação
Observações: Nada a assinalar.	

Tabela 1: Matriz de decisão - tabela - adaptada de Belo (2012)

3.4.2 Caracterização das Dimensões e da Tabela de Factos

Um dos passos mais importantes na criação de uma estrutura de dados é a definição das dimensões relevantes para o contexto da área de negócio que se pretende analisar. O objetivo é que não hajam dados desnecessários e que não haja uma mistura de informação no *data mart* desenvolvido. Assim, depois de definido o grão a ser utilizado na tabela de factos, é fundamental caracterizar os dados que serão acolhidos pelas dimensões e que vão fornecer contexto à tabela de factos. De acordo com os requisitos,

foi considerado fundamental a informação acerca do período do exercício, da rubrica e da empresa para categorizar os dados da estrutura. Estas tabelas foram designadas por "DimPeriodo", "DimRubrica" e "DimEmpresa", respetivamente.

Relativamente à dimensão "DimPeriodo", esta é uma dimensão temporal que contém os atributos "ano", "trimestre", "nome" e "nrMes". Esta é uma das tabelas essenciais que não pode faltar num *data warehouse*, porque permite localizar os dados no tempo. Isso possibilita ao utilizador compreender e analisar tendências ao longo de um determinado período, além de permitir que no *data warehouse* sejam armazenados os dados históricos, sem nunca perder a informação antiga. Um *data warehouse* sem informação temporal torna-se num sistema limitado na realização de análises históricas, uma vez que estas dependem da capacidade de comparar dados de diferentes períodos de tempo. Sem dados temporais, dificilmente é possível identificar padrões que possam ser usados para prever tendências futuras. A dimensão "DimPeriodo" tem como finalidade armazenar o mês, o número do mês, ano e trimestre em que o resultado das transações das rubricas foram efetuadas.

Quanto à tabela da dimensão "DimRubrica", esta armazena informação dos dados acerca das rubricas que constam no **SNC**. Para esta dimensão considerou-se importante integrar o código, nome da rubrica e as rubricas discriminadas, e, por fim, o tipo de rubrica que representa, ou seja, se é alusiva a uma despesa ou a uma receita. As rubricas discriminadas são categorias específicas dentro de uma rubrica principal, que, fornecem informação adicional mais detalhada. Por exemplo, a rubrica "gastos com pessoal" pode conter rubricas discriminadas, como "seguro de acidentes", "encargos sobre as remunerações', entre outros. Desta forma, esta dimensão visa armazenar a informação de cada rubrica facilitando a exploração e a consulta destas e identificar para cada uma os valores obtidos, dando assim contexto à tabela de factos.

Em relação à tabela da dimensão "DimEmpresa", esta requereu um tratamento diferente das dimensões anteriores, uma vez que se trata efetivamente de um grupo empresarial multinacional com um número elevado de empresas, que estão divididas em perímetros de acordo com a sua área de atuação. Cada empresa do grupo possui várias características, como o nome, a sede e o perímetro a que pertence. Para simplificar a estrutura e evitar redundância, foram criadas duas subdimensões que categorizam a dimensão empresa, nomeadamente, "DimPerimetro" e "DimLocal". A dimensão "DimPerimetro" pretende guardar a informação do nome do perímetro e a categoria da área que atua, isto é, retalho alimentar, tecnologia, vestuário, saúde, entre outros. Já a dimensão "DimLocal" visa armazenar a informação da cidade, país e continente onde a empresa está sediada. A principal razão para a criação destas subdimensões foi permitir guardar os dados de forma eficiente, evitando eventuais repetições de dados. Caso os atributos fossem armazenados apenas na tabela "DimEmpresa", seria necessário repetir a informa-

ção sobre o perímetro e o local de cada empresa. Para além disso, os atributos das subdimensões são informações secundárias. Posto isto, a tabela "DimEmpresa" contém os atributos nome da empresa e os identificadores únicos das subdimensões. Esta foi considerada uma dimensão com variação histórica, na medida em que pode ocorrer entradas e saídas de empresas no grupo ou, até mesmo, mudanças de perímetro. Assim, foi necessário armazenar o histórico das empresas numa tabela denominada de "DimEmpresa_Hist". Esta última contém os mesmos atributos da "DimEmpresa", à exceção do número de modificação e da data de atualização que permitem identificar quando é que foram feitas as alterações nas informações da empresa.

Em relação à tabela de factos "TF_Transacao", esta possui apenas uma medida quantitativa: o valor monetário do resultado de uma transação por rubrica, empresa e período do exercício. Esta medida é o principal dado analítico que os utilizadores pretendem acompanhar e analisar, pois permite avaliar o desempenho do negócio.

Resumidamente, o *data warehouse* é composto por uma tabela de factos e cinco dimensões, sendo que três dimensões são principais e duas são secundárias. A tabela "DimEmpresa_Hist" é uma tabela auxiliar que faz parte da dimensão "DimEmpresa".

3.4.3 Esquema Multidimensional

De seguida, apresenta-se o esquema multidimensional do *data warehouse*, de acordo com a caracterização da tabela de factos e dimensões realizada nas secções anteriores, com o propósito de facilitar a compreensão da estrutura de dados a ser implementada para o *data warehouse*. A Figura 11 ilustra um modelo de dados baseado num esquema em floco de neve, cuja tabela de factos ocupa uma posição central com três dimensões principais à sua volta, duas dimensões secundárias criadas para normalizar os dados e reduzir a redundância e uma dimensão para armazenar dados históricos.

A tabela de factos contém chaves estrangeiras que permitem criar um relacionamento com as dimensões "DimRubrica", "DimPeriodo" e "DimEmpresa", fornecendo informação adicional para contextualizar a medida de negócio a ser analisada. Cada registo nesta tabela representa o valor de uma transação de uma rubrica, efetuada por uma empresa, durante um determinado período do exercício. As dimensões representam atributos descritivos importantes para a análise do negócio, e cada uma delas possui uma chave primária única que a identifica.

No que concerne à dimensão "DimEmpresa", esta fragmenta-se em duas subdimensões que a caracterizam, resultando assim em duas chaves estrangeiras que se relacionam com as dimensões "DimLocal" e "DimPerimetro". Isto significa que um mesmo local pode abranger várias empresas sediadas e um mesmo perímetro pode incluir inúmeras empresas. Para além disso, a dimensão "DimEmpresa" é

uma tabela com variação o que implica que haja uma dimensão que armazene o seu histórico. Assim sendo, a dimensão "DimEmpresa_Hist" contém uma chave estrangeira que permite o relacionamento com a dimensão empresa e a chave primária passa a ser o número de modificação. Isto acontece porque uma empresa pode sofrer várias alterações ao longo do tempo.

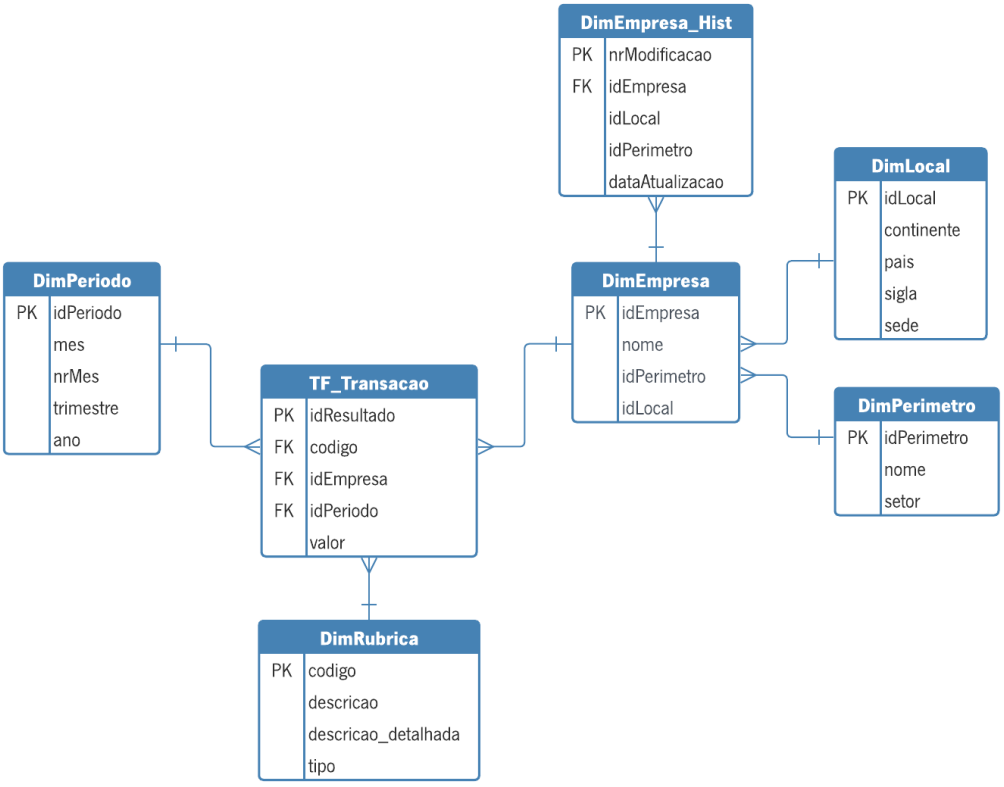


Figura 11: Esquema multidimensional do *data warehouse* desenvolvido

4 Um Data Warehouse para Consolidação de Contas

4.1 A Arquitetura do Sistema de Data Warehousing

4.1.1 Ferramentas Utilizadas

Para a implementação do sistema de *data warehousing*, foi essencial selecionar algumas ferramentas tecnológicas que pudessem ser o suporte e a base de todo o desenvolvimento do projeto. A ferramenta de **ETL** escolhida será usada como a principal ferramenta de trabalho, uma vez que possui os serviços essenciais para a extração, a transformação e o carregamento dos dados de forma automatizada no *data warehouse*. Atualmente, existem várias ferramentas de **ETL** *user-friendly* no mercado, que simplificam muito o trabalho de criação de um sistema de povoamento de um *data warehouse*. No entanto, não foi utilizado nenhuma delas. O sistema de povoamento foi implementado em *Python* e o *data warehouse* acolhido em *SQLite*. O *SQLite* é uma base de dados relacional de código aberto (Allen et al., 2010). A linguagem *Python* foi utilizada para executar todos os *scripts* produzidos no processo de **ETL** de uma só vez de maneira a tornar o processo automatizado. Por último, foi utilizada a ferramenta *Microsoft Power BI*, uma ferramenta de *business intelligence* que permite visualizar dados de forma interativa e dinâmica através de *dashboards* que facilitam a apresentação dos dados, tornando-os mais fáceis de interpretar e analisar.

4.1.2 Etapas de Desenvolvimento

Na Figura 12 está ilustrado o processo de desenvolvimento do sistema de *data warehousing* que foi implementado. O sistema foi desenvolvido em quatro fases. De referir:

1. Identificação das fontes de dados operacionais.
2. Processo de **ETL**.
3. Armazenamento no *data warehouse*.

4. Envio dos dados para uma aplicação de *business intelligence*.

Na primeira etapa são conhecidos os sistemas operacionais das fontes de dados. Neste caso, os dados necessários para o *data warehouse* estão disponíveis em três ficheiros **CSV**. De seguida, na segunda etapa é feito o processo de **ETL**, isto é, a extração dos dados das fontes, a sua transformação e o seu consequente carregamento no *data warehouse*. A extração dos dados das fontes é feita através da linguagem *Python*, que envia os dados para o sistema de gestão de base de dados *SQLite*. Neste sistema são realizadas todas as tarefas de limpeza e de transformação de dados, incluindo a criação de uma área de retenção para suporte à realização dessas tarefas, bem como o carregamento dos dados preparados no *data warehouse*. Na etapa seguinte, os dados transformados são armazenados no *data warehouse* e é feito o carregamento para as respetivas tabelas. O *data warehouse* foi criado também em *SQLite*. Por fim, a última etapa do sistema trata de enviar os dados diretamente para a ferramenta *PowerBI* através de uma conexão específica estabelecida com o *SQLite*.

