ESCOLA
SUPERIOR
DE TECNOLOGIA
E GESTÃO
POLITÉCNICO
DO PORTO

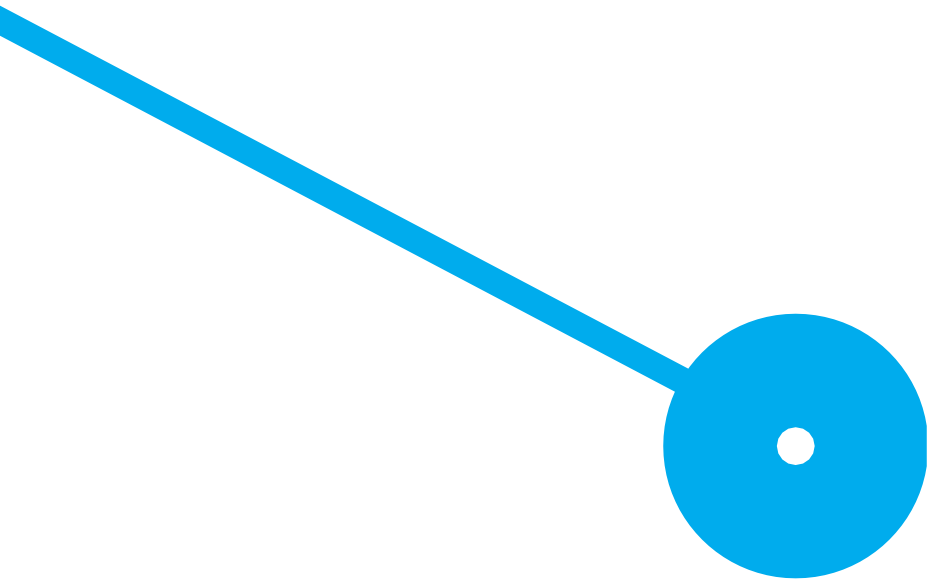
M

Mestrado
Engenharia Informática

Utilização de Catálogos de Dados no Desenvolvimento de *Dashboards*

Ana Catarina Moreira Henriques

10/2023



ESCOLA
SUPERIOR
DE TECNOLOGIA
E GESTÃO
POLITÉCNICO
DO PORTO

P.PORTO

M Mestrado
Engenharia Informática

Utilização de Catálogos de Dados no Desenvolvimento de *Dashboards*

Ana Catarina Moreira Henriques

Orientadores

Professor Doutor Bruno Oliveira

Professor Doutor Vasco Santos

Dissertação apresentada para cumprimento dos requisitos necessários à obtenção do grau de Mestre em 07/12/2023 pela Escola Superior de Tecnologia e Gestão do Instituto Politécnico do Porto.

10/2023

Declaração de Integridade

Eu, Ana Catarina Moreira Henriques, estudante nº 8170064, do Mestrado de Engenharia Informática da Escola Superior de Tecnologia e Gestão do Instituto Politécnico do Porto, declaro que não fiz plágio nem auto-plágio, pelo que o trabalho intitulado "Utilização de Catálogos de dados no desenvolvimento de *Dashboards*" é original e da minha autoria, não tendo sido usado previamente para qualquer outro fim. Mais declaro que todas as fontes usadas estão citadas, no texto e na bibliografia final, segundo as regras de referenciação adotadas na instituição.

Dedicatória e Agradecimentos

A realização desta tese foi possível através do contributo e apoio de várias pessoas às quais expresso o meu sincero agradecimento.

Em especial, queria agradecer ao meu orientador Professor Doutor Bruno Oliveira e ao meu coorientador Professor Doutor Vasco Santos pela orientação, aconselhamento e apoio ao longo deste percurso.

Não posso deixar de expressar a minha sincera gratidão aos meus pais e à minha irmã por todo o amor, apoio contínuo e incentivo ao longo desta jornada. Sem o encorajamento deles nada disto seria possível.

Por último, quero agradecer a todas as pessoas que se cruzaram comigo durante esta etapa e que de uma forma ou de outra incentivaram-me e ajudaram-me a cumprir os meus objetivos.

Obrigada a todos.

Resumo

O aumento constante da quantidade e diversidade de dados com que as organizações têm de lidar intensifica a importância e os desafios associados à gestão de dados. A transformação digital altera o comportamento dos utilizadores e impõe vários desafios às organizações que necessitam de recolher e utilizar dados mais diversificados para melhorar os serviços digitais, impulsionar a tomada de decisões e estimular a inovação. Com o aparecimento de novas abordagens para armazenar e explorar os dados, como *Data Lake* ou *Data Lakehouse*, os requisitos associados à construção e manutenção do catálogo de dados evoluíram e representam uma oportunidade para que as organizações possam desenvolver os seus processos de tomada de decisão. Assim, saber onde os dados estão armazenados e como podem ser utilizados melhora a capacidade de resposta às mudanças do negócio de forma a serem tomadas decisões mais sustentadas. Neste contexto, a utilização de catálogos de dados assume um papel cada vez mais relevante, já que são utilizados como uma ferramenta para encontrar, compreender e contextualizar os dados. Este trabalho explora a utilização de um catálogo de dados para apoiar a construção, validação e manutenção de *dashboards* no contexto de sistema de *Business Intelligence*.

Palavras-chave: Catálogo de Dados, *Business Intelligence*, *Dashboards*, Camada Semântica, Sistemas Analíticos, Data Lake

Abstract

The importance and challenges associated with data management become more significant as a result of the continuous increase in the quantity and variety of data that organizations must manage. Digital transformation changes user behavior and presents several challenges to organizations that need to collect and use more diverse data to improve digital services, boost decision-making, and stimulate innovation. With the emergence of new approaches to storing and exploring data, such as Data Lake or Data Lakehouse, the requirements associated with building and maintaining the data catalogue have evolved and represent an opportunity for organizations to develop their decision-making processes. Therefore, knowing where data is stored and how it can be used improves the ability to respond to business changes so that more sustained decisions can be made. In this context, the use of data catalogues assumes an increasingly relevant role, as they are used as a tool to find, understand, and contextualize data. This work explores the use of a data catalogue to support the construction, validation, and maintenance of dashboards in the context of a Business Intelligence System.

Palavras-chave: Data Catalog, *Business Intelligence*, *Dashboards*, Semantic Layer, Analytic Systems, Data Lake

Índice

Índice de Figuras

Siglas e Abreviaturas

BI	Business Intelligence
CD	Catálogo de Dados
DAX	Data Analysis eXpression Language
DBA	Administrador de Base de Dados
DL	Data Lake
DW	Data Warehouse
ETL	Extração, Transformação, Carregamento
KPI	Key Performance Indicators
ML	Machine Learning
ODRL	Open Digital Rights Language
PBI	Power BI
SK	Surrogate Key
JSON	JavaScript Object Notation

Esta página foi propositadamente deixada em branco.

CAPÍTULO 1

Contextualização e Motivação

1.1 Introdução

Os sistemas analíticos representam um ativo essencial para a análise das atividades empresariais. A capacidade que uma organização tem para analisar os seus dados e compreender o seu significado e contexto representa uma ferramenta poderosa que pode ser utilizada para apoiar os processos de tomada de decisão [1]. Durante muito tempo as organizações confiavam apenas na intuição para tomar as decisões. No entanto, à medida que a tecnologia avança, o mercado torna-se cada vez mais competitivo sendo necessário não só compreender a realidade organizacional, mas também mapeá-la com a realidade externa à organização de forma a tomar decisões mais consistentes e fidedignas [1].

A utilização de sistemas analíticos permite que as organizações compreendam o comportamento do mercado e, muitas vezes, antecipem mudanças que podem exigir alterações tanto táticas como estratégicas. Os sistemas analíticos capacitam as empresas a examinar dados históricos, tendências atuais e previsões futuras, oferecendo *insights* valiosos para orientar ações e estratégias de negócio [2]. Estes sistemas, também ajudam as organizações a tornarem-se mais ágeis e orientadas por dados, garantindo que as suas decisões sejam fundamentadas em informação sólida e não na intuição, levando a resultados mais consistentes e bem-sucedidos uma vez que a análise dos dados não só amplia o entendimento da organização sobre o seu desempenho interno mas também oferece uma visão precisa do cenário externo, o que é vital para a formulação de estratégias eficazes e para que as organizações se mantenham competitivas no ambiente de negócios atual.

Uma das formas mais acessíveis e simples para que as organizações analisem os seus dados é através da criação de relatórios e *dashboards*. Estes consistem numa combinação de diversos recursos visuais cuidadosamente apresentados num contexto específico, desempenhando a função vital de apresentar informação importante e relevante. Para construir estas representações visuais é crucial contar com os dados previamente preparados sendo também necessário criar medidas baseadas nos processos de negócio, perspetivas e eventos que precisam de ser analisados. As medidas tipicamente resultam de um cálculo utilizando fórmulas que podem estar dependentes do contexto de visualização, variando de acordo com a interação que o utilizador tem com os relatórios [3].

Isso permite uma exploração ágil e dinâmica dos dados uma vez que os utilizadores podem interagir de forma direta e imediata com os visuais, adaptando as análises de acordo com as suas necessidades e objetivos. Os relatórios e *dashboards* tornam-se assim ferramentas que possibilitam a compreensão eficaz dos dados e auxiliam na tomada de decisões informadas e estratégicas nas organizações [4].

Numa organização podem ser desenvolvidos vários relatórios/*dashboards* tipicamente contextualizados com as necessidades de diferentes departamentos e cenários. Sendo assim, neste contexto, utilizar e manter métricas de forma consistente representa um desafio considerável uma vez que a mesma medida pode ser utilizada em diferentes relatórios, em diferentes visualizações ou por outras medidas para a obtenção de cálculos derivados. A padronização e a conformidade nos dados e medidas são pilares fundamentais para garantir uma interpretação precisa e consistente dentro de um contexto de análise.

A utilização de factos conformes [5] é uma prática comum na modelação dimensional pois implica definir uma estrutura de dados uniforme e consistente, de modo que as medidas comparadas ou calculadas em conjunto tenham a mesma definição técnica. Um fator crucial neste processo é a definição da granularidade. A granularidade refere-se ao nível de detalhe dos dados recolhidos e armazenados. É uma descrição do quão específicos ou abrangentes são os dados. Encontrar o equilíbrio certo entre o nível de detalhe e agregação é essencial para atender às necessidades analíticas sem sobrecarregar o sistema com a quantidade excessiva de dados.

Um Catálogo de Dados (CD) desempenha um papel crucial na documentação e organização dos dados utilizados num sistema analítico [3]. Este, serve como repositório centralizado onde os metadados são armazenados contendo definições detalhadas dos objetos utilizados, dos seus relacionamentos, caminhos de exploração e avaliação da qualidade. Estes metadados são vitais para compreender e interpretar o contexto dos dados de forma eficaz. Para além disso, o CD atua como uma ponte entre os dados e o seu contexto quando enquadrados com

o negócio, ajudando a correlacionar os dados com o glossário empresarial de forma a permitir uma visão mais clara do significado e do contexto dos dados através da visualização, análise e monitorização de diferentes fontes de dados. Com isto, os dados são consultados de forma contextualizada melhorando a interpretação e utilização dos dados ao longo da organização [6].

Um Catálogo de Dados de medidas pode ser visto como um catálogo de dados especializado que concentra uma coleção de métricas de negócio específicas. Estas medidas são geridas pela equipa de *Data Engineers* e são projetadas para poderem ser utilizadas em várias fontes de dados, relatórios e *dashboards*. Ter um catálogo CD especializado para medidas permite uma fácil referência e utilização consistente dessas medidas em todo o ecossistema analítico da organização.

Neste projeto é explorada a utilização e aplicação de CD para suportar a construção, validação e manutenção de *dashboards* num sistema de *Business Intelligence*. Especificamente, será explorado a implementação de um Catálogo de Dados de medidas com o objetivo de apoiar a gestão e utilização eficiente de medidas no negócio. Na prática, o CD de medidas é apresentado para descrever semanticamente a composição, relacionamento, linhagem e representação visual das medidas de acordo com contextos específicos. Adicionalmente pretende-se evidenciar como o uso de metadados relacionados com as medidas pode melhorar o desenvolvimento de *dashboards* e relatórios proporcionando uma compreensão mais profunda e uma visualização mais eficaz das informações.

1.2 Apresentação e oportunidade do Tema

A utilização de um Catálogo de Dados (CD) representa uma estratégia fundamental na gestão eficaz dos dados dentro de uma organização. Os benefícios derivados da implementação de um CD são vastos e impactam positivamente vários aspetos da gestão organizacional [7]. A presença de um CD permite a valorização dos metadados associados aos conjuntos de dados. Os metadados, que compreendem informações sobre a origem, estrutura, significado e relacionamentos dos dados, ganham relevância e visibilidade, facilitando a compreensão e interpretação dos dados por parte dos utilizadores. Além disso, o catálogo de dados amplia as possibilidades de melhoria dos processos organizacionais não só a nível operacional, mas também a nível analítico [8].

A utilização de uma camada de modelação de dados, que incorpora os CD como parte integral, desempenha um papel essencial na representação e manutenção do conhecimento ao nível dos dados. Esta abordagem é independente de utilizadores específicos e possibilita aos gestores a utilização de todas as componentes do CD para compreender os dados e formular

novas consultas sobre os dados sem a necessidade de recorrer ao analista de dados ou ao administrador de bases de dados (DBA) [2]. Esta característica "*self-service*" é um elemento intrínseco aos sistemas de *Business Intelligence* mais tradicionais, e está cada vez mais presente no desenvolvimento de sistemas analíticos. Esta característica tem o propósito de capacitar os utilizadores casuais a explorar dados de forma intuitiva e promover o desenvolvimento de novas abordagens de análise, sem exigir a intervenção de utilizadores mais técnicos, como o administrador de bases de dados [3]. Neste contexto, o CD atua como uma infraestrutura de dados para orientar os gestores na descoberta e aproveitamento dos dados existentes.

O desenvolvimento de *dashboards* é uma etapa crucial na implementação de um sistema de *Business Intelligence*, pois é nesse ponto que os dados recolhidos e armazenados são transformados em informações significativas para apoiar a tomada de decisões. No entanto, esta tarefa é de elevada complexidade e especificidade, pois requer um mapeamento preciso entre os dados disponíveis e os requisitos específicos de tomada de decisão.

Ao utilizar um CD para orientar o desenvolvimento de *dashboards*, os profissionais de BI podem ter uma compreensão clara das fontes de dados, da sua estrutura, significado e relevância para os requisitos de tomada de decisão. Este aspeto contribui para uma seleção mais precisa dos dados necessários para alimentar os *dashboards* uma vez que os profissionais de BI podem utilizar o CD para identificar e selecionar os conjuntos de dados mais pertinentes e relevantes para atender às necessidades de análise resultando em *dashboards* mais precisos e eficazes que fornecem informações relevantes e acionáveis. Além disso, os *dashboards* muitas vezes envolvem diferentes perspetivas de análise. O CD pode ser usado para validar os dados nessas diferentes perspetivas (por exemplo, regras de agregação, métricas ou cálculos específicos). O que significa que os dados utilizados nos *dashboards* estão alinhados com as regras e critérios estabelecidos para garantir a precisão e a consistência das informações apresentadas.

As abordagens para a construção de sistemas analíticos têm evoluído neste sentido de priorizar a integração dos metadados com os ativos durante todo o processo de análise. Este avanço é notório na adoção te tecnologias como *Data Lakes* (DL) e *data lakehouses* que representam uma abordagem mais sem esquema para a gestão e análise de dados. Os *data lakes* e os *data lakehouses* são arquiteturas de armazenamento e processamento de dados que permitem a recolha e o armazenamento de uma ampla variedade de dados, estruturados e não estruturados, na sua forma bruta e original [9]. Ao integrar metadados de forma intrínseca a esses ativos de dados, as plataformas proporcionam uma visão abrangente e contextualizada, facilitando a compreensão e a interpretação dos dados ao longo de seu ciclo

de vida. Esta evolução reflete a compreensão crescente da importância da contextualização e da gestão eficaz dos metadados no ecossistema de dados. A integração dos metadados ao processo de análise impulsiona a eficiência operacional, a qualidade dos dados e, por fim, contribui para a tomada de decisões informadas e estratégicas nas organizações.

1.3 Motivação e Objetivos

Os relatórios/*dashboards* desempenham um papel crucial na apresentação e interpretação dos dados dentro de uma organização. Representam componentes de visualização de dados que são utilizados para apresentar medidas e *Key Performance Indicators* (KPIs) relevantes para a organização. Estes elementos visuais são projetados de acordo com as necessidades específicas da organização, visando facilitar a análise e a tomada de decisões informadas. A validação das medidas é um processo essencial no desenvolvimento e uso de relatórios. Garantir que os cálculos estão corretos e que os dados apresentados são precisos é fundamental para a confiança que os *decision makers* depositam nas informações apresentadas. A precisão dos dados é a base sobre a qual as decisões são tomadas, e qualquer imprecisão pode levar a escolhas equivocadas e, consequentemente, a impactos negativos nos resultados e na estratégia da organização.

Além disso, a monitorização contínua das medidas é necessária porque as necessidades e os objetivos das organizações estão sempre em evolução. Isto é à medida que a organização cresce, o foco estratégico pode mudar ou a organização pode ter de se adaptar a novos desafios e oportunidades, e por isso os requisitos para análise de dados também podem ser alterados. Portanto, é crucial que as medidas sejam validadas e monitorizadas ao longo do tempo para garantir que continuam a ser relevantes e calculadas da forma correta, alinhando-se sempre com as expectativas e necessidades atuais da organização.

A modelação dimensional desempenha uma etapa essencial no desenvolvimento de soluções de *Business Intelligence* (BI) e análise de dados. Durante esta etapa, ocorre a definição da estrutura da base de dados, incluindo dimensões, atributos, medidas e outros elementos essenciais. Vários aspetos relacionados com o processo de modelação e de identificação de requisitos não são incluídos de forma explícita na implementação do Data Warehouse (DW), já que, por exemplo, não é possível representar/diferenciar os vários tipos de medidas utilizando uma tabela relacional. No entanto, este tipo de informação é devidamente documentado e utilizado em diferentes fases do projeto. No caso das medidas, a documentação produzida durante o processo de modelação dimensional é utilizada na camada de BI para a construção de relatórios. Ao incluir este tipo de documentação num catálogo de dados, de forma que possa ser utilizada de forma ativa pelas ferramentas de BI [10], poderia auxiliar na compreensão e validação dos dados durante a análise.

Com este trabalho pretende-se explorar a implementação de um CD enriquecido com uma camada semântica que integra alguns dos princípios do processo de modelação dimensional especialmente orientado para a representação de medidas. Os metadados representados suportam a construção e manutenção de relatórios/*dashboards*, integrando o conceito de metadados ativos, ou seja, são utilizados como parte integrante do processo de construção e validação da camada de BI.

Pretende-se também contribuir para a literatura atual nas seguintes áreas:

a) Análise das principais técnicas para a construção de catálogos de dados;

b) Revisão das tecnologias para construção de catálogos de dados e como as principais ferramentas de BI tiram partido dessas tecnologias;

c) Implementação de um catálogo de dados com princípios semânticos especialmente orientado para suportar o desenvolvimento de *dashboards*/relatórios.

d) Implementar uma prova de conceito que possibilite a representação de metadados de acordo com os fundamentos da modelação dimensional. Tendo por base um caso de estudo, pretende-se avaliar o potencial da abordagem proposta para validar e monitorizar a construção de relatórios/*dashboards*.

1.4 Estrutura do documento

Este documento está organizado em oito capítulos cada um focado em aspetos específicos do tema em questão. O primeiro capítulo, Contextualização e Motivação, inicia-se com uma introdução do tema seguido da contextualização do mesmo e apresentação das oportunidades do tema. Neste capítulo também são delineadas as motivações e os objetivos. No segundo capítulo, Fundamentação Teórica, são apresentados os conceitos fundamentais relacionados com o tema. O capítulo seguinte, Revisão do Estado da Arte, consiste numa análise detalhadas do panorama atual de estudos, projetos e avanços relevantes relacionados com o desenvolvimento de Catálogos de Dados. No quarto capítulo, Definição de Medidas para Análise de Dados é apresentada a problemática da validação e monitorização de medidas. O capítulo 5, processo de modelação dimensional, apresenta uma visão detalhada do cenário em que o Catálogo de Dados foi implementado, incluindo os desafios enfrentados no processo. No capítulo Apresentação e Visualização dos dados são abordadas as etapas para o desenvolvimento de um relatório/*dashboard*. O capítulo 7 apresenta um componente para um Catálogo de Dados orientado para a caracterização de métricas e dos fundamentos teóricos da modelação dimensional. Por fim são apresentados os resultados obtidos, uma conclusão e o trabalho futuro.

CAPÍTULO 2

Fundamentação Teórica

Com a evolução dos sistemas de informação, o processo de tomada de decisão passou a fundamentar-se em dados de negócio que refletem a realidade organizacional. Nos dias de hoje, a quantidade de dados, tanto internos quanto externos às organizações, que impactam a tomada de decisões está em constante crescimento. Esse cenário demanda mecanismos cada vez mais sofisticados para extrair conhecimento valioso a partir desses dados, mantendo a sua consistência, precisão e atualização.

2.1 Dados, Informação e Conhecimento

Os dados são a base para a tomada de decisões dentro das organizações. São os elementos brutos, não processados e sem contexto. Representam uma fonte valiosa de informações que podem ser extraídas de uma variedade de fontes diretamente relacionadas ou não com os principais processos de negócio. No entanto, a mera recolha e armazenamento dos dados não agregam qualquer valor substancial. O verdadeiro valor dos dados emerge quando estes são submetidos a uma análise criteriosa e transformados em informações significativas [11].

A transformação dos dados em informações e conhecimento é um processo crucial. A contextualização dos dados é o primeiro passo para a sua transformação em informações. A partir desse contexto, os dados, começam a adquirir significado e relevância para o negócio. Quando os dados são organizados e estruturados de maneira adequada são transformados em conhecimento valioso. Esta evolução, representada na Figura 1, ilustra a trajetória desde dados até conhecimento.

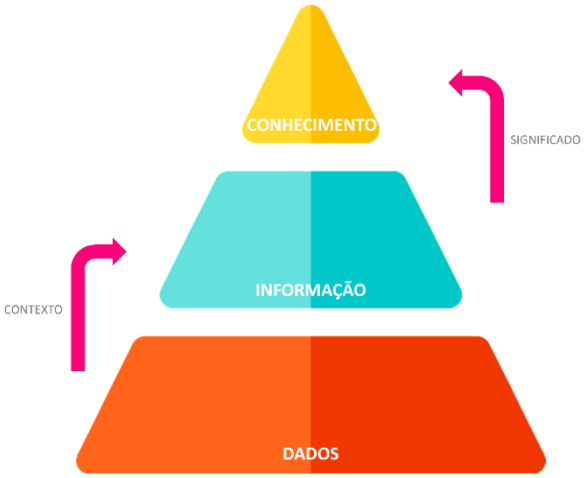


Figura 1- Pirâmide DIK (Data, Information, Knowledge) [1]

A interligação da informação e do conhecimento desempenha um papel fundamental na implementação e utilização eficaz dos sistemas de informação dentro das organizações. Estes sistemas são concebidos para aprimorar o processo de tomada de decisão. Em particular, os sistemas de apoio à decisão tiram partido das bases de dados organizacionais, organizando e agregando os dados de forma a produzir relatórios que são utilizados pelos gestores no processo de tomada de decisão.

Desta forma, os gestores têm a oportunidade de desenvolver conhecimento especializado em áreas específicas do negócio. A informação e o conhecimento assim adquiridos alimentam a capacidade dos gestores para tomar decisões informadas e estratégicas. Combinando dados, informação e conhecimento, as organizações podem aprimorar a sua eficiência operacional, otimizar estratégias e alcançar vantagem competitiva no mercado. Este ciclo de transformação contínuo desde dados brutos até conhecimento do negócio é essencial para a prosperidade e progresso das organizações.

