

Universidade do Minho
Escola de Engenharia
Mestrado em Engenharia Informática

Perfil de Especialização em Sistemas de Armazéns de Dados

Ano Letivo de 2024/2025

Sistema de Suporte à Decisão, para uma Loja de Eletrónica

António Silva
PG57867

David Teixeira
PG55929

Duarte Leitão
PG57872

João Pedro Pastore
PG55963

Vasco Faria
PG57905

Maio, 2025

Data da Receção	
Responsável	
Avaliação	
Observações	

Sistema de Suporte à Decisão, para uma Loja de Eletrónica

António Silva
PG57867

David Teixeira
PG55929

Duarte Leitão
PG57872

João Pedro Pastore
PG55963

Vasco Faria
PG57905

Maio, 2025

Resumo

No projeto desenvolvido para a *Perifericum*, estabelecimento especializado na venda de produtos eletrônicos em Braga, foi implementado um sistema de suporte à decisão baseado num *Data Warehouse* que revolucionou a caracterização de perfis de clientes e personalização de ofertas de produtos e serviços da empresa. Adotando a metodologia de Kimball, construímos um *Data Mart* de Vendas estruturado em cinco dimensões essenciais (Tempo, Cliente, Produto, Canal e Venda), utilizando o *Dimensional Fact Model* com ferramentas especializadas.

O levantamento de requisitos, combinando abordagens Goal-Driven, Supply-Driven e User-Driven, resultou em 33 requisitos estratégicos que guiaram o desenvolvimento. A arquitetura integra pipelines ETL automatizadas via *Apache NiFi*, consolidando eficientemente dados de vendas físicas e digitais.

O ecossistema analítico, potencializado por *Power BI*, *Pandas* e *Scikit-learn*, permitiu segmentar clientes através de algoritmos K-Means, identificando quatro perfis distintos. Implementamos sistemas de recomendação baseados em múltiplas abordagens de filtragem e desenvolvemos *dashboards* interativos para visualização de indicadores-chave.

Esta solução capacita a empresa para decisões estratégicas mais precisas, incrementando a fidelização e competitividade. Possíveis trabalhos futuros envolvem integração de dados em tempo real, expansão para novos *Data Marts* em Logística e Marketing, além do refinamento contínuo dos modelos preditivos, solidificando assim a presença da empresa nos mercados físico e digital.

Índice Geral

Resumo	3
1. Definição do Sistema	1
1.1. Contexto de Aplicação	1
1.2. Motivação e Objetivos do Trabalho	2
1.3. Análise da Viabilidade do Processo	3
1.4. Recursos e Equipa de Trabalho	4
1.5. Plano de Execução do Projeto	5
2. Levantamento e Análise de Requisitos	7
2.1. Método Adotado	7
2.2. Organização dos Requisitos Levantados	7
2.3. Análise e Validação Geral dos Requisitos	10
3. Modelação Dimensional de Dados	12
3.1. Apresentação da Abordagem Realizada	12
3.2. A Matriz de Decisão	13
3.3. Definição e Caracterização de Dimensões	14
3.4. Definição e Caracterização das Tabelas de Factos e Respetivos Grãos	18
3.5. Configuração dos Esquemas	20
3.6. As Vistas dos Agentes de Decisão	21
4. Arquitetura Geral do Sistema	23
4.1. Apresentação Geral	23
4.2. Fontes de Dados	25
4.3. Área de Preparação de dados	27
4.4. O Armazém de Dados	28
4.5. Exploração e Visualização de Dados	31
4.6. Aquisição de Conhecimento	31
5. O Povoamento de Dados	33
5.1. Apresentação da Abordagem Realizada	33
5.2. Mapeamento de Dados	34
5.3. Modelação do Sistema de Povoamento	37
5.4. Implementação do Sistema de Povoamento	42
5.5. Validação e Testes	43
6. Exploração e Análise de Dados	47
6.1. Organização Geral do Sistema de Dashboarding	47
6.2. Serviços de Exploração e Análise Implementados	48
7. Caracterização de Perfis de Clientes	53
7.1. Definição do Problema e Compreensão dos Elementos de Análise Envolvidos	53
7.2. Seleção e Preparação dos Dados	54
7.3. Identificação e Fundamentação da Técnica de Análise	55
7.4. Construção do Modelo de Análise	55
7.5. Validação do Desempenho do Modelo	58
7.6. Avaliação dos Resultados	60

8. Personalização de Ofertas de Produtos e Serviços	66
8.1. Definição do Problema e Compreensão dos Elementos de Análise Envolvidos	66
8.2. Seleção e Preparação dos Dados	66
8.3. Identificação e Fundamentação da Técnica de Análise	66
8.4. Construção do Modelo de Análise	67
8.5. Validação do Desempenho do Modelo	69
8.6. Avaliação dos Resultados	70
9. Conclusões e Trabalho Futuro	74
Referências	75
Lista de Siglas e Acrónimos	76

Índice de Figuras

Figura 1: Diagrama GANTT do Projeto	5
Figura 2: Matriz de Decisão	13
Figura 3: As dimensões do DM Vendas	14
Figura 4: Caracterização detalhada da dimensão Dim-Tempo.	15
Figura 5: Caracterização detalhada da dimensão Dim-Cliente.	15
Figura 6: Caracterização detalhada da dimensão Dim-Produto.	16
Figura 7: Caracterização detalhada da dimensão Dim-Canal.	16
Figura 8: Caracterização detalhada da dimensão Dim-Venda	17
Figura 9: Caracterização detalhada da tabela de factos TF-Venda.	18
Figura 10: Esquema Dimensional do <i>Data Mart</i> de Vendas da <i>Perifericum</i> .	20
Figura 11: Esquema geral da arquitetura do sistema de suporte à decisão da <i>Perifericum</i> .	23
Figura 12: Esquema Dimensional do <i>DataWarehouse</i> da <i>Perifericum</i> .	29
Figura 13: Esquema Lógico do <i>DataWarehouse</i> da <i>Perifericum</i> .	30
Figura 14: Modelo BPMN a nível do processo (1º nível de abstração)	38
Figura 15: Subprocesso BPMN para o carregamento da Dimensão Tempo	39
Figura 16: Subprocesso BPMN para o carregamento da Dimensão Produto	40
Figura 17: Subprocesso BPMN para o carregamento da Dimensão Cliente	40
Figura 18: Subprocesso BPMN para o carregamento tabela de factos Vendas	41
Figura 19: Agendamento automático do processo ETL com execução semanal.	42
Figura 20: Registo atualizado do campo <i>ultima_execucao</i> após execução do processo ETL	43
Figura 21: DimCliente – Registos Iniciais de Clientes	43
Figura 22: Clientes_Historico sem registos	43
Figura 23: Registos da loja online na TFVenda	44
Figura 24: Últimos registos da TFVenda	44
Figura 25: Antigo registo de informação de Nair Tavares	45
Figura 26: Informação atualizada na DimCliente	45
Figura 27: Novos registos de vendas inseridos na TFVenda	45
Figura 28: Registos extraídos com anomalias para a tabela de quarentena de vendas físicas	45
Figura 29: Registos extraídos com anomalias para a tabela de quarentena de vendas física	45
Figura 30: Registos extraídos com anomalias para a tabela de quarentena de vendas online	46
Figura 31: Dashboard: Distribuição de Clientes por Distrito, Sexo e Estado Civil	48

Figura 32: Dashboard: Profissões Mais Comuns entre os Clientes	48
Figura 33: Dashboard: Categorias de Produtos Mais Compradas e Marcas Mais Procuradas	49
Figura 34: Dashboard: Valor Médio de Compra por Período	49
Figura 35: Dashboard: Valor Médio de Compra por Grupo Etário	50
Figura 36: Dashboard: Valores Totais por Canal de Compra	50
Figura 37: Dashboard: Top 4 das Categorias de Marcas de Produtos mais Vendidos	51
Figura 38: Dashboard: Vendas por Dias da Semana	51
Figura 39: Dashboard: Percentagem de Desconto pelas Vendas Totais	52
Figura 40: Dashboard: Dispersão de Produtos por Percentagem de Descontos	52
Figura 41: Visualização 3D dos clusters dos clientes	59
Figura 42: Distribuição dos clientes por Cluster	60
Figura 43: Gráfico do Perfil de cada cliente por Cluster	61
Figura 44: Distribuição dos Valores de cada atributo por Cluster	63

Índice de Tabelas

Tabela 1: Agentes de Decisão e Vistas Correspondentes	21
Tabela 2: <i>source-to-target datamap</i> : Vendas Físicas → Dim-Produto	34
Tabela 3: <i>source-to-target datamap</i> : Vendas Online → Dim-Produto	34
Tabela 4: <i>source-to-target datamap</i> : clientes_logs → Dim-Cliente	35
Tabela 5: <i>source-to-target datamap</i> : Vendas Físicas → Dim-Tempo	35
Tabela 6: <i>source-to-target datamap</i> : Vendas Online → Dim-Tempo	36
Tabela 7: <i>source-to-target datamap</i> : Vendas Online → TF-Venda	36
Tabela 8: <i>source-to-target datamap</i> : Vendas Físicas → TF-Venda	37

1. Definição do Sistema

1.1. Contexto de Aplicação

Fundada em 2010 por João Silva Teixeira Faria Leitão, a *Perifericum* é uma loja especializada em tecnologia, situada na freguesia de S. Vitor, em Braga. Criada na sequência de uma mudança profissional do seu fundador — um ex-engenheiro informático —, a loja surgiu como forma de manter uma ligação ativa ao setor tecnológico, mas com total autonomia e proximidade ao cliente final.

Nos primeiros anos de atividade, a Perifericum centrou-se na reparação de computadores e na venda de acessórios básicos. Com o tempo, e acompanhando a evolução das preferências dos consumidores, a loja diversificou a sua oferta para incluir produtos de gaming, componentes de PC, smartphones, gadgets e periféricos. Esta expansão permitiu consolidar a sua reputação como uma referência local entre os entusiastas de tecnologia.

Apesar do reconhecimento conquistado, a loja enfrenta hoje desafios significativos num mercado cada vez mais competitivo. A presença dominante de grandes cadeias de retalho e a agressividade das lojas online — com preços baixos e entregas rápidas — colocam pressão sobre os comércios de proximidade. Estudos recentes apontam que 76% dos consumidores valorizam recomendações personalizadas e experiências de compra adaptadas aos seus perfis (McKinsey & Company, 2021). No entanto, a *Perifericum* ainda não dispõe de uma solução estruturada que permita conhecer melhor os seus clientes e personalizar as suas ofertas, o que limita o seu potencial de crescimento e diferenciação.

Ciente destas limitações, a loja deu recentemente um passo estratégico ao concluir a sua transição para o comércio eletrónico, procurando alargar o seu alcance e responder às novas exigências do consumidor digital. Inspirando-se em exemplos nacionais como a *GlobalData* e a *PCDiga*, que fortaleceram a sua presença online com plataformas de e-commerce e estratégias de marketing orientadas por dados, a *Perifericum* começou a explorar vendas digitais como forma de ampliar a sua área de influência para além de Braga.

Neste novo contexto de transformação, torna-se essencial estabelecer um conjunto de ferramentas que lhe permitam compreender melhor os seus clientes, personalizar a experiência de compra e tomar decisões comerciais mais informadas. A adoção de um sistema de suporte à decisão, assente na análise de dados e em técnicas de recomendação inteligente, surge assim como uma resposta estratégica para assegurar a sustentabilidade e competitividade da loja no cenário atual.

1.2. Motivação e Objetivos do Trabalho

Nos últimos anos, a *Perifericum* tem enfrentado dificuldades crescentes na compreensão das preferências dos clientes e na otimização das vendas. Apesar de oferecer uma vasta gama de produtos, a loja carece de um sistema estruturado de análise de dados, o que compromete a capacidade de personalizar a experiência de compra e fidelizar clientes. Esta lacuna tem-se traduzido numa perda de oportunidades de negócio e numa discrepância entre os resultados obtidos e os objetivos inicialmente definidos.

Atualmente, os registos de vendas na loja física são processados manualmente, o que impossibilita uma análise eficiente dos hábitos de consumo e impede a implementação de estratégias como as vendas cruzadas. Por exemplo, um cliente que adquire um rato gamer poderia beneficiar da recomendação de um teclado mecânico compatível — mas a loja não dispõe de um mecanismo automatizado que permita identificar e sugerir estas combinações. Adicionalmente, a ausência de uma visão consolidada do histórico de compras dificulta a criação de campanhas promocionais personalizadas e eficazes.

Perante este cenário, o Sr. João, com o apoio do seu ex-colega Sr. Gates, decidiu avançar com a implementação de um sistema de suporte à decisão, com o objetivo de melhorar a gestão de clientes e a experiência de compra. A solução assenta na conceção e desenvolvimento de um armazém de dados (*Data Warehouse*), que integrará históricos de transações e permitirá a aplicação de técnicas avançadas de análise de dados e aprendizagem automática (*machine learning*). Com esta infraestrutura, a loja poderá gerar recomendações personalizadas e melhorar significativamente os níveis de fidelização.

- Os principais objetivos do sistema são os seguintes:
 - **Caracterização de perfis de clientes:** através de técnicas de segmentação (*clustering*), pretende-se distinguir perfis como clientes leais, novos, ocasionais ou em risco, com base em dados demográficos, geográficos e históricos de compras. Esta informação permitirá definir estratégias comerciais específicas para cada segmento.
 - **Personalização de Ofertas de Produtos:** o sistema analisará padrões de consumo para segmentar clientes, oferecendo recomendações personalizadas de artigos. Por exemplo, clientes que frequentemente adquirem componentes de PC, como placas gráficas ou processadores, receberão sugestões proativas de itens complementares, como coolers, memórias RAM ou caixas com iluminação RGB.
 - **Melhoria da experiência de compra:** ao disponibilizar recomendações ajustadas aos perfis, a loja poderá oferecer uma experiência mais envolvente e personalizada, tanto na plataforma online como no espaço físico. Isto poderá incluir, por exemplo, sugestões semanais de produtos com base no histórico de cada cliente.
 - **Apoio à tomada de decisão estratégica:** o sistema irá incluir painéis interativos com indicadores relevantes, como produtos mais vendidos, tendências de mercado e desempenho de marcas. Esta visão global permitirá aos gestores tomar decisões mais informadas e alinhadas com os objetivos comerciais da loja.

Ao implementar este sistema, a empresa espera não apenas melhorar os seus processos internos e a relação com os clientes, mas também reforçar o seu posicionamento no mercado — num sector onde a valorização da experiência do cliente é cada vez mais decisiva.

1.3. Análise da Viabilidade do Processo

A implementação de um sistema de suporte à decisão para a *Periféricum* revela-se viável sob os pontos de vista técnico, financeiro e operacional, oferecendo benefícios claros em termos de eficiência, competitividade e experiência do cliente.

1.3.1. Viabilidade Técnica

O projeto proposto é tecnicamente viável. A loja dispõe dos equipamentos e infraestruturas necessários para suportar o sistema de suporte à decisão, incluindo servidores com capacidade adequada para o armazenamento e processamento de dados, bem como licenças ativas de software como *Power BI* e *MySQL*.

Supletivamente, a equipa técnica tem experiência comprovada em processos de integração de dados, análise preditiva e visualização, dominando ferramentas como *Apache NiFi*, *Pandas* e *Scikit-learn*. Não são identificadas limitações técnicas que comprometam os objetivos do projeto, pelo que este pode ser executado com os recursos atualmente disponíveis.

1.3.2. Viabilidade Financeira

A proposta destaca-se pela sua sustentabilidade financeira. Ao evitar custos variáveis de soluções cloud, a opção por uma infraestrutura local permite um investimento inicial controlado e previsível, com baixos custos operacionais a longo prazo.

O retorno do investimento será impulsionado pelo aumento do valor médio das vendas, resultado da personalização das ofertas e da segmentação eficaz dos clientes. A capacidade de recomendar produtos relevantes e potenciar vendas cruzadas deverá aumentar as taxas de conversão e a retenção de clientes, traduzindo-se num impacto direto nas receitas.

1.3.3. Viabilidade Operacional

A implementação será realizada de forma faseada, garantindo uma transição gradual que minimiza riscos e evita interrupções no funcionamento diário da loja. A usabilidade será uma prioridade: o sistema incluirá interfaces simples e intuitivas, permitindo a sua utilização por colaboradores sem formação técnica especializada.

As tarefas repetitivas, como recolha e análise de dados, serão automatizadas, reduzindo o esforço manual e libertando recursos para atividades de maior valor. O planeamento por fases permitirá ainda testar, ajustar e otimizar cada componente do sistema, garantindo uma adaptação contínua às necessidades da operação.

Em suma, a análise de viabilidade confirma que o projeto é exequível e vantajoso. A empresa ficará equipada com uma solução robusta, escalável e alinhada com a sua realidade, posicionando-se de forma mais competitiva no setor tecnológico — tanto no mercado local como nacional.

1.4. Recursos e Equipa de Trabalho

A execução do sistema de suporte à decisão da *Perifericum* exige uma combinação equilibrada de recursos humanos, tecnológicos e organizacionais.

1.4.1. Recursos Humanos

A equipa de trabalho será composta por profissionais internos com experiência nas áreas envolvidas, complementada, se necessário, com apoio externo:

- **Engenheiro de dados:** responsável pelos fluxos de integração (ETL), modelação do armazém de dados e gestão da base de dados.
- **Cientista de dados:** encarregue da análise exploratória, construção de modelos preditivos e validação dos resultados.
- **Analista de BI:** responsável pela construção de dashboards, definição de métricas relevantes e interação com os agentes de decisão.
- **Gestor de projeto:** assegura o planeamento, acompanhamento de tarefas, comunicação entre áreas e alinhamento com os objetivos estratégicos.

A carga de trabalho será distribuída em função de fases específicas do projeto, com momentos de maior concentração de esforço nas etapas de integração e validação dos modelos.

1.4.2. Recursos Tecnológicos

Os seguintes recursos já estão disponíveis e são considerados suficientes para o desenvolvimento do projeto:

- **Servidores locais** com capacidade para execução de ETL e modelação preditiva;
- **Licenças ativas do *Power BI***, já utilizadas noutras áreas da organização;
- **Infraestrutura MySQL** com instâncias dedicadas ao projeto;
- **Ambiente de desenvolvimento Python** com bibliotecas como Pandas e Scikit-learn configuradas.
- **Indyco Builder**, utilizado para a modelação conceptual do armazém de dados, garantindo coerência e rastreabilidade entre requisitos e estruturas de dados;
- **Microsoft Excel**, empregado na documentação detalhada dos elementos do *Data Warehouse*.

1.4.3. Recursos Organizacionais

O projeto conta com o apoio da direção da empresa, assegurando a utilidade prática dos indicadores e recomendações produzidas. A comunicação entre áreas será contínua, garantindo alinhamento estratégico e operacional.