Todas estas etapas, à exceção da última, foram implementadas usando vários *scripts* de **SQL**, que, por sua vez, estão interligados com um *script Python*. Este *script* permite atualizar os dados todos de uma vez. Assim, sempre que houver novas atualizações de dados para o *data warehouse*, o administrador do sistema apenas precisa de correr uma única vez os *scripts* desenvolvidos e, em seguida, atualizar as tabelas no *PowerBi*. A automatização deste processo garante uma gestão eficiente e ágil dos dados no *data warehouse*.

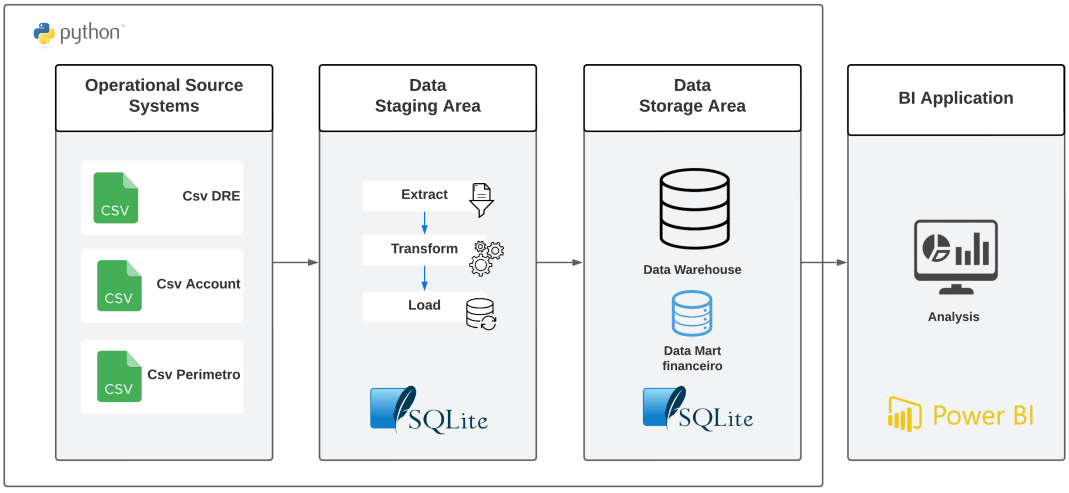


Figura 12: Etapas de desenvolvimento da arquitetura do sistema de *data warehousing*

4.2 Descrição das Fontes de Dados

Antes de implementar o sistema, teve-se que conhecer as suas fontes de informação para compreender que tipo de dados é que nelas estariam contidos, uma vez que as mesmas podem variar em termos de qualidade, confiabilidade e a acessibilidade dos dados que dispõem. A qualidade dos dados é um fator importante a ser considerado, uma vez que dados com baixa qualidade podem levar a conclusões incorretas. A confiabilidade dos dados está relacionada com a credibilidade da fonte e, por isso, é importante saber se os dados são atualizados regularmente e se a informação disponibilizada está atual. E, por último, a acessibilidade dos dados é importante para garantir que os dados possam ser facilmente encontrados e utilizados.

Assim, nesta etapa procurou-se conhecer o local no qual está guardada a informação necessária para o projeto, ou seja, tudo aquilo que se relaciona com as empresas do grupo, os seus gastos e receitas, num certo período de tempo. Nesse sentido, foram identificadas três fontes de dados internas, contendo ficheiros **CSV** como as principais e mais importantes para o contexto do problema a solucionar.

4.2.1 A Fonte "DRE"

Esta fonte diz respeito à demonstração do resultado do exercício, na qual estão armazenados os dados relativos ao desempenho financeiro de cada empresa, durante um determinado período do exercício. Todos os meses, a equipa recebe as demonstrações do resultado de cada uma das empresas, com os valores praticados no último mês. Este conjunto de dados será regularmente extraído para o sistema, para garantir a atualização contínua da informação. Assim, como se pode observar através da Tabela 2, cada linha da fonte original representa o valor correspondente de uma receita ou de um gasto realizado por uma empresa durante um determinado mês e ano. Cada uma dessas linhas é constituída pelos seguintes atributos:

- **COMPANY** - nome da empresa.
- **MONTH** - mês em que foi realizada a demonstração do resultado do exercício.
- **YEAR** - ano em que foi realizada a demonstração do resultado do exercício.
- **ACCOUNT_ID** - código único da rubrica.
- **AMOUNT** - valor monetário praticado relativo a uma rubrica.

Atributos	Exemplo	Tipo de Dados
COMPANY	Sabores da Vida	Varchar
MONTH	October	Varchar
YEAR	2015	Int
ACCOUNT_ID	P71100000	Varchar
AMOUNT	723 519.67	Float

Tabela 2: Dados da fonte “DRE”

4.2.2 A Fonte “Perimetro”

A fonte “Perimetro” contem os dados relativos às empresas que pertencem a um determinado grupo, ou seja, a um dado perímetro. Esta fonte possui todos os perímetros da empresa em questão. Porém apenas um deles é utilizado nesta dissertação. Estes dados não são atualizados com frequência, sendo revistos normalmente de ano em ano, a menos que hajam alterações no perímetro, tais como entradas ou saídas de empresas. Na Tabela 3 estão apresentados os atributos que integram uma linha de dados desta fonte. A sua descrição é a seguinte:

- **AREA_GEOGRAFICA** - continente a que pertence a empresa.
- **DIVISAO** - nome do perímetro.
- **PAIS** - sigla do país que a que pertence a empresa.
- **PAIS_DESCRICAO_DETALHADA** - nome do país.
- **SOCIEDADE** - nome da empresa.
- **SEDE** - nome da cidade a que pertence a empresa.
- **CATEGORIA** - setor de atividade em que atuam as empresas do perímetro.

Atributos	Exemplo	Tipo de Dados
AREA_GEOGRAFIA	Europa	Varchar
DIVISAO	MCC	Varchar
PAIS	PT	Varchar
PAIS_DESCRICAO_DETALHADA	Portugal	Varchar
SOCIEDADE	Sabores da Vida	Varchar
SEDE	Maia	Varchar
CATEGORIA	Retalho alimentar	Varchar

Tabela 3: Dados da fonte "Perimetro"

4.2.3 A Fonte "Account"

Nesta fonte pode-se encontrar as descrições das rubricas. Cada uma das suas linhas de dados descreve, de forma geral e detalhada, a rubrica de uma receita ou de um gasto, que é representado pelo seu código único. É uma fonte que raramente é atualizada, uma vez que as rubricas mantêm-se sempre as mesmas, exceto se houver mudanças no **SNC**. Esta fonte é importante, porque servirá para que, mais tarde, se possa realizar o mapeamento com a fonte "DRE". Na Tabela 4 é possível ver os três atributos de dados mais importantes, nomeadamente:

- **ACCOUNT_ID** - código único da rubrica.
- **DSC** - descrição detalhada da rubrica.
- **DSC_GERAL** - descrição geral da rubrica.

Atributos	Exemplo	Tipo de Dados
ACCOUNT_ID	P71100000	Varchar
DSC	Vendas de mercadorias	Varchar
DSC_GERAL	Vendas	Varchar

Tabela 4: Dados da fonte "Account"

4.2.4 Mapeamento de Dados

A conceção de um sistema de **ETL** incide sobre o mapeamento dos atributos dos dados de várias fontes para os atributos das tabelas do *data warehouse*. Desta forma, conhecidas as fontes e os dados

que as integram, é fundamental elaborar um mapeamento de dados. O mapeamento é um processo de identificação de dados, normalmente provenientes de várias fontes de dados, com a finalidade de correspondê-los às tabelas de dimensão e de factos no sistema de *data warehouse*. É uma etapa essencial para o desenvolvimento de um *data warehouse*, pois garante que os dados sejam integrados e organizados de forma eficiente.

A Tabela 5 apresenta uma visão geral do mapeamento de dados entre as fontes de dados e as tabelas de destino para o *data warehouse*, e os atributos associados em ambas. Além disso, são apresentados os tipos de dados de cada atributo nas fontes e os tipos de dados esperados no futuro sistema. Essa informação é importante para entender o formato de cada atributo, quer seja um número inteiro, decimal ou texto.

No que diz respeito à fonte de dados "DRE", esta é a principal fonte que fornece informações sobre os valores das rubricas realizadas pelas entidades num determinado mês e ano. Nesse sentido, os atributos "YEAR" e "MONTH" foram mapeados para a dimensão "DimPeriodo", visto que apresentam o período em que foi elaborada a demonstração dos resultados. É importante observar que esses atributos mantiveram o mesmo formato de dados, o que facilitou a integração. O atributo "COMPANY" foi associado à dimensão "DimEmpresa", com o mesmo formato de dados, a fim de obter o nome de todas as empresas dentro do perímetro. Por outro lado, o atributo "AMOUNT" da mesma fonte foi direcionado para a tabela de factos "TF_Transacao", uma vez que é a principal medida de negócio a ser analisada. Este apresenta um formato de número decimal, no entanto no *data warehouse* será convertido para um formato de número inteiro para simplificar futuras análises, já que habitualmente é mais conveniente trabalhar com valores arredondados.

Relativamente aos dados da fonte "Perimetro", estes estão relacionados com a caracterização do grupo, ao qual as entidades pertencem, assim como as suas localizações. Os atributos como "AREA_GEOGRAFICA", "PAIS" e "SEDE" foram mapeados para a subdimensão "DimLocal", enquanto que os atributos "DIVISAO" e "CATEGORIA" foram direcionados para a subdimensão "DimPerimetro". Todos os atributos desta fonte mantiveram o seu formato. Por fim, em relação à fonte de dados "Account", esta contém as informações das rubricas, incluindo o código único e as descrições para cada uma delas. Estes atributos foram mapeados para a dimensão "DimRubrica" e mantiveram os mesmos formatos de dados.

Em suma, o mapeamento de dados permite planear e compreender a passagem dos dados das fontes para o *data warehouse* e o tipo de alterações que serão necessárias. Na Tabela 5 é possível também verificar as alterações dos nomes dos atributos, tornando-os mais uniformes no *data warehouse*.

Fonte de Dados	Atributo Origem	Tipo de Dados Origem	Tabela Destino	Atributo Origem	Tipo de Dados Destino
DRE.csv	YEAR	Int	DimPeriodo	ano	Int
DRE.csv	MONTH	Varchar	DimPeriodo	mes	Varchar
DRE.csv	COMPANY	Varchar	DimEmpresa	nome	Varchar
DRE.csv	AMOUNT	Float	TF_Transacao	valor	Int
Perimetro.csv	AREA_GEOGRAFICA	Varchar	DimLocal	continente	Varchar
Perimetro.csv	PAIS_DESCRICAO_DETALHADA	Varchar	DimLocal	pais	Varchar
Perimetro.csv	PAIS	Char(2)	DimLocal	sigla	Char(2)
Perimetro.csv	DIVISAO	Varchar	DimPerimetro	nome	Varchar
Perimetro.csv	CATEGORIA	Varchar	DimPerimetro	setor	Varchar
Account.csv	ACCOUNT_ID	Varchar	DimRubrica	codigo	Varchar
Account.csv	DSC	Varchar	DimRubrica	descricao_detalhada	Varchar
Account.csv	DSC_GERAL	Varchar	DimRubrica	descricao	Varchar

Tabela 5: Mapeamento dos dados origem para as tabelas destino no *data warehouse*

4.3 A Implementação do Data Warehouse

Para implementar qualquer processo de povoamento no *data warehouse*, é imprescindível assegurar a sua existência prévia. Para isso, a implementação física do *data warehouse* foi realizada nos servidores e sistemas internos da empresa de maneira a garantir o controlo da infraestrutura, o acesso aos dados e a segurança.