2.2 Business Intelligence

Os sistemas de informação desempenham um papel crucial nas organizações, ajudando a gerir e aperfeiçoar as operações diárias e a orientar as estratégias futuras. Estes podem ser divididos em duas categorias principais: os que suportam a operacionalidade dos processos de negócio e os que facilitam a análise desses processos [12].

Os sistemas que suportam a operacionalidade dos processos de negócio estão voltados para a execução eficiente das tarefas diárias de uma organização. Estes sistemas automatizam e otimizam atividades rotineiras, como processamento de transações, gestão de inventário,

[1] Fonte da imagem: https://www.ontotext.com/knowledgehub/fundamentals/dikw-pyramid/

processamento de pedidos e gestão de recursos humanos. O objetivo é melhorar a eficiência e a eficácia das operações organizacionais. Por outro lado, os sistemas de *Business Intelligence* (BI) concentram-se na análise de dados para fornecer informações valiosas que suportem a tomada de decisões estratégicas. Os sistemas de BI alimentam-se dos dados disponíveis nas organizações e disponibilizam informação relevante para a tomada de decisão [13]. Estes sistemas são projetados para extrair, armazenar, analisar, partilhar e fornecer acesso aos dados de forma a apoiar os gestores na tomada de decisões informadas [14]. Os sistemas de BI recolhem e processam dados tanto internos quanto externos à organização e transformam-nos em informações úteis e acionáveis.

Os sistemas de BI utilizam dados históricos e atuais para criar perspetivas, métricas e análises que auxiliam na compreensão do desempenho passado e presente da organização. Permitem identificar tendências, padrões e perceções que podem orientar estratégias futuras, impulsionando a competitividade e inovação. Portanto, enquanto os sistemas que suportam a operacionalidade dos processos de negócio visam otimizar as atividades cotidianas, os sistemas de *Business Intelligence* são focados em fornecer informações estratégicas que orientem a tomada de decisões e impulsionem o crescimento e sucesso da organização. A combinação eficaz destes dois tipos de sistemas é essencial para o funcionamento eficiente e bem-sucedido de uma organização.

A implementação de um sistema de BI é um processo que compreende várias etapas e envolve uma complexidade significativa reunindo diferentes perfis e diferentes tipos de conhecimento [2]. Neste processo, os dados e a forma como estes se encontram estruturados geralmente são da responsabilidade do administrador de bases de dados (DBA). Por outro lado, o analista de dados desempenha um papel crucial, atuando como uma ponte entre os dados e os gestores. Os gestores representam um perfil de utilizadores que possuem um conhecimento aprofundado da lógica de negócio e das principais métricas a serem analisadas. No entanto, a estruturação dos dados e a forma como estes estão armazenados são geralmente complexas e, muitas vezes, técnicas, ficando sob a alçada do administrador de bases de dados. O analista de dados funciona como um tradutor, uma vez que consegue mapear a lógica de negócio com a lógica de armazenamento de dados. No entanto, este processo de tradução entre a lógica de negócio e a lógica de armazenamento de dados é naturalmente ineficiente e incomportável em organizações de grande escala. É por isso que este conhecimento deve ser mapeado de forma que qualquer utilizador o possa compreender.

Os modelos de dados são uma forma de representar este conhecimento. Por exemplo, os modelos dimensionais, quando vistos na perspetiva conceptual, representam os principais eventos de negócio, por meio de tabelas de factos, e a sua contextualização, através das

tabelas de dimensão. No entanto, é importante realçar que o modelo em si não é suficiente para representar todas as especificidades de um sistema pois este não se limita apenas ao modelo de dados analítico, mas também todas as características associadas ao armazenamento e extração dos dados nos sistemas fonte, assim como de todo o processo associado à sua integração no repositório analítico (conhecido como processo ETL) [15].

2.3 Modelação dimensional

A modelação dimensional [5] desempenha um papel fundamental nos sistemas de BI, fornecendo a estrutura necessária para organizar e representar os dados de forma eficiente, possibilitando a análise e criação de relatórios relevantes. A modelação dimensional, recorre a tabelas de facto e tabelas de dimensão para criar uma estrutura (tipicamente em estrela) que é altamente otimizada para consultas analíticas.

Diante de diferentes requisitos e pressupostos, é fundamental adotar soluções ajustadas e personalizadas levando em consideração a natureza do sistema em desenvolvimento. Os princípios da modelação dimensional estão alinhados com as características únicas dos sistemas de BI [5], [16]. A modelação dimensional leva em consideração que cada sistema de BI possui requisitos específicos, variando de acordo com a natureza do negócio, os objetivos de análise e as preferências da organização. Portanto, a estrutura dos dados é personalizada para atender a esses requisitos de forma eficaz e eficiente. Esta abordagem é direcionada para consultas que acedem a um grande volume de transações, suportando a preservação de dados históricos mesmo quando os sistemas operacionais realizam modificações ou exclusões de informação. A capacidade de manter os dados históricos é fundamental para análises comparativas ao longo do tempo, identificação de padrões e tendências, entre outros aspetos.

A modelação dimensional foi concebida com o propósito de apoiar a análise dos processos de negócio. Concentra-se em entender como é que esses processos são avaliados, como é que as métricas são aplicadas e como é que os dados são relacionados de forma a fornecer informações valiosas sobre o desempenho e a eficácia desses processos. O princípio central é claro: viabilizar a análise dos processos de negócio por meio de um processo de modelação que se foca na forma como estes processos são avaliados [12].

A modelação de dados é caracterizada por dois elementos-chave: Avaliação e Contexto, também denominados por Factos e Dimensões. A relação entre as tabelas de facto e as dimensões forma um esquema podendo ser um esquema em estrela (*Star Schema*) ou um esquema floco de neve (*Snow Flake*).

2.3.1 Dimensões

As dimensões desempenham um papel essencial na análise de dados, fornecendo contexto e uma perspetiva qualitativa da informação. Uma das principais funções das dimensões é a contextualização dos factos ou medidas dentro de um conjunto de dados de acordo com o grão definido. Por exemplo, numa base de dados de vendas, as dimensões poderiam ser cliente, produto, região geográfica, entre outros. As dimensões também fornecem através de hierarquias, caminhos de exploração de dados que podem ser explorados pelas ferramentas de BI permitindo que os utilizadores naveguem de forma intuitiva e eficiente pelos dados e façam análises mais detalhadas ou agregadas conforme necessário. Por exemplo, a dimensão "Calendário" pode ter hierarquias que vão desde o ano até ao mês, dia ou até mesmo hora.

As dimensões conferem contexto e significado aos factos de diferentes formas [12]:

- Atuam como filtros: as dimensões permitem a filtragem eficiente de consultas e relatórios, assim é possível direcionar as análises para conjuntos específicos de dados. Servem como critério para restringir os resultados destacando informações relevantes para a análise;
- Ordenam ou classificam informações: as dimensões também oferecem a capacidade de ordenar ou classificar as informações. Este fator é importante para apresentar os factos de forma organizada e compreensível facilitando a interpretação e a compreensão das informações;
- Contextualizam os factos: Ao acompanhar os factos, as dimensões, fornecem o contexto e significado aos relatórios. Isso significa que é possível não apenas entender o que aconteceu (os factos) mas também onde, quando, quem, e sob quais circunstâncias é que esses factos ocorreram, o que é fundamental para análises mais completas;
- Definem a organização e apresentação dos detalhes: Permitem que os dados sejam organizados de forma lógica e hierárquica possibilitando a criação de hierarquias, agrupamentos, subtotais e resumos que estruturam e tornam mais informativa a apresentação dos dados principais. Uma hierarquia é um atributo que pode ser um subconjunto de outro atributo possibilitando ao utilizador a visualização de dados mais detalhados (*drill-down*) ou menos detalhados (*roll-up*).

As dimensões são peças fundamentais na modelação dimensional. Para além de organizarem os dados também conferem relevância, contexto e clareza à análise sendo essenciais para a criação de relatórios coerentes e informativos. Estas tabelas, oferecem a flexibilidade e a estrutura necessária para explorar dados de forma significativa, possibilitando uma compreensão mais aprofundada do desempenho e das tendências do negócio. Quanto mais

descritivas forem as dimensões, maior a variedade de abordagens possíveis para a análise dos dados. É crucial salientar que as dimensões são essenciais para fornecer informações contextuais de contexto nos relatórios. Sem a presença e contribuição das dimensões, os relatórios perderiam a sua relevância e significado, pois é por meio do contexto proporcionado por estas que os dados se tornam interpretáveis e utilizáveis para a tomada de decisões por parte das organizações.

A dimensão temporal é uma dimensão típica à construção de um esquema dimensional. A inclusão da dimensão temporal permite a organização e categorização dos dados de acordo com períodos específicos, como dias, meses, anos e até mesmo horas, dependendo da granularidade desejada. Isso possibilita a análise de tendências, padrões e mudanças ao longo desses períodos, fornecendo uma visão temporal dos eventos e transições que ocorreram no negócio. Para além disso, a dimensão temporal possibilita a comparação de diferentes períodos, facilitando a identificação de mudanças ao longo do tempo e a avaliação do desempenho em relação a metas e objetivos estabelecidos. Sem esta dimensão seria difícil ou mesmo impossível analisar tendências, padrões e mudanças ao longo de diferentes períodos

A construção das dimensões engloba a atribuição *Surrogate Keys* (SK) que garantem a independência entre o mundo operacional e analítico. As *Surrogate Keys* são valores únicos gerados especificamente para servirem como identificadores exclusivos no DW. Esses identificadores não têm correlação direta com os identificadores usados no ambiente operacional o proporcionam uma camada de independência entre os sistemas operacionais e os sistemas analíticos.

2.3.2 Factos

A tabela de factos é o núcleo do modelo dimensional e representa os eventos gerados no âmbito de um processo de negócio. A tabela de factos captura as medidas que descrevem o processo sendo que estas medidas são elementos quantitativos ou métricas que fornecem informações essenciais sobre o desempenho ou características do processo. As medidas escolhidas devem ser cuidadosamente selecionadas para garantir que ofereçam um conjunto abrangente de informações relevantes para a análise do processo em questão sendo que devem refletir aspetos cruciais do processo de negócio em estudo, permitindo uma análise detalhada e precisa. As medidas, tipicamente representadas na tabela de factos, representam uma forma de quantificar um determinado facto relacionado com o processo de negócio. Podem incluir valores monetários, quantidades, taxas, contagens, percentagens, entre outras medidas que possam descrever e quantificar aspetos do processo de forma significativa.

As medidas podem ser aditivas, semi-aditivas e não aditivas. Medidas aditivas (*full additive*)

são aquelas cujas factos podem ser somados através de qualquer dimensão, produzindo um resultado com significado. Estas medidas são a base para a maioria das análises quantitativas e cálculos num DW. As medidas semi-aditivas são aquelas que podem ser agregadas apenas em algumas das dimensões, o que significa que a soma dos factos pode não produzir resultados com significado. Por exemplo, considerando o stock de um produto, a soma deste é uma medida semi-aditiva devido à natureza de *snapshots* que os stocks diários representam. Assim, a soma do stock do dia de ontem com o stock do dia de hoje não representa com o stock atual, pois cada valor de stock diário é uma *snapshot* num momento específico e não pode ser somado. Por fim uma medida não aditiva é um tipo de medida que não pode ser agregada. Isto significa que a agregação da medida em diferentes dimensões não possui um significado lógico ou não representa uma informação válida para análises ou interpretações precisas. Exemplos de medidas não aditivas seria uma média, uma taxa ou um rácio.

Um importante aspeto da tabela de factos é a granularidade. Cada linha presente na tabela de factos armazena um nível específico de detalhe no processo de negócio que é conhecido como granularidade. A granularidade define o quão detalhada é a informação em cada linha da tabela de factos. A granularidade pode ser classificada em três tipos: transacional, *snapshots* periódicos ou *snapshots* cumulativos [5]. A granularidade transacional envolve o nível mais detalhado da captura dos dados, onde as informações são registadas num nível de transações individuais. Cada ação, operação ou transação é registada e armazenada de forma individual, garantindo uma visão precisa e detalhada das atividades do sistema. Este tipo de tabela de factos representam um evento ocorrido num ponto instantâneo no tempo. As tabelas de factos com *snapshots* periódicos envolvem a captura de dados em intervalos específicos de tempo predefinidos. Em vez de registar cada transação individual, os dados são capturados num ponto no tempo, criando "instantâneos" das informações relevantes. Esses instantâneos fornecem uma visão geral do estado do sistema num momento particular, sendo úteis para análises e relatórios periódicos. Os *snapshots* são necessários para analisar o desempenho acumulado do negócio em intervalos de tempo regulares e previsíveis. Ao contrário das tabelas de factos transacionais, em que uma linha é carregada para cada ocorrência do evento, com o *snapshot*, é tirada uma “foto”, no final de um dia, semana ou mês, depois outra foto no final do próximo período e assim por diante. As tabelas de factos com *snapshots* cumulativos são utilizadas para mostrar a atividade de um processo que tem início e fim bem definidos. Por exemplo, no processamento de um pedido, um pedido passa por etapas específicas até que seja totalmente processado. Conforme as etapas para cumprir o pedido são concluídas, a linha associada na tabela de factos é atualizada.

A definição da granularidade no processo de modelação dimensional é um fator de extrema

importância devido ao seu impacto direto na precisão e utilidade do modelo resultante. Escolher a granularidade correta é essencial para garantir que o modelo represente adequadamente o processo em análise, capturando informações relevantes e evitando excesso ou falta de detalhe. Uma granularidade bem definida permite a agregação ou desagregação dos dados conforme necessário para análises específicas o que possibilita a visualização e a compreensão dos dados em diferentes níveis de detalhe. Para além disso, é necessário que a granularidade esteja alinhada com os objetivos e requisitos do negócio. Compreender profundamente o contexto do negócio e os processos subjacentes é fundamental para determinar a granularidade mais apropriada. A escolha certa contribui para a eficácia da análise de negócio e para o processo de tomada de decisões.

2.3.3 Técnicas de modelação

Após o processo de modelação dimensional é frequentemente associada uma topologia de esquema que basicamente caracteriza a forma como as tabelas de factos e as dimensões de um esquema dimensional estão organizadas.

O esquema dimensional em estrela é a abordagem mais comum. Este esquema é composto por uma tabela central chamada de "tabela de factos" e várias tabelas relacionadas denominadas por dimensões. A analogia da estrela surge da disposição visual dos elementos, onde a tabela de fatos está no centro e as dimensões rodeiam esta, criando uma aparência semelhante a uma estrela.

Uma característica essencial do modelo em estrela é a simplicidade e a facilidade de compreensão. A estrutura simples e clara facilita a consulta e a análise dos dados, tornando-o uma escolha popular para ambientes de BI e tomada de decisões. Este modelo tem algumas vantagens como a sua simplicidade associada à sua leitura e ainda ao reduzido número de junções necessárias para relacionar os factos com as respetivas dimensões.

Já o modelo floco de neve surge como uma variação do modelo em estrela. Este modelo consiste em dimensões que se relacionam diretamente com a tabela de factos (dimensões de primeira linha de análise), mas existem dimensões que se relacionam entre si (dimensões de segunda linha de análise), que surgem através da aplicação do processo de normalização ou através da utilização de *outriggers*.

2.4 Plataformas de dados

O cenário atual de dados é marcado por uma abundância de informações geradas a uma velocidade exponencial, provenientes de várias fontes e em diferentes formatos. Nesse contexto, a gestão eficaz desses dados tornou-se uma prioridade estratégica para as

organizações de todos os setores. Três abordagens têm se destacado como pilares fundamentais na estratégia de gestão de dados seguida por investigadores e profissionais da área: *Data Warehouse* [17], *Data Lake* [18] e o *Data Lakehouse* [19].

2.4.1 Data Warehouse

O *Data Warehouse* (DW) é um componente central no domínio de BI e representa uma arquitetura especializada para armazenar, gerir e analisar grandes volumes de dados. Um DW é tipicamente um repositório centralizado e integrado de dados recolhidos de várias fontes e armazenados num formato estruturado e otimizado para consultas. É tipicamente projetado tendo por base os princípios de modelação dimensional e fornece uma visão única e consolidada dos dados. A sua finalidade principal é oferecer uma única fonte de verdade para os dados de uma organização permitindo que os utilizadores executem consultas complexas de forma eficaz.

O processo de construção de um DW engloba a recolha de dados provenientes de diversas fontes operacionais, seguido pela integração, padronização e armazenamento centralizado desses dados de acordo com os requisitos específicos de tomada de decisão. Esta metodologia implica que a estrutura do DW seja pré-implementada antes de iniciar o carregamento dos dados no sistema (também conhecido como *schema-on-write*).

2.4.2 Data Lake

Na última década, testemunhamos um considerável aumento na quantidade e heterogeneidade dos dados que as organizações precisam de gerir impulsionando assim a necessidade de abordagens e tecnologias alternativas para lidar com novos cenários [9]. Com o avanço da tecnologia, as organizações começaram a lidar com uma quantidade massiva de dados, incluindo dados estruturados, semi-estruturados e não estruturados. Para além disso, esses dados eram provenientes de diversas fontes. Neste contexto, os *Data Lakes* [18] surgiram como uma solução, proporcionando a recolha, armazenamento e escalabilidade, principalmente em cenários cuja quantidade dados colocava em causa as abordagens existentes (predominantemente de origem relacional).

Os *Data Lakes* seguem uma abordagem *schema-on-read*, isto é, ao contrário ds DWs não existe uma estrutura pré-definida, proporcionando uma flexibilidade considerável para lidar com a evolução e diversidade dos dados. Esta característica é especialmente valiosa num ambiente em que a natureza e o formato dos dados pode variar significativamente. Esta flexibilidade não só facilita a gestão dos dados, mas também mostra-se complementar ao DW. O DW continua a desempenhar um papel fundamental nos sistemas de *Business Intelligence* uma vez que pode ser alimentado através de um subconjunto de dados existentes no DL. Esta

integração entre DL e DW é uma prática comum na implementação de sistemas de BI permitindo que as organizações extraiam o máximo valor dos seus dados [19] .

A ausência de um esquema pré-definido nos *Data Lakes* aumenta a flexibilidade e escalabilidade na gestão dos dados, resultando numa redução dos custos associados ao armazenamento desses dados [9]. No entanto, a flexibilidade adicional resulta em novos desafios para a gestão da qualidade dos dados. Enquanto nos DWs a qualidade dos dados era assegurada, em parte, pela existência de uma estrutura rígida (*schema*), a flexibilidade que os DLs proporcionam para lidar com mais dados e com estruturas cada vez mais heterogéneas, aumenta significativamente os desafios enfrentados no armazenamento e exploração dos dados [18].

2.4.3 Data Lakehouse

Os *Data Warehouses* marcaram a primeira geração de plataformas de dados, oferecendo estruturas organizadas para armazenamento e análise de dados. De seguida, os DLs evoluíram como a segunda geração, permitindo o armazenamento de uma ampla variedade de dados (estruturados e não estruturados) de forma mais flexível. A terceira geração surge com a introdução do conceito de *Data Lakehouse* [19].

Esta arquitetura combina elementos das gerações anteriores incorporando uma camada de metadados que suporta, por exemplo, a aplicação de transações (ACID) ou estruturas auxiliares nos arquivos que armazenam os dados. Ao aceder aos dados, o primeiro passo consiste em consultar a camada de metadados. Esta consulta é essencial para identificar quais objetos/arquivos que contêm as informações pretendidas, permitindo um acesso mais direcionado e eficaz aos dados necessários. Desta forma, a camada de metadados torna-se um componente crucial para a eficiência e para a integridade das operações realizadas [9].

2.5 Metadados

A perceção de metadados vai além da simples documentação textual. Abrange também declarações semânticas e sintáticas que possuem um potencial significativo. Os metadados não se limitam a um papel passivo de descrição, mas assumem uma função ativa e essencial na gestão e análise de dados em todas as ferramentas utilizadas [10].

Os metadados não são apenas etiquetas estáticas anexadas aos conjuntos de dados, mas também. Representam um sistema dinâmico de informações que oferecem *insights* sobre a estrutura, significado e contexto dos dados armazenados. A presença de declarações semânticas enriquece os metadados ao fornecer interpretações do significado subjacente dos dados, relacionamentos e ligações complexas entre eles. Por exemplo, os metadados podem

indicar onde é que os dados foram recolhidos, em que contexto foram gerados e até mesmo como é que estes se relacionam. Por outro lado, as declarações sintáticas descrevem a estrutura e a formatação dos dados auxiliando na interpretação correta dos mesmos.

Os metadados têm um papel ativo na gestão dos dados uma vez que ajudam na catalogação, organização e recuperação de informações tornando mais fácil localizar dados específicos quando necessário (fator importante no ambiente de *Big Data* onde a quantidade de informação pode ser elevada). Os metadados ajudam os utilizadores a entender a origem dos dados, a sua qualidade, como é que foram transformados e relacionados a outras informações. Assim, os metadados revelam-se como peças chave no processo de gestão e análise de dados, capacitando os utilizadores a explorar, entender e aplicar os dados de maneira eficaz e eficiente.

Os metadados associados a cada ativo de dados (tabelas, arquivos, bases de dados, modelos, definições e outras entidades relacionadas) desempenham um papel fundamental ao simplificar a identificação, avaliação e uso apropriado de cada ativo. Os metadados podem ser categorizados em diversas classes, sendo que um CD concentra-se principalmente em três delas [20]:

- **Metadados Técnicos ou Estruturais**: Este tipo de metadados oferecem uma visão detalhada de como os dados estão organizados. Abrange informação como a estrutura dos dados incluindo tabelas, colunas, tipos de dados, chaves primárias e estrangeiras, como também abrange restrições e validações aplicadas aos dados para garantir a sua integridade e consistência, entre outros. Este tipo de metadados permite ao utilizador compreender a forma e organização dos dados;
- **Metadados de Negócio**: Descrevem o significado e o contexto dos dados no contexto dos processos de negócio da organização. Compreende as descrições do negócio, comentários, anotações e outras informações relevantes que contextualizam os dados de acordo com as necessidades e objetivos do negócio. Os metadados de negócio são valiosos para garantir que os utilizadores entendam a relevância e o propósito de um ativo de dados em relação às operações da organização;
- **Metadados Operacionais**: Fornecem informações cruciais sobre as circunstâncias envolvidas na criação do ativo, bem como o seu histórico de utilização. Incluem dados sobre quando, como e por quem o ativo foi acedido, utilizado, atualizado ou modificado. Para além disso esses metadados devem especificar as permissões de acesso e utilização respondendo questões como a frequência de acesso a uma tabela por utilizadores específicos.

2.6 Catálogo de dados

Os catálogos de dados e as camadas semânticas têm tido algum destaque nos últimos anos na implementação de sistemas analíticos [21], [22]. A utilização de *Data Lakes* é uma forma de lidar com dados semi e não estruturados numa perspetiva *schema-on-read* [21]. Por isso, os dados armazenados num *Data Lake* não possuem um esquema definido, o que pode contribuir para uma degradação na forma como os dados são armazenados, resultando num "*data swamp*" [21], [22]. Por essa razão, a gestão de metadados tem assumido uma importância cada vez maior na implementação de *Data Lakes*, levando inclusivamente ao aparecimento de novas abordagens que consideram a camada de metadados como parte essencial da arquitetura do sistema (*Data Lakehouses*).

Os Catálogos de dados (CD) assumem-se como o padrão para a gestão de metadados no contexto de um sistema analítico [6]. Os CD centralizam e organizam as informações fundamentais referentes aos dados disponíveis, proporcionando uma visão unificada e facilitada das fontes, estrutura, metadados e relações entre os dados. Isto é, um Catálogo de Dados é um repositório de metadados sobre os ativos de dados existentes numa organização, incluindo conjuntos de dados, tabelas, bases de dados e outras fontes, como arquivos ou serviços. Fornece mecanismos para encontrar, compreender e aceder aos dados. A implementação de um catálogo de dados representa uma evolução significativa na eficiência e eficácia da gestão e utilização dos dados dentro de uma organização. Um CD atua como um repositório centralizado que cataloga dados disponíveis na organização, permitindo aos analistas pesquisar e encontrar os dados que necessitam de forma rápida e eficiente ao invés de perderem tempo a navegar por sistemas dispersos, documentos ou de dependerem de terceiros para obterem informações sobre o significado dos dados.