1.5. Plano de Execução do Projeto

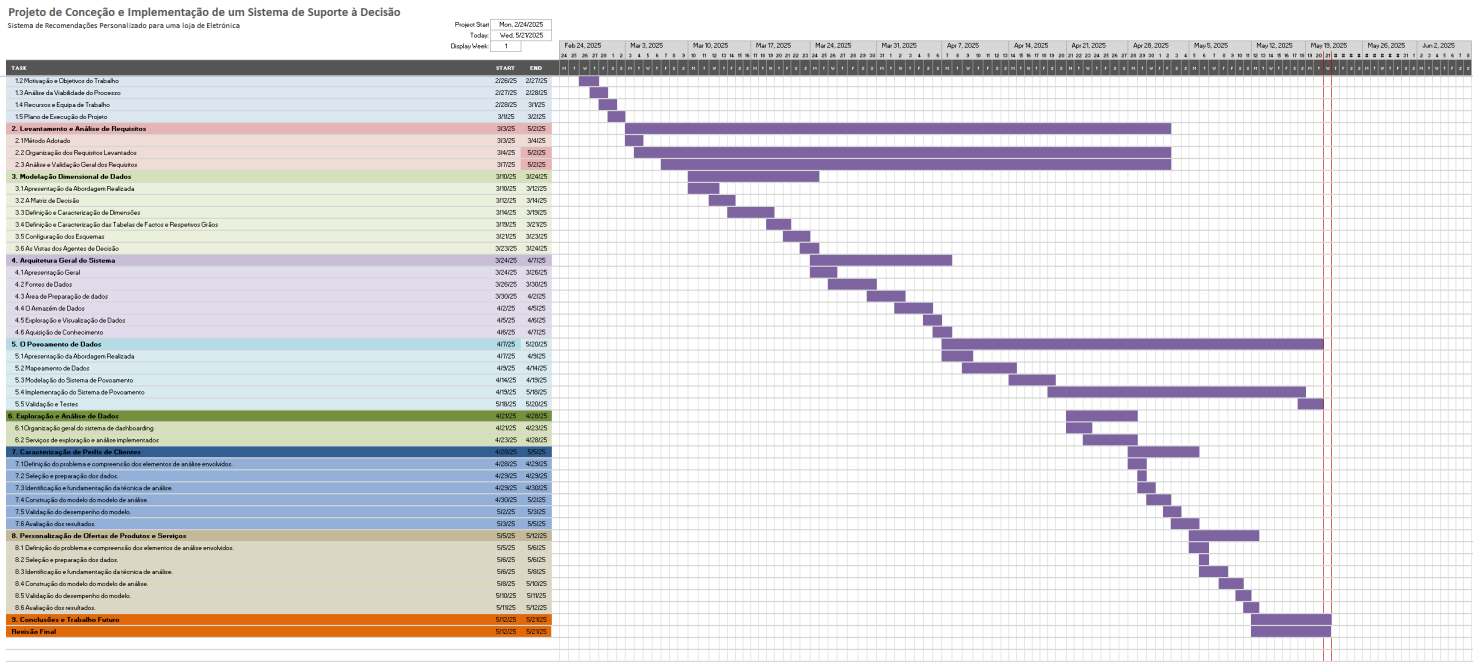


Figura 1: Diagrama GANTT do Projeto

O plano de execução do projeto está representado na figura acima, que organiza as tarefas por fases entre fevereiro e maio de 2025. As atividades foram distribuídas de forma lógica, tendo em conta a sua complexidade, interdependência e os diferentes intervenientes no projeto. O projeto está dividido em nove fases principais:

- **Definição do Sistema (24/02 a 03/03):** alinhamento inicial de objetivos, escopo e plano de ação.
- **Levantamento e Análise de Requisitos (03/03 a 02/05):** esta fase decorre ao longo de quase todo o projeto, refletindo a constante comunicação entre a equipa de desenvolvimento e os agentes de decisão. Foram identificados, ajustados ou removidos requisitos à medida que o projeto evoluía, o que permitiu alinhar continuamente a solução com as necessidades reais da loja.
- **Modelação Dimensional de Dados (10/03 a 24/03):** construção do modelo dimensional do armazém de dados que sustenta a solução analítica.
- **Arquitetura do Sistema (24/03 a 07/04):** definição técnica da infraestrutura e componentes do sistema.
- **Povoamento de Dados (07/04 a 20/05):** recolha, preparação e inserção de dados.
- **Exploração e Análise de Dados (21/04 a 28/04):** primeiros testes com os dados e validação de dashboards.
- **Caracterização de Perfis de Clientes (28/04 a 05/05):** construção de modelos analíticos para segmentação.
- **Personalização de Ofertas (05/05 a 12/05):** aplicação dos perfis para recomendar produtos e serviços.
- **Conclusões e Trabalho Futuro (12/05 a 21/05):** avaliação dos resultados e definição de propostas para fases seguintes.

É importante salientar que o povoamento de dados foi uma fase particularmente demorada devido à complexidade da ferramenta Apache NiFi utilizada para implementação do ETL. No entanto, as fases de caracterização de perfis de cliente, visualização de dados e personalização de ofertas de produtos e serviços foram executadas parcialmente em paralelo, mesmo com a pipeline ETL ainda em desenvolvimento. Isso foi possível graças à implementação de uma solução temporária baseada em *pandas*, que forneceu os dados necessários para essas fases, permitindo avançar com o desenvolvimento enquanto a solução ETL definitiva era finalizada.

Os prazos definidos asseguraram o cumprimento dos objetivos do projeto, oferecendo flexibilidade para adaptações e mantendo a colaboração contínua entre os diferentes atores envolvidos no desenvolvimento do sistema de suporte à decisão.

2. Levantamento e Análise de Requisitos

2.1. Método Adotado

Para identificar os requisitos do sistema de suporte à decisão da *Perifericum*, foi adotada uma abordagem mista que combina métodos *Goal-Driven*, *Supply-Driven* e elementos do *User-Driven* (Golfarelli, 2009). Esta estratégia permitiu alinhar o sistema aos objetivos estratégicos da loja, tirar partido dos dados já existentes e incorporar as expectativas dos agentes de decisão.

A componente *Goal-Driven* foi essencial para garantir que o sistema respondesse às metas previamente estabelecidas (Secção 1.2). Já a abordagem *Supply-Driven* permitiu explorar os dados operacionais disponíveis, garantindo que as decisões fossem suportadas por informação concreta e acessível, sem necessidade de recolha adicional intensiva. O método *User-Driven*, ainda que aplicado de forma moderada, foi utilizado para captar necessidades práticas, como a utilidade dos *dashboards* e a relevância das recomendações. Esta vertente contribuiu para a aceitação do sistema sem implicar custos excessivos ou um esforço adicional significativo.

O processo de levantamento incluiu entrevistas semiestruturadas com os principais intervenientes, centradas na experiência do utilizador. Também foi realizada uma análise documental dos registos de vendas mais recentes, já digitalizados, de forma a antecipar padrões de consumo. Por fim, um workshop de validação serviu para confirmar os requisitos recolhidos e ajustar a proposta às expectativas dos agentes de decisão.

A documentação gerada neste processo inclui atas das entrevistas e da sessão de validação.

2.2. Organização dos Requisitos Levantados

A identificação dos requisitos foi conduzida em três etapas principais: definição de metas com os agentes de decisão, análise dos dados operacionais disponíveis e validação participativa das necessidades. A partir destas ações, foi possível consolidar um conjunto claro de funcionalidades desejadas, diretamente relacionadas com a **área de decisão de Vendas**.

A estruturação dos requisitos centra-se nos dois objetivos essenciais da *Perifericum*: caracterizar perfis de clientes e personalizar ofertas de produtos. Para tal, recorreu-se à informação extraída dos registos de vendas e às orientações fornecidas pelo Sr. João e pelo Sr. Gates. A organização apresentada visa garantir que as decisões possam ser apoiadas por dados fiáveis e acionáveis, melhorando a eficácia comercial da loja.

2.2.1. A. Requisitos de Análise

Nº do Requisito	Data	Descrição	Tipo	Fonte	Revisor
R01	2025-03-11	Analisar vendas por período (mensal, trimestral, anual).	Análise	Equipa de Trabalho	Sr. João, Sr. Gates
R02	2025-03-11	Avaliar o impacto de feriados, fins de semana e sazonalidade no comportamento de compra.	Análise	Equipa de Trabalho	Sr. João, Sr. Gates
R03	2025-03-11	Identificar padrões de compra ao longo do tempo.	Análise	Equipa de Trabalho	Sr. João, Sr. Gates
R04	2025-03-11	Comparar performance de vendas entre loja física e loja online.	Análise	Equipa de Trabalho	Sr. João, Sr. Gates
R05	2025-03-12	Analisar o comportamento dos clientes por canal.	Análise	Equipa de Trabalho	Sr. João, Sr. Gates
R06	2025-03-12	Caracterizar o perfil dos clientes com base na faixa etária, sexo, localização, profissão, estado civil.	Análise	Equipa de Trabalho	Sr. João, Sr. Gates
R07	2025-03-12	Realizar uma análise RFM.	Análise	Equipa de Trabalho	Sr. João, Sr. Gates
R08	2025-03-12	Segmentar clientes com base no histórico de compras e preferências.	Análise	Equipa de Trabalho	Sr. João, Sr. Gates
R09	2025-03-13	Detetar variações de valores referentes aos clientes ao longo do tempo.	Análise	Equipa de Trabalho	Sr. João, Sr. Gates
R10	2025-03-13	Avaliar vendas por categoria de produto.	Análise	Equipa de Trabalho	Sr. João, Sr. Gates
R11	2025-03-13	Identificar produtos mais vendidos por canal, por período e por perfil de cliente.	Análise	Equipa de Trabalho	Sr. João, Sr. Gates
R12	2025-03-13	Avaliar impacto de descontos aplicados nos produtos.	Análise	Equipa de Trabalho	Sr. João, Sr. Gates
R13	2025-03-14	Apoiar estratégias de venda cruzada.	Análise	Equipa de Trabalho	Sr. João, Sr. Gates
R14	2025-03-14	Criar perfis de “combos” típicos por segmento de cliente.	Análise	Equipa de Trabalho	Sr. João, Sr. Gates
R15	2025-03-14	Identificar zonas geográficas com maior volume de vendas.	Análise	Equipa de Trabalho	Sr. João, Sr. Gates
R16	2025-03-14	Comparar comportamentos entre localidades.	Análise	Equipa de Trabalho	Sr. João, Sr. Gates

2.2.2. B. Requisitos Técnicos

Nº do Requisito	Data	Descrição	Tipo	Fonte	Revisor
-----------------	------	-----------	------	-------	---------

R17	2025-03-15	Povoamento semanal dos dados do <i>Data Warehouse</i> .	Técnico	Equipa de Trabalho	Sr. João, Sr. Gates
R18	2025-03-15	Preparação para escalabilidade e integração com outros Data Marts.	Técnico	Equipa de Trabalho	Sr. João, Sr. Gates
R19	2025-03-15	Tratamento de dados ausentes, nulos, duplicados e inconsistências entre dimensões e factos.	Técnico	Equipa de Trabalho	Sr. João, Sr. Gates
R20	2025-03-16	Aplicação de regras de negócio.	Técnico	Equipa de Trabalho	Sr. João, Sr. Gates
R21	2025-03-16	Validação de integridade referencial.	Técnico	Equipa de Trabalho	Sr. João, Sr. Gates
R22	2025-03-16	<i>Logging</i> e notificação de erros no ETL.	Técnico	Equipa de Trabalho	Sr. João, Sr. Gates
R23	2025-03-16	Armazenamento de dados históricos desde o início da operação (SCD Tipo 4 para clientes).	Técnico	Equipa de Trabalho	Sr. João, Sr. Gates
R24	2025-03-17	Otimização do modelo dimensional para consultas analíticas rápidas (<i>star schema</i>).	Técnico	Equipa de Trabalho	Sr. João, Sr. Gates
R25	2025-03-17	Agregações pré-calculadas e indexação para relatórios frequentes.	Técnico	Equipa de Trabalho	Sr. João, Sr. Gates
R26	2025-03-17	Capacidade de resposta a <i>dashboards</i> em tempo útil.	Técnico	Equipa de Trabalho	Sr. João, Sr. Gates
R27	2025-03-18	Integração de fontes de dados operacionais heterogéneas no DW via <i>pipeline</i> ETL.	Técnico	Equipa de Trabalho	Sr. João, Sr. Gates
R28	2025-03-18	Definição de uma área de preparação de dados (<i>staging area</i>) antes do carregamento no DW.	Técnico	Equipa de Trabalho	Sr. João, Sr. Gates

2.2.3. C. Requisitos Estratégicos e de Expansão

Número do Requisito	Data	Descrição	Tipo	Fonte	Revisor
R29	2025-03-19	Suporte à criação de ofertas personalizadas com base no comportamento de compra individual e padrões por grupo demográfico ou localização.	Estratégico	Equipa de Trabalho	Sr. João, Sr. Gates
R30	2025-03-19	Preparação para adição de novos canais.	Estratégico	Equipa de Trabalho	Sr. João, Sr. Gates
R31	2025-03-19	Integração futura com Data Marts de Marketing, Logística, Recursos Humanos.	Estratégico	Equipa de Trabalho	Sr. João, Sr. Gates

R32	2025-03-20	Evolução do modelo sem comprometer o desempenho (modelação <i>bottom-up</i>).	Estratégico	Equipa de Trabalho	Sr. João, Sr. Gates
-----	------------	--	-------------	--------------------	---------------------

2.3. Análise e Validação Geral dos Requisitos

A validação dos requisitos definidos nas secções anteriores foi conduzida de forma colaborativa entre os diferentes perfis da equipa de trabalho, garantindo que cada dimensão do projeto — técnica, analítica e estratégica — fosse devidamente representada e validada. Esta abordagem assegura que o sistema proposto esteja em conformidade com os objetivos organizacionais e operacionais da *Perifericum*, respondendo com eficácia às necessidades de apoio à decisão.

2.3.1. Validação por Perfil da Equipa

2.3.1.1. Gestor de Projeto

Validou a coerência global entre os requisitos estratégicos (Secção 2.2.3) e os objetivos definidos pela direção. Destacou a importância da escalabilidade do sistema, da preparação para novos canais e da capacidade de evolução do modelo de dados sem comprometer o desempenho. Confirmou também o alinhamento do cronograma de execução com os marcos definidos no plano do projeto.

2.3.1.2. Analista de BI

Validou os requisitos analíticos (Secção 2.2.1), sublinhando a relevância da segmentação de clientes, da análise por canal e do cruzamento entre perfis e preferências de consumo. Realçou o contributo dos *dashboards* na operacionalização dos resultados analíticos e a importância de dados fiáveis e consolidados no apoio à decisão.

2.3.1.3. Cientista de Dados

Confirmou que os requisitos técnicos e analíticos (Secção 2.2.1 e Secção 2.2.2) suportam a construção de modelos preditivos eficazes. Requisitos como o histórico de dados, a análise RFM e a segmentação por perfis garantem uma base sólida para análises avançadas. Destacou a necessidade de dados limpos e estruturados para garantir a qualidade dos modelos.

2.3.1.4. Engenheiro de Dados

Validou todos os requisitos técnicos (Secção 2.2.2), desde a integração de dados heterogéneos via ETL até à aplicação de regras de negócio, otimização de esquemas dimensionais e preparação do sistema para escalabilidade e performance. Enfatizou a importância da *staging area*, do controlo de qualidade dos dados e da monitorização com *logging* adequado.

2.3.2. Anomalias e Ações Corretivas

Foi detetada uma lacuna importante relacionada com a proteção de dados pessoais dos clientes. Para colmatar essa falha, foi introduzido o **Requisito R33** – Segurança e Privacidade dos Dados:

Número do Requi- sito	Data	Descrição	Tipo	Fonte	Revisor
R33	2025-03-20	Garantir a segurança e privacidade dos dados dos clientes, incluindo definição de perfis de acesso ao Data Warehouse por tipo de utilizador e conformidade com o RGPD e melhores práticas de segurança da informação.	Técnico/ Estraté- gico	Equipa de Trabalho	Sr. João, Sr. Gates

A análise geral dos requisitos evidencia uma forte adesão às necessidades dos diferentes elementos da equipa de projeto. O modelo proposto mostra-se tecnicamente sólido, analiticamente flexível e estrategicamente escalável. A única anomalia identificada foi corrigida com a adição de um novo requisito, garantindo que o conjunto final se encontra coerente, completo e validado para dar seguimento à fase de conceção detalhada e implementação do armazém de dados.

3. Modelação Dimensional de Dados

3.1. Apresentação da Abordagem Realizada

O processo de modelação dimensional foi realizado para estruturar o armazém de dados, com o objetivo de apoiar a análise de perfis de clientes e a personalização de ofertas de produtos. A modelação foi desenhada para organizar os dados de forma a suportar análises multidimensionais, facilitando uma subsequente implementação de um sistema responsável por segmentações detalhadas e recomendações baseadas em padrões de compra. A seguir, descreve-se o método, a notação e as ferramentas utilizadas.

Para orientar este trabalho, foi seguida a metodologia de modelação dimensional proposta por Ralph Kimball, que adota uma abordagem *bottom-up*. Esta estratégia privilegia a criação de Data Marts específicos para necessidades imediatas do negócio, que mais tarde podem ser integrados num DW mais abrangente. O processo iniciou-se com a identificação das principais necessidades analíticas, centradas em áreas como dados de clientes (e.g., demografia e localização), vendas (e.g., transações e volumes), produtos, tempo e canais de venda. A seleção destas áreas foi guiada pelos requisitos definidos (Secção 2.2) em conjunto com os agentes de decisão.

Com base nessa definição, foi desenvolvido um *Data Mart* inicial focado na análise de vendas. Este modelo permitiu responder rapidamente a perguntas relevantes, como “quais produtos são preferidos pelos clientes?” ou “qual é o canal mais popular de venda?”. A prioridade foi gerar valor analítico desde as fases iniciais do projeto, garantindo rapidez e impacto. Após a construção do modelo, a equipa realizou uma sessão de validação com os responsáveis pela loja, assegurando que os requisitos estavam bem representados. A abordagem adotada permite ainda a integração futura de novos *Data Marts*, mantendo a escalabilidade e a modularidade do sistema.

A escolha da metodologia de Kimball justifica-se pela sua simplicidade e eficácia em contextos orientados ao negócio. A estruturação dos dados em torno de factos e dimensões facilita a navegação e torna as consultas mais acessíveis. Adicionalmente, permite uma implementação incremental, o que é essencial num ambiente que exige agilidade e resultados práticos. Em contraste, a abordagem *top-down* de Bill Inmon, embora reconhecida pela sua consistência e rigor, envolve maior complexidade e tempo de desenvolvimento, tornando-se menos adequada para um projeto com objetivos de curto prazo como este.

Para representar o modelo dimensional, recorreu-se ao *Dimensional Fact Model* (DFM) (Golfarelli et al., 1998), proposto por Golfarelli et al. (1998), uma notação criada especificamente para modelação conceptual em armazéns de dados. Esta notação foi escolhida por permitir uma representação clara dos factos, dimensões e hierarquias, facilitando a compreensão tanto pela equipa técnica como pelos utilizadores de negócio. O DFM permite ilustrar métricas (como valores de vendas) e atributos que as descrevem (como categorias de produto), bem como definir hierarquias analíticas (como mês → trimestre → ano), o que o torna particularmente adequado para o projeto. Foi preferido em relação a notações como UML ou ORM, por ser mais direto e focado nas necessidades da análise multidimensional.

Quanto às ferramentas, a modelação foi apoiada pelo *Indyco Builder*, uma ferramenta especializada em esquemas dimensionais, utilizada para desenhar e documentar graficamente o modelo. A escolha recaiu na sua capacidade de facilitar a validação visual do modelo com os agentes de decisão. Complementarmente, recorreu-se ao *Microsoft Excel* para organizar os dados preliminares, construir matrizes de decisão e mapear os requisitos do projeto aos elementos do modelo. Esta combinação permitiu estruturar e validar o esquema de forma prática e eficaz.

O processo resultou num modelo dimensional claro e flexível, preparado para evoluir com o crescimento da organização. A abordagem adotada garante não só alinhamento com os objetivos analíticos atuais, mas também a capacidade de adaptação futura — um requisito essencial para apoiar decisões informadas num mercado dinâmico.

3.2. A Matriz de Decisão

A matriz de decisão foi desenvolvida para mapear os requisitos do sistema de suporte à decisão da *Perifericum* às estruturas multidimensionais do *Data Mart* de Vendas, garantindo que as necessidades dos agentes de decisão sejam atendidas de forma eficiente. Atua como uma ferramenta de planeamento que alinha os objetivos estratégicos do projeto às capacidades analíticas do sistema.

A matriz foi construída com base num *Data Mart* transacional, com periodicidade semanal, contendo cinco dimensões principais: Tempo, Cliente, Produto, Canal e Venda. Estas dimensões foram escolhidas por permitirem análises como “Quais clientes comprem mais determinados produtos?” ou “Quais itens são menos vendidos?”, fundamentais para o apoio à decisão na área comercial.

Caracterização de <i>Data Mart</i> de Vendas	
Identificação: Vendas	
Descrição Geral: Informação para suporte à tomada de decisão na área de vendas da "Perifericum", providenciando elementos de dados selecionados acerca das vendas de produtos eletrónicos nas lojas (física e online), para facilitar a caracterização de perfis de clientes e personalização de ofertas de produtos.	
Estrutura base	
Tabela de Factos>>	TF-Venda
<<Dimensões	
Tempo	✓
Cliente	✓
Produto	✓
Canal	✓
Venda	✓
Número Dimensões	5
Tipo	Transacional
Periodicidade	Semanal
Descrição	Transações comerciais de produtos eletrónicos.
Unidade estratégica	Incentivar as vendas de produtos eletrónicos. Definição e caracterização de perfis de vendas de produtos eletrónicos. Identificar e caracterizar nichos de mercado.
Utilizadores	Sr. João e Sr. Gates
Observações	
Nada a assinalar.	

Figura 2: Matriz de Decisão

Os agentes de decisão assumem papéis complementares. Sr. João, focado na gestão estratégica e na experiência do cliente, utiliza os insights para promover a fidelização e aumentar as vendas cruzadas.

Já o Sr. Gates, orientado para operações e marketing, apoia-se nas análises para otimizar campanhas promocionais.

A matriz de decisão assegura que ambos os perfis tenham acesso às informações relevantes para decisões eficazes em tempo útil. Ao mapear requisitos analíticos às estruturas de dados disponíveis, a matriz traduz as prioridades do negócio em capacidades práticas de análise. Com isso, consolida-se como um elemento essencial para o sucesso do sistema de suporte à decisão da *Perifericum*.