A construção do *data warehouse* foi feita através de um conjunto de *queries* em *SQLite* para realizar as suas operações, em particular, a criação das tabelas de dimensão, uma a uma, e a tabela de factos. Assim, a criação do *data warehouse* é feita de "fora para dentro", uma vez que se começa pela criação das dimensões secundárias, das dimensões principais e, por último, a tabela de factos. Esta abordagem permite que as dimensões sejam primeiramente povoadas, deixando para o final a povoação da tabela de factos, devido à necessidade de aplicar as regras de integridade referencial que foram estabelecidas.

Posto isto, começou-se por criar as tabelas das dimensões secundárias, nomeadamente, a dimensão "DimLocal" e a dimensão "DimPerimetro", respetivamente. Seguidamente, criou-se a tabela de dimensão "DimEmpresa" que interligou-se com as dimensões secundárias, através das suas chaves primárias, bem como a tabela da dimensão "DimEmpresa_Hist", que se interligou com a dimensão "DimEmpresa" através da chave primária. Depois, foi criada a tabela da dimensão "DimPeriodo" e "DimRubrica". Por fim, fez-se a criação da tabela de factos "TF_Transacao", estabelecendo-se as devidas regras de integridade com as tabelas de dimensão diretamente relacionadas. Após a criação de todas as tabelas do *data warehouse*, o sistema ficou preparado para ser povoado.

4.4 Implementação do Sistema ETL

O sistema de **ETL** é um processo usado para povoar os dados do *data warehouse* com os dados que forem extraídos das fontes de dados. Este é um processo fulcral que deve ser bem delineado, para que se possa garantir que os dados estejam disponíveis para a análise e que estejam de acordo com os requisitos definidos para a implementação do *data warehouse*. O sistema de povoamento envolve três fases diferentes de trabalho: extração, transformação e carregamento dos dados.

4.4.1 Extração dos Dados das Fontes

A primeira fase do processo de **ETL** envolve a extração dos dados das suas fontes originais. Após o estudo das três principais fontes de dados, estabeleceu-se uma conexão com cada uma dessas fontes por forma a se poder aceder aos dados nelas contidas. Os dados foram extraídos utilizando um *script Python*, que enviou os dados para um sistema de gestão de base de dados *SQLite*. Na Figura 13 é possível verificar todos os dados que foram extraídos das fontes "DRE", "Perimetro" e "Account", bem como os dados que foram selecionados para a etapa da transformação.

4.4.2 Transformação dos Dados

A fase seguinte do processo de **ETL** envolve a transformação dos dados. Esta é uma das fases mais complexas e cruciais do processo de **ETL**, uma vez que é nesta fase que se garantirá a qualidade e a veracidade dos dados. É muito natural que, ao se analisar os dados, se encontrem algumas inconsistências, erros, informação duplicada ou, valores omissos. De modo a resolver-se esses problemas, realizaram-se diversas tarefas de limpeza e de transformação numa zona de retenção de dados.

Uma área de retenção é uma componente essencial durante o processo de **ETL**, na qual se armazenam os dados em bruto antes de serem processados e carregados no *data warehouse*. Durante esta etapa, os dados são limpos, transformados e enriquecidos para que possam ser utilizados de forma consistente por outros processos posteriormente.

Para criar uma área de retenção foi necessário definir a estrutura das tabelas necessárias para armazenar os dados requeridos. Estrategicamente, optou-se por estruturá-las de igual forma à estrutura das tabelas do *data warehouse*, para permitir, mais tarde, um carregamento ágil e direto no sistema. No entanto, não foram estabelecidas ligações entre as tabelas referidas. Em todas as tabelas da área de retenção, foi criada uma chave primária incremental, que atribui um identificador único a cada novo registo. No entanto, na tabela "AR_Empresa" o caso é diferente. Se a chave primária fosse incremental,

duas ou mais empresas com o mesmo nome, poderiam ter identificadores únicos diferentes, após uma atualização no *data warehouse*. Assim, para garantir a unicidade dos dados e evitar dados duplicados, a solução adotada foi aplicar uma função *hash* no tratamento de dados da tabela "AR_Empresa". A função *hash* é determinística e atribui sempre o mesmo código para o mesmo *input*.

De seguida, fez-se uma seleção dos atributos relevantes para o *data warehouse* (Figura 13) e a sua consequente inserção nas tabelas temporárias da área de retenção, nomeadamente na tabela "AR_Empresa", "AR_Periodo", "AR_Rubrica", "AR_Perimetro", "AR_Local" e "AR_Transacao". Isto permite que apenas seja extraído o essencial e que não haja informação desnecessária. Para cada uma das tabelas realizou-se as transformações adequadas e necessárias. A limpeza de dados consistiu na identificação e correção de erros. Já a transformação dos dados consistiu na alteração do tipo de dados ou do seu conteúdo para que pudessem ajustar-se às necessidades do *data warehouse*. Para além disso, adicionou-se alguma informação aos dados para torná-los mais completos, como o atributo "nrMes" para a tabela "AR_Periodo" e o atributo "tipo" de conta para a tabela "AR_Rubrica".

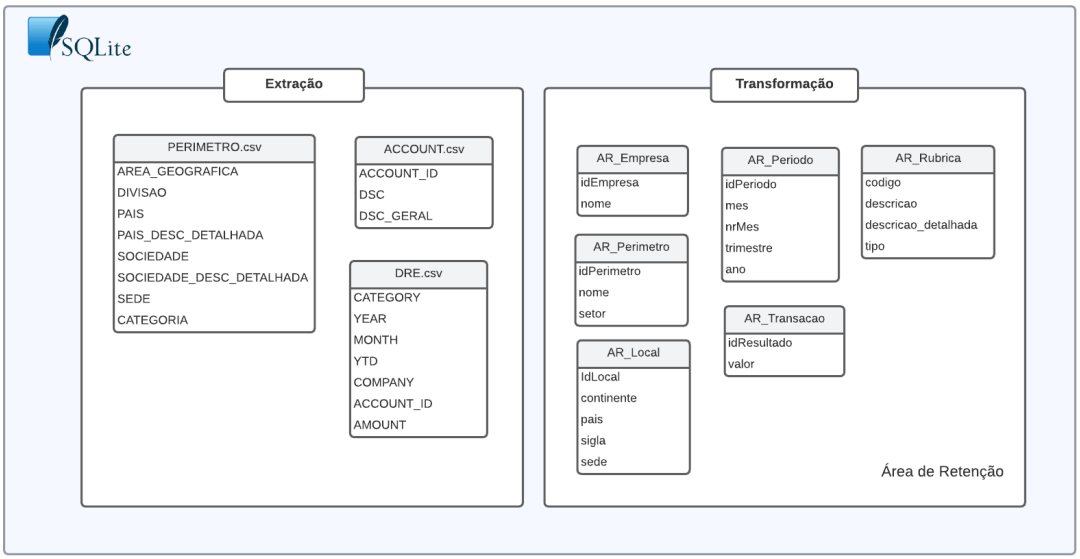


Figura 13: Ilustração da área de retenção desenvolvida em SQLite

4.4.3 O Carregamento dos Dados

O carregamento dos dados é realizado na última etapa do processo de **ETL**. Esta fase é fundamental para a inserção e manutenção dos dados no *data warehouse*, uma vez que garante que os mesmos estejam preparados para serem explorados e analisados pelas ferramentas de *business intelligence*. Uma vez criado o *data warehouse*, já com o formato de dados correto para cada atributo, é apenas necessário extrair os dados das tabelas temporárias da área de retenção para cada tabela de dimensão e tabela de

factos. Durante o processo de carregamento dos dados, o sistema verifica a existência de novos dados e, se houver, realiza o carregamento no *data warehouse*. Se os dados já estiverem presentes no *data warehouse*, o processo de povoamento não é realizado.

4.4.4 Esquema do Sistema de Povoamento ETL

Na Figura 14 está apresentado um esquema feito em **BPMN** do processo de povoamento de todo o sistema de *data warehousing*. O sistema inicia o processo com a extração das fontes de dados, em paralelo, utilizando um *script Python*. De seguida, para cada fonte, são selecionados apenas os atributos essenciais para o *data warehouse* e são inseridos numa área de retenção, na qual são realizadas todas as tarefas de limpeza e de transformação. Após esse processo, os dados são carregados para o *data warehouse*. É crucial manter uma sequência lógica aquando do carregamento de cada tabela, devido aos relacionamentos que existem entre as tabelas, de modo a preservar a integridade referencial dos dados. Embora a extração e a transformação ocorram em paralelo, o carregamento de cada tabela segue uma ordem específica, começando pelas tabelas de subdimensão, seguidamente pelas tabelas de dimensão e, por último, pela tabela de factos.

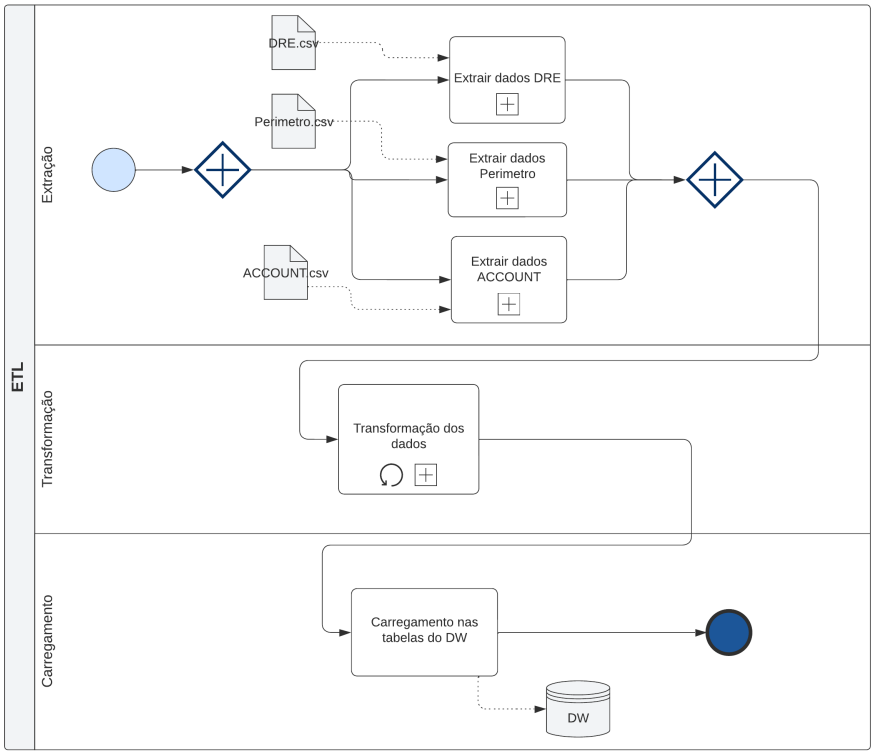


Figura 14: Esquema em BPMN do sistema de povoamento ETL

4.5 Povoamento das Dimensões e Tabela de Factos

De seguida, através de esquemas em **BPMN**, é explicado como é realizado o processo de povoamento para cada tabela de dimensão e de factos no *data warehouse*. O objetivo é garantir que sejam extraídos os dados necessários, que seja feito o seu tratamento conforme as suas necessidades e especificidades e, no final, que sejam povoados nas tabelas do *data warehouse*.

4.5.1 A Dimensão "DimPeriodo"

Para fazer o povoamento da dimensão "DimPeriodo" são extraídos dados da fonte "DRE" (Tabela 2), incluindo os atributos "YEAR" e "MONTH". Estes são inseridos na tabela "AR_Periodo", que foi criada na área de retenção. Durante o processo de transformação, os atributos extraídos passaram a ser designados por "ano" e "mes", tendo sido também criados novos atributos que incluem o número do mês designado por "nrMes" e o número do trimestre designado por "trimestre". O número do mês facilita a consulta e a análise aos dados, assim como o atributo "trimestre" que foi um dos requisitos solicitados. Além disso, foi realizado um processo de tradução do texto, originalmente em inglês, para português. Essa tradução garante que todos os dados estejam na mesma língua, uniformizando os dados e facilitando a sua compreensão nos processos de análise. Por fim, os dados da tabela temporária são inseridos na tabela "DimPeriodo" no *data warehouse*. Na Figura 15, é possível ver o processo de povoamento desta dimensão, desde a extração dos dados da fonte, até ao seu carregamento no *data warehouse*.