De forma a criar um Catálogo de Dados eficiente, é essencial seguir algumas práticas importantes de implementação e uso dos CD. Primeiramente, é fundamental definir de forma clara qual o propósito e o âmbito do CD. Nesta etapa é necessário identificar os tipos de dados que devem ser incluídos, o público-alvo e quais são os objetivos que o CD visa apoiar. Um fator importante a ter em atenção é a identificação dos interessados e estes devem ser envolvidos no processo de implementação de forma a garantir que o CD responda às necessidades e requisitos identificados. Outro aspeto a ser considerado é a utilização de padrões de metadados (padronização da nomenclatura) de forma a garantir a consistência e otimizar a localização e compreensão das informações. Estes padrões podem estabelecer cabeçalhos uniformes, descrições e atributos obrigatórios de forma a promover a coerência. Após identificar os ativos de dados a serem catalogados e determinar o nível de detalhe e de informações adicionais necessárias para cada ativo de dados o catálogo pode ser

desenvolvido

O processo de recolha e identificação de metadados pode recorrer a várias técnicas (como *Machine Learning*, Inteligência Artificial, anotações ou inferência semântica) de forma automatizar toda a recolha e catalogação dos dados, minimizando o esforço necessário para construir e manter o CD [23]. A procura manual de dados é um processo demorado e propenso a erros. Sem um CD, muitas vezes, é necessário recorrer a processos *ad-hoc*, o que pode resultar em desperdício de tempo e recursos valiosos. Com a existência de um catálogo, a pesquisa é agilizada, reduzindo a duplicação de esforços, facilitando a compreensão de dados e suportando a tomada decisão informada sobre quais os dados são mais apropriados para cada uma das diferentes análises.

Entre os benefícios do uso do CD estão:

- **Eficiência aprimorada na gestão dos dados**: Ao disponibilizar informações sobre os dados de forma organizada e acessível, o catálogo melhora a eficiência na manipulação e utilização dos dados;
- **Compreensão dos dados**: A inclusão detalhada de metadados oferece contexto adicional sobre os dados, tornando a sua compreensão mais fácil e completa para os utilizadores;
- **Redução de risco e erros**: Ao garantir que os analistas trabalham apenas com dados autorizados para serem utilizados para uma determinada finalidade e em conformidade com regulamentações e políticas de privacidade o catálogo mitiga o risco de erros;
- **Qualidade e agilidade na análise dos dados**: Ao permitir o acesso rápido e eficiente aos dados mais apropriados, os analistas, podem responder de forma ágil a problemas, desafios e oportunidades resultando em melhores análises e mais precisas.

A incorporação de um CD apresenta uma série de vantagens consideráveis, no entanto, é importante reconhecer e abordar os desafios inerentes ao uso desses catálogos sendo eles:

- **Manutenção**: Manter o catálogo de dados constantemente atualizado e preciso é uma tarefa desafiadora. Isso requer um compromisso contínuo com a revisão e atualização dos metadados para refletir com precisão as mudanças nos ativos de dados ao longo do tempo. A falta de manutenção adequada pode levar a informações desatualizadas e consequentemente à perda da confiabilidade no CD;
- **Interoperabilidade**: Integrar e conectar CD com diferentes tecnologias, sistemas e fontes de dados pode ser um processo complexo. Alcançar a interoperabilidade efetiva

entre essas plataformas exige esforços significativos de integração e padronização;

- **Acesso e Segurança**: Equilibrar a necessidade de acesso fácil aos dados com a segurança e conformidade adequada é um desafio crítico. O CD deve garantir que apenas os utilizadores autorizados tenham acesso a dados sensíveis e sigilosos, ao mesmo tempo em que oferece facilidade de utilização para aqueles que precisam dos dados para as suas tarefas. Isso envolve a implementação de políticas de segurança eficazes, controlo de acesso e auditoria para proteger os dados e garantir a conformidade com regulamentos e políticas internas.

Apesar das várias vantagens, é crucial reconhecer os desafios que surgem com a implementação e manutenção de CD. Superar estes obstáculos exige um compromisso contínuo com a qualidade, a inovação tecnológica e a colaboração entre os diversos *stakeholders* envolvidos [8]. Num cenário onde a quantidade e complexidade dos dados continuam a crescer exponencialmente, investir em CD robustos e eficazes torna-se não apenas uma vantagem competitiva, mas uma necessidade imperativa para as organizações que procuram contextualizar os seus dados e melhorar os vários aspetos relacionados com o armazenamento e exploração de dados.

Para ser considerado eficaz, o CD deve proporcionar funcionalidades robustas de pesquisa e descoberta como por palavras-chave e termos relacionados com o negócio (tipicamente através de glossários) para permitir que os utilizadores identifiquem conjuntos de dados relevantes de forma ágil. Os recursos de pesquisa em linguagem natural são particularmente importantes para os utilizadores não técnicos facilitando a interação com o CD. A capacidade de classificar os resultados da pesquisa com base na relevância e frequência de utilização também é extremamente útil. Além disso, o CD deve possibilitar que os utilizadores acrescentem informações técnicas por meio de etiquetas, anotações, termos de negócio e outros elementos que melhorem as capacidades de pesquisa e forneçam contexto adicional para uma recuperação de dados mais eficaz.

2.7 Principais Ferramentas Comerciais de Catálogos de Dados

A eficácia de um CD depende, muitas vezes, da capacidade dos utilizadores em compreender e utilizar as informações fornecidas, o que pode ser uma dificuldade nas organizações de grande dimensão, onde os utilizadores possuem diferentes níveis de habilidades e conhecimentos sobre os dados.

Dentro dos CD disponíveis, destacam-se exemplos de soluções proprietárias, tais como o *Atalion Data Catalog* , *Azure Data Catalog* [24] e *Google Cloud Data Catalog* [25]. Estas ferramentas possuem tecnologia e vocabulários específicos, e são otimizadas para as plataformas em que

estão inseridas. Na maior parte dos casos, os catálogos de dados são implementados considerando uma arquitetura e tecnologias e abordagem específica, o que limita a interoperabilidade e a flexibilidade dificultando a integração dessas soluções com outros sistemas e padrões semânticos estabelecidos.

Em alguns dos CD existentes, como é o caso do *Google Cloud Data Catalog* [25], *Atlan Data Catalog* [26], SAP *Data Intelligence platform* [27] e o *Collibra Data Catalog* [28], por vezes torna-se difícil fazer pesquisas bem-sucedidas dos metadados pois os utilizadores podem armazenar metadados utilizando um esquema flexível, que é fornecido pela maioria dos catálogos de dados empresariais de hoje em dia. No entanto, esta flexibilidade resulta em divergências entre os utilizadores quanto ao significado dos metadados e aos termos apropriados para descrevê-los uma vez que diferentes utilizadores podem discordar sobre o significado dos metadados e os termos que devem ser utilizados para descrevê-los. Portanto, os metadados podem ter nomes diferentes ou significados diferentes, dependendo dos utilizadores que os fornecem. Assim, enquanto a flexibilidade desses catálogos facilita o armazenamento inicial dos metadados, acaba por dificultar a sua posterior recuperação e utilização de forma eficaz.

Um CD deve ser abrangente na captura de metadados, de forma a representar todas as informações necessárias. Mas, além disso, deve ser compreensível para uma variedade de utilizadores, levando em consideração diferentes níveis de habilidade e interpretação. O objetivo é criar um ambiente no qual a captura e recuperação de metadados sejam facilitadas e eficazes para todos os utilizadores envolvidos. Na tabela estão presentes algumas das principais ferramentas de CD onde é descrito o que cada uma destas ferramentas faz, o que é que armazenam e o que difere em cada uma delas.

	Google Cloud Datacatalog	**Azure Data Catalog**	**Atlan Data Catalog**	**Collibra Data Catalog**
Funcionalidades	• Procurar entradas de dados às quais o utilizador tem acesso • Marcar entradas de dados com metadados • Proteção de dados confidenciais	• Torna os dados facilmente detetáveis e compreensíveis pelos utilizadores • Registo de fontes de dados	• Pesquisa e descoberta dos dados • Criação de glossários comerciais • Lidar com metadados em grande escala • Facilita a colaboração incorporada • Segurança e	• Centralização dos ativos de dados de toda a organização • Automatização da governança de dados • Criação de uma compreensão partilhada dos dados

			Privacidade dos dados	
Metadados	• Metadados técnicos • Metadados comerciais [29]	• Metadados descritivos (descrições, etiquetas, …) • Metadados estruturais (nomes de colunas, tipos de dados)	• Metadados ativos	• Metadados técnicos • Metadados Comerciais • Metadados de privacidade
Aceder ao catálogo	• Dataplex na consola do Google Cloud • Interface da linha de comando(CLI) gcloud • APIs do Data Catalog • Bibliotecas de cliente do Cloud	• Portal do catálogo de dados do Azure	• Plataforma	• Plataforma • API

Em resumo, os principais desafios associados às ferramentas comerciais de dados são:

- Representação descontextualizada dos vários conceitos: normalmente, estão focados em funcionalidades específicas não sendo possível representar os metadados de acordo com metodologias, modelos e conceitos da literatura. A maioria das soluções captura os metadados relacionados de acordo com um conjunto de entidades não relacionadas como *descrições*, *tags* ou políticas de acesso que acabam por ser baseadas em conceitos genéricos. Não é possível, por exemplo, tirar partido dos conceitos de modelação dimensional para caracterizar, relacionar e principalmente valdiar os vários conceitos.
- Dependência da arquitetura e tencologias associadas: Por exemplo o *Google Cloud Data Catalog* está integrado no ecossistema da *Google Cloud*. Este CD oferece metadados abrangentes e uma integração simples com outros serviços da plataforma, o que simplifica a gestão e a descoberta de dados. No entanto é orientado para o ecossistema *Google Cloud Plataform* e pode não ser a escolha ideal para organizações que utilizam outras plataformas de nuvem. Outro exemplo é o caso do *Azure Data Catalog*. Este CD é integrado com o Azure e oferece uma variedade de recursos de metadados, permitindo que os utilizadores documentem, encontrem e acedam facilmente a dados no ambiente Azure. Da mesma forma que o *Data Catalog da Google* está focado na plataforma *Google Cloud*, o *Azure Data Catalog* está focado na plataforma Azure e por isso não atende às necessidades definidas.

CAPÍTULO 3

Revisão do Estado da Arte

O desenvolvimento de sistemas analíticos tem passado por diversas transformações significativas devido ao aparecimento de novas abordagens para lidar com determinados problemas relacionados com a análise de dados [9]. O aparecimento de "*Big Data*" foi impulsionado por um aumento significativo dos dados utilizados até então para fins analíticos, contribuindo não só para um aumento do volume, mas também da diversidade de fontes de dados envolvidas. Estas mudanças têm sido impulsionadas por vários fatores. Primeiramente, o crescimento de dados proveniente de fontes diversas, como sensores, dispositivos IoT, redes sociais e outras fontes, requer uma abordagem mais sofisticada para catalogar e indexar esses dados. Além disso, a evolução das tecnologias de armazenamento e processamento de dados relacionadas com *Big Data*, como computação em nuvem e *Machine Learning* (ML), introduziu novos desafios e oportunidades [19].

Atualmente, num cenário onde o mercado é altamente competitivo, a recolha e a análise de dados tornou-se uma necessidade para as organizações. Os dados desempenham um papel importante uma vez que são um dos ativos mais importantes de uma organização. As organizações estão a dar cada vez mais importância aos dados e reconhecem que dados bem geridos e estruturados são essenciais para o sucesso organizacional. Com isto, há um interesse crescente em tornar os dados localizáveis, acessíveis, interoperáveis e reutilizáveis (Princípios FAIR) [30]. Os princípios FAIR têm como função estabelecer diretrizes claras para a gestão dos dados, garantindo que estes são facilmente localizáveis por aqueles que necessitam deles, acessíveis a todas as partes interessadas relevantes, interoperáveis para que possam ser utilizados em diferentes contextos e reutilizáveis para evitar a duplicação de

esforços e recursos facilitando assim o acesso e a reutilização de fontes de dados. Uma das formas de implementar estes princípios é através da utilização de CD [31]. Os catálogos são coleções organizadas de metadados projetados para simplificar a pesquisa e a compreensão dos dados sendo que, um CD, abrange metadados sobre várias dimensões, incluindo informações sobre os conjuntos de dados, como restrições de licença; detalhes de processamento, como métricas; informações para pesquisa, como termos de pesquisa relevantes; e até dados sobre as pessoas responsáveis pelos dados, como os proprietários [32]. A implementação de CDs oferece inúmeras vantagens. Como por exemplo, o mapeamento de grandes volumes de dados de forma organizada e acessível. Além disso, disponibilizam mecanismos que permitem aos utilizadores analisar o conteúdo, a qualidade e a utilidade das fontes de dados de forma mais eficaz [32], [33].

Os *Data Lakes* para armazenamento de dados popularizaram-se essencialmente devido à sua flexibilidade de armazenar dados heterogéneos e na sua flexibilidade que novas ferramentas e tecnologias forneceram aos analistas e às equipas que desenvolvem estes sistemas [34]. Com o decorrer dos anos, os CDs ganharam destaque como ferramentas essenciais para gerir metadados associados aos dados armazenados, principalmente no contexto de um *Data Lake* [35].

Os catálogos de dados modernos necessitam de ser capazes de lidar com uma variedade de formatos de dados, incluindo estruturados e não estruturados. Além disso, a interoperabilidade entre diferentes catálogos de dados e a capacidade de descobrir e aceder a dados de fontes externas são elementos-chave para enfrentar os desafios de exploração de dados em larga escala. Como discutido no capítulo anterior, a oferta tecnológica dos CD está fortemente associada a plataformas tecnológicas proprietárias, muitas vezes integradas como parte fundamental de soluções "*full-stack*". Isso significa que os CD estão a tornar-se numa parte essencial de um ecossistema de dados completo, que inclui armazenamento, processamento, análise e visualização de dados [8].

Recentemente foi discutido o conceito de “metadados ativos” o que pode representar alterações na gestão de metadados em aplicações de dados modernas [10]. Esta abordagem representa uma mudança significativa em relação à abordagem tradicional de metadados na qual, os dados, são armazenados estaticamente num repositório e acedidos manualmente uma ferramenta externa. A ideia principal desta nova abordagem é incorporar os metadados diretamente na plataforma de dados, transformando-os numa entidade ativa e dinâmica. Esta perspetiva contrasta com a abordagem “tradicional”, onde os metadados são tratados de forma estática e apenas como uma descrição passiva dos dados.

O conceito de metadados ativos no contexto de um CD potencia uma análise de dados mais

contextualizada, uma vez que os metadados podem ser utilizados para enriquecer a compreensão dos dados em tempo real. Este facto é particularmente valioso num ambiente de dados em constante mudança, no qual a interpretação precisa dos dados é essencial. Além disso, a gestão ativa de metadados pode melhorar a eficiência operacional, uma vez que os metadados estão disponíveis diretamente nos procedimentos de análise e transformação de dados, eliminando a necessidade de consultar um CD de metadados separado [10]. No contexto de um CD, estas tendências estão a impulsionar abordagens mais relevantes e disruptivas para representar e explorar metadados. Os CD estão a evoluir de forma a tornarem-se plataformas mais dinâmicas. Esta evolução é crucial especialmente devido às transformações recentes no cenário de dados, marcado pela crescente complexidade e aumento do volume de dados [10].

Várias pesquisas demonstram o potencial da utilização de um CD na prática. Um exemplo disse é o projeto de CD proposto por [36]. Neste projeto, os utilizadores podem armazenar metadados descritivos, incluindo informações relevantes, como o *Uniform Resource Identifier* que simplifica o acesso aos dados bem como o registo de informações relacionadas com o processo de criação desses dados. O CD proposto foi desenvolvido para lidar com grandes quantidades de dados obtidos a partir de medições em experiências de fusão modernas. Os resultados obtidos neste tipo de experiências frequentemente geram conjuntos de dados extremamente variados, que vão desde medições simples de valores escalares até sequências de vídeo em alta velocidade. Esta abordagem demonstra a importância da catalogação dos dados como uma ferramenta valiosa para organizar, aceder e entender informações complexas, especialmente em campos de pesquisa onde a heterogeneidade dos dados é uma característica comum.

Em [37] foi proposto um método para construir um CD utilizando uma base de dados de grafo (*graph database*) numa plataforma de partilha de dados. Os grafos são especialmente adequados para armazenar dados interconectados. O sistema proposto compreende a inclusão de metadados descritos pelo utilizador, o perfil dos dados e o esquema dos dados. Os metadados descritos pelo utilizador, como título e tamanho do arquivo são exemplos de informações que podem ser adicionadas ao catálogo. O perfil dos dados é elaborado através da extração de informações a partir de características e permite a definição de regras às quais os dados estão sujeitos. Já o esquema dos dados é aplicado a dados tabulares e refere-se se às informações da tabela nesses dados como o tipo de dados e o seu comprimento. Através do estudo realizado pelos autores de [37] foi possível constatar que a performance da estrutura proposta, baseada numa base de dados de grafo, é superior àquela em que o CD é criado através de uma base de dados relacional. Este fator destaca a eficácia e a eficiência desta

abordagem na organização e gestão de informação.

Em [38], os autores introduzem uma camada semântica que integra um CD semântico construído utilizando tecnologias padrão. Esta camada semântica é essencialmente composta por uma ontologia e por um grafo de conhecimento, oferecendo uma representação semântica abrangente de todos os recursos do *Data Lake*. Os recursos abrangem uma gama diversificada de elementos, incluindo documentos, conjuntos de dados e bases de dados, refletindo a heterogeneidade inerente ao ambiente de um *Data Lake*. A descrição semântica desses recursos fornece informações detalhadas sobre o conteúdo de cada recurso, a sua origem e as permissões do controlo de acessos associadas.

Um componente fundamental desta camada semântica é o suporte ao controlo de acessos que é fundamentado pela integração da ontologia *Open Digital Rights Language* (ODRL). A ODRL é uma linguagem semântica que permite a especificação de políticas de controlo de acesso de forma altamente descritiva. Esta abordagem possibilita a descrição granular do acesso aos recursos do *Data Lake*, mapeando quais são os recursos disponíveis, quem são os utilizadores autorizados e as ações permitidas a serem realizadas sobre cada um desses recursos. Por exemplo, esta abordagem permitiria especificar quem pode visualizar, modificar, partilhar ou remover um determinado recurso. Para além disso, estas políticas podem ser atualizadas e adaptadas de acordo com as necessidades em constante evolução de cada uma das organizações.

A utilização de grafos de conhecimento para a representação de metadados é uma tendência inovadora como evidenciado pelo estudo realizado em [39]. Neste artigo, os autores exploram a aplicação de grafos de conhecimento como uma abordagem eficaz para a descrição semântica e para a gestão de dados num contexto de *Data Lake*. A principal contribuição deste estudo reside na capacidade dos grafos de conhecimento em fornecer uma representação estruturada e semântica dos dados armazenados no DL. O que vai além da simples catalogação e organização dos metadados, pois permite uma compreensão mais profunda do significado e das relações entre os dados.

Os grafos de conhecimento podem capturar informações sobre conceitos, entidades e as suas interações, criando assim uma rede de conhecimento que enriquece a compreensão dos dados [40]. Ao adotar esta abordagem, a gestão de dados no DL é melhorada em vários sentidos [39]. Primeiramente, a capacidade de reutilização dos dados é amplamente beneficiada uma vez que os metadados semânticos permitem que os utilizadores identifiquem rapidamente dados relevantes, compreendam as suas características e relações, o que é fundamental para a reutilização eficaz de informações em diferentes contextos e aplicações. Os grafos de conhecimento fornecem uma base sólida para que esses algoritmos interpretem e explorem

os dados com mais precisão [40]. Os autores de [39] argumentam que os dados que carecem de descrições detalhadas sobre o seu significado e esquema têm um valor reduzido. Essa falta de contexto e informação semântica pode limitar a compreensão e a utilização eficaz dos dados. A abordagem proposta por estes autores integra os dados e os seus esquemas semânticos em grafos de conhecimento, permitindo a consolidação do conhecimento do domínio, incluindo regras e restrições de dados. Os autores afirmam que aplicações e interfaces de utilizador podem tirar proveito do conhecimento de forma eficaz, sem a necessidade de integrá-lo diretamente no seu código.

Existem diversos estudos que demonstram a eficácia da utilização de metadados para suportar a camada de apresentação de dados que geralmente se traduz em relatórios e *dashboards*. Em [41], é apresentado um grafo de conhecimento que é nada mais do que uma ontologia especializada. Esta ontologia é projetada para facilitar a descoberta de indicadores e a visualização de dados de forma mais significativa. Através da análise detalhada dos metadados a ontologia ajuda a identificar quais é que são os indicadores relevantes para um determinado contexto ou problema. Os autores também apresentam o desenvolvimento de uma aplicação que é capaz de analisar os metadados disponíveis, e através dessa análise permite construir e apresentar *dashboards* que estejam de acordo com os indicadores. Esta abordagem automatiza a criação de *dashboards* através da integração de metadados na camada de apresentação, agregando valor à interpretação dos dados. Com isto, este estudo, demonstra como os metadados e a utilização de uma ontologia bem elaborada podem ser utilizados de forma inteligente para aprimorar a apresentação dos dados por meio de relatórios e *dashboards*.

Já em [42] é introduzida uma abordagem inovadora que se concentra na personalização da exploração de dados, especialmente no contexto urbano. Para alcançar este objetivo, os autores desenvolvem uma ontologia especializada que oferece uma representação semântica dos indicadores-chave relevantes para análises urbanas. Esta ontologia é fundamental para estabelecer uma estrutura coerente e compreensível para os indicadores-chave, garantindo que a sua interpretação e aplicação sejam consistentes e precisas. Esta, ajuda também a estabelecer relações e contextos entre os diferentes indicadores, permitindo uma visão mais abrangente e integrada da informação. Além da ontologia, é também apresentada uma camada semântica que atua como um componente-chave no suporte à personalização do acesso aos dados urbanos. Esta camada é projetada para entender as preferências e as necessidades individuais dos utilizadores, permitindo a adaptação da apresentação dos dados de acordo com suas especificidades. Ao integrar a ontologia e a camada semântica, o sistema resultante é capaz de oferecer uma experiência personalizada aos utilizadores, possibilitando a

exploração dos dados de forma mais significativa e eficaz.

Ainda neste domínio foi apresentado um Catálogo de Classificação multirrótulo que utiliza aprendizagem máquina para a criação de um conjunto de rótulos para conjuntos de dados [43]. Os autores, propõem um catálogo online, baseado numa ontologia, que pode atribuir automaticamente diferentes descrições a cada conjunto de dados. Sendo que, as anotações geradas por este sistema seguem os princípios FAIR (Facilidade de Localização, Acessibilidade, Interoperabilidade e Reutilização) e TRUST (Transparência, Responsabilidade, Foco no Usuário, Sustentabilidade e Tecnologia) [44]. Isto significa que o sistema não organiza apenas os dados de forma eficaz, mas também os torna acessíveis, interconectados, e prontos para reutilização, enquanto se mantém transparente, responsável e centrado nas necessidades dos utilizadores, assegurando a sustentabilidade e aplicando tecnologias adequadas para alcançar estes objetivos.

No geral, todos os estudos mencionados anteriormente, evidenciam a importância crescente de incorporar metadados na camada de apresentação dos dados. Nestes estudos, é demonstrado como os metadados e a semântica podem ser poderosas ferramentas para melhorar a apresentação e a análise de dados, tornando a informação mais acessível e útil para os utilizadores finais. Também destacam a importância de adaptar as visualizações de dados às necessidades específicas, o que pode aumentar a eficácia e a utilidade das ferramentas de análise de dados.