3.3. Definição e Caracterização de Dimensões

As dimensões do *Data Mart* de Vendas da *Perifericum* foram definidas para suportar a área de decisão de vendas. Cada uma foi escolhida com base nas informações disponíveis nos registos de vendas e nas necessidades analíticas identificadas. A figura abaixo apresenta uma visão geral das dimensões, suas descrições e tipos:

Dimensões do Data Mart de Vendas			
Nr	Identificação	Descrição	Esquema (Tipo)
1	Tempo	Esta é a dimensão temporal. Acolhe todos os atributos que sustentem análises ao longo do tempo, como data, mês, feriado, etc.	Dim-Tempo (normal)
2	Cliente	Identificação e caracterização dos clientes de cada uma das lojas (física e online).	Dim-Cliente (com variação), Dim-Cliente-HST (Histórico)
3	Produto	Informação sobre o catálogo geral dos produtos eletrônicos à venda nas lojas.	Dim-Produto (normal)
4	Canal	Canal da venda	Dim-Canal (degenerada)
5	Venda	Identificador da venda	Dim-Venda (degenerada)

Figura 3: As dimensões do DM Vendas

A seguir, apresenta-se a caracterização individual de cada dimensão, incluindo definição, elementos de dados, hierarquias, tipo e justificação.

3.3.1. Dimensão Tempo (Tipo: Normal)

Representa o eixo temporal das transações, permitindo análises ao longo do tempo. Inclui atributos como data completa, mês, trimestre, ano, fim de semana e feriado. As hierarquias estruturam-se em níveis como Data → Mês → Trimestre → Ano.

Caracterização de dimensão							
Identificação	Dim-Tempo						
Descrição	Calendário do ano e seus atributos						
Tipo	Normal						
Dimensão	Aproximadamente 5549 registos						
Crescimento	Cresce à medida que as vendas serão registadas						
Atributos							
Nr	Identificação	Descrição	Chave (Tipo)	Domínio (Tamanho) [Range]	V/H/P	Variação	Exemplos
1	Datald	Data do calendário	S	Data			3/26/2026
2	Mês	Número do mês	N	Inteiro			12
3	Trimestre	Número do trimestre	N	Inteiro			4
4	Ano	Número do ano	N	Inteiro			2025
5	FimDeSemana	Indicação se é ou não um dia pertencente ao fim-de-semana	N	String(1)			S
6	Feriado	Indicação se é ou não um dia feriado	N	String(1)			N
Índices							
Nr	Identificação	Índice	Tipo				
1	Datald	Primário	Único, ordenado fisicamente (clustered) de forma crescente				
Hierarquia (Ramos)							
Nr	Identificação	Esquema					
1	H1	Datald -> Mês -> Trimestre -> Ano -> ALL					
2	H2	Datald -> Feriado -> ALL					
3	H3	Datald -> FimDeSemana -> ALL					
Perfis de Utilização							
Ambos os agentes de decisão.							
Observações							
Nada a assinalar.							

Figura 4: Caracterização detalhada da dimensão Dim-Tempo.

Esta dimensão é fundamental para identificar padrões de compra em períodos específicos e analisar comportamentos sazonais. Permite, por exemplo, verificar se há mais vendas em feriados ou fins de semana e apoiar campanhas promocionais baseadas em sazonalidade.

3.3.2. Dimensão Cliente (Tipo: Com Variação)

Representa os clientes da loja. Contém atributos como profissão, nome, estado civil, sexo, data de nascimento, concelho e distrito. A hierarquia Concelho → Distrito permite análises geográficas em diferentes níveis de agregação. Sendo uma dimensão com variação (de tipo 4), mantém o histórico de alterações numa tabela auxiliar (Cliente-HST).

Caracterização de dimensão							
Identificação	Dim-Cliente						
Descrição	Clientes da empresa e seus atributos						
Tipo	Com variação (Slowly Changing Dimension)						
Dimensão	Aproximadamente 1899 registos						
Crescimento	Cresce à medida que as vendas serão registadas						
Atributos							
Nr	Identificação	Descrição	Chave (Tipo)	Domínio (Tamanho) [Range]	V/H/P	Varição	Exemplos
1	Clienteld	Código interno para a identificação do cliente.	S	Inteiro			1
2	Profissão	Profissão do cliente.	N	String(75)	V/H/S		Engenheiro Informático
3	Nome	Nome do cliente.	N	String(75)			João Silva
4	Sexo	Sexo do cliente.	N	String(1)			F
5	Distrito	Distrito onde habita o cliente.	N	String(75)	V/H/S		Braga
6	Concelho	Concelho onde habita o cliente.	N	String(75)	V/H/S		Guimarães
7	EstadoCivil	Estado civil do cliente.	N	String(75)	V/H/S		Casado
8	DataNascimento	Data de nascimento do cliente.	N	Data			2/5/2003
Índices							
Nr	Identificação	Índice	Tipo				
1	Clienteld	Primário	Único, ordenado fisicamente (clustered) de forma crescente				
Hierarquia (Ramos)							
Nr	Identificação	Esquema					
1	H1	Clienteld ->Sexo -> ALL					
2	H2	Clienteld -> Concelho -> Distrito -> ALL					
3	H3	Clienteld ->DataNascimento -> ALL					
4	H4	Clienteld -> EstadoCivil -> ALL					
5	H5	Clienteld ->Profissão -> ALL					
Perfis de Utilização							
Ambos os agentes de decisão.							
Observações							
Nada a assinalar.							

Figura 5: Caracterização detalhada da dimensão Dim-Cliente.

Esta dimensão é essencial para a caracterização de perfis de clientes, permitindo segmentações baseadas em dados demográficos e geográficos. Suporta recomendações personalizadas como “produtos para jovens de Braga” ou “ofertas para professores”.

3.3.3. Dimensão Produto (Tipo: Normal)

Refere-se aos produtos vendidos, com atributos como marca e categoria. As hierarquias permitem análises agregadas por categoria ou marca.

Caracterização de dimensão							
Identificação	Dim-Produto						
Descrição	Produtos da empresa e seus atributos						
Tipo	Normal						
Dimensão	Aproximadamente 175 registos						
Crescimento	Cresce à medida que as vendas serão registadas						
Atributos							
Nr	Identificação	Descrição	Chave (Tipo)	Domínio (Tamanho) [Range]	V/H/P	Variação	Exemplos
1	Produtoid	Código interno para a identificação do produto.	S	Inteiro			1
2	Marca	Marca do produto.	N	String(75)			Logitech
3	Nome	Nome do produto.	N	String(75)			Logitech G305
4	Categoria	Categoria do produto.	N	String(75)			Periférico
Índices							
Nr	Identificação	Índice	Tipo				
1	Produtoid	Primário	Único, ordenado fisicamente (clustered) de forma crescente				
Hierarquia (Ramos)							
Nr	Identificação	Esquema					
1	H1	Produtoid ->Marca -> ALL					
2	H2	Produtoid ->Categoria -> ALL					
Perfis de Utilização							
Ambos os agentes de decisão.							
Observações							
Nada a assinalar.							

Figura 6: Caracterização detalhada da dimensão Dim-Produto.

Esta dimensão permite identificar padrões de consumo, como a preferência por produtos da categoria “Periféricos”. Também suporta recomendações baseadas em compras associadas, como sugerir teclados para quem compra ratos.

3.3.4. Dimensão Canal (Tipo: Degenerada)

Identifica se a transação ocorreu online ou na loja física. Trata-se de uma dimensão constituída por apenas um atributo, armazenada diretamente na tabela de factos, sem hierarquias.

Caracterização de dimensão							
Identificação	Dim-Canal						
Descrição	Canal das vendas da empresa						
Tipo	Degenerada						
Dimensão	Aproximadamente 2 registos						
Crescimento	Cresce à medida que as vendas serão registadas						
Atributos							
Nr	Identificação	Descrição	Chave (Tipo)	Domínio (Tamanho) [Range]	V/H/P	Variação	Exemplos
1	Canal	Canal da venda emitido pela loja ao cliente.	S	String(75)			Online
Índices							
Nr	Identificação	Índice	Tipo				
1	Canal	Primário	Único, ordenado fisicamente (clustered) de forma crescente				
Hierarquia (Ramos)							
Nr	Identificação	Esquema					
1	H1	Canal -> ALL					
Perfis de Utilização							
Ambos os agentes de decisão.							
Observações							
Nada a assinalar.							

Figura 7: Caracterização detalhada da dimensão Dim-Canal.

Apesar de simples, é crítica para distinguir padrões de compra por canal. Por exemplo, permite saber se um cliente prefere comprar online ou presencialmente, o que influencia estratégias de marketing e personalização de ofertas.

3.3.5. Dimensão Venda (Tipo: Degenerada)

Esta dimensão representa o identificador único de cada transação de venda, permitindo o rastreio individual de cada evento de venda no sistema.

É composta por um único atributo — o identificador único da venda. Não possui hierarquias ou atributos adicionais, sendo do tipo degenerada, ou seja, armazenada diretamente na Tabela de Factos.

Caracterização de dimensão							
Identificação	Dim-Venda						
Descrição	Identificador das vendas da empresa						
Tipo	Degenerada						
Dimensão	Aproximadamente 31748 registos						
Crescimento	Cresce à medida que as vendas serão registadas						
Atributos							
Nr	Identificação	Descrição	Chave (Tipo)	Domínio (Tamanho) [Range]	V/H/P	Variação	Exemplos
1	Venda	Identificador da venda emitido pela loja ao cliente.	S	Inteiro			1
Índices							
Nr	Identificação	Índice	Tipo				
1	Venda	Primário	Único, ordenado fisicamente (clustered) de forma crescente				
Hierarquia (Ramos)							
Nr	Identificação	Esquema					
1	H1	Venda -> ALL					
Perfis de Utilização							
Ambos os agentes de decisão.							
Observações							
Nada a assinalar.							

Figura 8: Caracterização detalhada da dimensão Dim-Venda

A dimensão Venda é essencial para garantir a unicidade de cada transação e possibilitar a navegação detalhada (*browsing*) entre eventos de venda. Embora não forneça contexto descritivo, é crucial para operações analíticas que exigem o rastreio de transações específicas, como auditorias, análises de outliers, ou investigações sobre compras incomuns. Por ser um identificador simples e direto, o tipo degenerado é apropriado.

3.3.6. Considerações finais

A escolha destas cinco dimensões assegura a cobertura das necessidades analíticas do projeto. As dimensões normais (Tempo e Produto) oferecem estabilidade para análises recorrentes; a dimensão com variação (Cliente) permite realizar o rastreio de mudanças ao longo do tempo; e as degeneradas (Canal e Venda) garantem a rastreabilidade da transação. Juntas, oferecem uma base sólida para os objetivos centrais da *Perifericum*.

3.4. Definição e Caracterização das Tabelas de Factos e Respetivos Grãos

A tabela de factos do *Data Mart* de Vendas foi concebida para suportar diretamente a área de decisão do projeto, conforme identificado na matriz de decisão. Esta tabela consolida as transações realizadas, agregando as informações essenciais para análises detalhadas e personalizadas. A figura seguinte apresenta uma visão geral da tabela de factos TF-Venda, incluindo atributos de dimensões, medidas e índices.

Povoamento	Realizado diariamente, entre a uma e as sete horas da manhã, iniciando-se, de preferência a sua execução às duas da manhã.					
Dimensão Inicial	1KR (234KB), após primeiro povoamento.					
Crescimento	10%/mês					
Período de dados	Os 2 últimos anos de vendas de produtos. Os dados de anos anteriores ficarão em arquivo.					
Atributos						
Dimensões						
Nr	Identificação	Chave	Tipo	Domínio	Descrição	Exemplos
1	Canal	S	D	String(75)	Identificação do canal da venda.	Loja
2	DataId	S	N	Data	Data em que a venda foi realizada.	2/3/2021
3	ClientId	S	V	Inteiro	Código interno para o cliente da "Perifericum".	213
4	ProdutId	S	N	Inteiro	Código interno para o produto da "Perifericum".	123
5	VendId	S	D	Inteiro	Identificação da venda.	123
Medidas						
Nr	Identificação	Domínio	Tipo (Função)	Descrição	Exemplos	
1	Quantidade	Inteiro	A (sum)	Número de exemplares vendidos do produto.	3	
2	PreçoUnitário	Decimal(19,2)	N	Preço unitário do produto.	20.21	
3	PorcentagemDesconto	Decimal(19,2)	N	Valor percentual do desconto efetuado.	30	
4	ValorTotal	Decimal(19,2)	A (sum)	Valor líquido da venda.	12	
Índices						
Nr	Identificação	Índice	Tipo			
1	(Canal, VendId, DataId, ClientId, ProdutId)	Primário	Único, ordenado fisicamente (clustered) de forma crescente			
2	ClientId	Secundário	Ordenado de forma crescente			
3	ProdutId	Secundário	Ordenado de forma crescente			
4	DataId	Secundário	Ordenado de forma crescente			
Perfis de Utilização						
Ambos os agentes de decisão.						
Observações						
Todos os valores considerados nos atributos medida são em Euros (€).						

Figura 9: Caracterização detalhada da tabela de factos TF-Venda.

3.4.1. Grão da Tabela de Factos

O grão da tabela TF-Venda corresponde à menor unidade de análise disponível: cada linha representa "a **venda** de um ou mais exemplares de um **produto**, a um **cliente** específico, num determinado **canal**, numa data concreta (**tempo**)". Este grão permite análises altamente granulares, essenciais para o escopo do projeto. Por exemplo, é possível identificar padrões de compra por cliente ou por período, permitindo a definição de estratégias individualizadas.

3.4.2. Atributos de Dimensão

A tabela de factos liga-se a três das cinco dimensões definidas anteriormente, através de chaves estrangeiras:

- **DataId**: Representa a data da venda, ligada à dimensão **Dim-Tempo**. Suporta análises temporais, como sazonalidade ou impacto de feriados.
- **ClientId**: Representa o cliente que efetuou a compra, ligada à dimensão **Dim-Cliente**. Permite segmentações detalhadas por perfil demográfico ou geográfico.
- **ProdutId**: Representa o produto vendido, ligada à dimensão **Dim-Produto**. Facilita a análise de preferências e padrões de consumo.

Estas chaves estrangeiras permitem análises multidimensionais, respondendo a questões estratégicas como: "Quais são os produtos mais vendidos a jovens do Porto no fim de semana?"

Além destas, a tabela inclui duas dimensões degeneradas:

- **Canal:** Representa o canal da transação (loja física ou online). Como se trata de um dado simples e diretamente associado ao facto, não está normalizado numa dimensão separada, sendo considerado uma **dimensão degenerada**, armazenada diretamente na tabela de factos.
- **Venda:** Identificador único da transação, utilizado como chave primária da tabela de factos. Este atributo também é considerado uma **dimensão degenerada**, pois fornece contexto transaccional (ex. número da fatura), mas não apresenta estrutura própria.

A presença destas dimensões degeneradas permite preservar a rastreabilidade das transações e a simplicidade estrutural sempre que não são necessários atributos descritivos adicionais.

3.4.3. Medidas e Funções de Agregação

As medidas foram escolhidas para capturar os dados quantitativos essenciais das transações. Cada medida está associada a um tipo e função de agregação, conforme descrito a seguir:

- **Quantidade:** Número de unidades vendidas. Tipo: Inteiro. Função de agregação: Soma (sum). Permite medir volumes de vendas por produto, cliente, canal ou tempo.
- **PreçoUnitário:** Preço por unidade. Tipo: Decimal(19,2). Não aditiva. Utilizado para cálculos derivados e análises de política de preços.
- **PercentagemDesconto:** Desconto percentual aplicado à venda. Tipo: Decimal(19,2). Não aditiva. Indica promoções ou campanhas ativas.
- **ValorTotal:** Valor líquido da venda. Tipo: Decimal(19,2). Função de agregação: Soma (sum). Permite avaliar o volume de faturação por diferentes dimensões.

As medidas aditivas (Quantidade e ValorTotal) suportam análises agregadas, como total de vendas por cliente ou por período. As medidas não aditivas (PreçoUnitário e PercentagemDesconto) oferecem contexto para interpretar variações de comportamento, influenciando estratégias de marketing e recomendação.

3.4.4. Justificação Geral

A tabela de factos TF-Venda foi desenhada como uma tabela transaccional, captando cada venda individual com alto nível de detalhe. Esta abordagem garante flexibilidade para análises cruzadas, suporte à segmentação de clientes e fundamentação sólida para recomendações personalizadas. As ligações com as dimensões permitem responder a perguntas críticas para a área de decisão, como:

- “Quais são os produtos mais comprados por clientes de uma determinada profissão?”
- “Em que períodos do ano se registam picos de vendas com desconto?”
- “Que tipo de produtos têm maior aceitação no canal online?”

A combinação entre o grão definido, as dimensões e as medidas assegura que o *Data Mart* está preparado para apoiar as decisões estratégicas da *Perifericum*.

3.5. Configuração dos Esquemas

O esquema global do armazém de dados da *Periféricum* foi concebido para responder às necessidades analíticas do projeto, estruturando a informação de forma organizada, acessível e eficiente. Neste momento, o armazém de dados é composto por um único *Data Mart* — o *Data Mart* de Vendas — configurado segundo um esquema em estrela (*star schema*), conforme ilustrado na Figura 10.

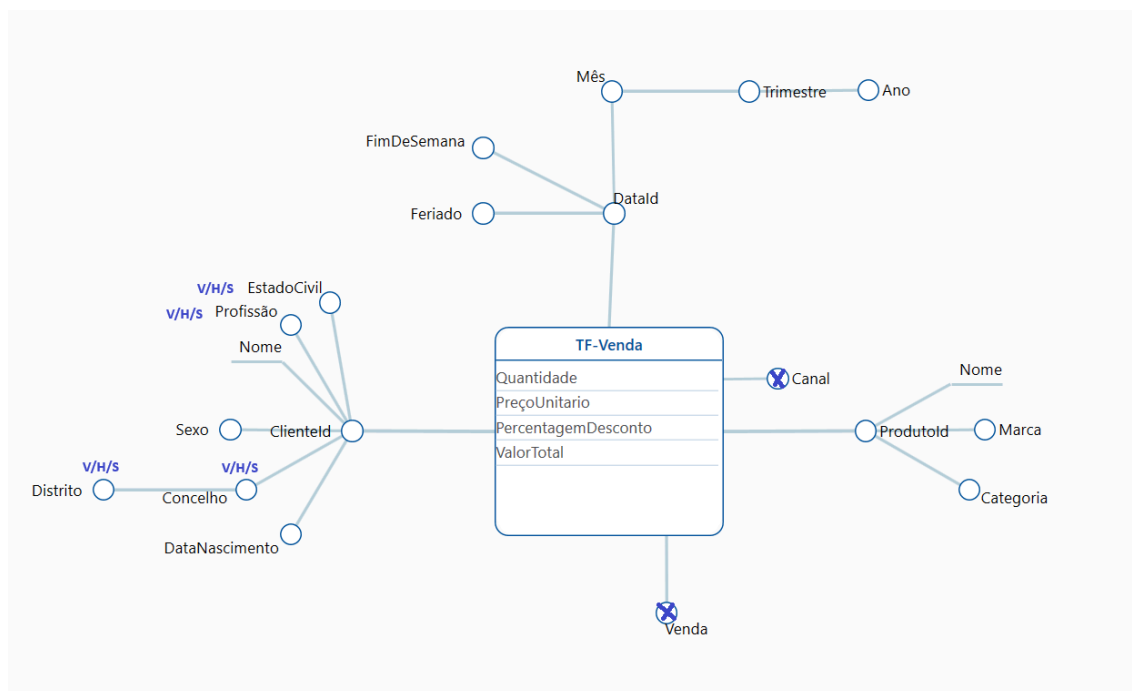


Figura 10: Esquema Dimensional do *Data Mart* de Vendas da *Periféricum*.

O esquema em estrela foi escolhido por ser uma estrutura simples, de alto desempenho (dada a estrutura não normalizada) e amplamente adotada em ambientes analíticos. A sua organização é centralizada numa tabela de factos — TF-Venda — que agrega as medidas quantitativas das transações, ligando-se diretamente a tabelas de dimensões que fornecem o contexto necessário para análises multidimensionais.

A tabela de factos TF-Venda contém os dados das transações de venda, nomeadamente as medidas Quantidade, PreçoUnitário, PercentagemDesconto e ValorTotal. Cada linha da tabela representa a venda de um ou mais exemplares de um produto, efetuada por um cliente, através de um canal específico, numa data concreta. Para contextualizar estas transações, a TF-Venda está ligada às seguintes dimensões: Dim-Tempo, que permite análises por período (mês, ano, etc.); Dim-Cliente, que fornece dados sociodemográficos e geográficos do cliente; e Dim-Produto, que descreve o produto vendido em termos de marca, nome e categoria.

Além destas, duas dimensões degeneradas são armazenadas diretamente na tabela de factos: o atributo Canal, que identifica se a venda foi feita online ou em loja física, e o identificador único da venda (Venda), que funciona como chave primária da tabela de factos e permite rastrear a granularidade da transação sem necessidade de atributos adicionais. Estas dimensões são consideradas degeneradas por não justificarem uma tabela de dimensão autónoma, dado que os seus atributos são atómicos e de valor descritivo direto.