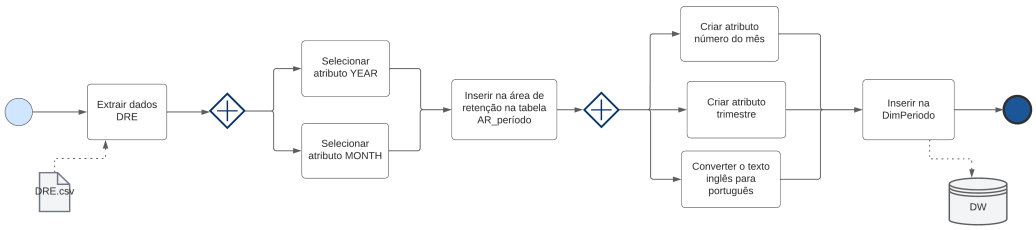


Figura 15: Esquema em BPMN do sistema de povoamento da dimensão "DimPeriodo"

4.5.2 A Dimensão "Rubrica"

O processo de povoamento da dimensão "DimRubrica" (Figura 16) inicia-se com a extração dos dados da fonte "Account", com a extração dos valores dos atributos "ACCOUNT_ID", "DSC" e "DSC_GERAL" que depois são inseridos na área de retenção na tabela "AR_Rubrica". Em seguida, durante o processo de transformação, esses atributos são renomeados para "codigo", "descricao_detalhada" e "descricao", respetivamente, sendo removidos os acentos e convertido o texto para maiúsculas. Neste processo de

povoamento, também é criado um atributo com o "tipo" de rubrica para acrescentar informação que permita aos utilizadores saber se cada rubrica corresponde a um gasto ou uma receita. Após estas ações terem sido realizadas, os dados são, então, inseridos na tabela "DimRubrica" no *data warehouse*.

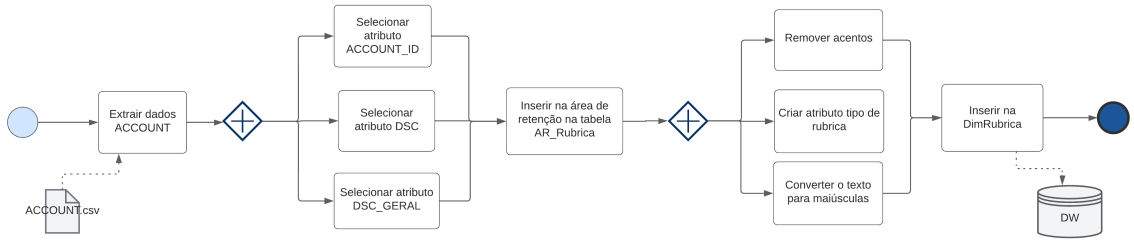


Figura 16: Esquema em BPMN do sistema de povoamento da dimensão "DimRubrica"

4.5.3 A Subdimensão "DimLocal"

A subdimensão "DimLocal" começa com a extração dos dados da fonte "Perimetro" inserindo os valores dos atributos "AREA_GEOGRAFIA", "PAIS_DESCRICAO_DETALHADA", "PAIS", "SEDE" na tabela "AR_local" na área de retenção. Estes atributos são renomeados para "continente", "pais" e "sigla", respetivamente. Seguidamente, estes dados passam por um processo de transformação, que envolve remover acentos, converter o texto para maiúsculas e substituir possíveis nulos pelo valor "desconhecido". Posteriormente, os dados são carregados no *data warehouse*, especificamente na tabela "DimLocal". Este processo está ilustrado na Figura 17.

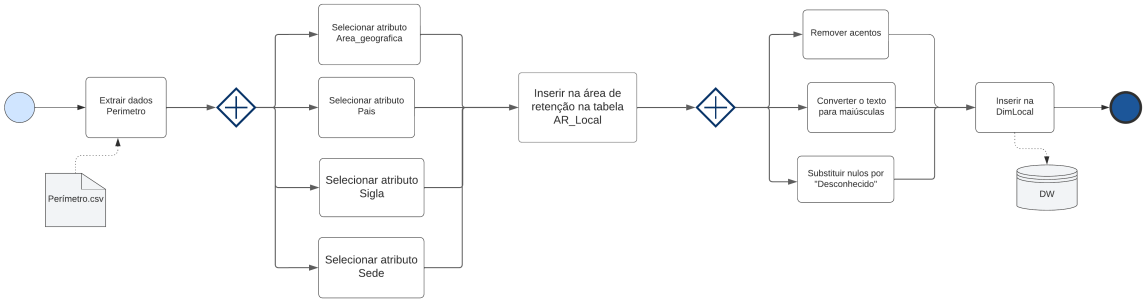


Figura 17: Esquema em BPMN do sistema de povoamento da dimensão "DimLocal"

4.5.4 A Subdimensão "DimPerimetro"

O processo de povoamento da dimensão "DimPerimetro" (Figura 18) inicia-se com a extração dos dados da fonte "Perimetro" inserindo os valores extraídos dos atributos "DIVISAO" e "CATEGORIA" na área de

retenção na tabela "AR_Perimetro". O atributo "CATEGORIA" é renomeado para "setor" e os dados são convertidos para letras maiúsculas. No final, os dados preparados são povoados na tabela "DimPerimetro" do *data warehouse*.

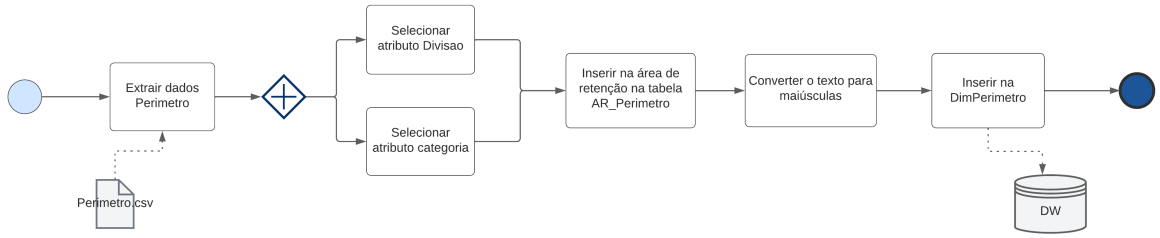


Figura 18: Esquema em BPMN do sistema de povoamento da dimensão "DimPerimetro"

4.5.5 A Dimensão "DimEmpresa"

No povoamento da dimensão "DimEmpresa", os dados do atributo 'COMPANY" são extraídos da fonte "DRE" e depois inseridos, na área de retenção na tabela "AR_Empresa". Para isso, foi necessário remover acentos e converter o texto para maiúsculas de todo os nomes das empresas. Por fim, os dados são carregados no *data warehouse* na tabela "DimEmpresa". Na Figura 19 pode-se verificar o processo de povoamento desta dimensão.

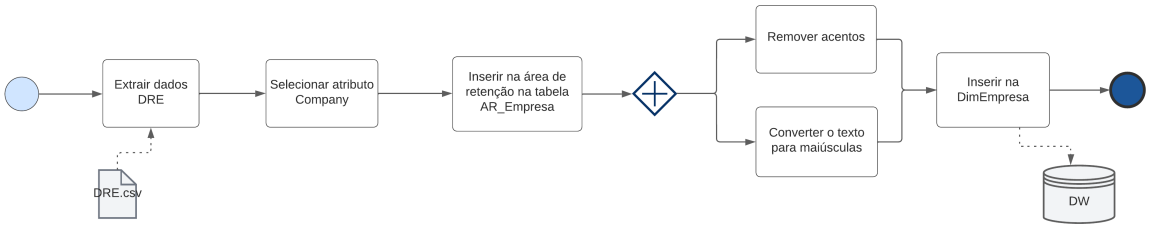


Figura 19: Esquema em BPMN do sistema de povoamento da dimensão "DimEmpresa"

4.5.6 A Dimensão "DimEmpresa_Hist"

A dimensão "DimEmpresa_Hist" é povoada com os dados históricos da dimensão "DimEmpresa". Na Figura 20, o processo começa com o sistema a verificar se os novos registos já estão na dimensão "DimEmpresa". Se não estiverem, o sistema insere-os na tabela. Porém, se os novos registos forem atualizações, o sistema transfere os registos desatualizados para a "DimEmpresa_Hist" e os registos atualizados são inseridos na dimensão "DimEmpresa".

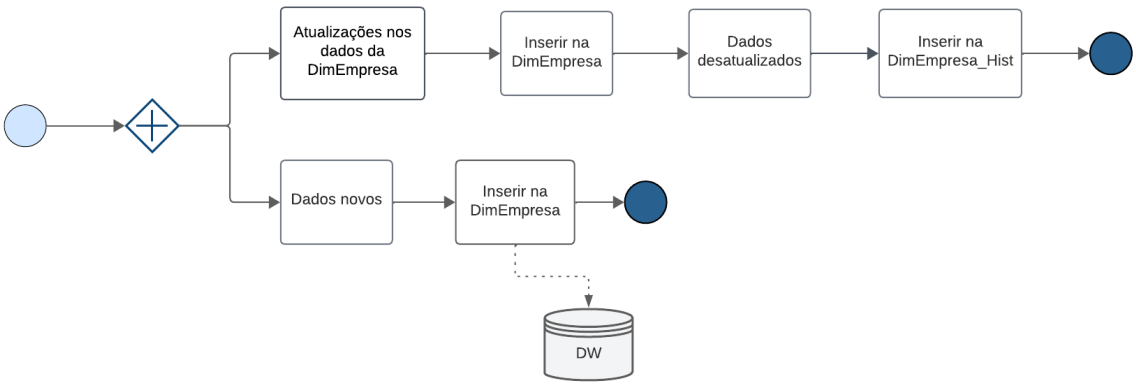


Figura 20: Esquema em BPMN do sistema de povoamento da dimensão "DimEmpresa_Hist"

4.5.7 A Tabela de Factos "TF_Transacao"

A tabela de factos "TF_Transacao" é a última tabela a ser povoada no *data warehouse*. Esta tabela contém apenas uma medida, o valor de transação. O sistema de povoamento inicia com a extração do atributo "AMOUNT" da fonte "DRE" que possui os valores de todas as transações realizadas. Depois, os dados são inseridos na tabela "AR_Transacao" na área de retenção. O atributo "AMOUNT" é renomeado para "valor" e o seu tipo é convertido para um número inteiro, porque os valores originais deste atributo são números decimais na fonte. Após este processo de transformação, os dados são carregados na tabela "TF_Transacao" no *data warehouse*.

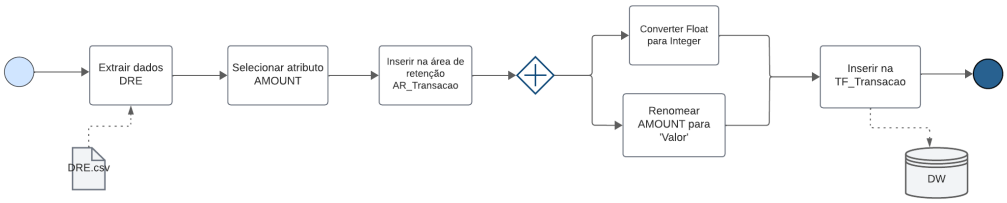


Figura 21: Esquema em BPMN do sistema de povoamento da tabela de factos "TF_Transacao"

4.6 Validação do Sistema de Povoamento ETL

Durante o processo de **ETL** foram realizadas várias ações de validação, com o propósito de garantir a integridade referencial, qualidade e consistência dos dados, bem como a *performance* do próprio sistema.