Resumidamente, a revisão do estado da arte consistiu em três áreas principais:

1º **Compreender a evolução da área dos dados**: Compreensão da evolução da área dos dados. O que inclui o estudo das tendências emergentes, avanços tecnológicos, isto é, analisar a forma como os sistemas analíticos têm evoluído;

2º **Compreender a composição/estrutura dos Catálogos de Dados:** Identificação das tendências associadas ao desenvolvimento dos Catálogos de Dados;

3º **Compreender como é que estão a ser desenvolvidos os Catálogos de Dados:** Estudo das técnicas e ferramentas utilizadas para criar Catálogos de Dados.

Na Tabela 1, encontram-se listados os artigos previamente referenciados, categorizados de acordo com a área de revisão do estado da arte à qual pertencem, o ano da sua publicação e no caso de artigos que abordem o tema do desenvolvimento de CD é indicada a abordagem utilizada para a representação de metadados.

Área	Ano	Artigo	Abordagem
1ª	2016	The FAIR Guiding Principles for scientific data management and stewardship [30]	
3º	2016	Data catalog project - A browsable, searchable, metadata system. Fusion Engineering and Design [36]	
3º	2017	From Data to City Indicators: A Knowledge Graph for Supporting Automatic Generation of Dashboards [41]	Ontologia
2º	2017	Data Catalogs Are the New Black in Data Management and Analytics [35]	
3º	2019	Using a Smart City ontology to support personalised exploration of urban data [42]	Ontologia especializada
1ª	2020	FAIR Enough? Enhancing the Usage of Enterprise Data with Data Catalogs [31]	
1º	2020	The TRUST Principles for digital repositories [44]	
3º	2020	Using semantic technologies to manage a data lake: Data catalog, provenance and access control [38]	Grafo de Conhecimento e Ontologia
2º	2020	Knowledge Graphs. Springer International Publishing [40]	
2º	2020	Introduction to Data Catalogs [32]	
2º	2020	The Data Catalog: Sherlock Holmes Data Sleuthing for Analytics [33]	
3º	2020	Building methods of intelligent data catalog based on graph database for data sharing platform [37]	Grafo
1º	2021	Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics [19]	
1º	2021	On data lake architectures and metadata management [34]	

2º	2021	Market Guide for Active Metadata Management [10]	
3º	2021	Using Knowledge Graphs to Manage a Data Lake [39]	Grafo de Conhecimento
1º	2022	Improving the documentation and findability of data services and repositories: A review of (meta)data management approaches [45]	
2º	2022	A catalogue with semantic annotations makes multilabel datasets FAIR [43]	Catálogo de Classificação
1º	2023	Data Lakehouse vs Data Warehouse vs Data Lake - Comparison of data platforms [9]	

Tabela 1- Área de Revisão do estado da arte e respetivos arquivos

CAPÍTULO 4

Definição de Medidas para Análise de Dados

Cada vez mais, as organizações, estão a adotar a prática de apoiar as suas decisões diárias com suporte em análises de *Business Intelligence* (BI) [46]. Este processo permite que quem toma as decisões obtenha uma visão mais completa do que está a acontecer ao nível organizacional. Para alcançar este propósito, os decisores geralmente recebem relatórios e representações visuais, por exemplo gráficos, que são baseados nos dados de negócio que estão armazenados no *Data Warehouse* (DW) tipicamente utilizando o modelo dimensional [5]. Nesta abordagem, as medidas relevantes para o negócio são armazenadas em tabelas de factos para que possam ser enquadradas de acordo com as diferentes dimensões de análise.

4.1 Caso de estudo

A criação de um DW é uma prática fundamental para a análise de dados permitindo que as organizações reúnam informações de várias fontes, as transformem e as organizem num ambiente centralizado facilitando a análise e a tomada de decisões baseadas em dados. Neste contexto, as motivações para a criação deste DW passam por melhorar a qualidade de serviço de Vendas e obter *insights* valiosos para aprimorar as estratégias comerciais. Com a criação do DW pretende-se também identificar sazonalidades/tendências/periodicidade de vendas por loja, cliente e produtos.

O esquema dimensional é composto por duas tabelas de factos:

- "*fact_sales_lines*": tabela de factos transacional que armazena as linhas de venda (grão) com as medidas: quantidade, valor, custo e margem (em euros).

- "*fact_sales_month*": tabela de factos agregada que armazena o total de vendas por mês, composta pelas medidas: número de clientes, preço médio de venda, quantidade e receita (em euros).

O diagrama da Figura 2 identifica as medidas importantes para o processo de negócio e apresenta o contexto dimensional em que essas medidas podem ser utilizadas. Os factos e as dimensões presentes neste diagrama podem ser combinados de várias formas respondendo a uma ampla gama de questões analíticas.

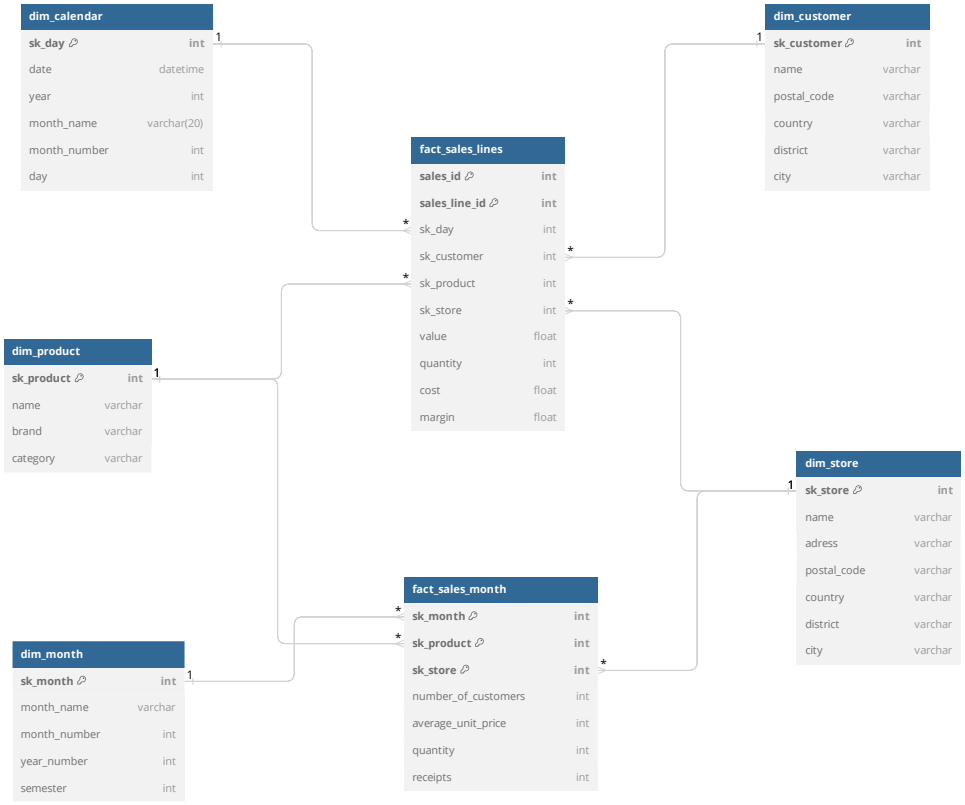


Figura 2- Sales Schema

Num contexto de um esquema em estrela, como o esquema apresentado na Figura 2, a maioria das consultas segue um padrão consistente. Essas consultas solicitam um ou mais factos juntamente com os atributos dimensionais que fornecem o contexto desejado. Os valores das dimensões também são utilizados para limitar o âmbito da consulta servindo como base para filtros ou restrições nos dados a serem procurados e agregados [12]. Embora os factos sejam armazenados num nível específico de detalhe (granularidade), podem ser agrupados em vários níveis de detalhe. Neste caso, a granularidade da tabela *fact_sales_line* é uma linha de venda já a tabela *fact_sales_month* é uma tabela de agregados mensais.

A medida "Quantidade" é uma das medidas presentes na tabela de factos *"fact_sales_lines"* que permite analisar as vendas em diferentes perspetivas, como por exemplo analisar o número de vendas por dia (dim_calendar), por loja (dim_store) ou até por cliente

(dim_customer). Neste caso, os factos podem ser resumidos somando os valores desta medida e agregados em função de todas as dimensões. Como tal, a medida é aditiva assim como a quantidade e o custo.

No contexto da modelação dimensional, para além das medidas aditivas, podem ainda existir medidas semi-aditivas e não aditivas. Num DW onde a tabela de factos *fact_stock_diario* representa o *stock* diário de todos os produtos existentes nas diferentes lojas, a medida *stock* do produto é uma medida semi-aditiva pois se considerarmos as dimensões "Produto", "Loja" e "Calendário", podemos somar o *stock* de vários produtos e de diversas lojas, mas não podemos agregá-lo sobre a dimensão "Calendário". Em cada consulta e relatório, os factos semi-aditivos devem ser utilizados com precaução. Ao somar um facto semi-aditivo, a consulta deve ser filtrada por uma linha única na dimensão não aditiva ou agrupada por linhas na dimensão não aditiva. Se o relatório for composto por subtotais ou totais gerais, as mesmas regras também devem ser seguidas dentro do próprio relatório - não somar através da dimensão não aditiva ao criar um total [12].

As medidas não aditivas não podem ser utilizadas para somar factos em nenhuma das dimensões. Na tabela *"fact_sales_month"* a medida "*average_unit_price"* é uma medida não aditiva uma vez que não faz sentido somar o preço médio de um produto em nenhuma das dimensões. No entanto, como o objetivo da modelação dimensional é facilitar consultas e análises, é preferível incluir medidas numéricas e aditivas na tabela de factos [15].

As medidas são também classificadas como elementares, agregadas ou derivadas [47]. Uma medida elementar representa uma medida (facto) na tabela de factos com o menor nível de detalhe possível. Por exemplo, no esquema dimensional apresentado, no caso da tabela de *"fact_sales_lines",* as medidas quantidade (número de unidades vendidas), valor (valor de venda em euros) e custo (valor de compra em euros) correspondem à granularidade da linha de vendas, ou seja, ao nível de detalhe mais granular. As medidas derivadas são obtidas a partir de medidas elementares. Isto é, são calculadas com base noutras medidas. No esquema apresentado na Figura 2, o cálculo da Margem, na tabela *"fact_sales_lines"*, é uma medida derivada uma vez que pode ser calculada como a diferença entre o "Valor" e o "Custo". Como esta medida é calculada a partir de outras medidas é então classificada como uma medida derivada. Ainda existem as medidas agregadas que são o resultado da aplicação de uma função de agregação. Por exemplo, a soma de toda a quantidade vendida num determinado período representa uma medida agregada.

As duas abordagens para a categorização de medidas são importantes para a construção de uma camada semântica capaz de suportar a representação de metadados num CD, uma vez que permite:

- Identificar as medidas que estão no nível mais granular de detalhe: As medidas elementares representam os dados no nível mais granular de detalhe e por isso não dependem de outras medidas para o seu cálculo e podem ser utilizadas para agregar e calcular novas medidas;
- Identificar as dependências entre medidas: A classificação das medidas desta forma também possibilita a identificação de dependência entre medidas. Este fator é importante, pois se uma das componentes necessárias para calcular uma medida derivada deixar de existir, a medida derivada perde a sua fundamentação e deixa de fazer sentido. Além disso, se as medidas utilizadas para calcular as medidas derivadas forem alteradas, o valor da medida derivada também será recalculado, dado que esta é calculada com base nas outras medidas. Por exemplo, no cálculo da margem, que depende do custo e do valor, se o custo for convertido para dólares enquanto o valor permanece em euros, a margem perde a sua coerência.
- Identificar a utilização de funções de agregação: Classificar as medidas desta forma também é útil para identificar a aplicação de funções de agregação. As funções de agregação são utilizadas para resumir os dados, geralmente agrupando-os em categorias. Este fator é importante para análises de alto nível e geração de relatórios. Por exemplo, somar as vendas mensais para obter as vendas anuais é uma forma de aplicar uma função de agregação. Esta categorização ajuda a distinguir entre as medidas que são usadas no seu nível mais granular e aquelas que são agregadas.
- Identificar a aplicação incorreta de funções de agregação: a categorização das medidas possibilita a identificação da aplicação errada de funções de agregação. Esta identificação é fundamental para garantir a correta aplicação das funções de agregação, especialmente em medidas onde estas funções não podem ser aplicadas. Por exemplo, é possível identificar se uma função de agregação foi indevidamente aplicada a uma medida que não é aditiva, evitando interpretações erradas dos dados.

Durante o processo de modelação dimensional, podem ser identificadas diversas medidas a partir dos requisitos de negócio. As métricas elementares estão tipicamente colocadas na tabela de factos de grão mais elementar. As restantes métricas podem ser integradas em tabelas de factos de acordo com a sua dimensionalidade. Como nem sempre se opta pelo desenho de esquemas derivados e agregados, algumas destas métricas são criadas na ferramenta de BI utilizada para a construção dos relatórios. Considerando o exemplo da Figura 2, as seguintes medidas poderiam ser criadas para suportar aspetos específicos dos relatórios ou *dashboards* utilizadas:

- **“Valor Vendas Totais”**: Esta medida é uma medida agregada que resulta da soma

do “Valor” das vendas. Fornece uma visão geral do valor total das vendas;

- **“Total de Unidades Vendidas”**: Medida agregada que resulta da soma da quantidade vendida;
- “**Valor Vendas do Mês Atual”** e **“Valor Vendas do Mês Anterior”**: Medidas que resultam da agregação da medida elementar “Valor”, mas ao longo da dimensão temporal “Calendário”. Consideram especificamente o mês atual e o último mês, permitindo a análise das vendas numa perspetiva temporal mais específica;
- **“Variação Mensal do Valor das Vendas”**: Medida derivada que calcula a variação do valor das vendas como a diferença entre o valor das vendas do mês atual e o valor das vendas do mês anterior:

  $$VariaçãoMensalValorVendas = VendasMêsAtual - ValorVendasMêsAnterior \quad (1)$$

- **“% de Variação do Valor das Vendas mensais”**: Medida não aditiva que corresponde à razão entre a Variação Mensal do Valor das Vendas e o Valor Total de Vendas do Mês Anterior. Esta proporção é importante para perceber o crescimento ou a redução percentual nas vendas em relação ao mês anterior.

  $$Variação ValorVendas\ (\%) = \frac{VariaçãoMensalValorVendas}{ValorVendasMêsAnterior} \times 100 \quad (2)$$

O processo de tipificação de medidas é importante para que a utilização das medidas seja realizada de forma a produzir resultados coerentes com os requisitos de negócio. Por exemplo, deve ser assegurado que não existem medidas calculadas através da soma de medidas não aditivas.

Com base nas medidas definidas anteriormente, poderá ser útil para a organização criar uma visualização com os valores de variação mensal do valor das vendas (1) de cada produto e uma segunda visualização com a evolução da percentagem de variação do valor das vendas mensais (2) associada a cada produto em cada dia. De notar que neste caso o valor desta medida está diretamente ligado ao valor da medida Variação mensal do valor das Vendas uma vez que a Medida % de Variação do valor das vendas mensais (2) é calculada com base na medida Variação Mensal do Valor das Vendas (1). Esta dependência requer cuidados especiais porque qualquer alteração na fórmula utilizada para calcular a Medida 1 afeta os valores da Medida 2. Supondo que por algum motivo a empresa compreendeu que do ponto de vista visual é melhor reportar a variação das vendas em milhares de euros e passa a utilizar a fórmula de cálculo da Medida 3. Esta mudança significa que a Medida 2 também deve ser alterada, caso contrário o gráfico apresentará valores incorretos uma vez que o valor da medida Variação Mensal do Valor das Vendas passaria a estar em milhares e o valor da medida Valor Vendas Mês Anterior (Medida também utilizada para calcular a Medida 2) não.

$$VariaçãoMensalVendas = (ValorVendasMêsAtual - ValorVendasMêsAnterior)/1000 \quad (3)$$

Portanto, ao analisar dados com centenas de medidas abrangendo diferentes áreas de negócio e com muitas medidas derivadas, é importante conhecer os relacionamentos entre cada uma das medidas e identificar os relatórios/gráficos em que são utilizadas para evitar erros deste tipo. Nestes casos, é importante garantir que alterações em qualquer uma das medidas não conduz a medidas incorretas nas áreas que dependem destas. A representação dos dados é outra consideração importante, pois uma mesma medida pode ser representada em diferentes formatos levando em consideração o contexto em que é utilizada. Existem também regras de negócio que podem afetar não apenas o cálculo da medida, mas também a forma como os atributos da dimensão são apresentados. Estas considerações são analisadas e avaliadas nas fases iniciais de desenvolvimento. Por exemplo, no processo de desenvolvimento da modelação dimensional [15], a identificação e a categorização das medidas ocorrem durante o desenho dos factos.

4.2 Utilização de Ferramentas de BI

Considerando as medidas apresentadas anteriormente, softwares de análise de dados, como o *Power BI* [2], podem ser utilizados para desenvolver relatórios que são essenciais para uma análise aprofundada dos dados do negócio. Os relatórios e os *dashboards* são duas ferramentas comuns para visualizar dados de forma clara e intuitiva. Estas ferramentas fornecem diversas visualizações para atender às necessidades analíticas. Gráficos de barras, pizza e gráficos de *gauge* são alguns exemplos de visualizações que podem compor relatórios e *dashboards*. Os relatórios e os *dashboards* podem ter significados diferentes dependendo da ferramenta ou âmbito em que são utilizados. Por exemplo, no Power BI (PBI) os *dashboards* são colocados no topo da hierarquia, ou seja, são construídos a partir dos relatórios desenvolvidos, representando uma narrativa por meio de visualizações. Neste caso, os relatórios incorporam a complexidade relacionada com a gestão e apresentação de dados resumidos. Os relatórios são compostos por páginas que, por sua vez, são compostas por recursos visuais configurados para apresentar dados de um conjunto de dados específico. Cada visual é composto por objetos de dados contendo informação (tabelas de dimensões e fatos) e as respetivas medidas utilizadas nos procedimentos de agregação.

As medidas consistem em cálculos realizados por meio de expressões matemáticas. Quando combinadas com filtros de visual, de página e de relatório, essas medidas permitem que a agregação de dados se ajuste de acordo com a interação do utilizador com os relatórios, permitindo uma exploração de dados mais rápida e dinâmica. As transformações nos processos de negócio podem implicar mudanças na composição das medidas, sendo que

[2] https://powerbi.microsoft.com/pt-pt/

essas medidas, podem ser partilhadas não apenas por múltiplas visualizações/relatórios, mas também podem ser utilizadas para criar outras medidas. Manter a exatidão das visualizações em relatórios de grande dimensão pode ser difícil, pois uma simples alteração pode exigir diversas adaptações, o que consome tempo e é propício a erros. Além disso, ao lidar com diversas medidas e esquemas dimensionais complexos, podem ocorrer erros e comprometer a precisão dos resultados apresentados nas visualizações.

Por exemplo, uma organização desenvolveu um relatório baseado no DW de Vendas (Figura 2) e foi necessário desenvolver em DAX (*Data Analysis eXpression Language*), no caso do Power BI, uma medida que calcula o Valor Total das Vendas de forma a atender aos requisitos identificados. Esta medida foi denominada de "ValorVendasTotal" e é calculada somando o facto "Valor" da tabela "fact_sales_lines".

$$ValorVendasTotal = SUM('fact\_sales\_lines'[Valor])$$

Esta medida é fundamental para desenvolver várias visualizações tais como: gráfico de linhas com a análise do valor total de vendas por mês, uma matriz que analisa o valor total de vendas por Família de produtos, marca, produto por mês e outra matriz que analisa o Valor Total de Vendas por Produto por Loja.

No gráfico que apresenta o Valor Total de Vendas por mês, há meses em que não foram registadas vendas, resultando na ausência de representação desses meses no gráfico. No entanto os decisores querem que todos os meses estejam representados no gráfico. Para atender a esta necessidade é necessário alterar a medida "ValorVendasTotal" e acrescentar um "+0" e assim todos os meses passam a ser representados no gráfico mesmo os meses em que não há o registo de vendas. Este tipo de especificidades fazem com que o desenvolvimento de novas medidas atinja proporções por vezes difíceis de monitorizar.

$$ValorVendasTotal = SUM('fact\_sales\_lines'[Valor])+0$$

O grande problema reside no facto de que a pessoa que realizou a alteração na medida "ValorVendasTotal" não tinha conhecimento das visualizações utilizam esta medida. Como resultado, não pôde avaliar devidamente o impacto dessa alteração. Após esta modificação, muitas das visualizações deixaram de funcionar corretamente como é o caso das visualizações que envolvem a utilização de uma matriz (análise do valor total de vendas por Família de produtos, marca e produto e a análise do Valor Total de Vendas por Produto por Loja). O problema está relacionado com a introdução do "+0" na medida, esta alteração resultou na geração de um produto cartesiano nas matrizes levando a ser apresentadas ao utilizador todas as combinações possíveis. Em casos em que o número de registos é substancial, a exibição das visualizações torna-se inviável, uma vez que os recursos do Power BI são esgotados. Numa situação como esta, a implementação de um CD que faça a monitorização das medidas

seria altamente benéfico pois, o CD, permitiria identificar de forma eficaz em que relatórios e visualizações é que uma medida específica, como “ValorVendasTotal”, está a ser utilizada. Este fator é crucial para avaliar o impacto de quaisquer futuras alterações e evitar problemas no funcionamento das visualizações devido a modificações inadvertidas.

CAPÍTULO 5

Processo de Modelação Dimensional

O processo de modelação dimensional é uma abordagem bastante conhecida e explorada para a conceção DWs. Através de um conjunto de etapas, são desenvolvidos modelos de dados relacionais que representam os dados operacionais para propósitos analíticos.

Através de um caso de estudo proposto procura-se apresentar uma análise aprofundada sobre a eficácia e importância da implementação de um Catálogo de Dados (CD) no contexto da validação e monitorização de medidas no processo de construção de relatórios/*dashboards*. A aplicabilidade do CD reflete-se na capacidade de representar e documentar as informações relacionados com os princípios fundamentais da modelação dimensional, o que permite uma compreensão abrangente dos dados, das suas características e dos seus relacionamentos. Ao alinhar-se com os princípios fundamentais da modelação dimensional, o catálogo possibilita a estruturação dos dados, favorecendo a criação de relatórios/*dashboards* coerentes e alinhados com as necessidades e objetivos definidos.

Neste capítulo, é apresentado um caso de estudo que servirá de base à construção do Catálogo de Dados. Com este capítulo pretende-se demonstrar os vários aspetos do processo de modelação dimensional, assim como o tipo de metadados que podem ser identificados utilizando um caso simples que é baseado num cenário real.

5.1 Caso de Estudo

A FootFusion (anonimização de uma empresa real cuja dimensão foi simplificada) é uma fábrica de produção de calçado reconhecida pela qualidade dos seus produtos. O seu modelo

de negócio envolve uma extensa rede de vendedores encarregados de distribuir os produtos diretamente para uma ampla gama de clientes que comercializam os sapatos nas suas lojas. Estes produtos são distribuídos no mercado nacional e no mercado internacional. Nos últimos anos, a empresa tem enfrentado desafios essencialmente relacionados com uma concorrência mais forte e a um mercado cada vez mais volátil e imprevisível, que a empresa tem tido dificuldades em se enquadrar. Em alguns casos a resposta a estas mudanças de mercado é lenta, o que faz com que alguns dos seus concorrentes se posicionem numa posição dominante do mercado. De forma a superar as dificuldades encontradas, é importante realizar uma análise aprofundada dos dados disponíveis no departamento de vendas, identificando as áreas de melhoria na sua estratégia de vendas.

Após uma reunião com os gestores, uma das análises prioritárias para a empresa passa por determinar quais são os produtos mais populares no mercado. Isso envolve identificar as preferências dos consumidores, tendências de compra e feedbacks dos clientes. Por exemplo, através da identificação de sazonalidades.