A escolha do esquema em estrela e da ligação direta entre a tabela de factos e as suas dimensões permite uma organização lógica e eficiente dos dados, otimizando a execução de consultas analíticas complexas. A estruturação neste formato reduz a complexidade das junções entre tabelas, melhorando

o desempenho de consultas e oferece elevada flexibilidade para análises agregadas ou detalhadas, essenciais para o objetivo da empresa.

Em termos de configuração global, o armazém de dados da *Perifericum* foi planeado com uma arquitetura modular e escalável (através da natureza *bottom-up*), permitindo a adição futura de novos *Data Marts* orientados a outras áreas funcionais da organização, como Inventário ou Marketing. A decisão de iniciar com um único *Data Mart* responde à necessidade de garantir uma implementação incremental e controlada, focada nas prioridades estratégicas identificadas, sem comprometer a capacidade de crescimento do sistema.

Em suma, a configuração atual do armazém de dados, assente num esquema em estrela centrado no *Data Mart* de Vendas, garante simplicidade e desempenho, servindo eficazmente os objetivos do projeto e criando uma base sólida para futuras expansões analíticas.

3.6. As Vistas dos Agentes de Decisão

No contexto empresarial da *Perifericum*, foram identificados dois agentes de decisão com responsabilidades distintas mas complementares: o Sr. João, responsável pela gestão estratégica e experiência do cliente, e o Sr. Gates, responsável pelas operações e pelo marketing. Com base na estrutura dimensional do *Data Mart* de Vendas, foram concebidas vistas específicas para cada agente, adaptadas aos seus objetivos analíticos e operacionais.

Ambas as vistas são construídas a partir da tabela de factos e das dimensões associadas, permitindo a exploração dos dados segundo diferentes perspetivas analíticas.

Agente de Decisão	Foco e Objetivo Principal	Dimensões Relevantes	Métricas de Interesse	Exemplos de Perguntas
<i>Sr. João</i> (<i>Gestão Estratégica e UX</i>)	Compreender o comportamento dos clientes e impulsionar a fidelização e as vendas cruzadas.	Cliente, Produto, Tempo, Canal	ValorTotal, Quantidade, PreçoUnitário, Porcentagem-Desconto	<ul style="list-style-type: none"> – Que categorias compram mais os clientes? – Qual a marca de produtos mais procurada pelos clientes? – Como varia o valor médio de compra por grupo etário?
<i>Sr. Gates</i> (<i>Operações e Marketing</i>)	Otimizar campanhas promocionais (descontos) e a performance dos canais de venda.	Tempo, Canal, Produto	ValorTotal, Quantidade, Porcentagem-Desconto	<ul style="list-style-type: none"> – Qual foi o desempenho das promoções durante o último fim de semana? – Qual canal gerou mais receita na campanha de Natal? – Que produtos têm melhor resposta promocional?

Tabela 1: Agentes de Decisão e Vistas Correspondentes

A vista associada ao Sr. João privilegia a compreensão do comportamento dos clientes. Permite segmentar clientes por perfil de consumo, analisar tendências de compra ao longo do tempo e avaliar o impacto da personalização de ofertas, contribuindo para estratégias de fidelização e vendas cruzadas.

Por outro lado, a vista atribuída ao Sr. Gates foca-se na eficiência operacional e no retorno das ações de marketing. Permite monitorizar o desempenho de campanhas promocionais, comparar resultados entre canais de venda e avaliar a resposta dos clientes a descontos aplicados.

Estas vistas estão diretamente alinhadas com o esquema dimensional, tirando partido das ligações entre a tabela de factos e as suas dimensões para possibilitar análises multidimensionais rápidas e relevantes. Ao fornecer respostas a perguntas operacionais e estratégicas através de *dashboards*, estas vistas garantem que os agentes de decisão tenham acesso à informação necessária para tomar decisões informadas, coerentes com os objetivos empresariais.

4. Arquitetura Geral do Sistema

4.1. Apresentação Geral

A arquitetura do sistema de suporte à decisão da *Perifericum* foi desenhada para potenciar a gestão de clientes e personalizar ofertas, enfrentando os desafios da competitividade no setor tecnológico. Assente num armazém de dados modular e numa camada analítica avançada, oferece escalabilidade, orientação a dados e suporte eficaz à tomada de decisões, tanto operacionais como estratégicas.

A solução estrutura-se em quatro áreas principais: Integração, Armazenamento, Análise Preditiva e Visualização de dados. Estas componentes articulam-se de forma coesa, permitindo transformar e consolidar dados dispersos em informação útil e acionável.

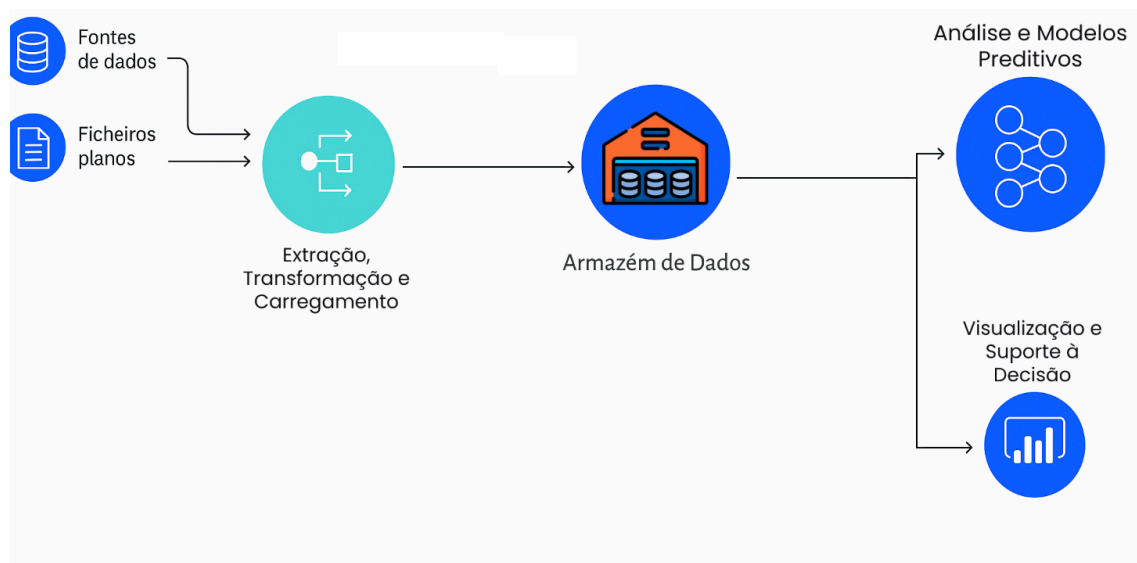


Figura 11: Esquema geral da arquitetura do sistema de suporte à decisão da *Perifericum*.

4.1.1. Integração de Dados

Esta camada é responsável pela extração, transformação e carregamento (ETL) dos dados provenientes das diversas fontes operacionais:

- **Apache NiFi**: Plataforma de ETL utilizada para consolidar dados brutos (estruturados, semi-estruturados ou não estruturados) em formatos coesos. É responsável por:
 - Processamento e extração em lote (batch), com periodicidade semanal.
 - Transformações (e.g., *source-to-target map*).
 - Carregamento incremental no *Data Warehouse*.

4.1.2. Armazenamento de Dados

A base de dados segue um modelo dimensional em estrela (*star schema*), desenhado para otimizar consultas analíticas eficientes e suportar histórico evolutivo dos clientes.

- **MySQL:** Armazena o *Data Mart* de Vendas com uma tabela de factos (TF_Venda) e dimensões (Dim_Cliente, Dim_Produto, Dim_Tempo, Dim_Canal e Dim_Venda).
- **SCD Tipo 4 – Histórico de Clientes:** A dimensão Dim_Cliente inclui uma tabela auxiliar, para suportar todas as mudanças ao longo do tempo.
- **Escalabilidade:** O modelo foi projetado para permitir a adição de novas dimensões (e.g., Dim_Localizacao) sem impactar o desempenho, graças à separação clara entre factos e dimensões.

4.1.3. Visualização e Suporte à Decisão

Apresenta os dados aos utilizadores finais (agentes de decisão) de forma visual e interativa.

- **Power BI:** Plataforma de Business Intelligence utilizada para criar *dashboards* com indicadores-chave (KPIs) como taxa de conversão, vendas por canal, valor médio por segmento.

4.1.4. Análise e Modelos Preditivos

Permite a aquisição de conhecimento a partir de padrões de consumo e perfis de clientes.

- **Pandas:** Manipulação e agregação de dados para análises descritivas e preparação dos dados para aprendizagem por máquina.
- **Scikit-learn:** Utilizado para clustering (e.g., K-means) de clientes com base em variáveis como idade, ticket médio e frequência de compra.

4.1.5. Fluxo Integrado

O sistema opera de forma cíclica e integrada:

1. A pipeline ETL (Apache NiFi) integra dados de diferentes fontes e carrega-os para o armazém de dados (MySQL).
2. Os dados são armazenados e organizados numa estrutura dimensional e os históricos de clientes são registados.
3. São executadas rotinas analíticas que produzem clusters e recomendações (scikit-learn e funções auxiliares).
4. Paralelamente, os dados provenientes do DW podem ser visualizados através de *dashboards* (Power BI), alimentados por consultas diretas ao *Data Warehouse*.

A arquitetura proposta permite à *Perifericum* transformar dados operacionais em vantagens estratégicas. A modularidade das áreas de trabalho, a integração com ferramentas de aprendizagem por máquina e a capacidade de expansão futura tornam esta solução adequada ao contexto atual e preparada para futuras exigências de crescimento e sofisticação analítica.

4.2. Fontes de Dados

O sistema de suporte à decisão integra múltiplas fontes de dados heterogêneas, provenientes das operações da loja, com diferenças em formato, estrutura e nomenclatura. Foram criadas e utilizadas tabelas de equivalência para mapear clientes e produtos, garantindo consistência no armazém de dados.

4.2.1. Produtos

- **Fonte:** Ficheiros JSON (`produtoscat_online.json`)
- **Número de Registos:** 175 produtos
- **Características:**

Atributo	Tipo de Dado
<i>nome</i>	String
<i>categoria</i>	String

Os produtos apresentam nomenclaturas padronizadas dentro de cada fonte de dados, porém podem variar entre fontes. Por exemplo:

- Fonte (Loja Física): “Logitech MX Master 3 Mouse”
- Fonte (Loja Online): “Rato Logitech MX Master”

A presente heterogeneidade pode resultar em diferentes nomes para o mesmo produto. Para resolver esta questão, foi utilizado o ficheiro `equivalencia_produtos.json`, que mapeia produtos provenientes de fontes heterogêneas para um mesmo identificador comum.

- **Categorias abrangidas (exemplos):** Periféricos, Áudio, Smartphones, Monitores

4.2.2. Clientes

- **Fonte:** Base de dados relacional (MySQL) clientes
- **Número de Registos:** 2000 clientes
- **Características:** Dados demográficos e de contato dos clientes.

Atributo	Tipo de Dado	Significado
<i>Nome</i>	VARCHAR(100)	Nome do cliente
<i>Profissão</i>	VARCHAR(100)	Profissão do cliente
<i>EstadoCivil</i>	VARCHAR(50)	Estado civil do cliente
<i>Sexo</i>	VARCHAR(10)	Sexo do cliente
<i>Distrito</i>	VARCHAR(100)	Distrito de residência do cliente
<i>Concelho</i>	VARCHAR(100)	Concelho de residência do cliente
<i>Telefone</i>	VARCHAR(20)	Nº de telefone do cliente
<i>Morada</i>	TEXT	Endereço de residência do cliente
<i>CódigoPostal</i>	VARCHAR(20)	Código postal do cliente
<i>DataNascimento</i>	VARCHAR(10)	Data de nascimento do cliente
<i>Email</i>	VARCHAR(150)	Endereço de correio eletrónico do cliente

4.2.3. Fontes Operacionais de Vendas

4.2.3.1. Loja Física

- **Período:** 2010 a 2020
- **Formato:** Ficheiro Excel (vendas_loja_fisica.xlsx)
- **Volume de Dados:** 20000 transações

Atributo	Tipo de Dado	Significado
<i>Data_Venda</i>	String	Data da realização da venda
<i>Produto</i>	String	Nome do produto vendido
<i>Categoria</i>	String	Categoria do produto
<i>Marca</i>	String	Marca do produto
<i>Quantidade</i>	Inteiro	Número de unidades vendidas
<i>Preco</i>	Decimal	Valor total da venda
<i>PercentagemDesconto</i>	Decimal	Percentagem de desconto aplicada
<i>Cliente_Nome</i>	String	Nome do cliente
<i>Concelho</i>	String	Concelho de residência do cliente
<i>Canal</i>	String	Canal de venda

4.2.3.2. Loja Online

- **Período:** 2021 a 2025
- **Formato:** Ficheiro JSON (vendas_loja_online.json)
- **Volume de Dados:** 20000 transações

Atributo	Tipo de Dado	Significado
<i>sale_id</i>	String	Identificador da venda
<i>date</i>	String	Data da realização da venda
<i>product_id</i>	String	Identificador do produto vendido
<i>name</i>	String	Nome do produto vendido
<i>category</i>	String	Categoria do produto
<i>brand</i>	String	Marca do produto
<i>price</i>	Decimal	Valor total da venda
<i>quantity</i>	Inteiro	Número de unidades vendidas
<i>discount</i>	Decimal	Percentagem de desconto aplicada
<i>customer_id</i>	String	Identificador do cliente
<i>email</i>	String	Email do cliente
<i>district</i>	String	Distrito de residência do cliente
<i>Channel</i>	String	Canal de venda

4.2.4. Desafios de Integração e Soluções

A diversidade das fontes apresenta desafios que foram resolvidos com tabelas de equivalência para garantir integração coerente no *Data Warehouse*.

Desafio	Solução
<i>Formatos heterogéneos (Excel, JSON, base de dados relacional)</i>	Pipeline ETL para conversão e integração de formatos
<i>Diferenças na nomenclatura dos produtos, ex. "Monitor 24 Polegadas HP" vs. "HP 24mh 24" Monitor"</i>	Tabelas de equivalência para alinhamento das denominações e IDs, ex. "Placa Gráfica NVIDIA RTX 3060" \equiv "NVIDIA GeForce RTX 3060 Graphics Card", com ProductID 67
<i>Estruturas distintas (tabelas planas vs. dados relacionais)</i>	Área de preparação de dados (<i>staging area</i>)
<i>Mapeamento de clientes entre fontes</i>	Tabelas de equivalência para nomes e identificadores únicos, ex. "João Silva" \equiv Clienteld 123

4.3. Área de Preparação de dados

A área de preparação de dados é responsável por transformar dados brutos, heterogéneos e inconsistentes em informação limpa, integrada e estruturada, pronta para ser carregada no armazém de dados. Este processo segue a arquitetura/pipeline ETL (Extração, Transformação e Carregamento), com o software Apache NiFi como motor central de integração.

4.3.1. Estrutura Geral do Processo

Os dados extraídos são inicialmente armazenados numa tabela RAW, mantendo o seu formato original. De seguida, são transformados e carregados em tabelas intermediárias RDY (e.g., produtosRDY, vendasRDY, etc.), onde já se encontram padronizados e prontos para posterior integração. A partir das tabelas RDY, os dados são direcionados para:

1. Tabelas de dimensões;
2. Tabela de factos;
- Quarentena (QUA), no caso de serem identificados registos com problemas ou marcas de validação (MV);
- Quarentena de clientes, especificamente para registos de clientes inválidos ou incompletos.

Existe ainda uma tabela de controlo, chamada *et_controlo*, que guarda a data da última execução do pipeline. Esta data é utilizada para extrair apenas os registos atualizados desde então (e.g., clientes com novos *logs*).

4.3.2. Etapas do Processo ETL

1. Extração

A fase de extração recolhe dados de quatro fontes principais (descritas na Secção 4.2).

Estas fontes apresentam diferenças relevantes ao nível da estrutura, da nomenclatura e dos identificadores (ex.: nomes vs. IDs), o que torna necessário um processo de normalização e integração cuidadoso.

2. Transformação

Nesta fase, o Apache NiFi aplica várias transformações críticas para garantir a integridade dos dados:

- **Padronização de formatos** (datas, códigos, unidades);
- **Limpeza de dados** incompletos, duplicados ou inválidos;
- **Mapeamento de identificadores e descrições**, através de tabelas de equivalência, como o ficheiro `equivalencia_produtos.json`, que associa produtos de fontes heterogéneas a um identificador comum.

3. Carregamento

Por fim, os dados transformados são inicialmente carregados nas tabelas de dimensões e, em seguida, integrados nas tabelas de factos do armazém de dados. Este processo é feito com carregamento incremental, processando apenas os dados novos ou modificados desde a última execução (controlada via *etl_controlo*). Desta forma, são asseguradas eficiência e escalabilidade, mesmo com o crescimento contínuo dos dados.

4.4. O Armazém de Dados

O armazém de dados desenvolvido baseia-se numa **arquitetura dimensional clássica**, que permite organizar, integrar e disponibilizar dados históricos relevantes para análise estratégica. Este armazém tem como objetivo centralizar a informação dispersa da empresa, otimizando a sua consulta e análise por parte dos agentes.

4.4.1. Componentes e Estrutura

O componente central do armazém é a **tabela de factos** (TF_Venda), onde são armazenados os principais indicadores quantitativos do negócio, como número de unidades vendidas, valor de venda, e descontos aplicados.

Este facto está rodeado por várias **tabelas de dimensão**, que descrevem os diferentes contextos da análise:

- Dim_Cliente: dados sociodemográficos e identificadores dos clientes;
- Dim_Produto: descrição, categoria e marca dos produtos;
- Dim_Tempo: hierarquias temporais (dia, mês, trimestre, ano);
- Dim_Canal: canais de venda utilizados (online, loja física);
- Dim_Venda: Identificador da transação associada.

Estas dimensões permitem ao utilizador analisar os factos segundo diversos eixos analíticos, promovendo a exploração dos dados em profundidade.

4.4.2. Esquema Conceptual

Descrito na Secção 3.5, o esquema segue uma arquitetura em estrela (*star schema*), favorecendo a simplicidade e o desempenho de consultas analíticas.

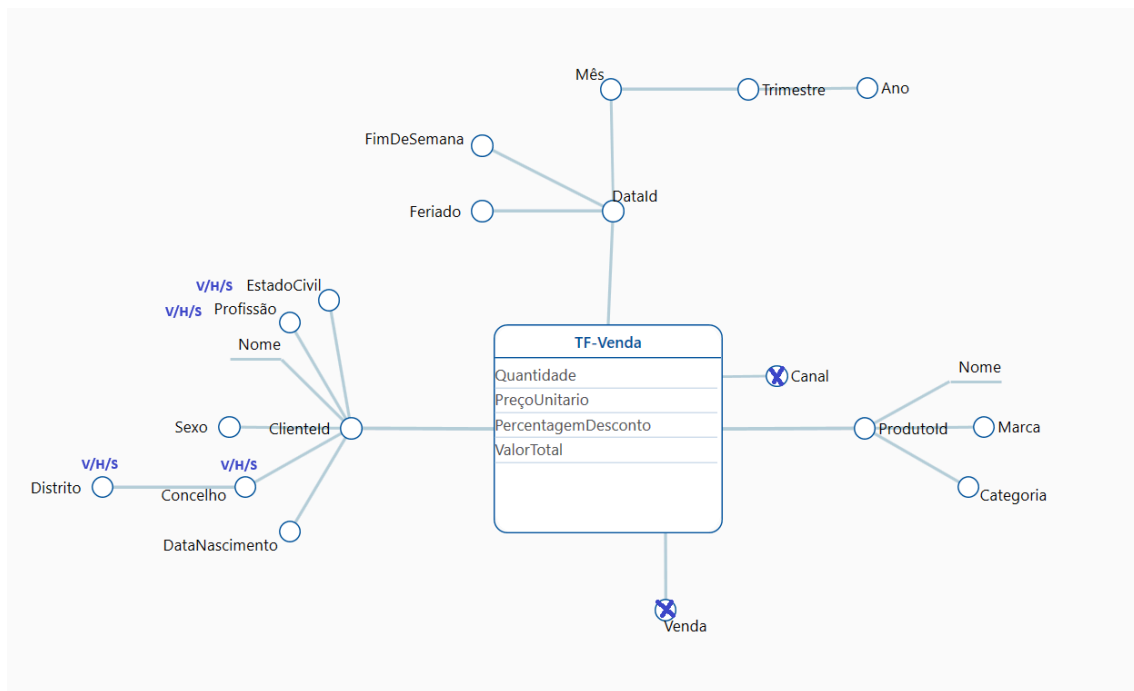


Figura 12: Esquema Dimensional do *DataWarehouse* da *Perifericum*.

4.4.3. Conversão para Esquema Lógico

A conversão para o modelo lógico formalizou a estrutura do armazém, com a definição dos tipos de dados, chaves e restrições.