As validações realizadas foram baseadas nas diretrizes elaboradas pelos autores Ferreira et al. (2010) que referem vários testes de validação que procuram garantir que todos os dados esperados sejam carregados, que o sistema **ETL** rejeite ou substitua os valores por defeito, que o carregamento dos dados e a *performance* seja eficiente e que o processo de **ETL** funcione corretamente. Para o desenvolvimento e teste do sistema, optou-se por utilizar um pequeno volume de dados. Como tal, apenas foram utilizados os dados relativos ao ano de 2015. Com esses dados, foram realizadas as seguintes validações:

- Comparou-se o número de registos entre os dados extraídos das fontes e os dados carregados para o *data warehouse*, garantindo que não houvesse duplicados ou dados em falta.
- Validou-se os atributos que possuem valores únicos, garantindo a sua unicidade.
- Procurou-se verificar os conteúdos de cada atributo, com fim de validar se o formato e o tipo de dados estavam corretos.
- Procurou-se testar a integridade referencial garantindo que as chaves estrangeiras e os relacionamentos entre as tabelas fossem respeitados.
- Garantiu-se que os dados com valores nulos fossem sempre convertidos para "desconhecido".
- Procurou-se garantir que todos os processos de transformação fossem aplicados.
- Efetuaram-se operações de consultas simples para validar a *performance* do *data warehouse*.

Posteriormente, realizou-se um segundo teste, carregando os dados relativos ao ano 2016. Este teste foi essencial para verificar se o sistema cumpria todas as especificidades necessárias para a transformação dos dados. Após a realização deste teste, é possível afirmar que o sistema de **ETL** estava preparado para os próximos carregamentos de dados.

4.7 Refrescamento do Data Warehouse

O refrescamento de um *data warehouse* envolve o processo de atualização dos dados armazenados no sistema de *data warehousing* a partir das suas fontes de dados. Este processo deve ser feito regularmente para manter a informação do *data warehouse* atualizada e disponível para consulta e análise dos dados. No entanto, escolher a melhor abordagem para a atualização dos dados depende muito das necessidades e dos requisitos do sistema e, principalmente, do negócio em questão. Assim, é necessário compreender qual será a melhor estratégia de refrescamento do *data warehouse*.

O carregamento diferencial é considerado a melhor abordagem, uma vez que apenas os dados novos ou modificados desde a última atualização do sistema são extraídos. Isto é, apenas os dados que foram alterados nas suas fontes são carregados para o *data warehouse*. Por exemplo, se no próximo carregamento houver um novo grupo de empresas, apenas é adicionado esse grupo, sem a necessidade de recarregar os dados das empresas já existentes no sistema. O mesmo se aplica para o período de tempo em que só é acrescentado dados novos caso ainda não esteja no sistema, por exemplo, o atributo ano. Este tipo de carregamento é particularmente útil em cenários com grandes volumes de dados e alterações frequentes nas fontes de dados, tornando o processo mais eficiente, reduzindo o tempo e os recursos envolvidos e, mantendo o *data warehouse* sempre atualizado.

Quanto ao agendamento do refrescamento dos dados, este deve ser planeado de acordo com as necessidades do negócio. Alguns sistemas podem exigir atualizações diárias, enquanto que outros podem funcionar bem com intervalos de tempo maiores. No caso deste sistema de *data warehousing*, optou-se por atualizar os dados no primeiro dia de cada mês, o que abrange apenas os dados relativos ao mês anterior. Esta escolha é especialmente relevante, uma vez que as demonstrações financeiras consolidadas são enviadas mensalmente para o sistema, o que demonstra a importância de alinhar o agendamento do refrescamento do *data warehouse* com as necessidades e processos de negócio. Além disso, à medida que o *data warehouse* cresce e evolui com novos dados, a estrutura de dados e os processos de refrescamento devem ser ajustados e otimizados para garantir o desempenho contínuo do sistema.

4.8 Visualização de Dados

Para aceder aos dados das tabelas do *data warehouse*, o acesso foi possível através de uma conexão direta entre o *PowerBI* e o *data warehouse*. Este tipo de conexão permite reduzir o tempo de acesso aos dados, bem como o torna mais eficiente, reduz o risco de erros humanos e garante uma execução consistente. Além de simplificar o acesso aos dados, também permite que os dados sejam atualizados automaticamente no *PowerBI*. Ou seja, sempre que ocorrer um novo carregamento de dados no *data warehouse*, as tabelas no *PowerBI* serão atualizadas automaticamente, assim como os *dashboards* criados. Este é um processo fundamental, uma vez que garante que os utilizadores tenham um acesso mais rápido aos dados mais recentes, permitindo uma maior eficiência e agilidade no processo.

De seguida, serão apresentados três *dashboards*, dois deles orientados para análise do desempenho financeiro da empresa e um outro para a previsão de vendas. O desenvolvimento destes *dashboards* teve como finalidade responder aos requisitos expostos anteriormente no capítulo 3. O processo de criação dos *dashboards* não levantou grandes problemas, uma vez que os seus requisitos foram bem definidos. Primeiramente, foi criado um *dashboard* direcionado para a análise dos resultados gerais do

grupo, para permitir a análise global do desempenho financeiro da empresa. De seguida, desenvolveu-se um *dashboard* especificamente orientado para a análise do volume de negócios do grupo. Por último, foi elaborado um *dashboard* que permitisse obter uma previsão de vendas da empresa para o ano seguinte.

De referir que, por questões de privacidade, os nomes das empresas do grupo foram alterados, sendo apresentados nos *dashboards* nomes de empresa fictícios.

4.8.1 Análise dos Resultados Gerais

Neste *dasboard* é analisado duas das principais medidas de desempenho, o resultado líquido e o **EBITDA**. O resultado líquido é uma métrica crucial para avaliar a rentabilidade e o desempenho financeiro da empresa. Esta reflete o valor líquido que a empresa efetivamente ganha após considerar todas as despesas e impostos, representando, portanto, o resultado das operações da empresa. Um resultado líquido positivo indica que a empresa obteve lucro, enquanto que um resultado líquido negativo indica prejuízo para a empresa.

Quanto ao **EBITDA**, este é uma medida financeira amplamente utilizada para avaliar a capacidade de gerar lucro e o desempenho da empresa ao longo do tempo. O valor do **EBITDA** representa o lucro operacional antes de se levar em consideração os efeitos das despesas financeiras, impostos, depreciações e amortizações. Quando este é positivo, significa que a empresa está a gerar lucro antes de considerar as despesas. Contudo, é muito possível que uma empresa obtenha um **EBITDA** alto, indicando um bom desempenho operacional, e, no final, obter um resultado líquido baixo, devido a valores significativos das despesas financeiras. Para além disso, o tamanho da empresa também afeta o resultado, uma vez que empresas maiores podem ter um resultado maior, devido a ter muitas operações. Na Figura 22 são apresentados os resultados obtidos nos anos 2015 e 2016.

Figura 22: Resultado líquido e EBITDA

Na Figura 23, está apresentado o número de empresas que obtiveram lucro ou prejuízo, por trimestre. Como se pode verificar através da Figura 23, em todos os trimestres houve sempre um maior número de empresas com lucro, destacando-se o quarto trimestre como o período com mais empresas que obtiveram

lucro. Isto permite ter uma visão geral acerca do grupo para ver se ele apresenta mais entidades lucrativas, pois se caso todos os trimestres houvesse mais empresas com prejuízo seria necessário repensar a estratégia de negócio.

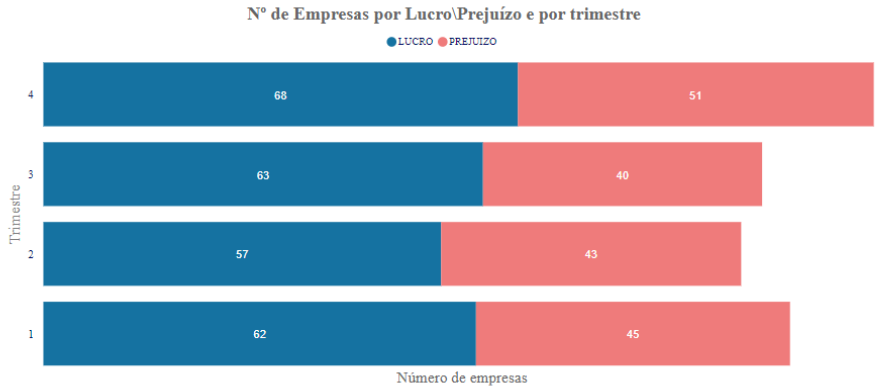


Figura 23: Gráfico do número de empresas que obtiveram lucro e prejuízo

Os gráficos seguintes (Figura 24) identificam as três principais empresas com os maiores prejuízos e lucros. Através da sua análise, é possível identificar quais as empresas que precisam de uma atenção especial, quer seja para melhorar o seu desempenho ou para aproveitar o seu potencial de crescimento. Como se pode observar pela Figura 24, as empresas que apresentam prejuízo nos anos 2015 e 2016 são a "Comida Orgânica", o "Jardim Florido" e a "Fabrica de Ideias". Para o mesmo período, as empresas que apresentam maior lucro são a "Easymobile", a "Magia do Saber" e o "Mercado Express".

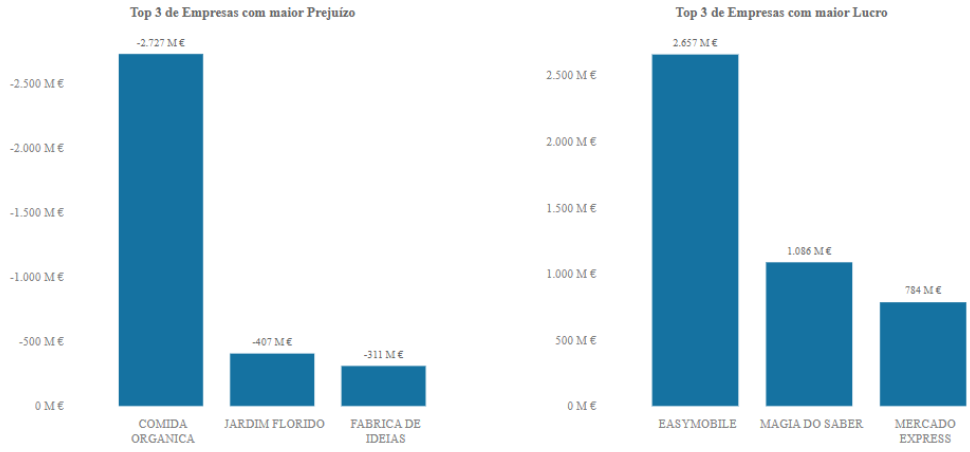


Figura 24: Gráfico do top 3 de empresas com maior lucro e maior prejuízo

De seguida, na Figura 25, apresenta-se um gráfico de evolução do resultado líquido ao longo dos meses, que evidencia picos altamente significativos nos meses de abril e maio e, posteriormente, entre

setembro e outubro. Neste gráfico é possível destacar alguns meses com um resultado líquido negativo, como fevereiro, junho, agosto e novembro, meses críticos que precisam de novas estratégias de negócio para potenciar o crescimento.

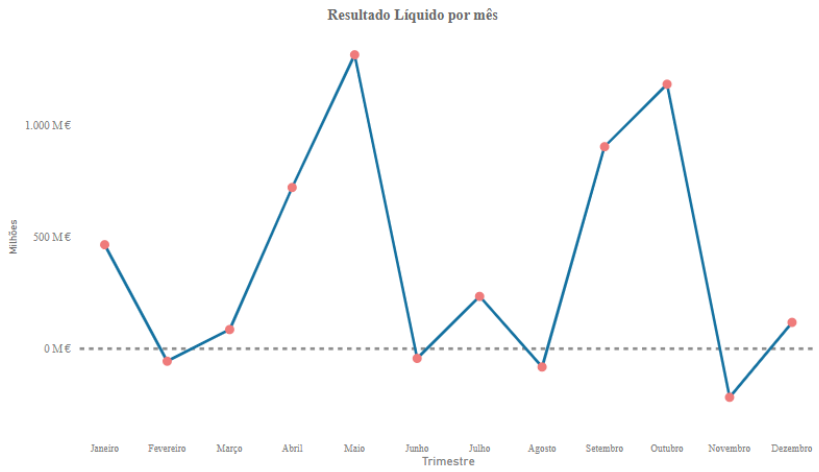


Figura 25: Gráfico de evolução do resultado líquido nos anos 2015 e 2016

Por fim, na Figura 26 apresenta-se o *dashboard*, com todos os gráficos descritos anteriormente, que fornece uma visão geral da situação financeira do grupo, com dados relativos aos anos de 2015 e 2016. Neste *dashboard* também é possível visualizar a distribuição das empresas sediadas por país, sendo Portugal o local com maior representatividade. O objetivo é obter uma visão abrangente do total de empresas que fazem parte do grupo e do local onde estão localizadas. Além disso, permite filtrar os dados por país, de modo a obter uma visão mais detalhada dos resultados das empresas por país.