Com base na análise dos dados a FootFusion pode tomar várias ações. Algumas dessas ações podem incluir:

- **Otimização da estratégia de stock**: Identificar os produtos mais populares e ajustar a produção, priorizando a fabricação dos itens com mais vendas e reduzindo a produção dos menos populares. Isso ajuda a garantir que o stock esteja alinhado com as preferências dos clientes.
- **Identificação de oportunidades de expansão**: Utilizar os dados das vendas para identificar regiões onde a empresa tem poucas vendas e, assim, encontrar oportunidades de expansão. Também é possível concentrar esforços nas regiões com maior potencial, investindo em marketing, vendas e logística para expandir a presença da empresa.
- **Identificação de sazonalidades**: Identificar padrões sazonais nos dados de vendas para ajustar o *stock*, aumentando a capacidade de produção durante os períodos onde as vendas são mais frequentes e reduzindo-a nos períodos de menor atividade.
- **Ajuste de preços**: Analisar os dados para identificar produtos com preços desalinhados em relação aos concorrentes. Ajustar os preços adequadamente para maximizar as vendas e a margem de lucro, garantindo competitividade no mercado.

As estratégias decorrentes da análise detalhada dos dados de vendas oferecem uma oportunidade valiosa para melhorar a estratégia de vendas da FootFusion.

5.2 Objetivos

A implementação do Data Mart de Vendas tem como principal objetivo fornecer informações valiosas e pertinentes para os *decision makers*, oferecendo suporte crucial nas análises essenciais para o funcionamento eficaz do negócio. Este Data Mart será desenvolvido para atender às necessidades do departamento comercial, sendo responsável pela gestão estratégica das vendas. Os principais utilizadores deste Data Mart serão os diretores do departamento, proporcionando-lhes uma visão detalhada e precisa do desempenho das vendas e auxiliando na tomada de decisões estratégicas respondendo a questões como: Quais são os principais produtos vendidos que contribuem para uma maior receita de vendas?; Quais são os principais padrões de compra dos clientes, como recorrência, valor médio de compras, entre outras?; Qual é o período do ano onde a faturação é mais elevada?; Qual é a margem de lucro associada a diferentes produtos, entre outras. A criação deste Data Mart suportará o processo de negócio da venda de produtos.

5.3 Descrição do processo de negócio

Associados à atividade da FootFusion, existem diversos processos de negócio que suportam diversas áreas/departamentos, como é o caso da produção de calçado e da sua comercialização. Na Figura 3, está representado o diagrama BPMN global dos processos de negócio da FootFusion sendo que este processo é composto por um conjunto de subprocessos. O processo descreve o momento em que um cliente realiza uma encomenda até à entrega dos produtos e a finalização da venda.

Este processo inicia-se quando o cliente realiza uma encomenda a um vendedor solicitando a aquisição de produtos específicos. O vendedor, ao receber o pedido da encomenda, comunica-o à manufatura iniciando assim o processo de produção. A manufatura, de seguida, adquire os materiais necessários, produz os produtos solicitados, realiza o devido embalamento e envia os produtos para o armazém. Ao receber os produtos produzidos, o armazém assume a responsabilidade de comunicar à transportadora para efetuar a distribuição dos produtos aos clientes. A transportadora, por sua vez, realiza a entrega dos produtos ao cliente. Quando os produtos são entregues à transportadora o vendedor recebe o aviso da entrega e emite a fatura. Após a entrega, o cliente recebe os produtos encomendados e verifica a qualidade dos produtos recebidos. Se houver algum problema com a encomenda, é emitida uma nota de crédito ao cliente se não o processo termina.

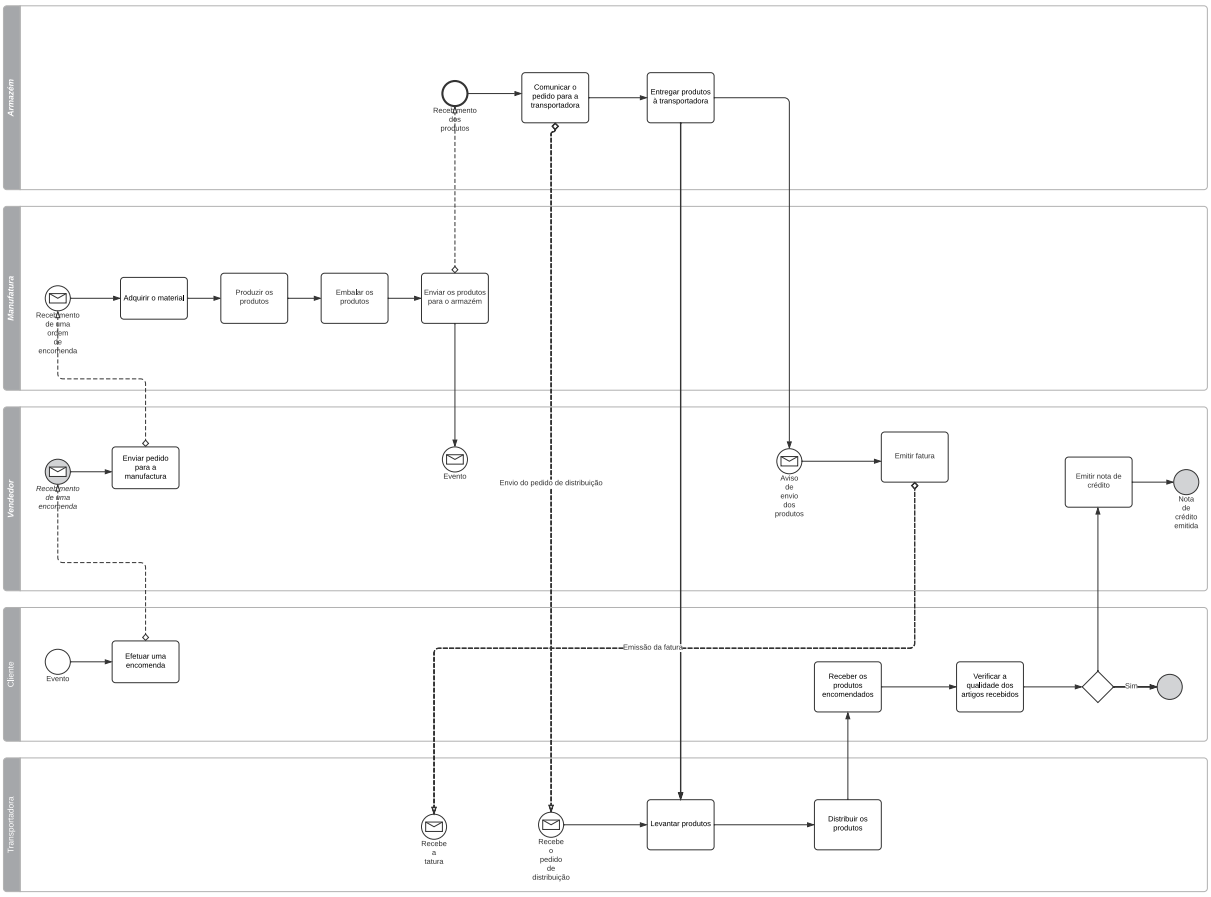


Figura 3- Processo de negócio da FootFusion

5.4 Lista de *queries* dos clientes do Data Mart de Vendas

Neste caso o processo de negócio que será trabalhado é o processo de faturação para isso será desenvolvido um *Data Mart* de Vendas. O *Data Mart* de vendas é um subconjunto de um *Data Warehouse* focado em armazenar e disponibilizar informações relacionadas com a área de vendas. Este, pode conter uma ampla gama de dados relacionados às vendas, incluindo informações sobre artigos, clientes, regiões, preços, volume de vendas e muito mais. Os principais clientes deste *Data Mart* são o diretor do departamento Comercial e os responsáveis pela estratégia de vendas na organização.

De seguida, são apresentados alguns exemplos de consultas que podem ser feitas ao *Data Mart*, alinhadas com as preocupações atuais do diretor comercial da FootFusion:

1. Qual é o valor de faturação por mês?
2. Qual é o valor de faturação por mês e por produto?
3. Qual é a diferença do valor de Faturação comparado com o valor do ano anterior?
4. Qual é o número de unidades vendidas por produto?
5. Qual é o número de unidades vendidas por produto por região?
6. Qual é o valor de faturação por cliente, por região e por produto?

7. Qual o valor de crescimento do número de clientes comparado com o ano anterior por região?
8. Qual o valor da margem por produto e por região?
9. Qual é a margem de lucro por marca?
10. Quais são os clientes com um maior volume de faturação?
11. Quais são os 5 vendedores com um maior valor de vendas?
12. Qual é o valor médio de vendas por cliente, região e vendedor?
13. Qual é o número de transações por região?

5.5 Modelo dimensional: Aplicação do método dos 4 passos de Kimball

Para o desenvolvimento do *Data Warehouse*, foi adotado o método dos 4 passos de *Kimball* [5]. O primeiro passo deste método passa pela identificação e caracterização das áreas de negócio relevantes para a organização. Cada área de negócio representa um conjunto específico de atividades e operações que desempenham um papel crucial no funcionamento da empresa.

Cada processo de negócio possui as suas próprias particularidades, desafios e requisitos de exploração. Compreender profundamente essas particularidades é fundamental para o sucesso do projeto de implementação do *Data Warehouse*, pois permite a criação de uma estrutura de dados que atenda às necessidades analíticas e estratégicas de cada área de negócio. Neste caso o processo de negócio analisado é o processo do registo de Vendas dos produtos (emissão de uma fatura). Na Figura 4, encontra-se representado o diagrama BPMN representativo do processo de emissão de um documento de venda. Em primeiro lugar é necessário identificar qual é o documento que vai ser emitido (por exemplo fatura ou nota de crádito) seguido da identificação do cliente para o qual o documento vai ser emitido. De seguida, é necessário verificar quais são as documentos a transformar. Após a identificação das documentos a informação é validada e 0 documento é emitido.

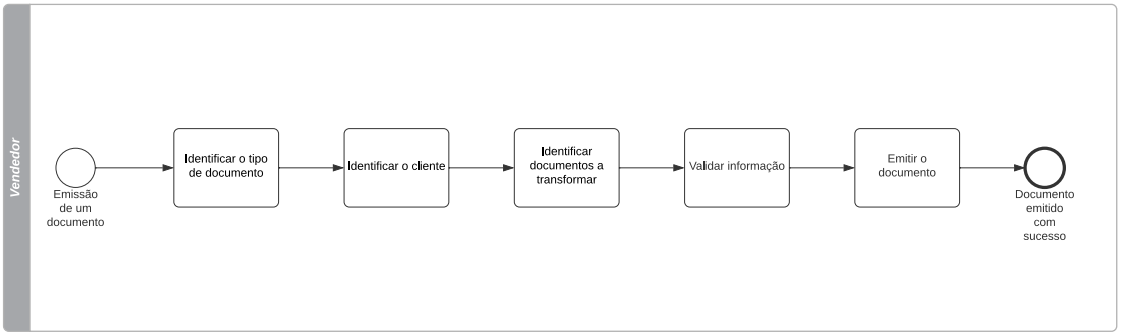


Figura 4- Diagrama BPMN do processo de emissão de um documento de venda

Após a identificação do processo de negócio, é crucial definir o grão. O grão representa o nível de granularidade ou detalhe dos dados que serão armazenados na tabela de factos. Estabelecer claramente o grão é fundamental para garantir a consistência e a precisão das informações ao

longo da modelação e implementação do DW. Neste contexto, o grão é a linha de venda. Isto significa que cada linha na tabela de factos representa a quantidade vendida de um produto a um cliente, por um vendedor numa determinada data.

Após a definição do grão, procede-se com a identificação das perspetivas de análise, também conhecidas como dimensões. As dimensões são essenciais para contextualizar os factos, fornecendo o caráter qualitativo da informação armazenada. Em geral, as dimensões respondem a questões fundamentais que são cruciais para compreender os eventos registados na tabela de fatos, incluindo "Quem?", "Quando?", "O quê?", "Onde?", "Como?" [48]. Foram identificadas as seguintes dimensões:

- **Artigo (“O quê?”)**: Através da dimensão Artigo podem ser analisadas informações relacionadas com a marca, modelo, categoria, características e outros atributos essenciais dos produtos.
- **Calendário (“Quando?”)**: A dimensão Data é fundamental para a análise temporal das vendas. Permite a segmentação das informações de vendas com base em períodos específicos, como mês, trimestre ou ano;
- **Cliente (“Quem?”**): A dimensão Cliente pode ser utilizada para analisar as informações relacionadas com os clientes;
- **País (“Onde?”)**: A dimensão País oferece informações relacionadas às regiões geográficas, permitindo análises com base em localizações geográficas o que inclui detalhes sobre diferentes países ou áreas geográficas onde as vendas são realizadas. A análise desta dimensão é crucial para identificar padrões de vendas em diferentes regiões, adaptar estratégias conforme as características de cada mercado e otimizar operações globais;
- **Região (Onde?):** Na dimensão região são apresentadas todas as regiões de Portugal. Através desta dimensão é possível analisar, por exemplo, a distribuição das vendas pelo território nacional**;**
- **Vendedor (“Quem?”**): A dimensão Vendedor pode ser utilizada para analisar as diferentes vendas de cada um dos vendedores;
- **Documento**: A dimensão documento representa os tipos de documentos existentes como por exemplo Fatura (FA) ou Nota de Crédito (NC).

Estas dimensões fornecem uma base sólida para a análise detalhada e para a tomada de decisões informadas no âmbito das operações de venda, contribuindo para o sucesso e a eficácia dos negócios. Cada dimensão oferece uma perspetiva única e importante para a compreensão abrangente do desempenho e melhoramento contínuo das estratégias de vendas.

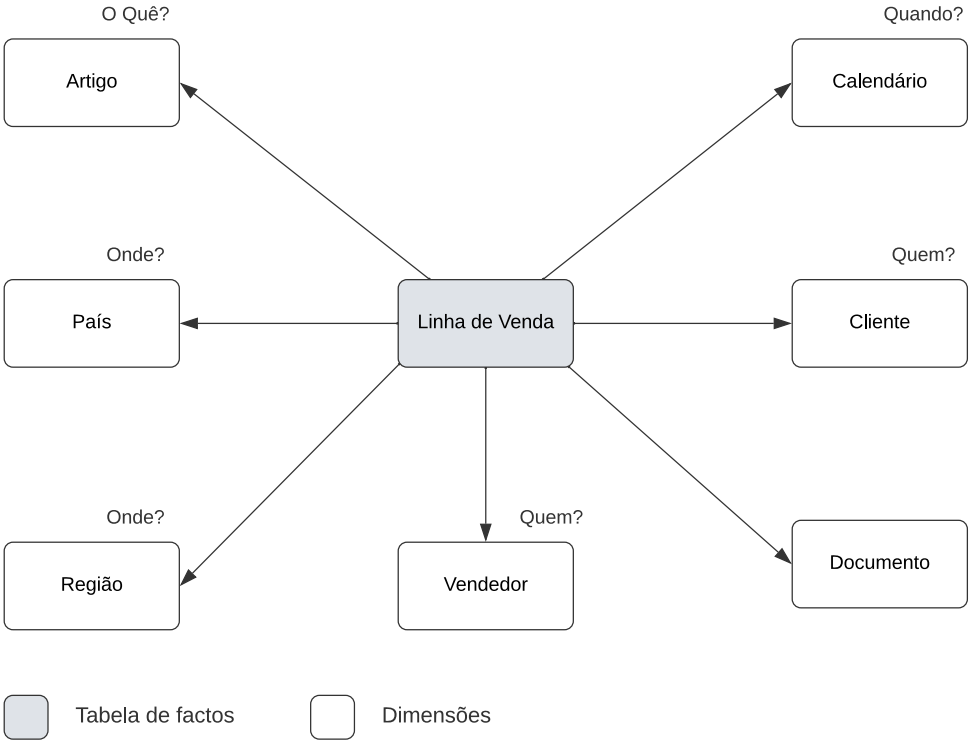


Figura 5- Dimensões do Data Mart de Vendas [49]

Por fim, no processo de desenvolvimento do Data Mart de Vendas, são identificados os factos, que representam as medidas de negócio essenciais para a análise e avaliação do desempenho da organização. É crucial que essas medidas estejam declaradas de acordo com o grão, alinhando-se com o processo de negócio que está a ser modelado.
Neste caso, foram identificadas as seguintes medidas ou factos:

- **Quantidade** (Aditivo): Representa a quantidade vendida do produto na venda;
- **ValorEur** (Aditivo): Refere-se ao valor em euros correspondente a cada linha de venda;
- **DescontoEur** (Aditivo): Representa o valor em euros dos descontos aplicados em cada linha de venda, proporcionando *insights* sobre a política de preços e desconto;
- **PreçoPadrão** (Aditivo): Indica o preço padrão de custo dos produtos envolvidos em cada linha de venda.

Estas medidas de negócio são importantes para a avaliação do desempenho de vendas, permitindo análises detalhadas e estratégicas sobre a quantidade de produtos vendidos, a faturação gerada, os descontos concedidos e os preços padrão de custo praticados. Estas medidas, são todas medidas elementares e são a base para a tomada de decisões informadas e estratégicas no âmbito das operações de vendas da organização, contribuindo para a eficácia das estratégias comerciais e o alcance de metas e objetivos estabelecidos. Ao estarem alinhadas

com o grão definido, essas medidas fornecem insights valiosos para impulsionar o sucesso e o crescimento do negócio.

Em resposta às questões de negócio relacionadas com o volume de vendas, tais como: Qual é a quantidade de unidades vendidas por produto? (Q4) E também, qual é a quantidade de unidades vendidas por produto e por região? (Q5) foi identificada a medida “Quantidade”. Por outro lado, a medida “ValorEur” foi identificada para responder a questões que envolvem o valor da faturação, como: Qual é o valor total de faturação por mês? (Q1). E, adicionalmente, qual é o valor de faturação mensal por produto? (Q2) Além disso, permite avaliar a variação do valor de faturação em relação ao ano anterior (Q3) e determinar o valor de faturação por cliente, por região e por produto (Q6). Também possibilita identificar os principais clientes com maior volume de faturação (Q10) e os cinco vendedores que alcançaram um maior valor de faturação (Q11).

Através das medidas "ValorEur" e "PreçoPadrão", é possível calcular a margem de lucro. O que permite responder a perguntas como: Qual é a margem de lucro por produto e por região? (Q8) E ainda, qual é a margem de lucro por marca? (Q9). Na tabela de factos foram selecionadas as medidas descritas anteriormente porque se alinham com os objetivos da tabela de factos, mas as restantes continuam a fazer parte do catálogo e podem ser criadas num esquema derivado/agregado (não é o caso) ou na própria ferramenta de BI

5.6 Matriz e arquitetura BUS do DW

Na Tabela 2, está representada a matriz BUS de alguns dos processos de negócio da FootFusion. A matriz BUS representa esquematicamente o DW e é utilizada para criar, documentar e apresentar a arquitetura do DW. Nesta matriz estão presentes as dimensões que fazem parte do *Data Mart* para a análise e compreensão abrangente das operações de vendas como também estão presentes as dimensões do processo de Gestão de Recursos Humanos e o processo de Compras.

Processos de Negócio/ Dimensões	**Artigo**	**Data**	**Cliente**	**Departamento**	**Fábrica**	**Fornecedor**	**Funcionário**	**País**	**Região**	**Secção**	**Vendedor**	**Documento**
Vendas	X	X	X					X	X		X	X
Gestão de Recursos Humanos		X		X	X		X			X		
Compras	X	X				X		X				

Tabela 2- Matriz BUS do DW

5.7 Desenho do Esquema Dimensional

O desenho do esquema dimensional é feito com base na aplicação do método dos 4 passos de Kimball [5]. O modelo lógico representado na Figura 6 apresenta um esquema em estrela que surge como resultado da aplicação do processo de modelação dimensional. Este modelo integra as várias dimensões apresentadas que fornecem contexto para as várias questões analíticas apresentadas. Estas dimensões são Calendário (que responde à questão "*When*"), Cliente e Vendedor ("*Who*"), Artigo ("*What*"), e por fim, Região Local e País ("*Where*").

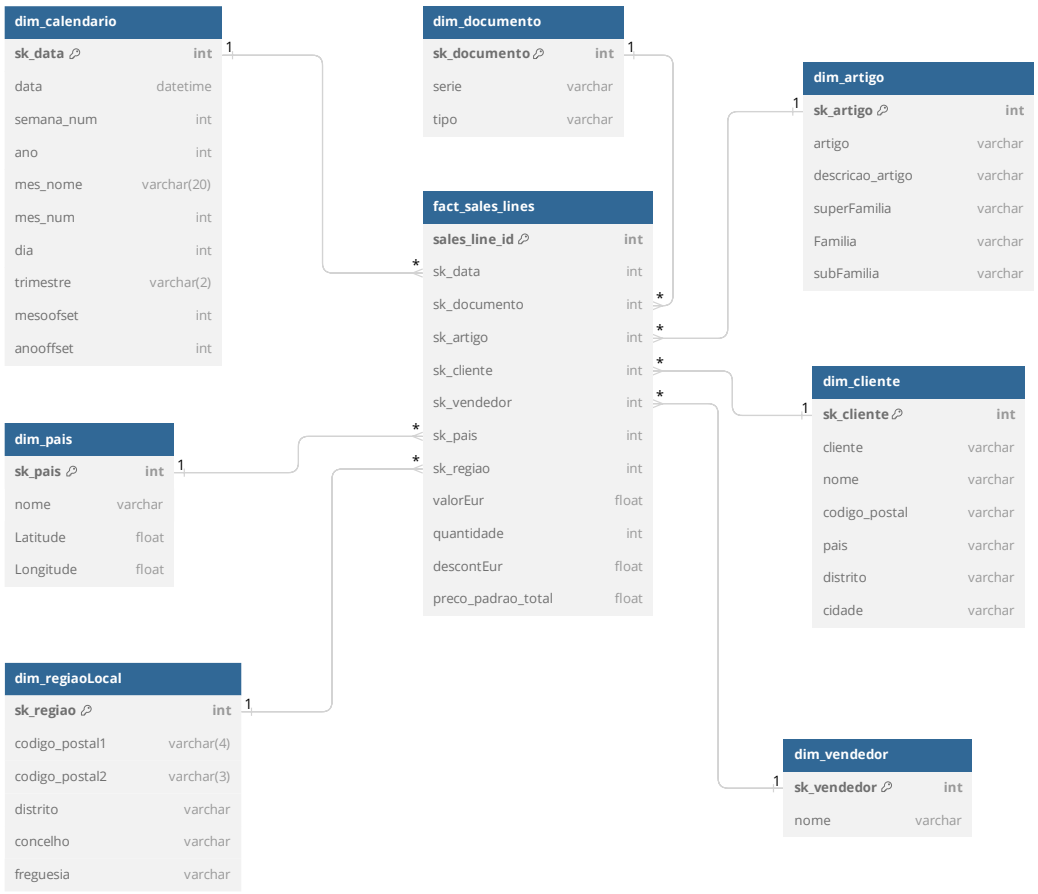


Figura 6 - Estrela de Vendas

5.8 Desenho da matriz de validação (Métricas vs *Queries*)

Na Tabela 3, é apresentada uma matriz que engloba tanto as dimensões como as medidas. O objetivo desta matriz passa por analisar todas as consultas realizadas no levantamento de requisitos e verificar se foram identificadas as dimensões e as medidas necessárias para responder a todas as consultas identificadas sendo que, as medidas que não fazem parte da tabela de factos têm de ser implementadas na ferramenta de BI. A matriz suporta a validação do modelo, garantindo que o modelo abrange os requisitos de negócio. Para a construção da matriz é necessário colocar nas colunas as questões de negócio identificadas no ponto 5.4 e nas linhas é necessário identificar as dimensões e as medidas que são necessárias paras responder a cada uma das questões.