1. **Identificação de entidades e atributos:** cada dimensão (não degenerada) foi modelada como uma tabela relacional, com os seus atributos devidamente tipificados (ver secção Secção 3.3).
2. **Chaves primárias e estrangeiras:** as chaves primárias das dimensões (`ClienteId`, `ProdutoId`, `DataId`) foram utilizadas como chaves estrangeiras na tabela `TF_Venda`, garantindo a integridade referencial.
3. **Conversão de dimensões degeneradas:** os atributos diretamente associados ao facto — como o número da venda (`Dim_Venda`) ou o canal da venda (`Dim_Canal`) — foram tratados como dimensões degeneradas e incorporados diretamente na tabela de factos `TF_Venda`, uma vez que não requerem uma tabela de dimensão própria. Esta abordagem simplifica o modelo e evita operações de junção desnecessárias.
4. **Tipificação de dados:**
 - **Textos:** `varchar`
 - **Números:** `int`
 - **Datas:** `date`
 - **Valores monetários:** `decimal`
5. **Regras de integridade:** restrições de integridade referencial asseguram que não existam factos sem correspondência nas dimensões.
6. **Desnormalização controlada:** o modelo em estrela mantém as tabelas de dimensão desnormalizadas, otimizando o desempenho das consultas analíticas.
7. **Criação da tabela `Dim_Cliente_Historico` (SCD Tipo 4):** Para suportar a gestão de mudanças históricas nos dados dos clientes, foi criada a tabela `Dim_Cliente_Historico`, seguindo o padrão *Slowly Changing Dimension* (SCD) Tipo 4. Esta tabela armazena o histórico completo das alterações

dos atributos dos clientes, mantendo versões anteriores dos registos, enquanto a tabela Dim_Cliente contém apenas os dados atuais.

4.4.4. Utilidade dos Componentes

- **Tabela de factos:** armazena os dados quantitativos a analisar (vendas), funcionando como o núcleo da análise.
- **Tabelas de dimensão:** fornecem contexto descritivo, permitindo segmentar e explorar os factos sob diversas perspetivas.
- **Modelo em estrela:** garante simplicidade estrutural e rapidez nas consultas.
- **Conversão lógica e integridade:** asseguram consistência e fiabilidade dos dados armazenados.
- **Desnormalização:** melhora o desempenho e reduz a complexidade das consultas analíticas.

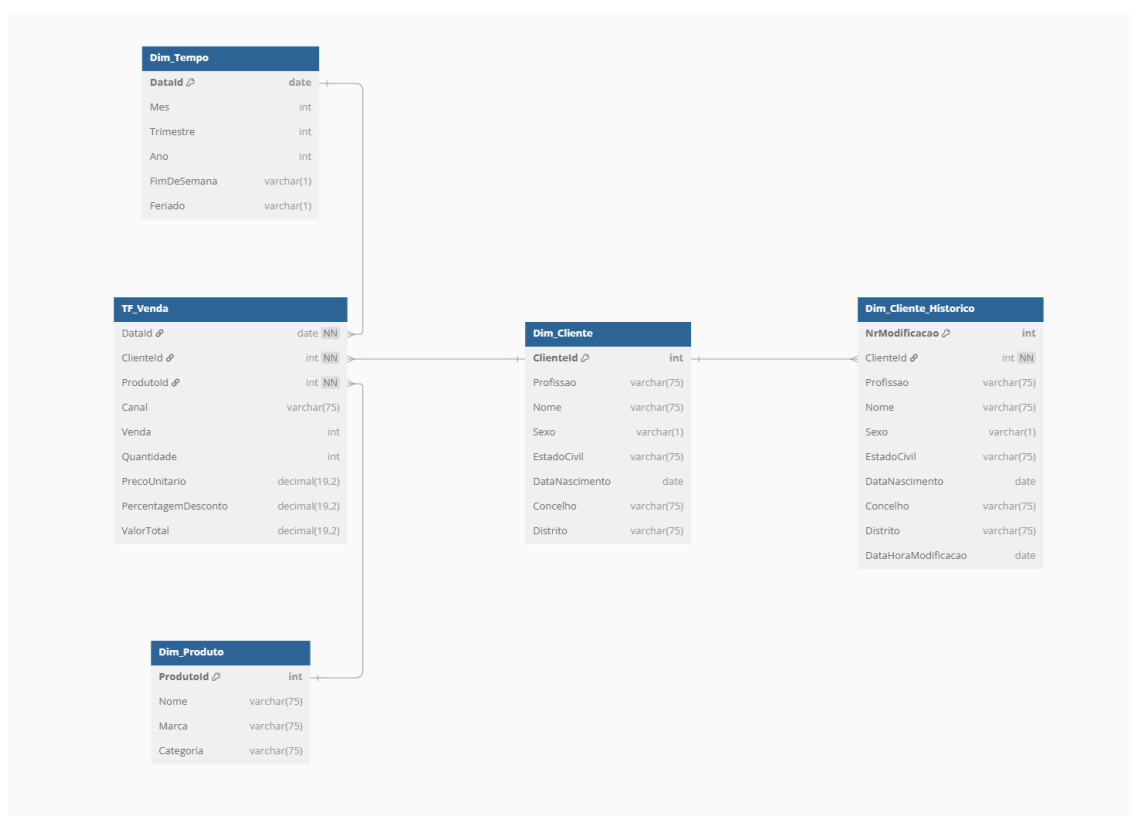


Figura 13: Esquema Lógico do DataWarehouse da Perifericum.

O modelo lógico garante robustez, desempenho e flexibilidade no suporte à análise multidimensional de dados. A clara separação entre factos e dimensões permite agregar indicadores, aplicar filtros e efetuar análises temporais ou comportamentais com facilidade.

4.5. Exploração e Visualização de Dados

A conceção e estruturação dos dashboards para o sistema foram orientadas por uma lógica centralizada nas vistas e responsabilidades dos principais agentes de decisão: o Sr. João, com um **foco estratégico** e de gestão de clientes, e o Sr. Gates, com **responsabilidades operacionais** e de marketing (Secção 3.6). Esta diferenciação funcional esteve na base do desenho dos dashboards, permitindo que fosse acedido de forma intuitiva e eficiente às análises mais relevantes para as respetivas funções.

Para garantir alinhamento entre os dashboards e os objetivos de negócio, o processo iniciou-se com uma identificação clara das decisões típicas tomadas por cada perfil e das métricas que melhor as poderiam suportar. A partir daí, foi definida uma estrutura modular de dashboards, agrupando visualizações segundo eixos analíticos distintos: **Dashboards para Gestão estratégica e experiência do cliente** e **Dashboards para Operações e Marketing**.

Para garantir que os painéis refletissem sempre a realidade atual do negócio, foi estabelecida uma ligação direta entre o Power BI e o armazém de dados. Esta abordagem permite que as consultas realizadas nas visualizações sejam executadas em tempo real sobre o motor de base de dados (MySQL), assegurando que as métricas apresentadas correspondem aos dados mais recentes carregados no sistema.

A sincronização dos dados é assegurada por processos de ETL semanalmente, orquestrados pelo Apache NiFi. Estes processos garantem que os dados transacionais provenientes da loja física e da plataforma online são extraídos, transformados e carregados de forma incremental para o armazém de dados, respeitando as regras de qualidade e integridade definidas na área de preparação. Assim, cada nova semana de operação é refletida automaticamente nos dashboards, sem necessidade de intervenção manual.

4.6. Aquisição de Conhecimento

A aquisição de conhecimento representa uma etapa central na construção de um sistema de suporte à decisão orientado para a personalização da experiência do cliente e otimização das estratégias comerciais. A sua implementação tem como objetivo sustentar os processos de caracterização de perfis de clientes (Secção 7) e de personalização de ofertas de produtos (Secção 8), permitindo que a loja se torne mais proativa, eficiente e centralizada no cliente.

4.6.1. Arquitetura do Sistema de Aquisição de Conhecimento

O sistema de aquisição de conhecimento será suportado por uma infraestrutura já delineada no *Data Warehouse*, que integra dados de múltiplas fontes, devidamente processados através de pipelines ETL. A partir desta base sólida, o sistema vai articular-se em três níveis principais:

1. Extração de Conhecimento via Segmentação de Clientes

A partir dos dados demográficos, comportamentais e transacionais disponíveis no DW, o sistema aplica técnicas de clustering (nomeadamente K-Means) para segmentar os clientes em grupos homogéneos. Estes clusters refletem padrões distintos de consumo, como frequência de compra, tipos de produtos adquiridos, canais preferenciais e montantes gastos. Foram identificados perfis como “Cliente dispendioso e volátil” ou “Comprador frequente e diversificado”, cuja análise permite fundamentar ações personalizadas.

2. Geração de Conhecimento para Recomendação Personalizada

Com base na segmentação realizada, o sistema aplica algoritmos de recomendação, distinguindo entre clientes novos (sem histórico) e existentes. Para os primeiros, utiliza-se uma abordagem baseada em popularidade (demographic filtering), enquanto para os segundos recorre-se a técnicas mais sofisticadas como content-based filtering e collaborative filtering. Em casos específicos, combina-se ambas as estratégias numa abordagem híbrida.

3. Ciclo Iterativo de Aprendizagem e Atualização

O sistema está desenhado para funcionar de forma contínua e adaptativa. A cada nova interação ou transação registada, os dados são incorporados no DW e alimentam os modelos analíticos, permitindo reavaliar perfis, ajustar recomendações e refinar as estratégias comerciais. A análise da evolução temporal do comportamento de compra (tendência de gastos, variação mensal, entre outros) garante a atualização constante do conhecimento e melhora a precisão das previsões.

Em termos de **Enquadramento Tecnológico e Funcional**, a plataforma de aquisição de conhecimento articula-se com ferramentas como Python (para análise e modelação preditiva) e Scikit-learn (para clustering e filtragem).

Este sistema permite, assim, transformar dados limpos do armazém em conhecimento acionável, promovendo uma atuação mais eficaz da empresa na captação, retenção e fidelização dos seus clientes — alinhando a estratégia empresarial com as expectativas individuais de cada consumidor.

5. O Povoamento de Dados

5.1. Apresentação da Abordagem Realizada

O processo de integração de dados para o povoamento do DW foi desenvolvido com base na metodologia clássica ETL (*Extract, Transform, Load*), amplamente utilizada em sistemas de apoio à decisão. Esta abordagem permite estruturar o fluxo de dados desde as fontes de origem até ao armazém de dados final, assegurando consistência, qualidade e rastreabilidade.

5.1.1. Frequência e Estratégia de Carregamento

A integração é realizada em lote (*batch*), com uma frequência semanal, e segue uma lógica de carregamento incremental — ou seja, apenas os dados novos ou modificados desde a última execução são processados. Este mecanismo incremental garante eficiência e escalabilidade, evitando o reprocessamento desnecessário de grandes volumes de dados. A data da última execução é registada numa tabela de controlo (*etl_controlo*), sendo utilizada como referência para identificar os dados a extrair.

5.1.2. Ferramenta de Implementação

A implementação prática deste processo ETL foi concretizada com recurso ao Apache NiFi, uma ferramenta robusta de orquestração de pipelines de dados. O NiFi foi utilizado tanto na extração como no encaminhamento dos dados transformados para o DW. A sua interface gráfica permitiu modelar, monitorizar e automatizar os fluxos de integração de forma eficaz e escalável.

5.1.3. Representação do Processo

Para documentar e comunicar o funcionamento do processo de integração, foi adotada a notação BPMN (*Business Process Model and Notation*). Esta notação gráfica facilita a visualização e o entendimento dos diferentes passos do fluxo de dados, promovendo a transparência e a partilha de conhecimento entre as equipas técnicas e funcionais.

5.1.4. Lógica de Integração

A lógica de integração baseia-se no carregamento incremental dos dados relevantes, extraídos das fontes e encaminhados para os componentes de transformação. Estas transformações, por exemplo, incluem normalizações e validações. Os dados transformados são então carregados primeiro nas tabelas de dimensões e, posteriormente, nas tabelas de factos do armazém de dados, seguindo a ordem lógica da modelação dimensional.

5.2. Mapeamento de Dados

O mapeamento de dados desenvolvido foi essencial para garantir a correta transição dos dados operacionais para o *Data Warehouse*, permitindo estruturar a informação de forma a apoiar a análise de vendas.

Além das fontes principais de dados (Secção 4.2), foram utilizados dois ficheiros adicionais em formato .csv para apoiar o processo de transformação e integração de dados:

- *equivalencia_produtos.csv*: utilizado para mapear os nomes dos produtos provenientes das vendas de diferentes canais, de forma a disponibilizar informações complementares, essenciais para o povoamento da dimensão DimProduto;
- *equivalencia_clientes.csv*: permite associar os nomes ou identificadores dos clientes presentes nas fontes de vendas físicas e online.

De forma a unificar estas informações no armazém de dados, foi necessário definir um *source-to-target datamap*, que identificasse como cada campo das fontes deveria ser tratado, transformado e carregado no DW.

Origem (vendas_loja_fisica.xlsx)	Transformações	Destino (Dim-Produto)
<i>Produto</i>	Pesquisa no ficheiro <i>equivalencia_produtos.json</i> para obter <i>Produtold</i> , remover prefixo “P” e converter para inteiro	<i>Produtold</i>
<i>Produto</i>	Sem transformação	Nome
<i>Categoria</i>	Sem transformação	Categoria
<i>Marca</i>	Sem transformação	Marca

Tabela 2: *source-to-target datamap*: Vendas Físicas → Dim-Produto

Origem (vendas_loja_online.xlsx)	Transformações	Destino (Dim-Produto)
<i>product_id</i>	Remover prefixo “P” e converter para inteiro	<i>Produtold</i>
<i>product_id</i>	Pesquisa no ficheiro <i>equivalencia_produtos.json</i> para obter nome do produto através de “Nome_Excel”	Nome
<i>category</i>	Sem transformação	Categoria
<i>brand</i>	Sem transformação	Marca

Tabela 3: *source-to-target datamap*: Vendas Online → Dim-Produto

Origem (Tabela relacional clientes_logs)	Transformações	Destino (Dim-Cliente)
Email	Pesquisa no ficheiro equivalencia_clientes.csv para obter id do cliente através de "Email"	Clienteld
EstadoCivil	Sem transformação	EstadoCivil
Profissão	Sem transformação	Profissão
Nome	Sem transformação	Nome
Sexo	Sem transformação	Sexo
Concelho	Sem transformação	Concelho
Distrito	Sem transformação	Distrito
DataNascimento	Convertida para formato DATE	DataNascimento

Tabela 4: source-to-target datamap: clientes_logs → Dim-Cliente

Origem (vendas_loja_fisica.xlsx)	Transformações	Destino (Dim-Tempo)
Data_Venda	Convertida para formato DATE	DataId
Data_Venda	Lookup no ficheiro feriados_nacionais2010-2025.csv com base em Data_Venda ("S" ou "N")	Feriado
Data_Venda	DAYOFWEEK(Data_Venda) Verifica se o dia da semana é sábado (6) ou domingo (7); retorna "S" ou "N"	FimDeSemana
Data_Venda	Derivado de "Data_Venda"	Mês
Data_Venda	Determinado com base no mês (1-3 = T1, 4-6 = T2, 7-9 = T3, 10-12 = T4)	Trimestre
Data_Venda	Derivado de "Data_Venda"	Ano

Tabela 5: source-to-target datamap: Vendas Físicas → Dim-Tempo

Origem (vendas_loja_online.xlsx)	Transformações	Destino (Dim-Tempo)
<i>date</i>	Sem transformação	DataId
<i>date</i>	Lookup no ficheiro feriados_nacionais2010-2025.csv com base em date ("S" ou "N")	Feriado
<i>date</i>	DAYOFWEEK(date) Verifica se o dia da semana é sábado (6) ou domingo (7); retorna "S" ou "N"	FimDeSemana
<i>date</i>	Derivado de "date"	Mês
<i>date</i>	Determinado com base no mês (1-3 = T1, 4-6 = T2, 7-9 = T3, 10-12 = T4)	Trimestre
<i>date</i>	Derivado de "date"	Ano

Tabela 6: *source-to-target datamap*: Vendas Online → Dim-Tempo

Origem (vendas_loja_online.xlsx)	Transformações	Destino (TF-Venda)
<i>VendaId</i>	auto_increment	VendaId
<i>date</i>	Sem transformação	DataId
<i>product_id</i>	Remoção prefixo "P" e converter para inteiro	ProdutoId
<i>channel</i>	Renomeação para "Canal"	Canal
<i>customer_id</i>	Renomeação para "ClientId", Remoção prefixo "C"	ClientId
<i>price</i>	Renomeação para "ValorTotal"	ValorTotal
<i>price & quantity</i>	Cálculo: price / quantity	PreçoUnitario
<i>quantity</i>	Renomeação para "Quantidade"	Quantidade
<i>discount</i>	Conversão para decimal (valor de 0 a 1), Renomeação para "PorcentagemDesconto"	PorcentagemDesconto

Tabela 7: *source-to-target datamap*: Vendas Online → TF-Venda

Origem (vendas_loja_fisica.xlsx)	Transformações	Destino (TF-Venda)
<i>Vendald</i>	auto_increment	Vendald
<i>Data_Venda</i>	Convertida para formato DATE	DataId
<i>Produto</i>	Pesquisa no ficheiro equivalencia_produtos.json para obter Produtold do produto através de "Nome_Excel"	Produtold
<i>Canal</i>	Sem transformação	Canal
<i>Cliente_Nome</i>	Pesquisa no ficheiro equivalencia_clientes.csv para obter Clienteld do produto através de "Nome"	Clienteld
<i>Preço</i>	Renomeação para "ValorTotal"	ValorTotal
<i>Preço & Quantidade</i>	Cálculo: Preço / Quantidade	PreçoUnitario
<i>Quantidade</i>	Sem transformação	Quantidade
<i>Porcentagem_Desconto</i>	Conversão para decimal (valor de 0 a 1)	PorcentagemDesconto

Tabela 8: *source-to-target datamap*: Vendas Físicas → TF-Venda

5.3. Modelação do Sistema de Povoamento

O modelo desenvolvido para o sistema de integração de dados foi concebido com base na notação BPMN e inspirado no artigo “*Using BPMN for ETL Conceptual Modelling: A Case Study*” (Oliveira et al., 2021), representando de forma abstrata e generalizada o pipeline de povoamento do armazém de dados. O foco esteve na clareza e simplicidade, priorizando a eficiência do processo e o cumprimento das dependências entre tabelas.

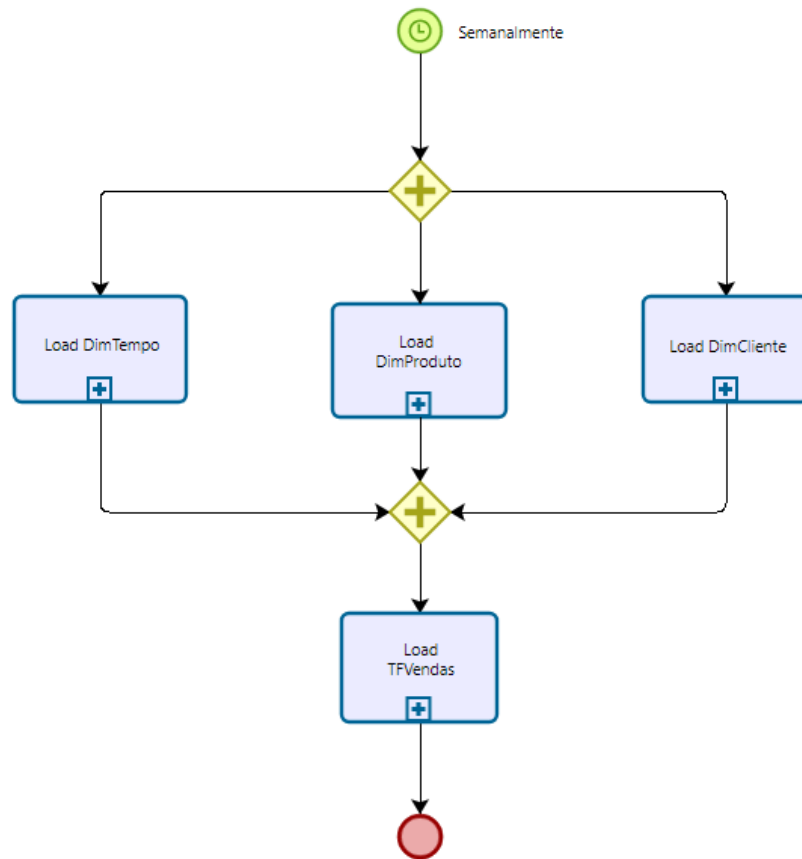


Figura 14: Modelo BPMN a nível do processo (1º nível de abstração)

O processo é composto essencialmente por duas fases:

1. **Carregamento das dimensões**
2. **Carregamento da tabela de factos**

A figura acima indica que o carregamento das diferentes tabelas dimensionais é iniciado através de atividades paralelas, permitindo que estas operações ocorram de forma independente e simultânea. Este paralelismo é justificado pela inexistência de dependência direta entre estas dimensões durante o processo de povoamento. Termina com uma *gateway* de sincronização, garantindo que todas as dimensões estejam devidamente carregadas antes do início do carregamento da tabela de factos.