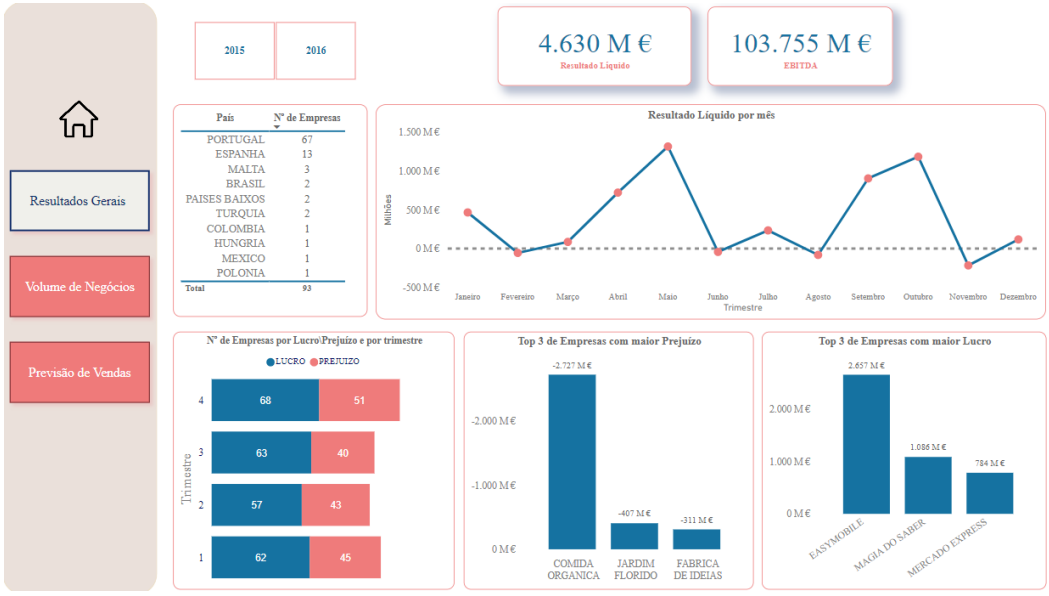


Figura 26: *Dashboard* de análise de resultados gerais do grupo

4.8.2 Análise do Volume de Negócios

Na Figura 27 está apresentado o valor das métricas da receita líquida, volume de negócios e margem de lucro. Enquanto que a receita líquida diz respeito à receita total da empresa após a dedução de todos custos e despesas, refletindo efetivamente o valor que a empresa ganhou, o volume de negócios é relativo ao valor total das vendas da empresa, sem deduzir quaisquer custos ou despesas. Assim, a receita líquida trata-se de uma medida mais precisa do desempenho financeiro, enquanto que o volume de negócios é uma medida que pode ser usada para comparar empresas de diferentes setores.

A margem de lucro é uma métrica financeira que mede a quantidade de lucro que a empresa obteve em relação às suas receitas. Se a empresa apresentar uma margem de lucro negativa, indica que teve um prejuízo, o que significa que a empresa gastou mais do que recebeu. Portanto, quanto maior for a margem de lucro, maior será a proporção do lucro em relação à receita. Deste modo, consegue-se ver que a diferença entre o volume de negócios e a receita líquida não é muito significativa, tendo existido poucas deduções de despesas. Para além disso, por cada venda realizada em 2015 e 2016, obteve-se uma margem de lucro de 4%.

Figura 27: Receita líquida, volume de negócios e margem de lucro

Seguidamente, na Figura 28, pode-se observar a percentagem de valor por cada categoria de receita, como vendas, prestações de serviços, rendimentos e ganhos financeiros, ganhos de investimento e outros rendimentos. A maior percentagem de rendimento são as vendas, seguindo-se de outros rendimentos e de prestações de serviços. Apesar das vendas normalmente serem a maior fonte de receita para as entidades do grupo, existem entidades que apenas prestam serviços a outras entidades.

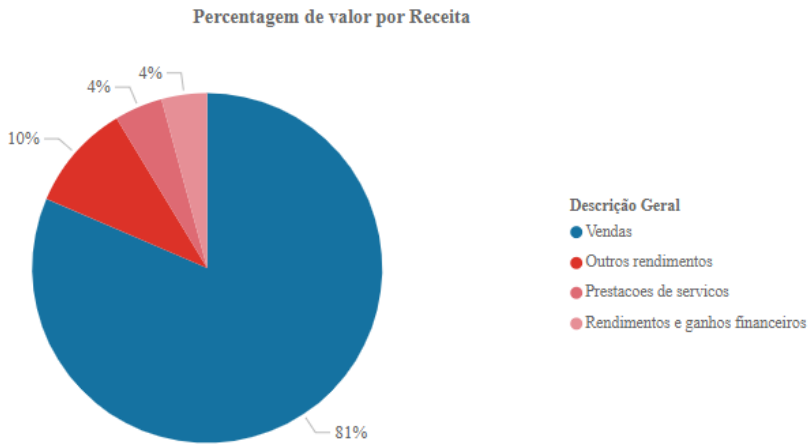


Figura 28: Gráfico de percentagem de valor por receita

Na Figura 29 pode-se ver o "top" três de empresas com maior volume de negócios, que foram "Chef em Casa", a "Comida Organica" e a "Comida com Prazer". Um facto interessante a observar é que nem sempre as empresas com maiores volumes de negócios conseguem apresentar lucros no final do período do exercício. Por exemplo, a empresa "Comida Organica", apesar de ser uma das empresas com maior volume de negócios, também apresenta o maior prejuízo (Figura 24). Isto significa que as receitas da empresa provavelmente não são suficientes para compensar todas as despesas.

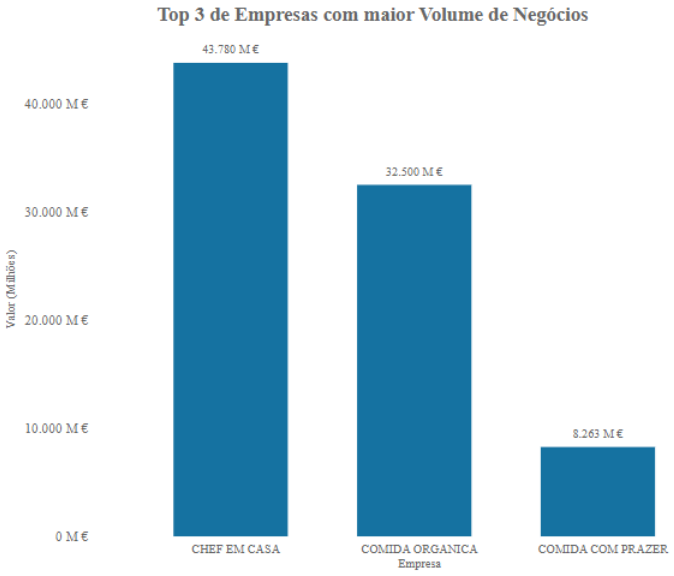


Figura 29: Gráfico de top 3 de empresas com maior volume de negócios

Por fim, na Figura 30 está apresentada a evolução das receitas e o custo de vendas ao longo dos

meses. A finalidade é analisar em que alturas do ano é que os custos de vendas são maiores à receitas de vendas. Através do gráfico (Figura 30) pode-se verificar que ao longo do período as receitas são, geralmente, maiores que o custo de vendas, destacando os meses de agosto a outubro como aqueles que têm maior receita de vendas. Já nos meses de junho e de novembro, a receita não compensa significativamente os gastos.

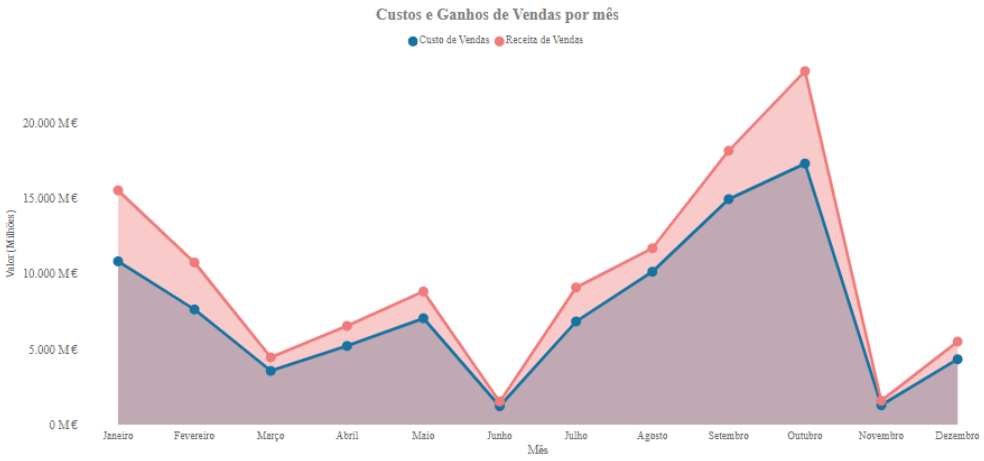


Figura 30: Gráfico de evolução dos custos e ganhos de vendas por mês

Na Figura 31 pode-se ver o *dashboard* do volume de negócios do grupo para os anos 2015 e 2016. Neste *dashboard* está incorporada uma tabela com as empresas todas do grupo, que pode ser filtrada para visualizar a informação relativa a cada empresa. No entanto, o objetivo principal é obter uma visão geral do volume de negócios do grupo. Sempre que houver novo refrescamento no *data warehouse*, o *dashboard* é atualizado com novos dados.

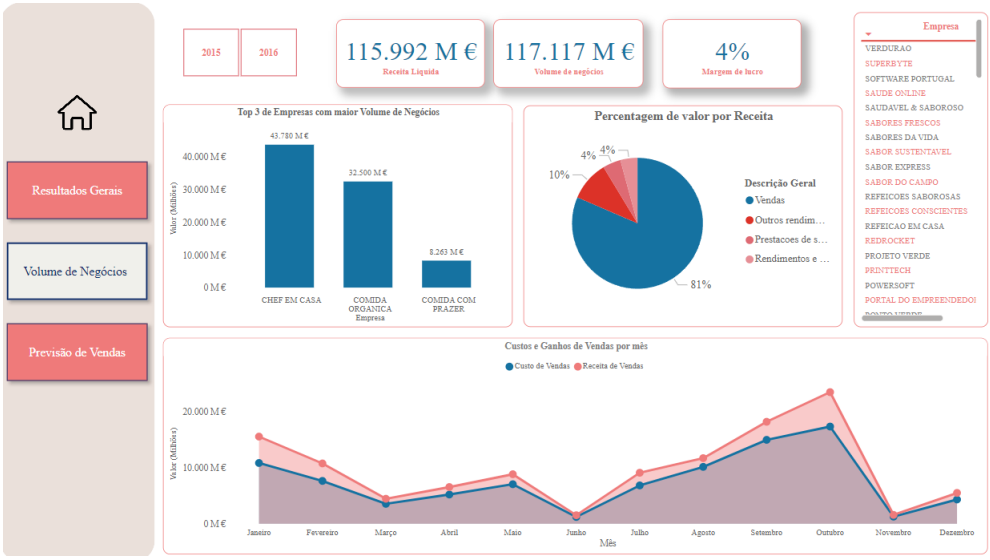


Figura 31: *Dashboard* de análise de volume de negócios

4.9 Previsão de Vendas

A previsão de vendas é um processo complexo que envolve diversas técnicas e métodos para estimar futuras receitas de vendas. Trata-se de um desafio importante e fulcral para as empresas, uma vez que uma previsão deve permitir ajudar a tomar decisões estratégicas, com base em informação histórica e confiável. Na Figura 32 pode-se observar uma previsão de vendas realizada para os meses seguintes do ano 2017. Esta previsão desempenha um papel crucial no planeamento financeiro, permitindo estimar as despesas e as receitas, bem como calcular o lucro esperado. Além disso, ajuda a empresa a tomar decisões estratégicas, tais como a definição de novos objetivos, o desenvolvimento de novos produtos ou de serviços e, até uma possível expansão para novos mercados. Também permite uma melhoria no desempenho da empresa, identificando oportunidades de crescimento e áreas de trabalho que precisam de melhorias.