	TIPO	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13
Dimensão Artigo		**X**	**X**		**X**	**X**	**X**		**X**	**X**				
Dimensão Calendário			**X**	**X**										
Dimensão Cliente							**X**				**X**		**X**	
Dimensão Documento													**X**	
Dimensao Pais						**X**	**X**		**X**					
Dimensao Regiao Local						**X**		**X**						**X**
Dimensão Vendedor												**X**	**X**	
Medidas Agregadas														
Volume de Vendas	**SUM**	**X**	**X**	**X**			**X**				**X**	**X**		
Total Quantidade	**SUM**				**X**	**X**								
Nº Clientes	**COUNT**							**X**						
Nº Transações	**COUNT**													**X**
Medidas Derivadas														
Margem									**X**	**X**				
Média de Vendas													**X**	
…														

Tabela 3- Matriz de validação

CAPÍTULO 6

Apresentação e Visualização de Dados

A visualização de dados desempenha um papel fundamental na análise e na compreensão de dados complexos. Normalmente, a visualização de dados é realizada através de gráficos representando várias perspetivas sobre os dados considerando o objetivo de cada um. A importância da visualização de dados reside no fato de que, em um mundo cada vez mais orientado por informações, a capacidade de traduzir números e estatísticas em representações visuais claras, significativas e interativas é essencial para tomar decisões informadas e comunicar eficazmente resultados.

As ferramentas de visualização de dados são cada vez mais um componente de uma ferramenta de BI, o que faz com que sejam construídas com um estilo de arquitetura específico ou considerando uma abordagem que unifica a integração com os restantes componentes, como é o caso do modelo de dados ou do processo de povoamento.

O *Power BI* (PBI) é um desses exemplos, disponibilizando um conjunto de elementos visuais e interativos bastante poderosos que tiram partido de um modelo de dados relacional, que privilegia a utilização de esquemas em estrela, possibilitando não só a otimização do desempenho, mas também da própria interatividade existente entre as diferentes visualizações no contexto do relatório [50]. As dimensões funcionam como base para os filtros para as consultas, podendo também ser utilizadas para a especificação de agrupamentos em hierarquia para a identificação de vários níveis de detalhe.

O PBI possibilita a criação de relatórios altamente personalizados adaptados aos diferentes casos de utilização e requisitos específicos. O *design* e o processo de desenvolvimento do relatório devem fluir naturalmente do processo de modelação de dados, visto que as medidas, relações e dimensões do modelo agora são utilizados para visualizar e analisar dados. A escolha dos tipos

de visualização desempenha um papel significativo na interpretação dos dados. O PBI oferece uma ampla gama de opções visuais como mapas, diagramas de árvore, gráficos de barras, de colunas, entre muitas outras opções.

Cada relatório é elaborado levando em consideração o domínio do negócio, tipicamente associados a cada *Data Mart* desenvolvido. Qualquer tipo de informação pode ser representado visualmente num relatório desde que seja relevante e acrescente valor ao negócio em questão. O desenvolvimento de relatórios envolve uma série de etapas que abrangem a extração de dados de uma ou mais fontes, a sua transformação de acordo com requisitos de visualização e do modelo de dados subjacente e ainda a configuração das visualizações mais adequadas para a representação de informação.

A elaboração de um relatório compreende um conjunto de etapas essenciais. Em primeiro lugar, é necessário conectar com a fonte de dados. Sendo que, existem inúmeras opções de acesso a fontes de dados. Neste caso a fonte de dados é o *Data Mart* que foi apresentado anteriormente e que está armazenado no *SQL Server*.

Uma vez estabelecida a conexão com a fonte de dados, é necessário realizar as devidas transformações nos dados para garantir a sua adequação. Estas transformações são realizadas através do *Power Query.* Neste caso, não foi necessário efetuar transformações nem limpar os dados porque o *Data Mart* foi construído através de um processo de Extração, Transformação e Carregamento (ETL). No entanto, é importante referir, que ao importar os dados é necessário validar sempre se o tipo de dados de cada atributo é o correto sendo que, é neste passo, que se alteram os tipos de dados caso seja necessário.

Após concluir o tratamento dos dados, é necessário proceder à criação do modelo de dados com os devidos relacionamentos. A primeira etapa consiste na identificação das dimensões criadas e da(s) tabela(s) de factos. No contexto atual e como referido anteriormente foram identificadas as seguintes dimensões: Cliente, Vendedor, Calendário, Pais, Região Local, Artigo. Depois de identificar as dimensões foi identificada a tabela de factos neste caso vendas, pois é a informação que vai ser analisada. Quando estão identificadas as dimensões e os fatos criam-se os relacionamentos entre as tabelas. É importante salientar que os relacionamentos devem ser criados com base num único campo em cada tabela (Figura 7).

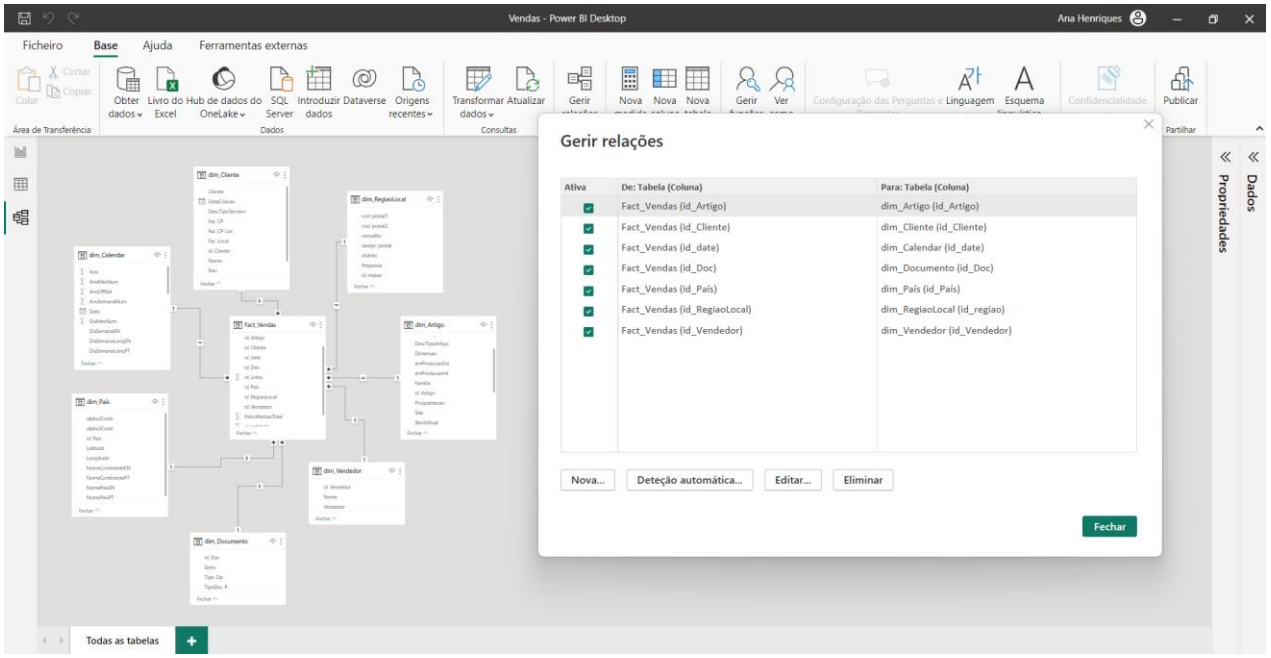


Figura 7- Esquema em estrela do conjunto de dados existente

Depois do modelo de dados estar criado este pode ser melhorado através da definição de Colunas Calculadas e Medidas, alinhadas com as questões de negócio que necessitam de ser respondidas e que foram identificadas anteriormente, utilizando a linguagem DAX (*Data Analysis eXpression Language*). O DAX é uma linguagem composta por uma variedade de funções, agregações, operadores e constantes que podem ser incorporados numa fórmula ou expressão permitindo o cálculo e o retorno de um ou mais valores [51]. Esta linguagem permite a criação de novas informações facilitando, por exemplo, o desenvolvimento de *KPI´s* que proporcionam uma perspetiva diferenciada sobre os dados analisados. A linguagem DAX possui uma vasta gama de funções que estão classificadas em várias categorias incluindo Data e tempo, Inteligência de tempo, Filtro, Informação, Lógica, Estatística, entre outras. É no momento da criação das medidas que é necessário ter em atenção o tipo de cada um dos factos (aditivo, semi-aditivo e não aditivo) de forma a não aplicar funções erradas garantindo a precisão das informações apresentadas aos utilizadores.

Para responder às questões de negócio identificadas no capítulo 5 é necessário criar várias medidas, como por exemplo:

- **Faturação** (Medida 1): Para o cálculo da faturação é necessário efetuar a agregação do facto ValorEur da tabela Fact_Vendas. A mesma medida pode ser utilizada em diferentes contextos tais como: análise do total das vendas por mês, por produto, por cliente, por região, entre outros e por isso a mesma medida será utilizada em diferentes visuais. Neste caso, a aplicação da função SUM é possível dada a natureza aditiva do facto. O resultado da medida será do tipo inteiro e por isso não terá casas decimais

$$Faturação = SUM(Fact\_sales\_line[Valor]) \quad (1)$$

- **Faturação LY** (Medida 2): Cálculo da Faturação do ano anterior ao ano em análise. Esta medida é uma medida derivada uma vez que é calculada com base na Medida 1.

$$Faturaçã oLY = CALCULATE([Faturação], SAMEPERIODLASTYEAR(dimCalendar[Data])) \quad (2)$$

- **Diferença do valor da Faturação** (Medida 3) De forma a analisar o crescimento do negócio é necessário calcular a diferença do Volume de Vendas. Esta medida também é uma medida derivada uma vez que é calculada através da Medida 1 e da Medida 2.

$$DiferençaValorFaturação = [Faturação] - [FaturaçãoLY] \quad (3)$$

- **% da diferença do valor de faturação** (Medida 4): Outra medida que é necessário criar é a percentagem do valor de faturação do ano em análise comparado com o valor da faturação do ano anterior sendo assim é necessário criar uma medida derivada (uma vez que é criada com base noutras duas medidas) e não aditiva em que o formato do resultado será uma percentagem e tem duas casas decimais.

$$\%DiferençaValorFaturação = DIVIDE([Faturação], [FaturaçãoLY]) \quad (4)$$

- **Custo** (Medida 5): Cálculo do valor total do custo. Esta é uma medida aditiva uma vez que pode ser aplicada uma função de agregação à mesma.

$$Custo = SUM(Fact\_sales\_line[PrecoPadraoTotal]) \quad (5)$$

- **Margem** (Medida 6): Medida derivada uma vez que é construída com base em duas medidas faturação e custo.

$$Margem = [Faturação] - [Custo] \quad (6)$$

- **% Margem** (Medida 7): Medida derivada uma vez que é construída com base em duas medidas margem e Faturação

$$\% Margem = DIVIDE\ ([Margem], [Faturação]) \quad (7)$$

- **Transações** (Medida 8): Cálculo do número de transações efetuadas

$$Transações = COUNTROWS[Fact\_sales\_line] \quad (8)$$

- **NúmeroClientes** (Medida 9): Cálculo do número de transações efetuadas

$$NúmeroClientes = DISTINCTCOUNT(Fact\_sales\_line\ [isk\_cliente] \quad (9)$$

Como foi possível perceber, a criação de um relatório é uma tarefa complexa que envolve a definição e utilização de várias medidas em diferentes visuais. Muitas vezes, estas medidas estão dependentes umas das outras o que pode criar alguns desafios na gestão e manutenção de todas as medidas. À medida que o negócio evolui, podem ser necessárias alterações no negócio que

podem implicar mudanças na composição das medidas. No entanto, identificar o impacto dessas alterações em todas as medidas e nas diversas visualizações e relatórios que as utilizam pode ser uma tarefa árdua uma vez que, uma medida pode ser utilizada não apenas por várias visualizações/relatórios, mas também podem ser utilizadas para criar outras medidas. Manter a correção das visualizações, principalmente em grandes relatórios, torna-se difícil porque é necessário identificar qual o impacto das alterações uma vez que uma simples mudança pode exigir várias adaptações e identificar em que medidas, visuais e relatórios a medida que tem de ser alterada está a ser utilizada. É um processo muito demorado, custoso e propício a erros o que compromete os resultados apresentados nas visualizações.

A medida Faturação LY (2) e a Medida Percentagem da Diferença do Valor da Faturação (4) são exemplos de medidas derivadas que foram criadas com base na medida Faturação (1). No entanto, uma mudança na Medida Faturação (1) pode ter um impacto significativo nas medidas derivadas da mesma e, consequentemente nas visualizações que a utilizam estas medidas. Se a medida Faturação (1) for alterada para apresentar apenas os dados do ano de 2023, isso terá um efeito direto nas medidas derivadas uma vez que em todas as medidas só serão utilizados dados de 2023. Por exemplo, a medida Faturação LY (2), que calcula a Faturação do ano anterior, não será capaz de utilizar os dados de 2022 para essa análise. O que leva a resultados imprecisos uma vez que os dados do ano anterior estão ausentes. É por estas razões que é extremamente necessário identificar qual o impacto que as alterações vão provocar uma vez que através da análise das dependências das medidas é possível concluir se a medida pretendida pode ser alterada ou se é necessário criar medidas novas para obter a informação correta.

Outra necessidade de rastrear e gerir medidas é no contexto de agregação dos dados. É crucial compreender a natureza das medidas, pois determina se estas podem ou não ser submetidas a funções de agregação. A utilização de funções de agregação de forma inadequada pode resultar em resultados imprecisos e incoerentes, comprometendo a qualidade das análises. A medida % margem (7) é um exemplo de uma medida não aditiva. Isso significa que a esta medida não pode ser aplicada uma função de agregação da mesma forma que estas funções são aplicadas em medidas aditivas, como por exemplo a soma de vendas ou o custo. Imaginando o seguinte cenário: a medida % margem (7) calcula o quociente entre o valor da margem com o valor da faturação. Agregar esta medida não faria sentido uma vez que as percentagens não são aditivas e por isso o resultado não representaria um valor coerente o que podia levar a conclusões erradas. Portanto, a importância de identificar medidas não aditivas está principalmente na preservação do significado correto dos dados.

Para organizar toda a informação das visualizações, dos relatórios, das medidas como: perceber em que visuais as medidas estão a ser utilizadas, de que tipo é que estas são, qual é a fórmula de cálculo de cada uma, quais são as suas dependências, quantas casas decimais é que o

resultado deve de ter é um processo complicado e demorado. Com isto, é necessário criar uma estratégia que facilite todo este processo e que seja de fácil consulta de forma que no momento em que seja necessário alterar uma medida ou perceber se uma medida pode ou não ser utilizada num visual específico seja fácil identificar em que visuais é que uma determinada medida está a ser utilizada, quais são as suas dependências, qual o tipo da medida, entre outros fatores e assim, analisar de forma rápida e simples quais as consequências de uma alteração, por exemplo, de forma a reduzir possíveis erros para não ser apresentada informação errada aos diferentes utilizadores. É por estas razões que será desenvolvido um grafo de conhecimento onde todas estas informações serão representadas de forma clara e precisa.

Por fim, depois de identificar e criar todas as medidas são então identificadas as visualizações mais adequadas para a construção do relatório que responda a todas as questões de negócio identificadas. O aspeto visual apelativo é um dos ingredientes para a criação de *dashboards*/relatórios. A quantidade de informação, a sua utilidade e relevância para os principais perfis de tomada de decisão, representam um aspeto crítico que determinam a aplicabilidade e a importância destes componentes de visualização no contexto dos processos de tomada de decisão empresarial. Além disso, as *dashboards*/relatórios necessitam de balancear todos estes elementos (qualidade visual, relevância, utilidade e quantidade de informação) e suportar requisitos operacionais, táticos e estratégicos [7]. É por estas razões que é fundamental compreender as necessidades das organizações ao nível do processo de tomada de decisão, as especificidades dos seus processos de negócio e proceder ao seu mapeamento com os dados, transformações e cálculos que são necessários para que seja possível atingir *dashboards*/relatórios de qualidade.

Na Figura 8, é apresentada uma das páginas elaboradas no contexto do relatório de vendas. Neste relatório é possível analisar as vendas em diversas perspetivas e, adicionalmente, é possível analisar os clientes, incluindo quem são os compradores mais e menos frequentes e que produtos é que é que cada um destes clientes compra. Na página do relatório que está presente na Figura 8 que foi devidamente estruturada, é possível analisar diversos indicadores de *performance* (*KPI's*) que incluem elementos cruciais como Faturação, Margem, Custos, Transações e Clientes. Através da análise destes *KPIs*, é possível obter uma visão abrangente do estado atual da organização. Para além disso, existe a possibilidade de comparar os valores de faturação ao longo dos diferentes meses com os números de faturação do ano anterior ao período em análise. O que possibilita uma avaliação sólida da evolução do desempenho ao longo do tempo. No que diz respeito à análise geográfica, é possível analisar a Faturação por país, o que ajuda a compreender como estão a ser as vendas nas diferentes regiões. Além disso, existem *insights* detalhados sobre o desempenho de cada vendedor, permitindo identificar os contribuintes mais produtivos. Para uma análise mais aprofundada, o utilizador tem a possibilidade de explorar

a hierarquia de produtos, incluindo informações sobre Marca, Família, SubFamília e Artigo. Isso possibilita uma compreensão mais granular dos produtos que estão a impulsionar o desempenho das vendas. Além de todas as funcionalidades enumeradas anteriormente, existe a flexibilidade de aplicar filtros para refinar ainda mais o relatório, permitindo que os utilizadores extraiam informações específicas de acordo com suas necessidades e objetivos de análise.

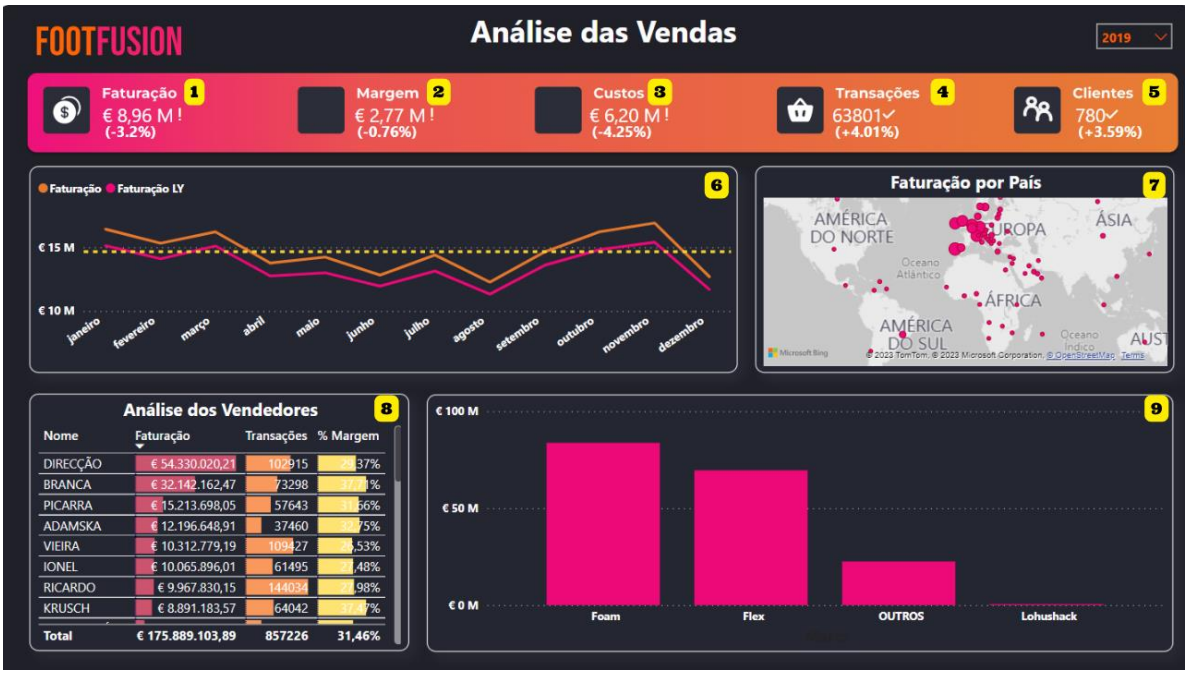


Figura 8- Relatório Análise de Vendas

Na Tabela 4, são apresentadas as medidas, tabelas e atributos envolvidos na criação de cada uma das visualizações (Na Figura 8, foi associado um número a cada uma das visualizações para ser mais simples identificar cada uma das visualizações criadas).

Visualização	Medidas	Dimensões	Atributos
1	Faturação Faturação LY	Dim_calendario	ano
2	Margem Margem LY	Dim_calendario	ano
3	Custos Custos LY	Dim_calendario	ano
4	Transações Transações LY	Dim_calendario	ano
5	NúmeroClientes NúmeroClientesLY	Dim_calendario	ano

6	Faturação Faturação LY	Dim_calendario	trimestre mes_nome dia
7	Faturação	Dim_pais	nome
8	Faturação Transações % Margem	Dim_vendedor	nome
9	Faturação	Dim_produto	superFamilia família subfamília artigo

Tabela 4- Identificação das visualizações, medidas, dimensões e atributos envolvidos na criação de cada visualização

CAPÍTULO 7

Catálogo de dados para *Business Intelligence*

O conceito de Catálogo de Dados (CD) desenvolvido nesta dissertação baseia-se nos fundamentos da modelação dimensional de forma a armazenar e representar os metadados de forma eficaz e coerente. A sua estrutura é desenhada de forma a potenciar o desenvolvimento de relatórios/*dashboards* através da representação das entidades identificadas, dos seus relacionamentos e propriedades.

Mais especificamente, é apresentado um componente para um CD orientado para a caracterização de métricas considerando os fundamentos teóricos da modelação dimensional especialmente orientados para a caracterização das medidas [5] [12]. O conceito de metadados ativos são um importante pilar desta abordagem, uma vez que a ideia passa por incorporar os metadados nas ferramentas existentes, de forma que possam ser utilizados para melhorar ou corrigir a análise de dados. É importante realçar que o catálogo de medidas representa um componente do CD e pode ser estendido para suportar outros aspetos relacionados com a gestão de dados no contexto das aplicações de BI, como controlo de acessos, rastreabilidade ou qualidade de dados.

A criação de relatórios/*dashboards*, é a fase final do ciclo de desenvolvimento de um esquema dimensional. Este ciclo inicia-se com a identificação dos requisitos e termina quando um conjunto de relatórios personalizados é disponibilizado para atender às necessidades dos decisores. Durante esta etapa, os requisitos, que foram previamente identificados, são traduzidos em ferramentas específicas, ocultando a complexidade do *design* e configuração da base de dados.

Um dos desafios encontrados passa por manter a documentação conceptual inicial sincronizada com a respetiva de implementação. Ou seja, a ideia passa por utilizar a caracterização do modelo

para validar a aplicação das medidas na componente de BI e permitir que as medidas/métricas desenvolvidas diretamente na ferramenta de BI possam ser integradas com a documentação realizada numa fase inicial do projeto.

De forma a testar a integração do CD foi utilizado o Power BI (PBI). A comunicação entre o catálogo de dados e o PBI é suportado por um ficheiro .pbit, extraído do PBI. O ficheiro .pbit representa o modelo do relatório, contendo a estrutura, *layout* e configurações do relatório. Através do processamento deste ficheiro, os dados armazenados no catálogo podem ser integrados no PBI, permitindo a criação de uma base para a construção dos relatórios. Da mesma forma, os dados podem ser extraídos do ficheiro pbit e cruzados com os dados existentes no CD, permitindo a validação da utilização das medidas e da sincronização da documentação presente no CD [52]. Por exemplo, com a informação representada no CD, é possível identificar as medidas agregadas ou derivadas, validar se uma função de agregação está a ser aplicada de forma correta ou identificar medidas que são necessárias para o desenvolvimento de uma visualização.