Após o término do carregamento das dimensões, o processo continua com o carregamento da tabela de factos TF-Venda. A estrutura do BPMN assegura esta ordem lógica e evita problemas de integridade referencial no DW.

5.3.1. Carregamento Dim-Tempo

O carregamento da dimensão Dim-Tempo é realizado através de um subprocesso bem definido, representado na figura abaixo:

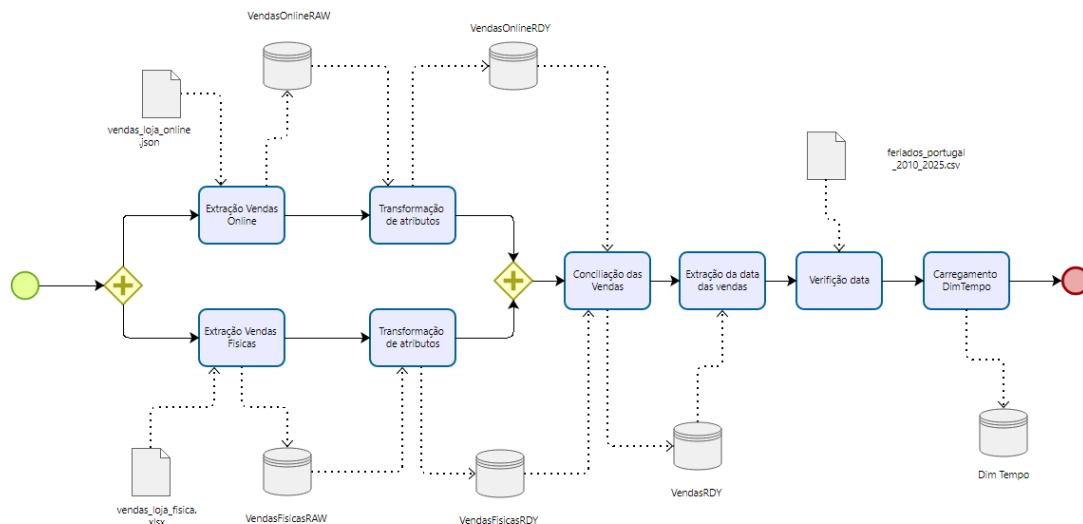


Figura 15: Subprocesso BPMN para o carregamento da Dimensão Tempo

Este subprocesso inicia-se com a extração das vendas provenientes de duas fontes distintas:

- vendas_loja_online.json (vendas online), que são armazenadas na base de dados na área preparação VendasOnlineRAW.
- vendas_loja_fisica.xlsx (vendas físicas), que são armazenadas na base de dados na área de preparação VendasFísicasRAW.

Após a extração, os dados passam por um processo de limpeza e transformação, que inclui:

- Padronização dos nomes dos atributos (por exemplo, renomear “date” para “DataId”).
- Conversão de tipos de dados (ex.: datas em texto para formato DATE).
- Remoção de atributos irrelevantes.

Os dados transformados são então armazenados nas áreas prontas para integração, VendaOnlineRDY e VendasFísicasRDY. Estas áreas intermediárias (RAW → RDY) são fundamentais para isolar o tratamento específico de cada origem e garantir que os dados estejam em formato compatível e consistente antes da sua integração.

Após o pré-processamento, os dados convergem para uma atividade de conciliação de vendas, que unifica ambas as fontes numa visão consolidada, armazenada em VendasRDY.

A partir das VendasRDY, é realizada a extração da data das vendas, ou seja, são identificadas todas as datas únicas em que ocorreram transações. Estas datas são então submetidas a uma etapa de verificação, onde é consultada uma fonte externa com os feriados nacionais (feriados_portugal_2010_2025.csv) para identificar se a data corresponde a um feriado.

Após esta validação, os dados são carregados na dimensão Dim-Tempo. Esta dimensão é enriquecida automaticamente através de triggers na base de dados, que calculam e preenchem atributos derivados como:

- Trimestre (Trimestre)
- Se a data é um fim de semana (FimDeSemana)

5.3.2. Carregamento Dimensão Produto

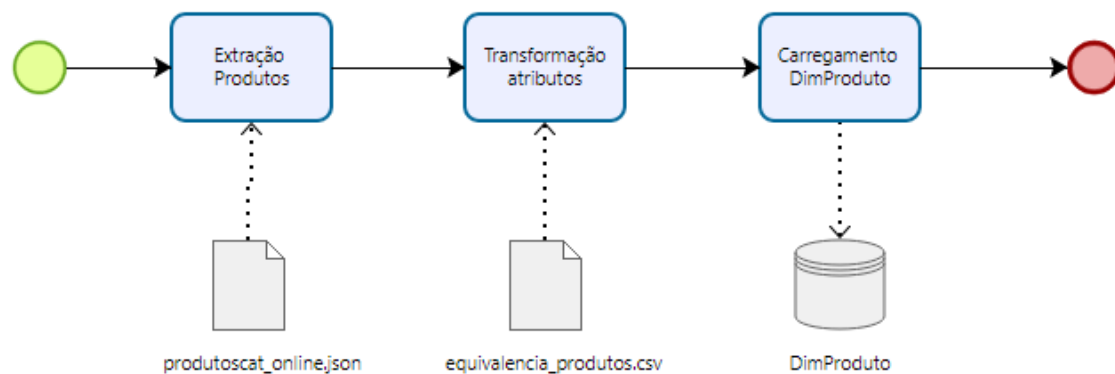


Figura 16: Subprocesso BPMN para o carregamento da Dimensão Produto

O processo de integração da Dimensão Produto inicia-se com a extração dos dados de produtos, provenientes do ficheiro `produtoscat_online.json`. Esta fonte contém as informações brutas de produtos disponibilizados na loja online.

Seguidamente, os dados extraídos são submetidos a uma transformação de atributos, que tem como objetivo padronizar e enriquecer os dados para garantir consistência e integridade. Durante essa fase, é utilizada a tabela de equivalência de produtos (`equivalencia_produtos.csv`), que permite realizar mapeamentos entre diferentes códigos e nomenclaturas de produtos, associando-os a um modelo comum. Isto é essencial para consolidar variações de nomes, sendo que produtos em loja online e física têm nomes diferentes.

Após a transformação, os dados padronizados são carregados na Dimensão Produto (Dim-Produto), que passa a conter todos os atributos relevantes de cada produto, já com os nomes uniformizados, categorias hierarquizadas e códigos equivalentes devidamente tratados.

5.3.3. Carregamento Dimensão Cliente

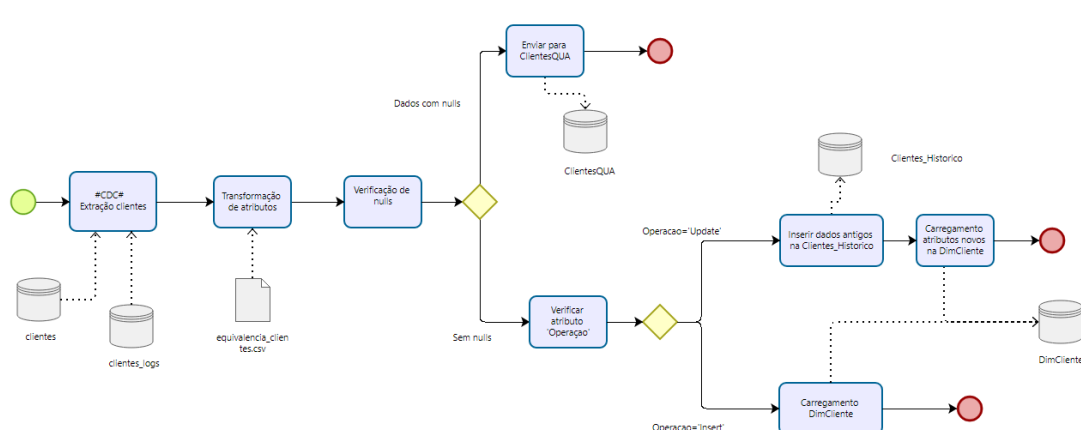


Figura 17: Subprocesso BPMN para o carregamento da Dimensão Cliente

O subprocesso de Carregamento da Dimensão Cliente tem como objetivo realizar a inserção ou atualização de registros na tabela de dimensão DimCliente, garantindo a integridade e o histórico das informações dos clientes.

O processo inicia a extração de dados da tabela `clientes_logs`, que contém os registros de alterações (logs) realizados na tabela `clientes`. Os logs são utilizados para identificar os dados novos ou modificados que precisam ser processados, através de um Change Data Capture (CDC).

Em seguida, os dados extraídos passam por uma etapa de transformação, na qual são padronizados e enriquecidos com base em uma tabela de equivalência (`equivalencia_clientes.csv`). Durante essa transformação, é verificado se todos os campos obrigatórios estão preenchidos; caso algum campo esteja vazio ou inválido, o registro é direcionado para uma tabela de quarentena (`ClientesQUA`), impedindo o seu carregamento na dimensão até que os dados sejam corrigidos. Essa etapa assegura que apenas informações completas e padronizadas sejam carregadas na dimensão.

Após a transformação, o processo verifica o tipo de operação registrada nos logs, utilizando o atributo “Operacao”, que pode ser:

- **Insert:** Indica um novo cliente. O registro é considerado novo e é carregado diretamente na tabela `Dim-Cliente`.
- **Update:** Indica que houve uma alteração num cliente existente. Antes de atualizar a dimensão, os dados anteriores são armazenados na tabela `Clientes_Historico`, preservando o histórico das alterações. Finalmente, os atributos atualizados são carregados na `Dim-Cliente`.

5.3.4. Carregamento Tabela de factos Vendas

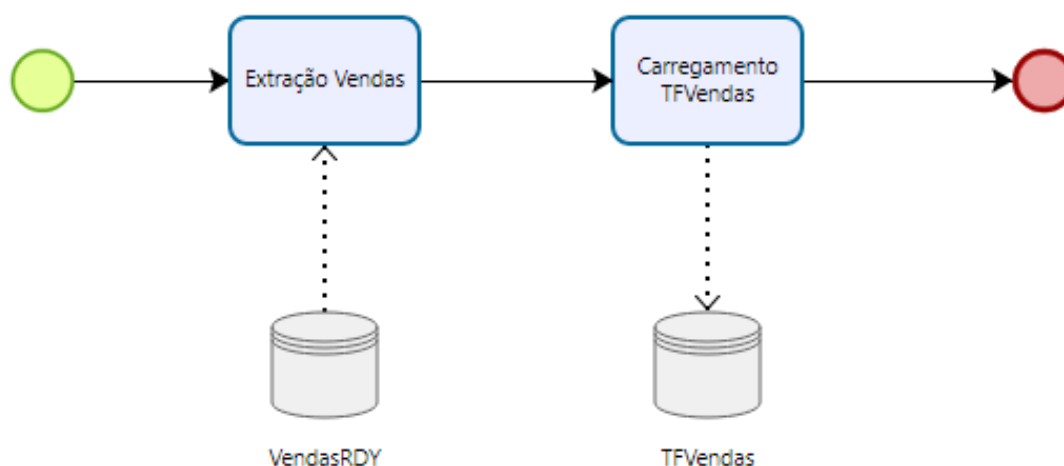


Figura 18: Subprocesso BPMN para o carregamento tabela de factos Vendas

O processo tem início com a extração dos dados provenientes da tabela `VendasRDY`. Esta tabela já foi previamente preparada num subprocesso, no qual os dados de vendas foram validados, tratados e deixados prontos para carregamento.

Os dados extraídos são então carregados na tabela de factos `TFVendas`.

5.4. Implementação do Sistema de Povoamento

A implementação do sistema de povoamento foi inspirada na modelação BPMN previamente definida, integrando diversas fontes de dados e tecnologias de processamento para garantir a extração, transformação e carregamento (ETL) dos dados.

Desde o início, foi incorporado um mecanismo de controlo através da tabela *etl_controlo*, que armazena a data da última extração realizada com sucesso. Esta tabela é consultada no início de cada execução para garantir que apenas os dados modificados ou adicionados após a última extração sejam processados. Com isso, o sistema evita retrabalho, melhora o desempenho e assegura que o processo ETL seja incremental e eficiente.

As fontes de dados utilizadas incluíram ficheiros CSV, Excel, JSON e bases de dados relacionais, com os dados sendo ingeridos inicialmente numa *staging area*. O ETL foi implementado utilizando a ferramenta Apache NiFi, em conjunto com SQL para tratamento e carregamento intermediário dos dados nas tabelas da *staging area*.

Foram utilizados diversos lookups durante o processo de transformação para validação e complementação de atributos, assegurando que os dados fossem padronizados e enriquecidos conforme necessário.

A validação dos dados incluía principalmente a verificação de campos nulos em atributos obrigatórios. Registos com dados incompletos ou inconsistentes eram automaticamente direcionados para tabelas de quarentena, garantindo que apenas dados válidos fossem inseridos nas tabelas finais. Foram criadas tabelas específicas de quarentena para diferentes domínios, como *VendasFisicasQUA*, *VendasOnlineQUA* e uma para clientes.

Para garantir a correta sequência das operações e evitar conflitos de integridade referencial ou problemas de concorrência, foram introduzidos mecanismos de espera (*waits*) entre as etapas do processo. Por exemplo, após a execução dos inserts na tabela *DimCliente*, o processo aguarda a sua conclusão antes de iniciar os updates, assegurando que todos os dados estejam previamente carregados.

Além disso, durante o carregamento entre camadas — como da tabela *ProdutosRAW* para a tabela *ProdutosRDY* — são aplicadas pausas que garantem que as dependências entre tabelas sejam respeitadas, evitando falhas decorrentes de execuções em paralelo ou dados ainda não disponíveis. Estas estratégias aumentam a robustez do pipeline e contribuem para a consistência dos dados ao longo do processo ETL.

O povoamento foi projetado como um processo automatizado, com execução semanal (uma vez por semana), garantindo a atualização periódica dos dados analíticos com o mínimo de intervenção manual. O processo é disparado automaticamente e abrange a execução sequencial das etapas de extração, transformação e carregamento.

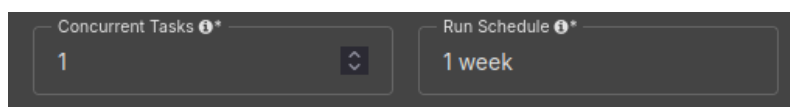


Figura 19: Agendamento automático do processo ETL com execução semanal.

Durante a implementação, identificaram-se desafios de performance no Apache NiFi, especialmente ao lidar com grandes volumes de dados e múltiplas operações de lookup.

5.5. Validação e Testes

O objetivo deste capítulo é validar a correta implementação da Dim-Clientes como uma dimensão com variação (*Slowly Changing Dimension* Tipo 4) no processo ETL. Utilizou-se uma base de dados com triggers para deteção de updates e deletes, sendo a base de dados relacional a fonte principal de verdade.

A arquitetura foi desenhada para manter a tabela Dim-Cliente sempre com os dados atuais dos clientes e armazenar na tabela *Clientes_Historico* os registos históricos de alterações.

5.5.1. Primeira Fase

Numa primeira fase, foi executado (pela primeira vez) o processo ETL, com o objetivo de carregar os dados atuais dos clientes e os primeiros registos de vendas. Como ainda não foram realizadas atualizações, este teste serve como ponto de partida para validação estrutural e correta separação das tabelas.

No final de cada processo ETL, o campo *ultima_execucao* na tabela *etl_controlo* é automaticamente atualizado com a data e hora do acontecimento, o que permite monitorizar o estado do processo e garantir que as execuções ocorrem conforme o planeamento.

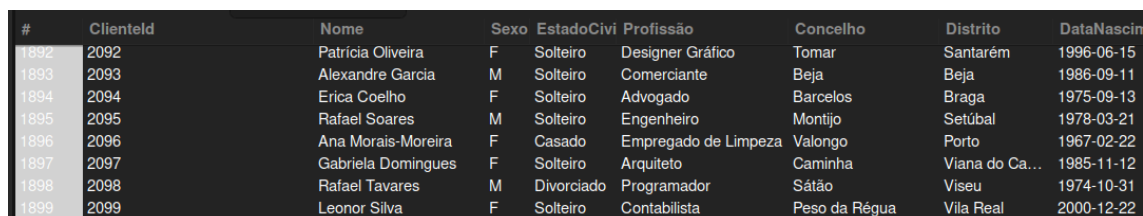


#	processo	ultima_execucao
1	clientes_log_etl	2025-05-19 18:25:20

Figura 20: Registo atualizado do campo *ultima_execucao* após execução do processo ETL

Foram analisadas as seguintes tabelas:

– Dim-Cliente



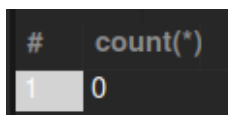
#	Clienteld	Nome	Sexo	EstadoCivi	Profissão	Concelho	Distrito	DataNascim
1892	2092	Patrícia Oliveira	F	Solteiro	Designer Gráfico	Tomar	Santarém	1996-06-15
1893	2093	Alexandre Garcia	M	Solteiro	Comerciante	Beja	Beja	1986-09-11
1894	2094	Erica Coelho	F	Solteiro	Advogado	Barcelos	Braga	1975-09-13
1895	2095	Rafael Soares	M	Solteiro	Engenheiro	Montijo	Setúbal	1978-03-21
1896	2096	Ana Morais-Moreira	F	Casado	Empregado de Limpeza	Valongo	Porto	1967-02-22
1897	2097	Gabriela Domingues	F	Solteiro	Arquiteto	Caminha	Viana do Ca...	1985-11-12
1898	2098	Rafael Tavares	M	Divorciado	Programador	Sátão	Viseu	1974-10-31
1899	2099	Leonor Silva	F	Solteiro	Contabilista	Peso da Régua	Vila Real	2000-12-22

Figura 21: DimCliente – Registos Iniciais de Clientes

A imagem mostra os dados carregados na tabela DimCliente após a execução inicial do ETL. Verifica-se que os dados mais recentes e ativos dos clientes estão corretamente representados, sem duplicações ou inconsistências.

– Clientes_Historico

Query: `SELECT count(*) FROM DW.Clientes_Historico;`



#	count(*)
1	0

Figura 22: Clientes_Historico sem registos

A imagem acima confirma que, numa primeira execução, não existem registos históricos. Verifica-se, porque nenhuma atualização foi feita até este ponto, o que é o comportamento esperado para uma dimensão SCD Tipo 4 no carregamento inicial.

– TFVenda

9458	9940	2021-10-08	1149	116	2	130.05	10.00	260.10	Online
9459	9941	2022-03-18	1149	89	3	32.65	4.00	97.94	Online
9460	9942	2022-12-16	1149	39	5	10.40	10.00	52.02	Online
9461	9943	2023-05-14	1149	175	3	1040.44	5.00	3121.32	Online

Figura 23: Registos da loja online na TFVenda

31745	33440	2016-03-29	2099	83	1	217.10	25.00	217.10	Loja Física
31746	33441	2018-01-09	2099	69	1	34.50	0.00	34.50	Loja Física
31747	33442	2019-01-08	2099	22	2	17.25	10.00	34.50	Loja Física
31748	33443	2019-10-02	2099	111	1	40.31	0.00	40.31	Loja Física

Figura 24: Últimos registos da TFVenda

A tabela TFVenda apresenta os registos de vendas associadas aos clientes.

5.5.2. Segunda fase

Após a execução inicial e a validação estrutural, foi conduzida uma segunda fase de testes com o objetivo de simular a evolução natural dos dados ao longo do tempo. Nesta etapa:

- Foram adicionados novos registos de clientes na base de dados relacional de origem.
- Inseriram-se também novos registos de vendas, associados tanto a clientes previamente existentes quanto a novos clientes.
- Atualizaram-se alguns registos de clientes já presentes na dimensão DimCliente, com alterações em determinados atributos.

Para validar o comportamento do processo ETL em relação a alterações nos dados de clientes, foi efetuada uma modificação no registo da cliente Nair Tavares, presente na tabela *clientes*:

Nair Tavares, Farmacêutico, Casado, Feminino, Viana do Castelo, Arcos de Valdevez, 22/09/1976, nair.tavares@gmail.com, C2045

Foi então executada a seguinte query **UPDATE**:

```
UPDATE FonteClientes.clientes
SET Profissão = 'Comerciante',
    Sexo = 'Masculino',
    Distrito = 'Coimbra',
    Concelho = 'Arganil',
    Morada = 'Alameda Freitas, 98\n5897-865 Sabugal',
    CódigoPostal = '7224-893',
    DataNascimento = '1974-01-03'
WHERE Email = 'nair.tavares@gmail.com';
```

Após a execução do processo ETL, os resultados observados foram os seguintes:

- A tabela *Clientes_Historico* passou a conter o registo anterior da cliente, mantendo o histórico da versão original.