Existem vários métodos de previsão de vendas. A escolha do método mais adequado deve ser feita com base na quantidade de dados disponíveis, nas necessidades da empresa e no contexto específico do problema em questão. Uma vez que apenas foram fornecidos dados relativos aos anos de 2015 e 2016 para o desenvolvimento do *data warehouse*, este acaba por fornecer pouca informação histórica. Por essa razão, o modelo **ARIMA** foi considerado o mais adequado.

4.9.1 O Modelo ARIMA

Para a realização da previsão de vendas para o ano seguinte, 2017, foi utilizado o modelo **ARIMA**. Este modelo foi escolhido por ser um método simples de compreender e de implementar, e por ter a capacidade de lidar com tendências e sazonalidades, o que o torna adequado para modelar séries temporais com padrões, já que os dados das vendas dos anos 2015 e 2016 apresentam tendências ao longo do tempo.

A aplicação deste modelo foi feita através de uma implementação específica, em *Python*. No processo de implementação, foi utilizada a biblioteca *"pmdarima"* (Smith et al., 2017), que possui várias funções que permitem executar modelos **ARIMA** em *Python*. A função *"auto_arima()"* da biblioteca tem como propósito explorar várias combinações possíveis entre as ordens p, d e q e escolher o modelo que melhor se ajusta aos dados com base num conjunto de critérios, como o **AIC** ou o **BIC**. Estes critérios são duas medidas de qualidade usadas para avaliar o modelo. Quanto menor for o valor de ambos, melhor será modelo.

A função, com base em vinte e quatro observações, encontrou os melhores parâmetros para o modelo ARIMA(1,2,3) e para o modelo SARIMA(1,0,0)12 que resultou do modelo SARIMAX. Os resultados obtidos com esta função estão apresentados na Tabela 6. O modelo SARIMAX é uma combinação entre o modelo

ARIMA não sazonal e o modelo **SARIMA** sazonal, sendo utilizado para modelar uma série temporal. Esta combinação de modelos é frequentemente utilizada quando se está a lidar com séries temporais, que apresentem tanto padrões sazonais como tendências não sazonais. Neste caso, o modelo **ARIMA** apresenta um componente autorregressivo de ordem a 1 (AR), um componente de diferenciação com ordem a 2 (I) e um componente de média móvel de ordem a 3 (MA). Já o modelo **SARIMA** apresenta um componente autorregressivo sazonal de ordem a 1 (SAR), nenhuma diferenciação sazonal (DS) e nenhuma média móvel sazonal (SMA). Para além disso, sugere uma sazonalidade anual (doze meses).

Resultados SARIMAX	
Melhor modelo	ARIMA(1,2,3)(1,0,0) [12]
Número de Observações	24
AIC	1041.854
BIC	1048.400

Tabela 6: Resultados do melhor modelo ARIMA

4.9.2 O Resultado Final

Após a aplicação do modelo, foi gerado um gráfico que nos permite comparar os dados previstos de vendas e os dados correspondentes a anos anteriores. Na Figura 32 está apresentada a previsão de vendas mensais para o ano 2017. Através da sua análise, pode-se verificar que as previsões de vendas para 2017 seguem padrões sazonais semelhantes aos de anos anteriores. Notavelmente, a tendência sazonal aponta para aumentos nas vendas durante os meses de abril a maio e, posteriormente, entre julho e outubro. No entanto, o modelo implementado prevê que os meses de fevereiro, junho e novembro continuarão a registar vendas baixas, mantendo, assim, a consistência dos dados históricos.

Em suma, o resultado obtido mostra que o modelo é eficaz ao se manter próximo da realidade observada em anos anteriores. O modelo também permite auxiliar na capacidade de prever comportamentos futuros, proporcionando uma base sólida para o desenvolvimento de estratégias e de processos de tomada de decisão sustentados.

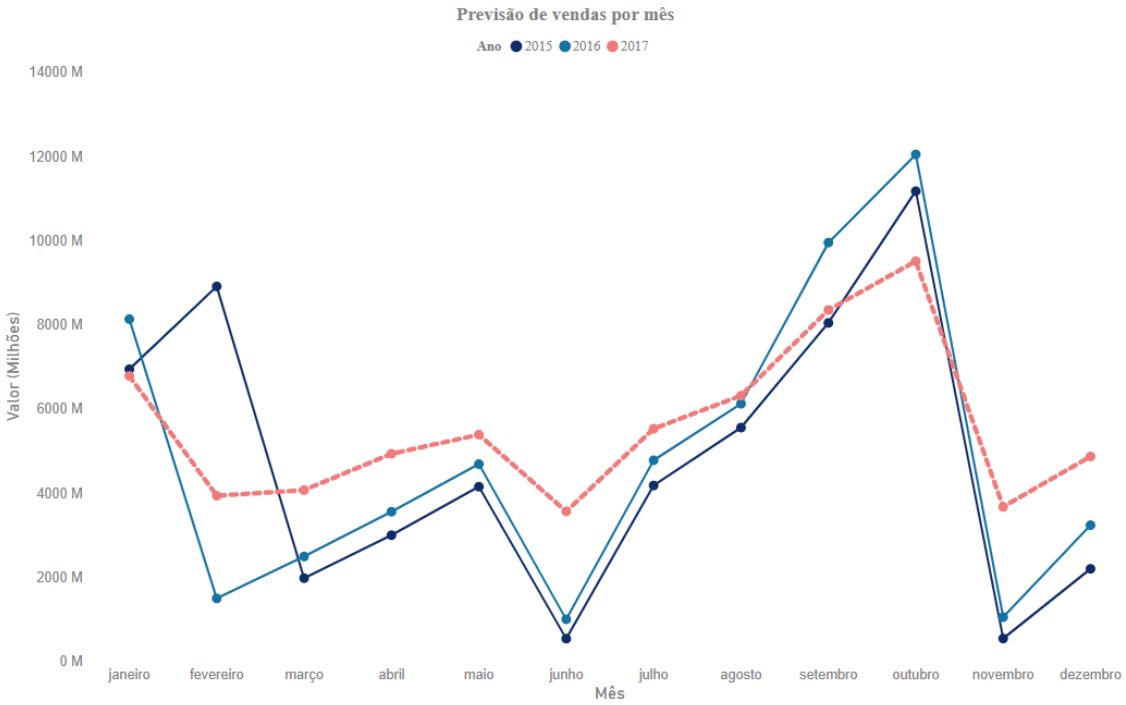


Figura 32: Gráfico de previsão de vendas mensais para o ano 2017

5 Conclusão

5.1 Conclusões

Num mundo cada vez competitivo, os dados têm trazido mudanças significativas ao mercado empresarial, em termos gerais, e às empresas, em particular. Com o desenvolvimento das tecnologias da informação e da comunicação, as empresas geram dados muito mais rapidamente e recolhem dados em grandes volumes. Desta forma, as empresas recorrem aos dados para que possam tomar decisões estratégicas mais fundamentadas, rápidas e eficientes que contribuam oportunamente para a sua vantagem competitiva. Por esta razão, as empresas procuram soluções eficazes que permitem guardar dados importantes para que sejam transformados em informação útil com o intuito de serem retirados *insights* valiosos para o negócio. Os sistemas de *business intelligence* surgem, assim, como um papel fulcral na obtenção de conhecimento para dar suporte à tomada de decisão, uma vez que fornecem ferramentas que permitem extrair, transformar e armazenar grandes quantidades de dados.

Nesta dissertação, foi desenvolvido um sistema de *data warehousing* para facilitar os processos de análise de informação na área da consolidação de contas. Começou-se por fazer um levantamento de requisitos que o sistema deveria ser capaz de responder, para compreender quais as necessidades dos agentes de decisão e o que se esperava que fosse alcançado no final. De seguida, com base nesses requisitos, foi realizada a modelação multidimensional de um *data warehouse*, o que permitiu esboçar o sistema através de um esquema em formato de estrela, com as tabelas de dimensão interligadas a uma tabela de factos. Para isso foi necessário caracterizar o *data warehouse* definindo o grão da tabela de factos e das suas tabelas de dimensão. Seguidamente, foram conhecidas as fontes de informação e foi realizado o mapeamento dos dados das fontes de informação para as tabelas do *data warehouse*.

Após esta etapa, foi desenvolvido um processo de **ETL**, para fazer a extração dos dados das fontes para uma área de retenção e consequente povoamento do *data warehouse*. Nessa área de retenção foram executadas todas as transformações de dados necessárias para garantir a sua consistência e qualidade. No final, realizou-se o carregamento dos dados no *data warehouse*. Posteriormente, os dados do *data warehouse* foram explorados por uma ferramenta de visualização de dados, alimentando, assim,

os *dashboards* desenvolvidos, com informação para análise. Além disso, realizou-se uma previsão de dados através do modelo **ARIMA** que possibilita ainda sustentar as tomadas de decisão.

Por fim, este trabalho foi um culminar de aprendizagens ao nível técnico de ferramentas e de conhecimento de uma área financeira. Uma das dificuldades sentidas ao longo do trabalho foi o fornecimento de poucos dados para a realização do sistema, a limitação imposta pelo anonimato dos dados e, ainda, a falta de acesso a ferramentas de trabalho, particularmente, de **ETL** e de sistemas de gestão de base de dados. Por essa razão foi necessário recorrer a outras alternativas, como *SQLite* e *Python* para realizar o processo de **ETL** e ao mesmo tempo automatizá-lo. No entanto, permitiu conhecer ainda mais e pôr em prática estas ferramentas que são úteis para o mercado de trabalho. Além disso, esta área financeira tornou-se um desafio para quem não tem muitos conhecimentos acerca da mesma, tendo sido necessário estudá-la para estruturar o sistema e poder realizar análises financeiras. Conclui-se que, o trabalho realizado cumpriu com os objetivos inicialmente propostos. Tratou-se de um trabalho desafiante e enriquecedor no âmbito dos sistemas de *data warehousing*.

5.2 Trabalho Futuro

Após terminado este trabalho, verificou-se, que algumas melhorias poderiam ser realizadas através da manutenção e otimização do sistema. Por exemplo, o processo de **ETL** que se desenvolveu pode ser melhorado, pois tendo em conta a extração de mais dados ao longo do tempo, possivelmente será necessário adicionar mais condições de transformação aos dados. Com a obtenção de mais informação e dados históricos no *data warehouse*, poderão surgir novas necessidades de negócio e, consequentemente, novos requisitos que o sistema deve responder no final. Por essa razão, o sistema de *data warehousing* deve ser sempre adaptado às novas solicitações para ser bem sucedido. Além disso, também poderão ser feitas novas análises aos dados com a realização de novos *dashboards* que apresentem outro tipo de gráficos relevantes. Quanto à previsão de vendas, à medida que hajam mais dados históricos disponíveis, o modelo fica mais enriquecido e com maior probabilidade de prever eficazmente. No entanto, poderá ser adotado outro tipo de modelo de previsão mais complexo e eficaz, dependendo dos objetivos da organização. Finalmente, é importante acompanhar o sistema de *data warehousing*, com a finalidade de detetar possíveis anomalias ou erros que deverão ser tratados de imediato.