Os metadados gerados e criados para dar suporte ao CD são armazenados utilizando um modelo de dados em grafo. O CD pode ser dividido em duas camadas: Camada de Aplicação e a Camada do Modelo [52]. Na camada de aplicação pode ser representada a configuração do PBI descrevendo como é que os dados são definidos e relacionados, incluindo informações sobre as visualizações, relatórios e fórmulas DAX utilizadas. As medidas desenvolvidas devem ser representadas juntamente com as funções DAX e com os modelos relacionais utilizados para as calcular. A camada do modelo permite a identificação dos principais conceitos e dos princípios de modelação dimensional, como tabelas de facto, dimensões, medidas, entre outros.

7.1 Caracterização do Conhecimento

O Modelo de Dados é suportado por um grafo de propriedades que representa as instâncias criadas de acordo com os conceitos baseados nos fundamentos da modelação dimensional e que estão diretamente relacionados com o caso de aplicação. O grafo nesta camada é uma representação estruturada do conhecimento utilizado no processo de modelação dimensional para a caracterização de medidas e também dos principais conceitos relacionados com a sua utilização. Este grafo de conhecimento identifica as entidades de domínio que precisam de ser representadas assim como o seu relacionamento:

- **Factos**: Representa as tabelas de factos do DW;
- **Dimensões**: Representa as tabelas de dimensão do DW;
- **Atributos**: Engloba os atributos presentes nas tabelas de factos e dimensões do DW;
- **Medidas**: Representa as medidas existentes sejam elas parte integrante do DW ou

criadas na ferramenta de BI;

- **Funções**: Funções de agregação utilizadas para desenvolver medidas;
- **Relatório**: Identificação do relatório, neste caso em específico desenvolvidos em *Power BI*;
- **Páginas**: Identificação das páginas do relatório;
- **Visualizações**: Compreende as diferentes visualizações que compões um relatório.

Após a identificação das entidades, é necessário identificar os relacionamentos que podem existir entre cada uma das entidades e que na prática representam regras na construção deste modelo de alto nível:

- As dimensões são compostas por atributos;
- As tabelas de factos possuem relacionamento com as dimensões;
- As tabelas de factos possuem medidas;
- As tabelas de dimensão podem ter relacionamentos com uma ou mais tabelas de factos ou com outra dimensão;
- As medidas desenvolvidas na ferramenta de BI só estão relacionadas com as dimensões no caso de ser aplicado ou filtro ou uma função de agregação;
- As medidas podem ser agrupadas (*group by*) por uma ou mais dimensões;
- As medidas podem ser elementares, derivadas e de agregação sendo que as funções de agregação podem ser o SUM, COUNT, MAX, entre outras;
- Um relatório é composto por páginas que por sua vez são compostas por visualizações;
- As visualizações podem ser criadas através de medidas e de atributos das dimensões sendo que uma visualização tem de ter sempre um relacionamento com pelo menos uma medida;
- Uma visualização pode ser filtrada pelos atributos das dimensões

Na Figura 9, é apresentado um esquema que representa os relacionamentos que podem existir entre cada uma das entidades definidas.

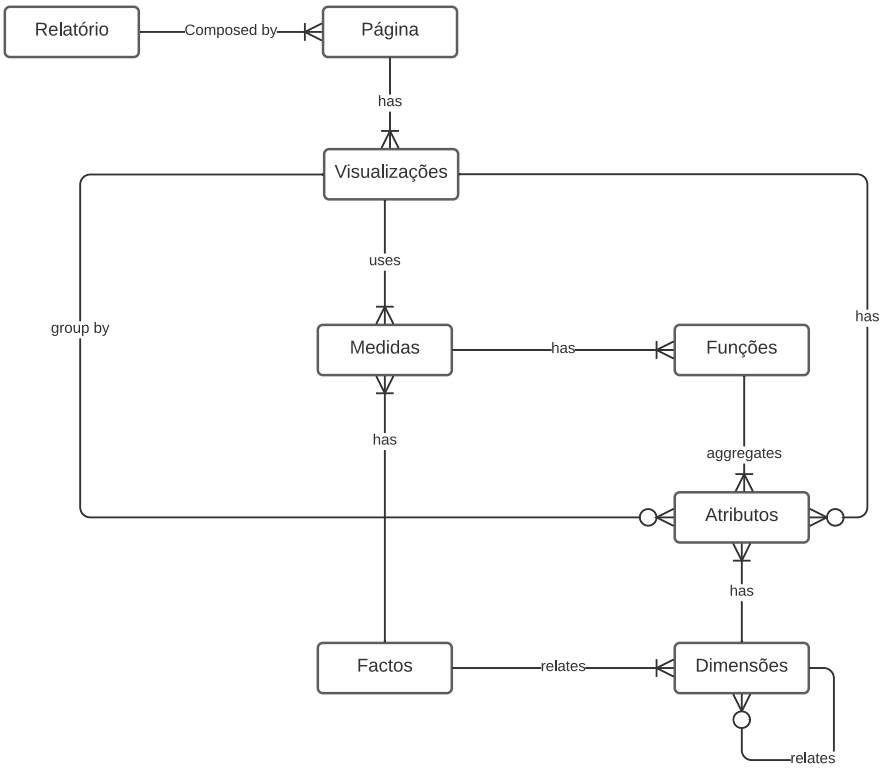


Figura 9- Relacionamentos entre as diferentes entidades

Depois de identificadas todas as regras é então necessário desenvolver o CD, de forma a apoiar a representação das medidas e o contexto relacionado. O CD desenvolvido descreve entidades, propriedades e relacionamentos interligados. Consiste no seguinte:

- **Nós**: Representam as entidades identificadas no contexto do domínio em questão. Para expressar claramente o propósito de cada nó, são atribuídas várias etiquetas. Os nós podem representar artefactos de dados, como tabelas de dimensão (etiquetadas como "*Dimension*") e tabelas de factos (etiquetadas como "*Fact*"). Adicionalmente, são representados os seus atributos ("*Attribute Label*") e em particular as medidas ("*Measure Label*"). As medidas podem ser anexadas à tabela de factos ou podem representar medidas criadas no conceito da ferramenta de BI. De forma a ser possível entender como é que as medidas são calculadas e utilizadas funções de agregação ("*Function Label*") e expressões ("*Expression Label*") também são também representadas utilizando nós específicos. Além disso, são representados conceitos relacionados à visualização de dados utilizados pelas ferramentas de BI. Foram identificados relatórios ("Relatório") e as suas visualizações ("Visualização"). As relações entre estes conceitos permitem identificar os relatórios e respetivas visualizações que utilizam medidas específicas (permitindo a rastreabilidade das medidas);
- **Relacionamentos:** Representam a associação existente entre as entidades/nodos: têm um tipo, uma direção e propriedades. O rótulo "Relates" é utilizado para associar nós

com a *label* "*Fact*" a nós com a *label* "Dimension", e o rótulo "*has*" fornece uma forma de descrever um relacionamento mais forte entre conceitos. Por exemplo, artefactos de dados (factos e dimensões) e as suas partes, ou seja, campos e, mais especificamente, medidas. A etiqueta "*has*" também é utilizada para associar visualizações a relatórios e funções ou expressões específicas a medidas ou visualizações para descrever cálculos específicos. A etiqueta "uses" descreve uma dependência mais fraca entre os nós e é útil para descrever como os nós dependem uns dos outros. Pode ser utilizado para descrever a dependência de uma determinada medida de outras medidas (para medidas derivadas), a dependência de visualizações de medidas específicas ou a dependência de algumas funções/expressões em campos ou medidas específicas. Também estão representados algumas etiquetas específicas para descrever medidas ou atributos que podem ser utilizados num cálculo específico: "*aggregates*" para descrever a agregação de dados, "*groupby*" para descrever restrições de agrupamento ou "*filter*" para descrever alguma seleção aplicada aos dados que será usado em um cálculo específico.

- **Propriedades**: Representam pares de chave-valor. São utilizadas nos nós para armazenar dados adicionais. Neste caso as propriedades dos nós são utilizadas para descrever o tipo de medidas ("elementar", "agregada" ou "derivada") e a categoria de cada uma das medidas ("aditiva", "não aditiva" ou "semi-aditiva") - as propriedades de cada um dos nós foram omitidas da Figura 10. As propriedades do nó são muito importantes pois é através destas que identificamos o tipo e a categoria de cada uma das medidas e assim é possível identificar, por exemplo, se uma medida não aditiva está a ser agregada, se uma medida depende de outra ou não, se a medida é elementar e por isso não pode ser analisada num grão inferior ao identificado, entre outros.

Para além das regras definidas anteriormente existem regras adicionais que se deve ter em consideração de acordo com as propriedades do nó:

- Nunca pode ser aplicada uma função de agregação a uma medida não aditiva;
- Uma medida semi-aditiva não pode ser agregada em todas as dimensões;
- Quando uma medida é derivada as partes que a compõe nunca podem ser eliminadas;
- Uma medida pode ter uma ou mais medidas derivadas

O grafo de propriedades representado na Figura 10 descreve um subconjunto dos conceitos relacionados ao caso de estudo descrito. O subconjunto dos conceitos apresentado concentra-

se nas medidas e em como estas podem ser representadas no modelo de dados do grafo. Existem medidas criadas no contexto de tabelas de factos (a relação "*has*" com os nós "*Fact* " permite a sua identificação) e medidas criadas no contexto de uma ferramenta específica de BI, neste caso, o Power BI. A aditividade das medidas é expressa através das propriedades do nó, permitindo identificar medidas a que não podem ser aplicadas a função SUM. Por exemplo, numa operação de agregação. Este fator é particularmente útil se algum dos relatórios utilizar estas medidas para agregar os dados, permitindo a identificação de erros que podem comprometer a perceção dos dados apresentados aos utilizadores. O grafo também pode representar as limitações na agregação através do relacionamento "não agregável", proporcionando outro aspeto de validação às medidas utilizadas nos relatórios.

O tipo de cada medida é classificado como "elementar" se estiver no nível de maior detalhe existente estando representado não só nas tabelas de factos como em medidas derivadas que possuem um filtro sem qualquer agregação, "agregado" se representar uma agregação de factos e "derivado" se for calculado a partir de outras medidas. Essas definições são úteis para identificar dependências entre medidas caso exista alguma alteração numa das medidas ou se pretenda eliminar uma medida que faça parte de uma medida derivada. Se uma das partes que compõe a medida derivada for eliminada essa medida deixa de fazer sentido.

Todos os fatores apresentados representam um problema real para os desenvolvedores do *Power BI*, pois uma medida pode ser utilizada em diversas visualizações e pode ser a origem de dezenas de outras medidas. Em cenários complexos pode ser difícil gerir todas estas dependências e prever o impacto de uma simples alteração numa fórmula de uma medida. A dependência também é preservada usando funções e expressões, descrevendo medidas e manipulando cálculos, revelando dependências importantes não apenas entre medidas, mas também entre atributos dimensionais (por exemplo, envolvendo filtros). Estes cálculos podem ser expressos entre medidas e visualizações. Para além da rastreabilidade das medidas, é possível validar a utilização das medidas quando utilizadas no contexto de um relatório, permitindo por exemplo compreender se as medidas estão a ser utilizadas de acordo com a sua definição original.

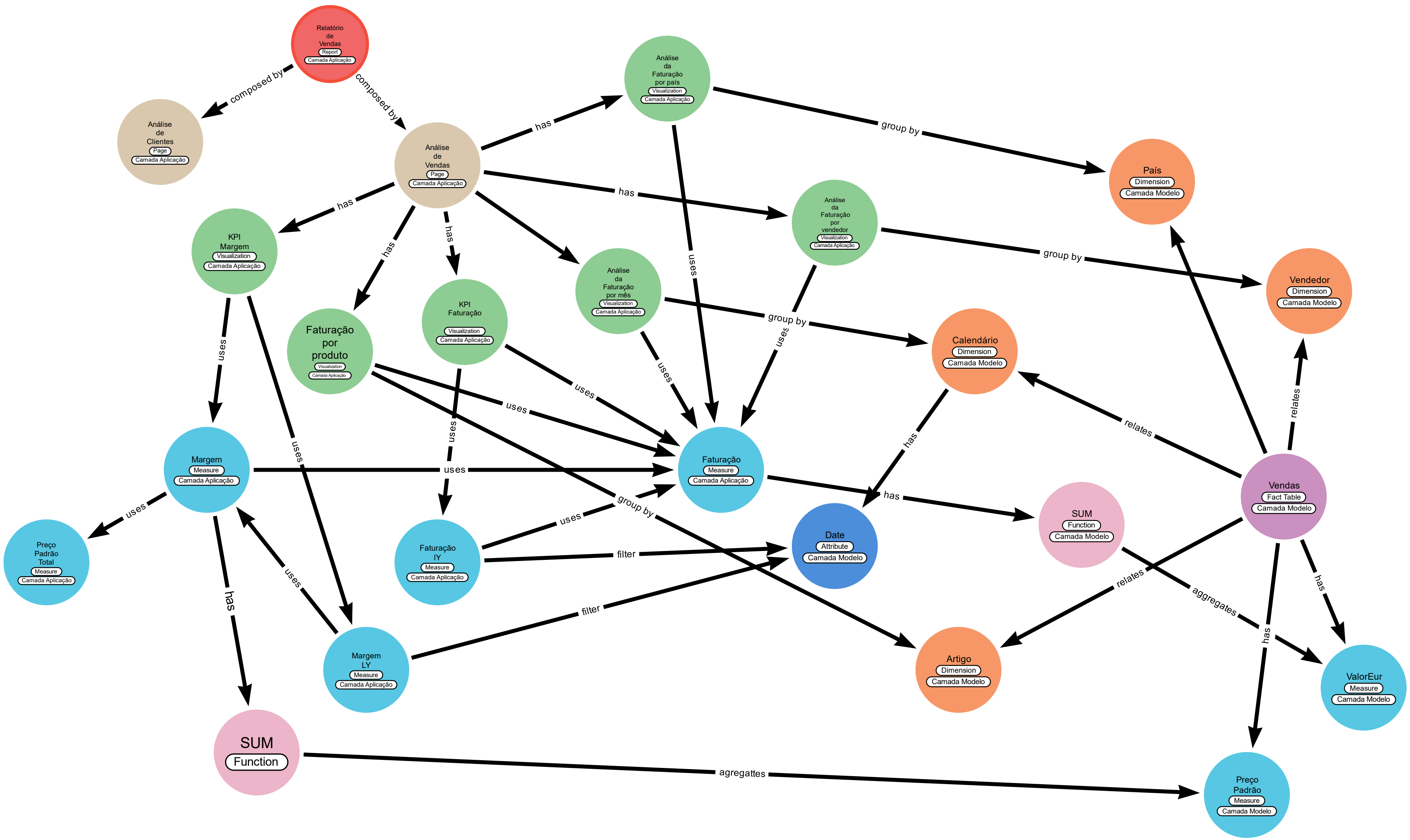


Figura 10- Excerto do grafo

7.2 Aplicação Prática

Como mencionado anteriormente, um relatório criado no PBI pode ter visualizações, filtros e medidas específicas sobre os dados. As medidas, podem ser desenvolvidas com base em cálculos que são necessários para suportarem a análise. As medidas podem ser utilizadas para resumir, agregar ou calcular valores específicos, como foi explicado anteriormente. Todos os artefactos desenvolvidos num projeto do PBI são armazenados dentro de um ficheiro pbit, que é utilizado como um arquivo de modelo que contém o modelo de dados e as consultas do relatório PBI. As medidas e o esquema do modelo podem ser extraídos, analisados e vinculados como no modelo de grafo apresentado na Figura 10.

Para suportar a implementação das camadas, foi implementada uma base de dados em *Neo4j* e foram desenvolvidas em *Cypher* (linguagem utilizada no *Neo4j*) um conjunto de regras de validação e monitorização para o caso de estudo apresentado. Estas regras, visam identificar agregações incorretas, seja identificando medidas não agregáveis, em algumas dimensões, ou agregações erradas (como a utilização da função SUM em medidas não aditivas). Por exemplo, a medida % *da diferença do valor de faturação* (Medida 4) não pode ser utilizada com o operador SUM, pois não faz sentido adicionar valores percentuais. Além disso, o grafo, permite a monitorização, que é na verdade um problema para os utilizadores do PBI. Assim, através do grafo, é possível compreender o impacto quando há alterações em alguma das medidas, como por exemplo em medidas relacionadas ou nas visualizações e relatórios associados.

Uma questão relevante seria, por exemplo, identificar as medidas que utilizam (derivam) da medida Faturação. Esta medida agrega o facto ValorEur da tabela *Fact_sales_line* como é descrito no capítulo 6.

$$Faturação = SUM(Fact\_sales\_line[Valor]) \quad (1)$$

Para isso, foi criada uma expressão Cypher que vai buscar as medidas (*label "measure*") que se relacionam com a medida Faturação (nó que tem a etiqueta "*measure*" e a propriedade nome é "Faturação") em vários "níveis" e não só as que estão diretamente relacionadas com esta medida.

```
MATCH(m:Measure{nome:'Faturação' }) <-[:uses*]-(o:Measure) return m,o
```

Na Figura 11, são identificadas as medidas que estão relacionadas com a medida Faturação, através do relacionamento denotado como "*uses*", o que significa que as medidas apresentadas estão dependentes da Medida Faturação para serem calculadas. Neste caso as medidas que estão diretamente dependentes da medida Faturação são a Margem, Faturação LY, Diferença do Valor da Faturação e a % da diferença do valor da Faturação. Na imagem também é apresentado o nó Margem LY uma vez que esta medida é dependente da Medida Margem para ser calculada.

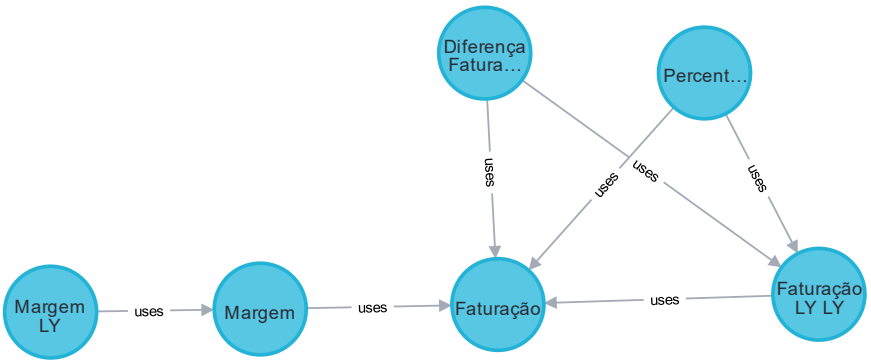


Figura 11- Medidas que estão conectadas à medida Faturação

De seguida são apresentadas algumas expressões Cypher também desenvolvidas para validar as medidas:

- *MATCH(m:Measure{category:'non-additive'}) return m* (1)
- *MATCH(m) <-[:aggregates*]-(f:function) return m, f* (2)
- *MATCH(m:Measure{cattegory:'non-additive'}) <-[:aggregates*]-(f:function) return m, f* (3)

Na expressão 1, são identificadas as medidas a que a função de agregação SUM não pode ser aplicada. Para isso é necessário identificar as medidas do tipo não aditivo uma vez que a estas medidas não podem ser aplicadas funções de agregação. Na expressão criada é necessário identificar os nós com a *label "Measure"* e que têm a propriedade *"category"* definida como *"non-additive".* Através do grafo também é possível identificar que funções de agregação são aplicadas a cada uma das medidas para isso é criada uma expressão (2) onde são procurados os nós com o relacionamento *"aggregates"* com nós do tipo "*function"*. Também é possível identificar se está a ser aplicada uma função de agregação a uma medida não aditiva (3). Neste caso, é necessário identificar as medidas que possuem a "*category*" definida como "*non-additive*" o que indica que são medidas não aditivas. A consulta procura então por relacionamentos de agregação ("*aggregate*") que ligam essas medidas às funções de agregação.

De forma a monitorizar as medidas, é possível criar uma expressão *Cypher* que identifique as visualizações que utilizam uma medida específica, neste caso a medida "Faturação" de forma a analisar o impacto que uma alteração numa medida causa. Para isso a expressão deve identificar o nó com a *label "Measure"* e que a propriedade nome seja "Faturação". Depois são identificados os relacionamentos do tipo "uses" que vão do nó "*Measure*" para os nós "*Visualization*"

```
MATCH(m:Measure{nome:"Faturação"}) <-[:uses]-(v:Visualization) return m,v
```

Na Figura 12, é possível observar todas as visualizações que necessitam da medida faturação para serem criadas. Neste caso, as visualizações que necessitam desta medida para serem criadas são: KPI Faturação (1), Análise da Faturação por mês (6), Análise da Faturação por país (7), Análise dos Vendedores (8) e Análise da faturação por produto (9). Todas estas visualizações

foram identificadas no capítulo 6.

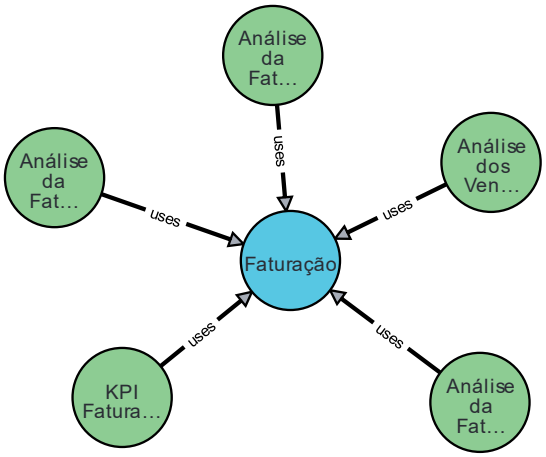


Figura 12- Grafo com as visualizações que utilizam a medida Volume de Vendas

Se fosse necessária a identificação das visualizações que fazem parte de uma determinada página e que medidas é que são necessárias para criar essas visualizações poder-se-ia usar uma expressão semelhante a:

```
MATCH(m: Measure) <-[:uses*]-(v:Visualization) <-[:has*] -
(p:Pagina{nome:'Análise Vendas'}) return m,v,p
```

Através da expressão mencionada anteriormente são identificadas as visualizações que compõe a página “Análise Vendas” do relatório que foi apresentado no capítulo 6. Este relatório é composto por duas páginas sendo uma delas a página “Análise Vendas”. É possível identificar as visualizações através do relacionamento denotado como “*has*” e são também identificadas as medidas que são necessárias para criar as visualizações da página através do relacionamento denotado como “*uses*”. Na Figura 13, é apresentado o nó “Análise Vendas” que representa a página do relatório. Através do relacionamento “has” são apresentados os nós que representam as visualizações que constituem esta página. Por fim, são também apresentadas as medidas que são necessárias para criar cada uma das visualizações presentes na página “Análise Vendas”.

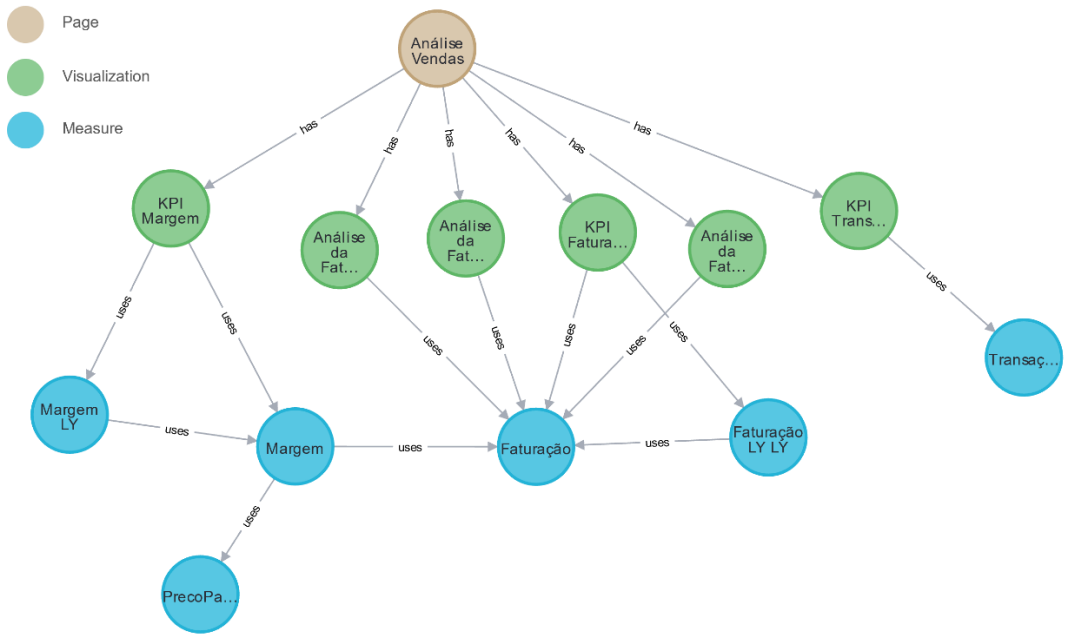


Figura 13- Visualizações e Medidas que compõe a página Análise Vendas

Todas as consultas apresentadas anteriormente podem ser encapsuladas numa API que pode ser disponibilizada para permitir a integração com ferramentas de BI, como é o caso do PBI. A API foi desenvolvida em *Python*, com o auxílio da biblioteca *Neo4j* e *Flask*[3]. A biblioteca Neo4j é utilizada para a conexão e manipulação dos dados na base de dados Neo4j. O uso do *Flask*, *framework* de desenvolvimento *web* para *Python*, facilita a criação de *endpoints*, tornando o acesso às consultas seguro e de fácil utilização. Esta *framework* foi utilizada para fins demonstrativos devido à sua simplicidade de utilização.