#	Histo	Clienteld	Nome	Sexo	EstadoCivil	Profissao	Concelho	Distrito	DataNascimento	Modificao	DataHoraModificaçã
1	1	2045	Nair Tavares	F	Casado	Farmacêutico	Arcos de Valde...	Viana do C...	1976-09-22	Profissao de "Farmacêutico" para "Comerciante"; ... 8431-866 Pombal" para "Alameda Freitas...	2025-05-19 18:19:23

Figura 25: Antigo registo de informação de Nair Tavares

– A tabela *DimCliente* foi atualizada com os novos dados, refletindo a versão mais recente do cliente

1846	2044	Sofia Soares	F	Solteiro	Psicólogo	Boticas	Vila Real	1988-05-10
1847	2045	Nair Tavares	M	Casado	Comerciante	Arganil	Coimbra	1974-01-03
1848	2046	Ricardo Alves	M	Casado	Rececionista	Condeixa-a-Nova	Coimbra	1987-02-08

Figura 26: Informação atualizada na DimCliente

Foi ainda possível validar que novos registos de vendas foram corretamente processados e inseridos na tabela TFWenda.

#	DataId	Clienteld	Produtolc	Quantidade	PrecoUnitari	PercentagemDescont	ValorTotal	Canal	VendaId
51186	2024-06-30	409	117	1	54.92	10.00	54.92	Online	53186
51187	2024-06-30	458	161	2	34.18	0.00	68.35	Online	53187
51188	2024-06-30	600	127	3	14.16	0.00	42.47	Online	53188
51189	2024-06-30	795	74	3	65.77	50.00	197.31	Online	53189
51190	2024-06-30	809	108	1	137.83	10.00	137.83	Online	53190
51191	2024-06-30	911	48	2	119.00	10.00	238.00	Online	53191
51192	2024-06-30	967	65	2	187.12	50.00	374.23	Online	53192
51193	2024-06-30	994	49	3	90.58	0.00	271.75	Online	53193

Figura 27: Novos registos de vendas inseridos na TFWenda

5.5.3. Tabelas de Quarentena

Como parte do processo de controlo de qualidade e monitorização de integridade dos dados, foram também utilizadas as tabelas *ClientesQUA*, *VendasFisicasQUA* e *VendasOnlineQUA*. Estas tabelas têm como objetivo registar ocorrências inesperadas ou incoerências detetadas durante a execução do processo ETL, permitindo um acompanhamento mais preciso e rastreável de anomalias.

A tabela *ClientesQUA* armazena registos de clientes que não puderam ser corretamente inseridos na dimensão principal *DimCliente*, devido a inconsistências ou falhas de validação. Entre os motivos mais comuns estão valores em falta.

#	id	Clienteld	Nome	Profissao	EstadoCivi	Sexo	Distrito	Concelho	Telefone	Morada	CódigoPostal	Email	DataN
96	96	2017	Carolina Matias	NULL	Casado	NULL	NULL	Braga	NULL	Praca Ana Miranda, 38 1123-667 Fafe	8114-277	carolina.matias@yahoo.com	1991-06
97	97	2036	Fabiana Pires	Psicólogo	NULL	Feminino	NULL	Chaves	NULL	Travessa Pereira, 24 0227-116 Vila Nova de Foz Côa	2840-839	fabiana.pires@hotmail.com	1968-05
98	98	2038	Helena Costa	Contabilista	Solteiro	NULL	NULL	NULL	950116621	R. Nogueira, 57 5946-816 Povoa de Varzim	6843-959	helena.costa@yahoo.com	1984-05
99	99	2090	Isabela Cruz	Estudante	Casado	NULL	Coimbra	Figueira da Foz	992206198	NULL	1917-851	isabela.cruz@hotmail.com	2008-05
100	100	2091	Beatriz da Tavares	Técnico de Mark...	Divorciado	Feminino	NULL	NULL	981536459	Travessa de Coitinho, 37 4220-918 São João da Madeira	7580-171	beatriz.tavares@hotmail.com	...

Figura 28: Registos extraídos com anomalias para a tabela de quarentena de vendas físicas

A tabela *VendasFisicasQUA* recolhe vendas oriundas do canal físico que foram rejeitadas durante o processo de carregamento na tabela TFWenda.

#	Data_Venda	Cliente_Nome	Produto	Quantidade	Percentagem_Descont	Preço	Canal	Erro
1	02/02/2013			3	10.00	421.72	Loja Física	Valores a Null
2				5	NULL	211.81	Loja Física	Valores a Null
3	31/03/2016	Erika Costa	Teclado Corsair K55		5.00	NULL	Loja Física	Valores a Null
4			Placa Gráfica Gigabyte RTX 3060 Ti	3	NULL	114.81	Loja Física	Valores a Null
5	21/07/2016	Alice-Bianca Cruz	SSD Samsung 980 Pro 2TB	1	10.00	NULL		Valores a Null
6	27/01/2017		Impressora Canon PIXMA G6020	1	0.00	NULL	Loja Física	Valores a Null
7	11/04/2019		Smartphone Google Pixel 6		25.00	NULL	Loja Física	Valores a Null
8	09/11/2015		Smartphone Oppo Reno 6		NULL	NULL	Loja Física	Valores a Null
9	05/04/2011	Júlia Brito	Placa Gráfica Gigabyte RTX 3060 Ti	3	NULL	98.12	Loja Física	Valores a Null

Figura 29: Registos extraídos com anomalias para a tabela de quarentena de vendas física

De forma análoga, a tabela *VendasOnlineQUA* armazena registos de vendas do canal online que não cumpriram os critérios necessários para serem integrados na tabela de factos. Estes registos permitem auditar o processo e identificar padrões de falhas específicas no canal online.

#	Data	Clientid	Email	ProdutoId	Quantidade	PercentagemDescont	Preco	Canal	Erro
1	2024-04-06	C124	isaac.amorim@sapo.pt	NULL	1	25.00	58.99	Online	Valores a Null
2	2022-08-14	C127	jaime.sousa@gmail.com	NULL	NULL	NULL	NULL	Online	Valores a Null
3	2022-09-17	NULL	NULL	P044	5	46.00	242.00	NULL	Valores a Null
4	2022-06-06	NULL	NULL	NULL	NULL	NULL	NULL	Online	Valores a Null
5	2021-07-01	C125	NULL	P099	2	21.00	255.82	Online	Valores a Null
6	2023-10-27	NULL	ema.freitas@hotmail.com	NULL	NULL	NULL	NULL	Online	Valores a Null
7	2022-03-09	NULL	mara.anjos@hotmail.com	P007	1	21.00	137.74	NULL	Valores a Null
8	NULL	C100	leandro.lourenço@outlook.com	NULL	NULL	NULL	NULL	Online	Valores a Null
9	2021-07-04	C168	afonso.carvalho@yahoo.com	NULL	1	0.00	648.19	Online	Valores a Null

Figura 30: Registos extraídos com anomalias para a tabela de quarentena de vendas online

A existência destas três tabelas de quarentena permite um processo de ETL mais robusto e auditável, oferecendo visibilidade total sobre falhas e facilitando a sua correção em ciclos posteriores de carregamento.

5.5.4. Conclusões Finais

Os testes realizados confirmam que a arquitetura ETL implementada está a funcionar corretamente, garantindo o correto carregamento inicial dos dados e a separação adequada entre dados atuais e históricos, conforme definido pela dimensão SCD Tipo 4. A atualização automática do campo *ultima_execucao* permite um controlo eficiente das execuções, facilitando a extração incremental dos dados.

Adicionalmente, a validação com alterações reais nos dados dos clientes demonstrou que o processo ETL identifica e regista corretamente as modificações, preservando o histórico e atualizando a dimensão principal de forma consistente. A utilização das tabelas de quarentena reforça a robustez do processo, assegurando o tratamento adequado de dados inconsistentes e facilitando a monitorização e correção de anomalias.

Em suma, o sistema ETL está preparado para suportar o crescimento e a evolução dos dados ao longo do tempo, assegurando integridade, auditabilidade e eficiência.

6. Exploração e Análise de Dados

6.1. Organização Geral do Sistema de Dashboarding

Com base nas vistas analíticas previamente definidas para os agentes de decisão da *Perifericum* (Secção 3.6), a organização geral do sistema de dashboarding foi estruturada em torno das necessidades específicas de cada perfil, garantindo clareza, rapidez na interpretação e alinhamento com os objetivos operacionais e estratégicos da empresa.

No caso do Sr. João, responsável pela gestão estratégica e pela experiência do cliente, os dashboards foram concebidos com o objetivo de aprofundar o conhecimento sobre os perfis e comportamentos dos consumidores. A organização dos dashboards reflete a diversidade de análises necessárias à definição de estratégias de fidelização, personalização e segmentação de campanhas. Entre os dashboards desenvolvidos para este perfil destacam-se:

- **Distribuição de clientes por distrito** – permite mapear a base de clientes por localização geográfica.
- **Distribuição por sexo e estado civil** – possibilita identificar diferenças comportamentais relevantes entre segmentos demográficos.
- **Profissões mais comuns entre os clientes** – útil para segmentações de marketing mais personalizadas e identificação de nichos.

No caso do Sr. Gates, cuja responsabilidade recai sobre as operações e o marketing da empresa, os dashboards foram organizados com foco na monitorização da performance comercial e na avaliação contínua da eficácia das campanhas promocionais. A estrutura escolhida privilegia uma leitura rápida e comparativa dos principais indicadores, permitindo uma atuação ágil face às dinâmicas do mercado e ao comportamento dos consumidores.

Entre os dashboards mais relevantes para este perfil encontram-se:

- **Valores totais por canal de compra** – permite comparar o desempenho entre os canais físico e digital, apoiando decisões sobre alocação de recursos e estratégias omnicanal.
- **Valores totais por categorias e marcas de produtos** – oferece uma visão sobre quais os segmentos com maior volume de vendas, contribuindo para a gestão de portefólio e campanhas direcionadas.
- **Porcentagem de desconto face às vendas totais** – essencial para avaliar o impacto direto das promoções no volume de vendas e na rentabilidade das ações de marketing.

Esta organização garante que o Sr. Gates tenha acesso a informação clara e acionável, ajustada ao seu papel operacional, permitindo otimizar continuamente as campanhas e alinhar as decisões com os objetivos comerciais da empresa.

6.2. Serviços de Exploração e Análise Implementados

Os dashboards foram construídos com foco na simplicidade de utilização, rapidez de exploração e riqueza analítica, permitindo aos utilizadores filtrar e cruzar dimensões relevantes como Tempo, Cliente, Produto e Canal. Abaixo descrevem-se os principais dashboards desenvolvidos, a sua utilidade e a forma como servem os objetivos de análise de cada agente.

6.2.1. Dashboards para Gestão Estratégica e Experiência do Cliente

– Distribuição de Clientes por Distrito, Sexo e Estado Civil

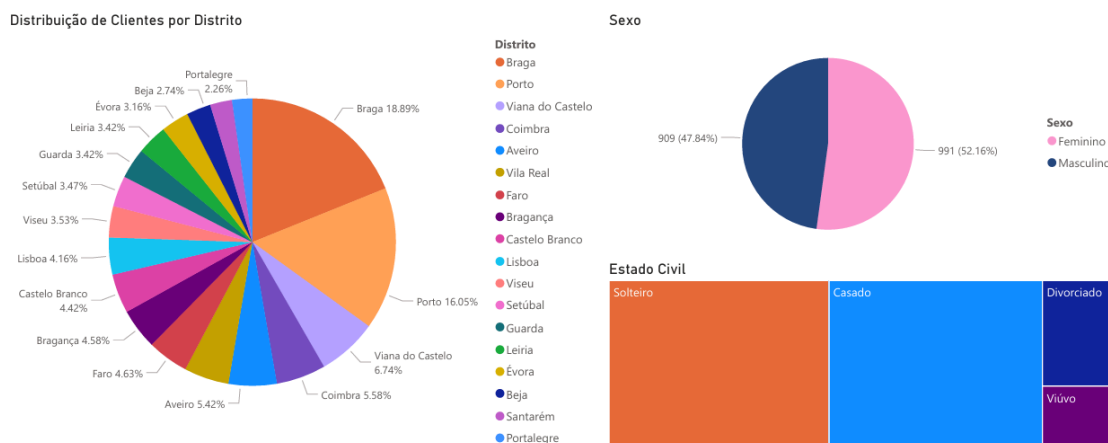


Figura 31: Dashboard: Distribuição de Clientes por Distrito, Sexo e Estado Civil

Permite identificar a composição demográfica e geográfica da base de clientes. Este dashboard facilita segmentações geográficas e socioculturais, úteis para campanhas localizadas ou personalizadas. É um dashboard descritivo com apoio à segmentação estratégica.

– Profissões Mais Comuns entre os Clientes

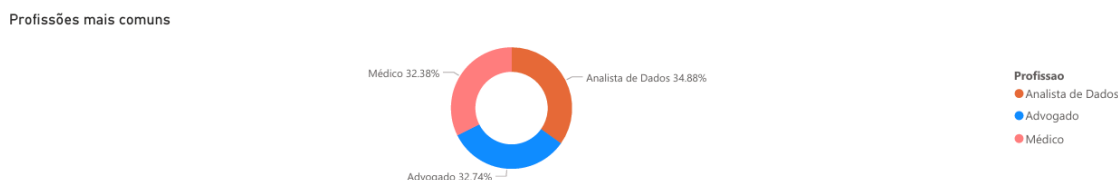


Figura 32: Dashboard: Profissões Mais Comuns entre os Clientes

Ajuda a compreender o perfil profissional predominante, permitindo identificar segmentos socioeconómicos com maior presença. É útil para afinar campanhas direcionadas e otimizar a comunicação. Funciona como suporte à análise de público-alvo.

– Categorias de Produtos Mais Compradas e Marcas Mais Procuradas

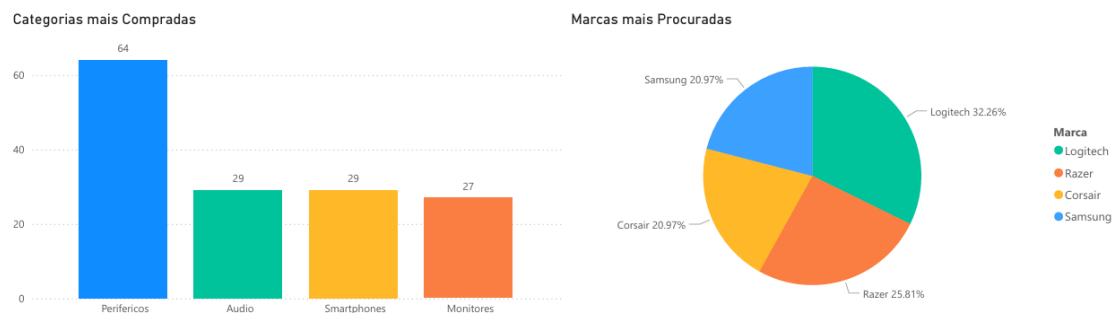


Figura 33: Dashboard: Categorias de Produtos Mais Compradas e Marcas Mais Procuradas

Revela preferências de consumo, permitindo orientar o planeamento de stock, parcerias com fornecedores e estratégias de cross-selling. Este dashboard combina análise descritiva com apoio à decisão comercial.

– Valor Médio de Compra por Período

Valor Médio de Compra por Período

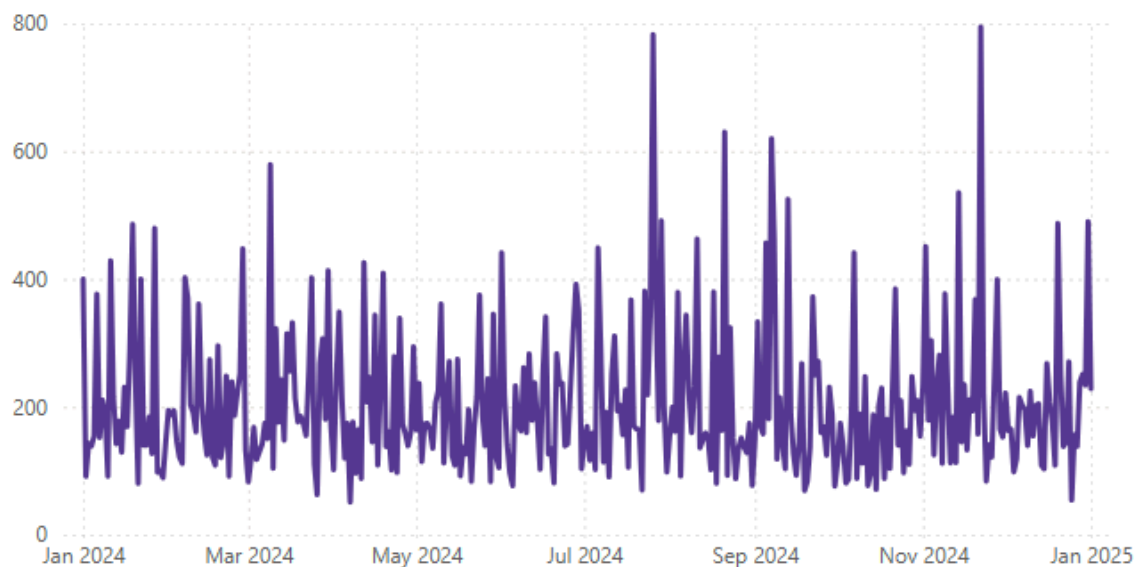


Figura 34: Dashboard: Valor Médio de Compra por Período

Fornece uma visão temporal do comportamento de compra, permitindo identificar tendências e ciclos de consumo. Suporta decisões estratégicas baseadas em sazonalidade ou impacto de campanhas em períodos específicos.

– Valor Médio de Compra por Grupo Etário

Valor Médio de Compra por Grupo Etário

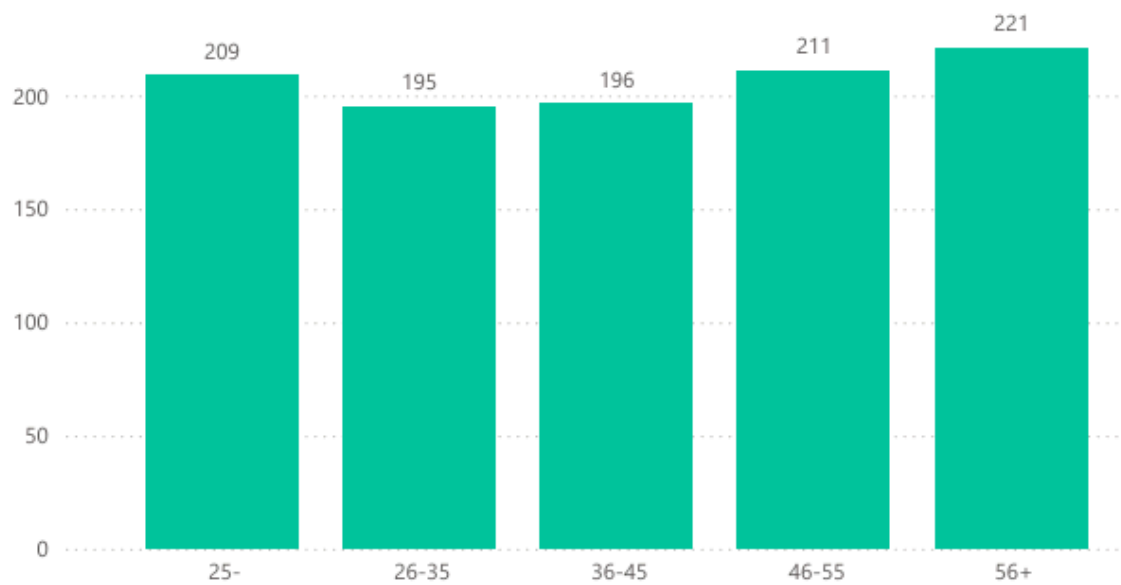


Figura 35: Dashboard: Valor Médio de Compra por Grupo Etário

Cruza idade com comportamento de compra, identificando o poder de compra e preferências por faixa etária. Suporta ações de personalização de ofertas e marketing segmentado.

Estes dashboards apoiam o Sr. João na criação de estratégias centradas no cliente, melhorando a retenção e a experiência através de decisões baseadas em dados reais e atualizados.