Bibliografia

Henrique Silvério Ferreira Abreu. Sistema de business intelligence numa empresa do setor têxtil. Master's thesis, Universidade do Minho, 2021.

Elsa Adriana e Maciel Silva. Consolidação de Contas na Quantal Group SA. Master's thesis, Universidade de Coimbra, 2018.

Seddiq Q. Abd Al-Rahman, Ekram H. Hasan, e Ali Makki Sagheer. Design and implementation of the web (extract, transform, load) process in data warehouse application. *IAES International Journal of Artificial Intelligence*, 12:765–775, 6 2023. ISSN 22528938. doi: 10.11591/ijai.v12.i2.pp765-775.

Grant. Allen, Michael. Owens, e Michael. Owens. *The definitive guide to SQLite*. Apress, 2010. ISBN 9781430232254.

Ladjel Bellatreche. Techniques d'optimisation des requêtes dans les data warehouses. In *6th International Symposium on Programming and Systems ISPS 2003 (ISPS 2003)*, pages 81–98, Alger, Algeria, May 2003.

O. Belo. Modelação Dimensional de Dados. Textos da Lição de Síntese, Provas de Agregação, Departamento de Informática, Escola de Engenharia, Universidade do Minho, 2012.

Hélder Daniel Borges. Exploração de Séries Temporais em Processos de Previsão de Vendas. Master's thesis, Universidade do Minho, 2015.

Surajit Chaudhuri e Umeshwar Dayal. An Overview of Data Warehousing and OLAP Technology. *SIGMOD Rec.*, 26(1):65–74, mar 1997. ISSN 0163-5808. doi: 10.1145/248603.248616. URL `https://doi.org/10.1145/248603.248616`.

Surajit Chaudhuri, Umeshwar Dayal, e Vivek Narasayya. An Overview of Business Intelligence Technology. *Commun. ACM*, 54:88–98, 08 2011. doi: 10.1145/1978542.1978562.

Min Chen, Shiwen Mao, e Yunhao Liu. Big Data: A survey. In *Mobile Networks and Applications*, volume 19, pages 171–209. Kluwer Academic Publishers, 4 2014. doi: 10.1007/s11036-013-0489-0.

Raquel Santos Costa. Previsão de Vendas Aplicada a Perfis de Alumínio. Master's thesis, Universidade de Aveiro, 2015.

Sérgio António Ramos Costa. Sistema de Business Intelligence como suporte à gestão estratégica. Master's thesis, Universidade do Minho, 2012.

Filipa Couto e Luis Dias. Definição de KPI e implementação de soluções BI para alavancar o processo de tomada de decisão na Indústria Corticeira. Master's thesis, Faculdade de Engenharia da Universidade do Porto, 2023.

Nadine Côrte-Real. *Big Data & Analytics*. Influência, 2022. ISBN 9789896235505.

Thomas H. Davenport. How strategists use "big data" to support internal business decisions, discovery and production. *Strategy and Leadership*, 42:45–50, 7 2014. ISSN 10878572. doi: 10.1108/SL-05-2014-0034.

Diário da República - 1.ª série — N.º 173. Sistema de Normalização Contabilística - SNC. https://www.cnc.min-financas.pt/snc.html, 2009. Acedido em 25 de outubro de 2023.

Daniel R Dolk. Integrated model management in the data warehouse era. *European Journal of Operational Research*, 122(2):199–218, 2000. ISSN 0377-2217. doi: 10.1016/S0377-2217(99)00229-5.

Cristina Dutra e Aguiar Ciferri. Distribuição dos Dados em Ambientes de Data Warehousing: O Sistema WebD2W e Algoritmos Voltados à Fragmentação Horizontal dos Dados, 2002.

João Ferreira, Miguel Miranda, António Abelha, e José Machado. O Processo ETL em Sistemas Data Warehouse. *INForum 2010*, 01 2010.

Raquel Marques Ferreira. Previsão na Área Farmacológica - Modelos Estatísticos vs Deep Learning. Master's thesis, Universidade do Minho, 2017.

Luís Pedro Novais Freitas. Desenvolvimento de um sistema de Business Intelligence com um algoritmo de recomendações. Master's thesis, Universidade do Minho, 2021.

Amir Gandomi e Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35:137–144, 2015. ISSN 02684012. doi: 10.1016/j.ijinfomgt.2014.10.007.

Nuno Miguel Martins Garcia. Consolidação de contas: potenciação do ERP na preparação das demonstrações financeiras consolidadas. Master's thesis, Instituto Superior de Contabilidade e Administração de Coimbra, 2019.

Guido L. Geerts. A Design Science Research methodology and its application to accounting information systems research. *International Journal of Accounting Information Systems*, 12:142–151, 6 2011. ISSN 14670895. doi: 10.1016/j.accinf.2011.02.004.

Matteo Golfarelli, Dario Maio, e Stefano Rizzi. The Dimensional Fact Model: A Conceptual Model for Data Warehouses. *Int. J. Cooperative Inf. Syst.*, 7:215–247, 06 1998. doi: 10.1142/S0218843098000118.

J. Han e M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, Burlington, 3rd edition, 2011.

Jiawei Han, Jian Pei, e Hanghang Tong. *Data mining: Concepts and Techniques*. Morgan Kaufmann, 2022.

Alan Hevner, Salvatore T. March, Jinsoo Park, e Sudha Ram. U-CARE View project Modeling Customer Churn View project. *Management Information Systems Quarterly*, 28:75–, 03 2004.

W.H. Inmon. *Building The Data Warehouse (4th Ed.)*. Wiley India Pvt. Limited, 2005. ISBN 9788126506453.

Ralph Kimball e Margy Ross. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. John Wiley & Sons, Inc., USA, 2nd edition, 2002. ISBN 0471200247.

Milan Kubina, Michal Varmus, e Irena Kubinova. Use of Big Data for Competitive Advantage of Company. *Procedia Economics and Finance*, 26:561–565, 2015. ISSN 22125671. doi: 10.1016/s2212-5671(15)00955-7.

Yaokun Lin. Model Selection with AIC & BIC. https://machinelearningabc.medium.com/model-selection-with-aic-bic-10ac9dac4c5a, 2021. Acedido em 25 de outubro de 2023.

José Fernando Pereira Magalhães. Abordagem semântica para a integração de dados em Big Data Warehouses. Master's thesis, Universidade do Minho, 2019.

Francisco Peres Marques. Investimentos financeiros e consolidação: exemplos no grupo Visabeira. Master's thesis, Faculdade de Economia da Universidade de Coimbra, 2019.

Solomon Negash e Paul Gray. Business Intelligence. In *Handbook on Decision Support Systems 2*, volume 13, page 423, 01 2003.

Celina M. Olszak. Business Intelligence Systems for Innovative Development of Organizations. In *Business Intelligence Systems for Innovative Development of Organizations*, volume 207, pages 1754–1762. Elsevier B.V., 2022. doi: 10.1016/j.procs.2022.09.233.

Ahmed Oussous, Fatima Zahra Benjelloun, Ayoub Ait Lahcen, e Samir Belfkih. Big Data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences*, 30(4):431–448, 10 2018. ISSN 2213-1248. doi: 10.1016/j.jksuci.2017.06.001.

João Pedro Fernandes Pais. Métodos de Previsão. Master's thesis, Faculdade de Economia da Universidade do Porto, 2017.

José Luís Pereira e Marco Costa. Decision support in big data contexts: A business intelligence solution. In *New Advances in Information Systems and Technologies*, volume 444, pages 983–992. Springer Verlag, 2016. ISBN 9783319312316. doi: 10.1007/978-3-319-31232-3_93.

Fotios Petropoulos et al. Forecasting: theory and practice. *International Journal of Forecasting*, 38(3): 705–871, 2022. ISSN 0169-2070. doi: 10.1016/j.ijforecast.2021.11.001.

Xuedi Qin, Yuyu Luo, Nan Tang, e Guoliang Li. Making data visualization more efficient and effective: a survey. *VLDB Journal*, 29:93–117, 1 2020. ISSN 0949877X. doi: 10.1007/s00778-019-00588-3.

Jorge Manuel Quintaneiro, Bruno Ricardo, e Gomes Martins. Demonstração de Resultados (DR). Master's thesis, Instituto Superior de Engenharia de Coimbra, 2000.

Villarejo Ramos, Angel F, Cabrera Sánchez, e J.-P. Factors affecting the adoption of Big Data analytics in companies. *RAE - Revista De Administracao De Empresas*, 59(6):415–429, 2019. doi: 10.1590/S0034-759020190607.

Jayanthi Ranjan. Business intelligence: Concepts, components, techniques and benefits. *Journal of Theoretical and Applied Information Technology*, 9:60–70, 01 2009.

Ana Cristina dos Santos Carvalho Ribeiro. A importância da consolidação de contas nas autarquias. Master's thesis, Escola Superior de Tecnologia e Gestão de Viseu, 2010.

Célia Andreia Santos Ribeiro. Porque razão os pequenos grupos não consolidam contas? Estudo exploratório. Master's thesis, Faculdade de Economia, Universidade do Porto, 2017.

Diogo Xavier Teixeira Ribeiro e Santos Morais. A Importância da Contabilidade na Análise Financeira das Empresas: Estudo de um Caso. Master's thesis, Instituto Superior de Gestão, 2021.

María Teresa Rodríguez, Sérgio Nunes, e Tiago Devezas. Telling stories with data visualization. *NHT 2015 - Proceedings of the 2015 Workshop on Narrative and Hypertext - co-located with HT 2015*, pages 7–11, 9 2015. doi: 10.1145/2804565.2804567.

Maribel Yasmina Santos e Isabel Ramos. Business Intelligence: tecnologias da informação na gestão de conhecimento. In *Business Intelligence: tecnologias da informação na gestão de conhecimento*. FCA - Editora de Informática, Lda, 2006.

Carlos Sezões e José Oliveira. *Business intelligence*. Negócio electrónico. Sociedade Portuguesa de Inovação, Porto, 2006.

Taylor G. Smith et al. pmdarima: Arima estimators for Python, 2017. URL `http://www.alkaline-ml.com/pmdarima`. Acedido em 26 de outubro de 2023.

J Sreemathy, K Naveen Durai, E Lakshmi Priya, R Deebika, K Suganthi, e PT Aisshwarya. Data integration and etl: A theoretical perspective. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 1655–1660, 2021. doi: 10.1109/ICACCS51430.2021.9441997.

E. Stellwagen e Len Tashman. ARIMA: The Models of Box and Jenkins. *Foresight: Int. J. Appl. Forecast.*, pages 28–33, 01 2013.

Elena Geanina Ularu, Florina Camelia Puican, Anca Apostu, Manole Velicanu, et al. Perspectives on Big Data and Big Data Analytics. *Database Systems Journal*, 3(4):3–14, 2012.

Oana Velcu-Laitinen e Ogan M. Yigitbasioglu. The use of dashboards in performance management: Evidence from sales managers. *International Journal of Digital Accounting Research*, 12:39–58, 2012. ISSN 15778517. doi: 10.4192/1577-8517-v12_2.

Carlo. Vercellis. *Business intelligence: data mining and optimization for decision making*. Wiley, 2008. ISBN 9780470511381.

Jingting Wang e Bao Liu. Design of ETL Tool for Structured Data Based on Data Warehouse. *ACM International Conference Proceeding Series*, 10 2020. doi: 10.1145/3424978.3425101.

Vitor Wilher. Modelos SARIMA. https://analisemacro.com.br/economia/comentario-de-conjuntura/modelos-sarima/, 2022. Acedido em 25 de outubro de 2023.

Lamia Yessad e Aissa Labiod. Comparative study of data warehouses modeling approaches: Inmon, Kimball and Data Vault. In *2016 International Conference on System Reliability and Science (ICSRS)*, pages 95–99, 2016. doi: 10.1109/ICSRS.2016.7815845.