Para fins de teste foram criados 5 endpoints:

- **getall**: Retorna todos os nós do grafo. Fornece todas as entidades que estão representadas no grafo e repetivas propriedades e relacionamentos (Endpoint de validação);
- **getnonadditivemeasures**: Endpoint que identifica todas as medidas não aditivas (Endpoint de validação);
- **getderivedmeasures**: Retorna as medidas derivadas de determinada medida isto é, através deste endpoint é possível identificar as medidas que estão dependentes de determinada medida para serem calculadas (Endpoint de validação);
- **getmeasurevisualizations**: Este endpoint tem como função retornar as visualizações que necessitam de determinada medida para serem criadas (Endpoint de monitorização);
- **getpagemeasuresandvisualizations:** Identifica todas as medidas e visualizações que são utilizadas numa página (Endpoint de monitorização).

[3] https://flask.palletsprojects.com/

Para incorporar os dados do PBI no grafo é necessário extrair as informações do ficheiro .pbit. O formato de arquivo .pbit está associado ao PBI. Este ficheiro, contém informações e recursos necessários para criar um relatório no Power BI. Os principais elementos deste ficheiro são (Figura 14):

- **Conexões de dados**: O arquivo .pbit contém informações sobre as fontes de dados às quais o relatório está conectado. Isso inclui detalhes sobre as fontes de dados, como tipo de conexão, localização da fonte de dados, credenciais, configurações de atualização automática e qualquer informação necessária para ser possível a conexão às fontes de dados.
- **Esquema do modelo de dados**: O arquivo .pbit armazena a estrutura do modelo de dados subjacente ao relatório, incluindo tabelas, colunas, relacionamentos entre tabelas e medidas.
- **Visualizações e gráficos**: As visualizações criadas são armazenadas no arquivo .pbit. Incluindo informações sobre a formatação, configuração e layouts das visualizações.
- **Metadados:** Os metadados contêm informações sobre as tabelas, colunas e medidas, como descrições, formatação condicional e configurações específicas de visualização.
- **Medidas e cálculos personalizados**: Fórmulas DAX usadas para criar medidas personalizadas e colunas calculadas também são parte do arquivo .pbit.
- **Regras de segurança**: O ficheiro inclui configurações de segurança relacionadas com o modelo de dados, como definições de funções de segurança, filtros de linha e outras restrições de acesso aos dados.

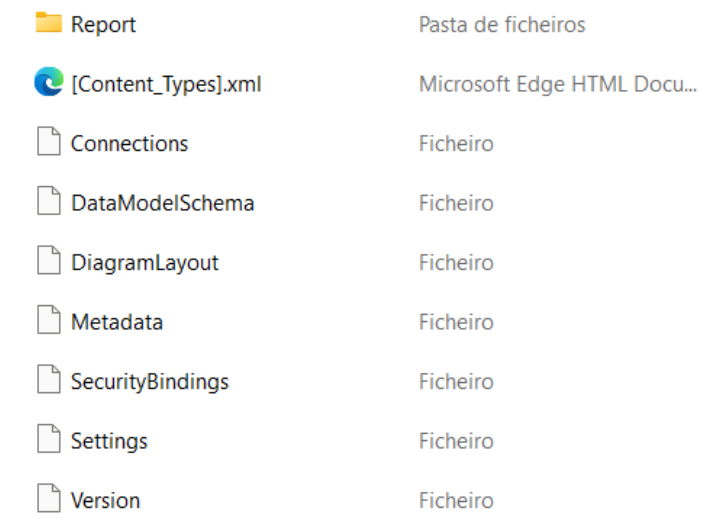


Figura 14- Estrutura do ficheiro .pbit

Analisando ao pormenor o ficheiro DataModelSchema, ficheiro *json* (JavaScript Object Notation) onde é armazenada toda a informação referente ao modelo de dados. Este ficheiro compreende

dados como as tabelas, atributos, tipos de dados, relacionamentos existentes entre cada uma das tabelas, medidas desenvolvidas no PBI e respetiva fórmula de cálculo bem como o tipo de dados associado a cada uma das colunas e medidas. O ficheiro inicia a sua estrutura apresentando todas as tabelas presentes no relatório especificando as colunas e colunas calculadas, o tipo de dados de cada uma das colunas e são também identificadas as hierarquias. Além disso, são também armazenadas as consultas necessárias para a criação de cada uma das tabelas e também são apresentadas as medidas de cada uma das tabelas bem como a respetiva fórmula de cálculo. Na Figura 15, está presente um excerto deste ficheiro onde é apresentada a tabela dim_artigo bem como um dos atributos desta tabela nomeadamente o id_artigo. Na Figura 15, é também representado o relacionamento entre a tabela fact_sales_lines e a tabela dim_artigo bem como o relacionamento entre a tabela fact_sales_lines e dim_cliente.

Todos os componentes descritos anteriormente compõem a estrutura de um arquivo .pbit, e cada um desempenha um papel fundamental na representação dos dados, do modelo e das configurações do relatório. Este ficheiro contém as definições e o estado do relatório, mas não contém os dados em si. Para possibilitar a integração dos dados do PBI, foi utilizado um script capaz de ler o ficheiro .pbit, permitindo assim, a criação das correspondências entre as informações extraídas do ficheiro .pbit e os conceitos do grafo. Este processo possibilita uma integração eficiente e eficaz dos dados e das configurações do relatório do PBI no contexto do grafo desenvolvido, sendo assim possível validar e monitorizar a construção de relatórios/*dashboards*.

```
{
  ...
    "tables": [
      {
        "name": "dim_Artigo",
        "lineageTag": "5991ef55-f225-40c2-a932-b3936b460cba",
        "columns": [
          {
            "name": "id_Artigo",
            "dataType": "int64",
            "sourceColumn": "id_Artigo",
            "formatString": "0",
            "lineageTag": "f4d8ff55-7998-4664-98da-00619e5ba8e1",
            "summarizeBy": "none",
            "annotations": [
              {
                "name": "SummarizationSetBy",
                "value": "Automatic"
              }
            ]
          },
           ...
    ]

"relationships": [
      {
        "name": "6bc98592-a503-49dd-b890-4a7e3af4467d",
        "fromTable": "Fact_sales_lines",
        "fromColumn": "id_Artigo",
        "toTable": "dim_Artigo",
        "toColumn": "id_Artigo"
      },
      {
        "name": "76ea0f48-fbf2-4aa9-909e-3a8c6db8b83b",
        "fromTable": "Fact_sales_lines",
        "fromColumn": "id_Cliente",
        "toTable": "dim_Cliente",
        "toColumn": "id_Cliente"
      },


       ...
]
```


Figura 15- Excerto do ficheiro DataModelSchema

CAPÍTULO 8

Discussão de resultados

A evolução dos sistemas analíticos tem impulsionado a crescente utilização de Catálogos de Dados (CD). É possível comprovar este facto através da análise da literatura na qual se observa um aumento no número de artigos e publicações dedicadas ao tema. O interesse neste tema reflete a importância da utilização dos CD para compreender os dados dentro das organizações. Através da revisão da literatura é evidente que existem diversas abordagens para o desenvolvimento de CD. Entre elas está a utilização de taxonomias, ontologias e grafos de conhecimento que têm emergido como estratégias promissoras para potenciar a organização e a representação semântica dos dados. A diversidade de abordagens evidencia a adaptabilidade na construção de CD, permitindo que estes sejam ajustados de acordo com as características e necessidades específicas de cada contexto organizacional. Neste sentido, é necessário considerar não apenas a eficácia técnica das abordagens, mas também a sua viabilidade.

Foram identificadas várias ferramentas comerciais de CD tais como *Google Cloud Data Catalog*, *Azure Data Catalog* ou *Atlan Data Catalog*. Contudo, um dos principais desafios enfrentados passa pelo facto de que a maioria destas ferramentas são pagas o que implica dificuldades na avaliação do cumprimento dos requisitos identificadas por parte destas ferramentas. Além disso, outra consideração importante é que muitas destas ferramentas são orientadas a uma plataforma específica e por isso apresentam várias limitações.

Grande parte das ferramentas existentes concentra-se predominantemente na captura de metadados associados ao conjunto de dados como origem, estrutura, significado e relacionamentos dos dados. Sendo assim, é notável a limitação destas ferramentas em representar os metadados de acordo com os fundamentos da modelação dimensional. Esta falta

de capacidade restringe a utilidade da utilização dos CD em ambientes em que a modelação dimensional é fundamental como na construção de relatórios/*dashboards* em sistemas de BI.

A representação inadequada dos metadados em conformidade com os fundamentos da modelação dimensional pode impactar negativamente a eficiência e a eficácia na criação de relatórios e análises, limitando a compreensão dos dados e a capacidade de extrair *insights* valiosos. Portanto, é importante que as ferramentas de CD evoluam para incorporar essa capacidade de representação de metadados de acordo com os fundamentos da modelação dimensional, a fim de atender adequadamente às necessidades dos utilizadores e garantir uma visão abrangente e precisa dos dados.

Após a identificação dos requisitos essenciais para a implementação adequada do CD, identificação das entidades que necessitam de ser representadas, relacionamentos que podem existir entre cada uma das entidades, e a definição das propriedades necessárias para cada nó foi possível desenvolver um grafo de conhecimento que representa alguns dos conceitos da modelação dimensional. O grafo é a representação estruturada das entidades, relacionamentos e propriedades necessárias para garantir a eficácia da aplicação dos princípios da modelação dimensional.

O CD desenvolvido está orientado para a representação de medidas, representando um subconjunto específico dos conceitos da modelação dimensional. A construção do grafo, que suporta o CD, foi influenciada pela experiência em Power BI, o que estabelece certas limitações na aplicabilidade do caso de estudo uma vez que não foi possível estender a análise para outros modelos de ferramentas. Dessa forma, a análise e desenvolvimento do grafo foram direcionados principalmente para atender às necessidades específicas do PBI. Neste momento, é necessário extrair os dados do arquivo .pbit e alimentá-lo na API de forma manual, mas a API foi desenvolvida com a perspetiva de uma futura integração automatizada.

.

CAPÍTULO 9

Conclusão e Trabalho Futuro

Nos últimos anos, as organizações têm-se deparado com um aumento significativo na quantidade de dados que necessitam de ser processados e analisados para apoiar os seus processos de tomada de decisão. Além dos dados organizacionais, normalmente suportados por uma arquitetura de dados centralizada, há uma necessidade crescente de integrar dados externos que possam complementar e contextualizar a realidade organizacional. Com isto, o número de fontes de dados externos cresceu exponencialmente. Sendo assim, a informação provém de diversas origens como sensores, redes sociais entre outras. Este diversificado conjunto de fontes gera um desafio adicional: a necessidade de armazenar e gerir eficazmente estes dados.

A catalogação dos dados e as camadas semânticas desempenham um papel crucial na facilitação da compreensão e controlo dos dados. Estas estratégias fornecem contexto e significado aos dados, possibilitando a sua correta exploração pelos utilizadores. Ao fornecer uma estrutura que organiza e descreve os dados de forma compreensível a procura pelo entendimento dos conjuntos de dados, dados disponíveis é facilitada melhorando a eficiência e a eficácia. Esta contextualização dos dados não ajuda apenas os utilizadores a interpretarem os dados de forma adequada, mas também oferece controlo sobre o uso e a interpretação dos dados garantindo precisão e consistência nas análises.

No contexto específico de criação de relatórios/*dashboards,* a aplicação destas práticas revela-se particularmente benéfica. A catalogação dos dados e contribui para tornar o processo de *design* e implementação de relatórios/*dashboards* mais simples e mais controlado. Para além disso, estas práticas também simplificam o processo de manutenção e controlo de todo o

processo de visualização e exploração dos dados.

Foi apresentado um CD que aborda um subconjunto de componentes de um sistema analítico, nomeadamente na gestão e aplicação de medidas num contexto de relatórios/*dashboards* para BI. Este subcomponente oferece uma estrutura que ilustra como as medidas podem ser documentadas e conectadas ao contexto subjacente como tabelas de facto/dimensões e relatórios/visualizações preservando propriedades importantes que restringem o seu uso para fornecer *insights* comerciais. Assim, foi implementado um grafo de propriedades que atua como uma base de dados de metadados. Este grafo de propriedades assume um papel fundamental, uma vez que apoia o processo de desenvolvimento de relatórios/*dashboards* ajudando a garantir a correção dos dados e permitindo o controlo e a rastreabilidade dos mesmos.

Os metadados gerados e criados para dar suporte ao CD são armazenados utilizando um modelo de dados em grafo. O CD pode ser dividido em duas camadas: Camada de Aplicação e a Camada do Modelo. Na camada de aplicação pode ser representada a configuração do Power BI, ferramenta utilizada neste caso de estudo. Nesta camada, é descrito como é que os dados são definidos e relacionados, incluindo informações sobre as visualizações, relatórios e fórmulas DAX utilizadas. As medidas desenvolvidas devem ser representadas juntamente com as funções DAX e com os modelos relacionais utilizados para as calcular. A camada do modelo permite a identificação dos principais conceitos dos princípios de modelação dimensional, como tabelas de facto, dimensões, medidas, entre outros. O CD de dados desenvolvido foi influenciado pela utilização do Power BI e possui funcionalidades limitadas uma vez que se concentra principalmente na organização e representação das métricas sendo que no futuro pode ser explorada a aplicabilidade do catálogo noutros modelos de ferramenta.

A expansão das áreas de cobertura para metadados representa um importante caminho a ser explorado como trabalho futuro. Ao aumentar a abrangência dessas áreas, é possível proporcionar uma base sólida na procura e descoberta de informações essenciais. Isso traduz-se numa maior eficiência na análise de dados, uma vez que os profissionais podem encontrar e compreender mais facilmente os dados. Ao expandir as áreas de cobertura é possível, por exemplo, fornecer controlo de acesso e regras de privacidade para os dados. Este fator é fundamental especialmente num cenário em que a proteção de a segurança dos dados são cada vez mais importantes. Assim, é possível implementar políticas de acesso garantido que apenas as partes autorizadas têm acesso aos dados.

Além disso, também é possível contribuir para o rastreamento de dados útil para pipelines ETL. A rastreabilidade dos dados é fundamental para garantir a qualidade dos dados permitindo que as organizações identifiquem problemas nas *pipelines* e tomem medidas corretivas de forma eficaz aumentando assim a confiabilidade e a consistência dos dados. Todos os aspetos

referidos anteriormente podem estar integrados em grafos de conhecimento que podem ser expandidos com taxonomias e ontologias trazendo novas capacidades relacionadas com a representação e exploração de dados. Um dos exemplos pode ser a utilização de mecanismos de inferência.

Esta página foi propositadamente deixada em branco.

Bibliografia

[1] F. V. PRIMAK, *Decisões com bi (business intelligence)*. 2008.

[2] R. Sherman, *Business Intelligence Guidebook: From Data Integration to Analytics (1st ed.). Morgan Kaufmann Publishers Inc.* 2014.

[3] P. Alpar and M. Schulz, “Self-Service Business Intelligence,” *Business and Information Systems Engineering*, vol. 58, no. 2, pp. 151–155, 2016, doi: 10.1007/s12599-016-0424-6.

[4] Leão da Silva and Airton Pereira, “POWER BI PARA TOMADA DE DECISÕES ESTRATÉGICAS: ANÁLISE DE INDICADORES-CHAVE DE DESEMPENHO (KPIS).,” *REVISTA FOCO*, 2022.

[5] R. Kimball and M. Ross, *The Data Warehouse Toolkit, The Definitive Guide to Dimensional Modeling* . 2013.

[6] D. Wells, *An Introduction to Data Catalogs: The Future of Data Management.* 2019.

[7] X. Zhang, K. Gallagher, and K. Gallagher, “BI APPLICATION: DASHBOARDS FOR HEALTHCARE,” in *AMCIS 2011 Proceedings*, 2011.

[8] A. I. to D. C. T. F. of D. Management, “An Introduction to Data Catalogs: The Future of Data Management,” 2019.

[9] Mariusz Kujawski, “Data Lakehouse vs Data Warehouse vs Data Lake - Comparison of data platforms,” https://medium.com/@mariusz_kujawski/data-warehouse-data-lake-and-data-lakehouse-comparison-of-data-platforms-842f0288b71.

[10] Guido De Simoni., “Market Guide for Active Metadata Management,” 2021.

[11] C. ZINS, “Conceptual approaches for defining data, information, and knowledge,” *Journal of the American society for information science and technology*, 2007.

[12] C. Adamson, *Star Schema - The complete reference*. 2010.

[13] D. Turban, E., Sharda, R., and Delen, “Decision Support and Business Intelligence Systems (9th eds),” New Jersey, USA: Pearson Education, Inc., 2010.

[14] L. G. Burton, B., “Business Intelligence Focus Shifts From Tactical to Strategic,” 2006.

[15] R. Kimball and M. Rosse, *The Kimball Group Reader - Relentlessly Practical Tools for Data Warehousing and Business Intelligence*. Wiley, 2016.

[16] R. Kimball, “Kimball Dimensional Modeling Techniques”.

[17] H. N. Benkhaled, D. Berrabah, and F. Boufares, “Data Warehouses and Big Data,” *International Journal of Organizational and Collective Intelligence*, vol. 10, no. 3, pp. 1–13, 2020, doi: 10.4018/ijoci.2020070101.

[18] Bill Inmon, *Data Lake Architecture: Designing the Data Lake and Avoiding the Garbage Dump*. Technics Publications; 1st edition, 2016.

[19] M. Armbrust, A. Ghodsi, R. Xin, M. Zaharia, and U. Berkeley, “Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics,” *Cidr*, 2021.

[20] “https://www.oracle.com/big-data/data-catalog/what-is-a-data-catalog/.”

[21] N. Miloslavskaya and A. : Tolstoy, *Big Data, Fast Data and Data Lake Concepts. In: 7th Annual International Conference on Biologically Inspired Cognitive Architectures (BICA 2016)*. 2016.

[22] I. Suriarachchi and B. Plale, : “Crossing Analytics Systems: A Case for Integrated Provenance in Data Lakes. In: 12th IEEE International Conference on e-Science (e-Science 2016), Baltimore, MD, USA,” 2016.

[23] D. Wells, “An Introduction to Data Catalogs: The Future of Data Management,” 2019.

[24] “O que é o catálogo de dados do azure,” https://learn.microsoft.com/pt-pt/azure/data-catalog/overview.

[25] “https://cloud.google.com/data-catalog/docs?hl=pt-br.”

[26] atlan, “The new way for data-driven teams to discover, understand, trust, and collaborate on data assets. https://atlan.com/.”

[27] SAP, “ SAP Business Technology Platform: Turn data chaos into data value with data intelligence. https://www.sap.com/products/data-intelligence.html.”

[28] “Collibra Data Catalog https://www.collibra.com/us/en/products/data-catalog.”

[29] “Tipos de metadados Google Cloud Data Catalog https://cloud.google.com/data-catalog/docs/concepts/metadata?hl=pt-br.”

[30] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, and Baak A, “The FAIR Guiding Principles for scientific data management and stewardship.,” 2016.

[31] Labadie C, Legner C, Eurich M, and Fadler M, “ FAIR Enough? Enhancing the Usage of Enterprise Data with Data Catalogs. IEEE 22nd Conference on Business Informatics (CBI),” 2020.

[32] Wells D., *Introduction to Data Catalogs*. 2020.

[33] O’Neil BK and Fryman L, *The Data Catalog: Sherlock Holmes Data Sleuthing for Analytics.* 2020.

[34] P. Sawadogo and J. Darmont, “On data lake architectures and metadata

management," Jul. 2021, doi: 10.1007/s10844-020-00608-7.

[35] E. Zaidi, G. De Simoni, R. Edjlali, and Alan D. Duncan, "Data Catalogs Are the New Black in Data Management and Analytics," 2017.

[36] Stillerman J, Fredian T, Greenwald M, and Manduchi G., " Data catalog project - A browsable, searchable, metadata system. Fusion Engineering and Design," 2016.

[37] Choi MY, Moon CJ, and Jung SJ, "Building methods of intelligent data catalog based on graph database for data sharing platform. ICIC Express Letters, Part B: Applications, vol 11," 2020.

[38] H. Dibowski, S. Schmid, Y. Svetashova, C. Henson, and T. Tran, "Using semantic technologies to manage a data lake: Data catalog, provenance and access control. CEUR Workshop Proceedings," pp. 65–80, 2020.

[39] H. Dibowski and S. Schmid, "Using Knowledge Graphs to Manage a Data Lake," *Informaitk 2020, Lecture Notes in Informatics (LNI)*, no. January, pp. 41–50, 2021.

[40] D. Fensel *et al.*, "Knowledge Graphs. Springer International Publishing. ," 2020.

[41] E. Blomqvist *et al.*, "From Data to City Indicators: A Knowledge Graph for Supporting Automatic Generation of Dashboards. ," 2017.

[42] D. Bianchini, V. de Antonellis, M. Garda, and M. Melchiori, *Using a Smart City ontology to support personalised exploration of urban data (discussion paper).* . 2019.

[43] Kostovska A, Bogatinovski J, Džeroski S, Kocev D, and Panov P, "A catalogue with semantic annotations makes multilabel datasets FAIR," 2022.

[44] Lin D, Crabtree J, Dillo I, Downs RR, Edmunds R, and Giaretta D, "The TRUST Principles for digital repositories."

[45] Rezník T, Raes L, Stott A, De Lathouwer B, Perego A, and Charvát K, " Improving the documentation and findability of data services and repositories: A review of (meta)data management approaches.," 2022.

[46] N. Caseiro and Arnaldo Coelho, " The influence of Business Intelligence capacity, network learning and innovativeness on startups performance," *Journal of Innovation & Knowledge* , pp. 139–145, 2019.

[47] B. Oliveira *et al.*, "A Measure Data Catalog for Dashboard Management and Validation." [Online]. Available: https://orcid.org/0000-0002-5555-4567

[48] L. Corr and J. Stagnitto, *Agile Data Warehouse Design, DecisionOne Press*. 2011.

[49] T. M. CONNOLLY and C. E. BEGG, *Database systems, a practical aproach to design, implementation, and management*. 2005.

[50] "Compreender o esquema em estrela e a importância para o Power BI," https://learn.microsoft.com/pt-pt/power-bi/guidance/star-schema.

[51] "Aplicar s noções básicas de DAX no Power BI Desktop."

[52] D. Oliveira, “ScienceDirect Towards a Data Catalog for Data Analytics-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0) Peer-review under responsibility of the scientific committee of the International Conference on Industry Sciences and Computer Sciences Innovation,” 2022. [Online]. Available: www.sciencedirect.comwww.elsevier.com/locate/procedia1877-0509