6.2.2. Dashboards para Operações e Marketing

– Valores Totais por Canal de Compra

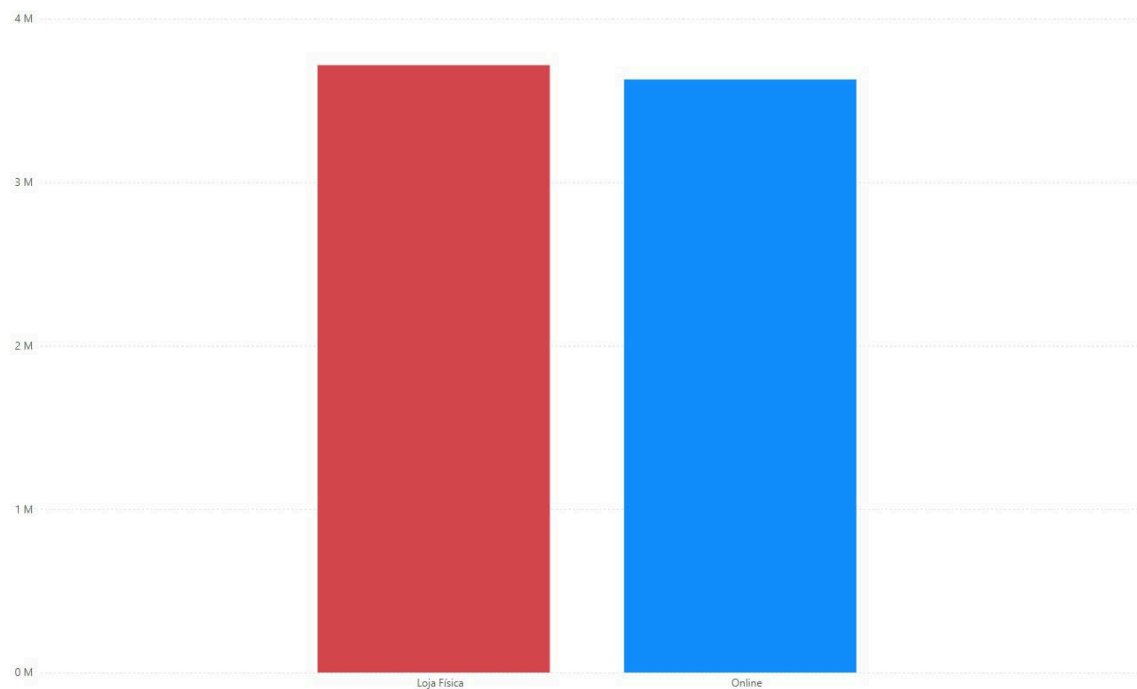


Figura 36: Dashboard: Valores Totais por Canal de Compra

Permite analisar o desempenho de cada canal (online, físico), ajudando a identificar os mais rentáveis e os que necessitam de reforço. Suporta decisões de alocação de recursos e estratégias multicanal, essenciais para otimização operacional e crescimento sustentado.

– Categorias de Marcas de Produtos mais Vendidos

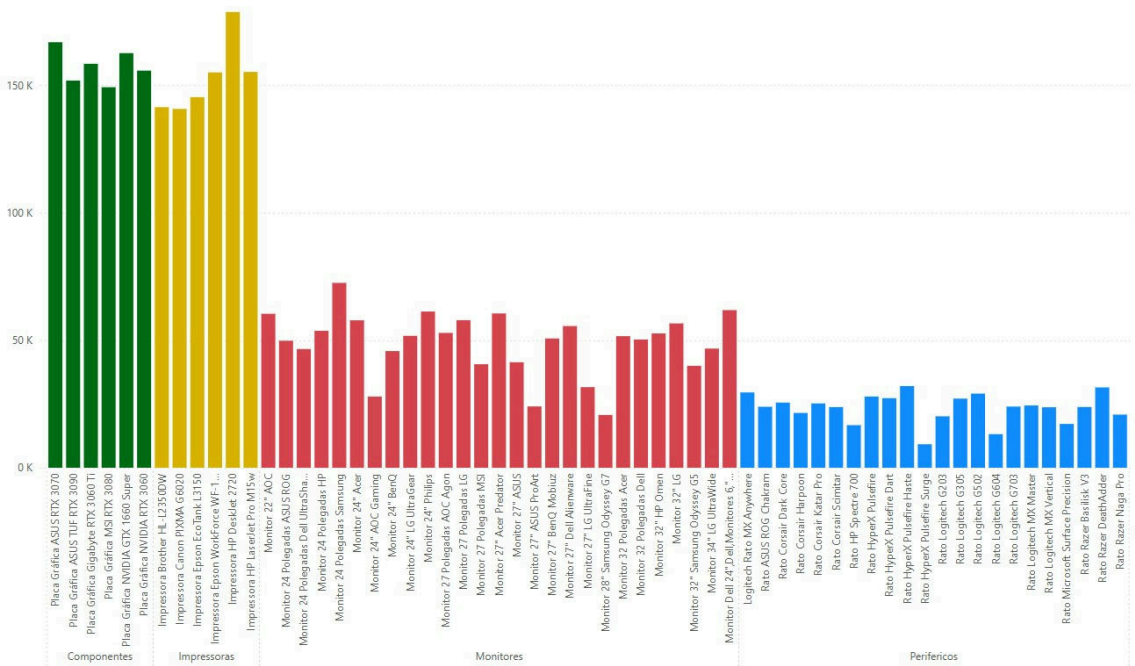


Figura 37: Dashboard: Top 4 das Categorias de Marcas de Produtos mais Vendidos

Oferece insights sobre o portfólio de produtos, destacando quais categorias e marcas geram mais receita. Auxilia no planeamento de promoções e gestão de relações com fornecedores. É um dashboard chave para decisões comerciais e de marketing.

– Vendas por Dias da Semana

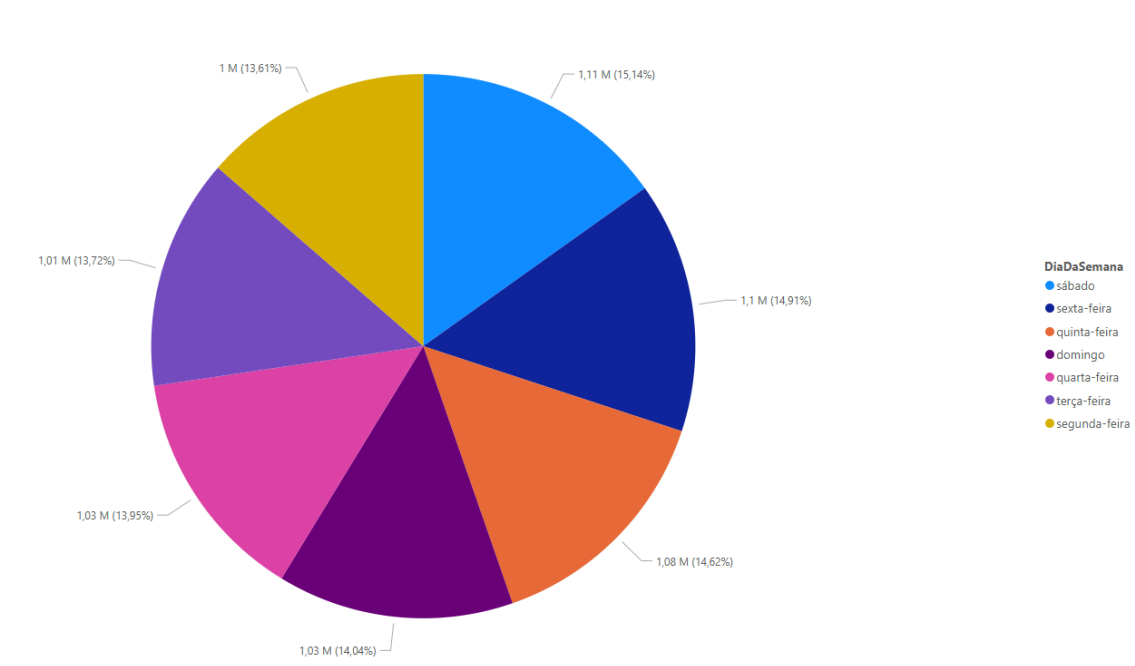


Figura 38: Dashboard: Vendas por Dias da Semana

Evidencia padrões de consumo ao longo da semana, permitindo ajustar horários de atendimento, campanhas promocionais e logística. É especialmente útil para maximizar eficiência operacional e rentabilidade diária.

– **Porcentagem de Desconto nas Vendas Totais**

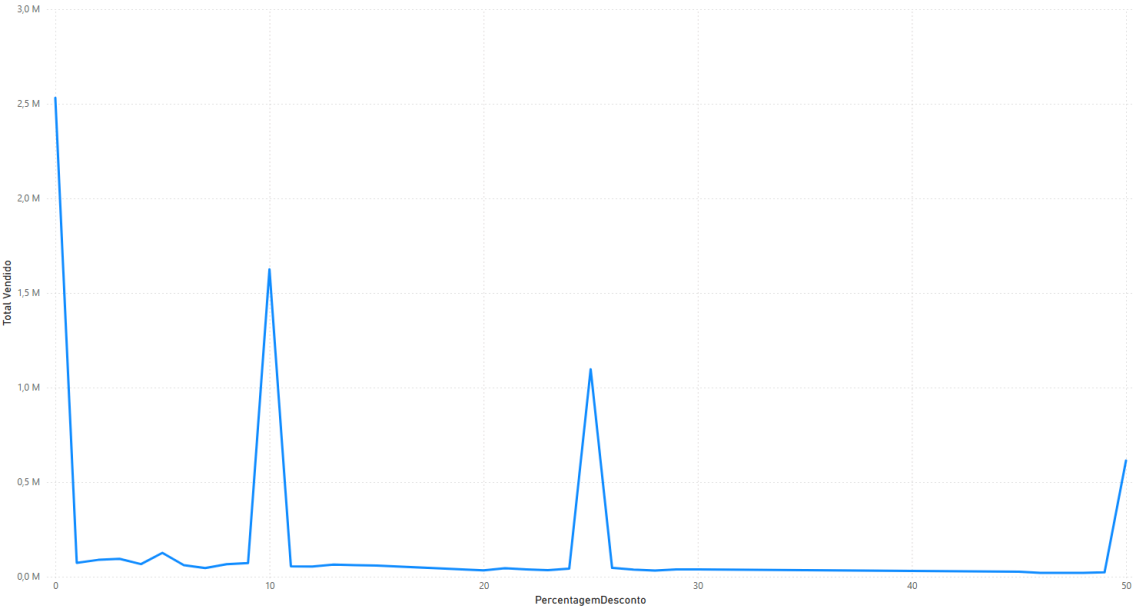


Figura 39: Dashboard: Porcentagem de Desconto pelas Vendas Totais

Avalia o impacto dos descontos nas vendas, ajudando a perceber se as promoções estão a gerar retorno significativo. Suporta decisões sobre políticas de desconto, margens e estratégias de pricing.

– **Dispersão de Produtos por Descontos**

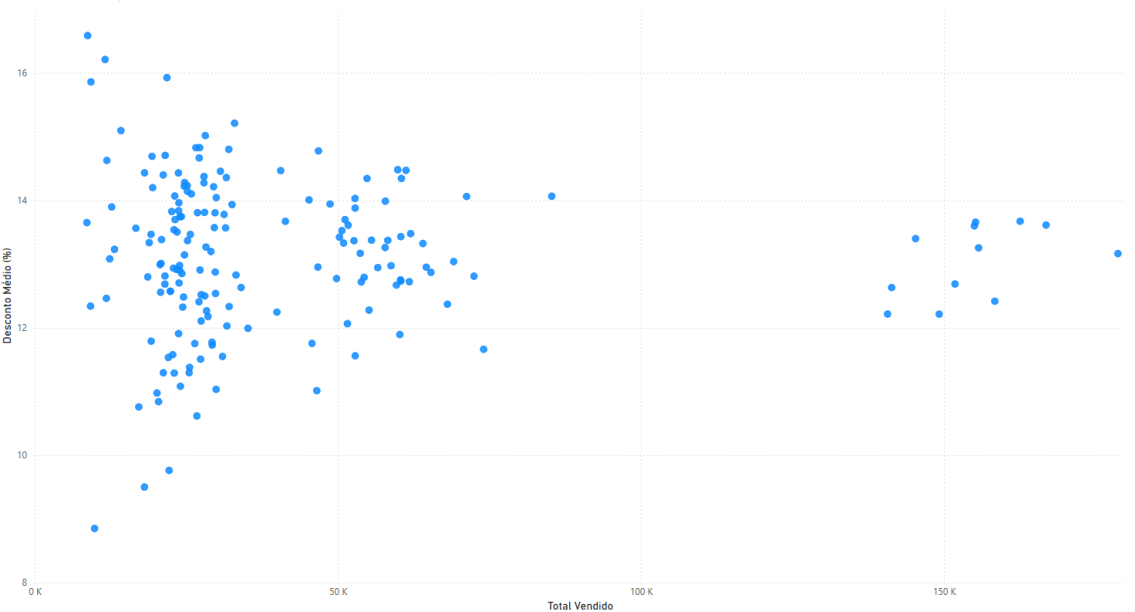


Figura 40: Dashboard: Dispersão de Produtos por Porcentagem de Descontos

Visualiza como os produtos em geral respondem a níveis variados de desconto. Permite identificar quais as promoções que mantêm valor sem necessidade de desconto. É útil para calibrar campanhas promocionais com maior precisão.

7. Caracterização de Perfis de Clientes

7.1. Definição do Problema e Compreensão dos Elementos de Análise Envolvidos

A limitação estratégica significativa culmina num dos principais problemas da *Perifericum*: a ausência de um sistema estruturado para segmentação e análise de perfis de clientes. Esta lacuna compromete seriamente a capacidade da loja em personalizar ofertas e implementar estratégias eficazes de fidelização. Como consequência direta, observam-se três problemas principais: perda sistemática de oportunidades de negócio (especialmente em vendas cruzadas), campanhas de marketing excessivamente genéricas e programas de fidelização pouco segmentados e, portanto, pouco eficientes.

O cerne da solução proposta reside na segmentação dos clientes em grupos homogêneos, utilizando três eixos principais de análise.

Em primeiro lugar, os Dados Demográficos permitirão categorizar os clientes por faixas etárias (como Adolescentes de 12-17 anos ou Jovens Adultos de 18-25 anos, calculados a partir da data de nascimento), por localização geográfica (distrito e concelho) e por profissão - variável particularmente relevante pois sabemos que estudantes, profissionais de TI e outros grupos apresentam padrões de consumo distintos.

Em segundo lugar, os Dados Comportamentais oferecerão insights valiosos sobre a frequência de compra (se mensal, semanal ou esporádica), o valor médio gasto por transação e as preferências de produtos, estas últimas determinadas através da análise histórica das categorias de produtos adquiridos.

Finalmente, as Métricas de Valor do Cliente permitirão classificar os consumidores em grupos estratégicos como Clientes VIP (com alta frequência e volume de compras), Clientes em Risco (com diminuição recente no engajamento) e Clientes Novos (ainda sem histórico consolidado).

7.1.1. Fontes de Dados

A base desta análise repousa sobre três pilares fundamentais de informação. A **Tabela de Factos** (TF-Venda) constitui o repositório central das transações históricas, contendo dados cruciais sobre produtos adquiridos, valores transacionados e datas das operações. Complementarmente, a Dimensão **Dim-Cliente** armazena os dados demográficos essenciais, desde idade até localização e profissão. Por fim, a Dimensão **Dim-Produto** oferece o detalhamento necessário sobre categorias e marcas dos itens comercializados, permitindo cruzar informações sobre preferências específicas dos diferentes segmentos identificados.

7.1.1.1. Variáveis-Chave

Variável	Tipo	Descrição
<i>Idade</i>	Numérico	Derivada da data de nascimento (DataNascimento).
<i>Localização</i>	Categórico	Distrito e concelho (Distrito, Concelho).
<i>Frequência de Compra</i>	Numérico	Número de transações nos últimos 12 meses.
<i>Valor Total Gasto</i>	Numérico	Soma do valor das compras (ValorTotal).
<i>Categoria Favorita</i>	Categórico	Categoria mais comprada (Categoria em Dim-Produto).

7.1.2. Métodos e Técnicas Propostas

– 1. Análise Exploratória de Dados (EDA)

O processo analítico seguirá uma abordagem metodológica rigorosa em três fases consecutivas. A primeira etapa consistirá numa Análise Exploratória de Dados (EDA), que permitirá identificar padrões básicos, distribuições e correlações preliminares entre variáveis. Esta fase é crucial para compreender a qualidade dos dados e orientar as abordagens subsequentes.

– 2. Machine Learning (Clustering)

Na segunda fase, aplicaremos técnicas de Machine Learning, especificamente algoritmos de Clustering, com destaque para o método K-means. Esta escolha fundamenta-se na sua comprovada eficácia para problemas de segmentação com características numéricas, além da sua relativa simplicidade de implementação e interpretação. O número ótimo de clusters será determinado através do método do cotovelo, complementado pela análise do Silhouette Score.

– 3. Validação dos Clusters

A fase final dedicar-se-á à Validação dos Clusters obtidos, utilizando não apenas métricas estatísticas como o mencionado Silhouette Score, mas também através de uma avaliação qualitativa da coerência empresarial dos grupos identificados. Esta dupla validação - quantitativa e qualitativa - garantirá que os resultados sejam tanto estatisticamente robustos quanto comercialmente relevantes para a *Perifericum*.

7.2. Seleção e Preparação dos Dados

A maior parte da seleção e preparação dos dados foi realizada na etapa de ETL, o que reduziu significativamente a carga de trabalho nesta fase. Inicialmente, o tratamento de valores ausentes era feito aqui, mas foi posteriormente transferido para o ETL, otimizando o processo e a limpeza dos dados.

A análise exploratória através dos gráficos — percentagem de compras por distrito, compras online versus físicas e distribuição por profissão — forneceu insights valiosos que orientaram o processo de feature engineering.

Dessa forma, o foco principal desta etapa passou a ser o feature engineering (Secção 7.4.2), a normalização das variáveis (feature scaling, Secção 7.4.5) e a redução de dimensionalidade (dimensionality reduction, Secção 7.4.6), preparando os dados adequadamente para os modelos de machine learning subsequentes.

7.3. Identificação e Fundamentação da Técnica de Análise

Para orientar a fase de análise e modelação dos dados no presente projeto, foi adotada a metodologia **SEMMA** (Sample, Explore, Modify, Model, Assess). Esta metodologia foi selecionada por se adequar de forma eficaz ao contexto atual do projeto, no qual os dados já se encontram limpos, tratados e estruturados, provenientes do data warehouse, ou seja, a parte do *Sample* já foi tratada previamente.

De seguida fizemos uma pequena exploração dos dados mais centralizada para futuro *feature engineering* e não tanto como análise de dados feita já previamente no relatório. Em termos de modificações, criamos as novas variáveis que achamos importantes para o clustering e para recomendações e depois destas modificações implementamos o K-Means para fazer os clusters. No final, a partir do Silhouette Score e de uma validação quantitativa e qualitativa determinamos a qualidade destes clusters.

Em conclusão, a metodologia SEMMA foi escolhida por se adequar a projetos com dados já preparados, permitindo iniciar a análise com a extração de amostras e a exploração estatística. A sua estrutura sequencial — que inclui a modificação, modelação e avaliação dos dados — favoreceu a construção de modelos mais robustos.

7.4. Construção do Modelo de Análise

A construção do modelo de análise começou pelo seu carregamento e por uma visualização superficial. Foi também feita uma análise dos tipos de dados carregados.

7.4.1. Análise de dados

Foi feita uma análise mais detalhada de métricas relevantes para avaliar a sua relevância para a construção dos perfis de clientes.

Começamos por avaliar a distribuição das compras por distrito. Verificou-se que os distritos do Porto e Braga apresentam os maiores volumes de compras, destacando-se significativamente em relação aos demais. Apesar de se observar que os distritos do Porto e Braga concentram um maior volume de compras, os restantes distritos apresentam valores bastante próximos entre si, sem diferenças expressivas. Por este motivo, a variável “distrito” não contribui significativamente para a diferenciação entre perfis de clientes e, consequentemente, para a personalização de ofertas de produtos e serviços.

De seguida, analisamos a percentagem de compras online e a percentagem de compras em loja física. Embora os valores sejam globalmente próximos, a variação entre clientes é suficiente para influenciar padrões de comportamento. Assim, esta variável foi considerada relevante para a construção dos clusters, uma vez que permite identificar preferências no canal de compra que, combinadas com outras variáveis, contribuem para uma segmentação mais rica e significativa.

A profissão do cliente também foi analisada com o objetivo de identificar uma possível relação entre a ocupação profissional e os hábitos de compra. No entanto, verificou-se que não existe uma profissão com destaque significativo em termos de volume de compras, sendo que as percentagens estão distribuídas de forma bastante homogênea entre as diferentes categorias. Por este motivo, concluiu-se que a profissão não contribui de forma relevante para a diferenciação entre perfis de clientes, e foi por isso excluída da construção dos clusters.

7.4.2. Feature Engineering

Com os dados carregados podemos obter novas features que permitem obter mais informação sobre os hábitos de compra dos clientes.

Tendo em conta que pode ser aplicado um desconto ao produto, calculamos o **preço final** do produto. Esta feature vai ser utilizada mais tarde para criar features relacionadas com os gastos do cliente.

A criação da feature **idade** permite captar variações no comportamento dos clientes relacionadas à sua fase de vida. Esta variável é importante para identificar segmentos com necessidades e preferências distintas.

A feature **percentagem de compras online** é importante para entender o comportamento de compra dos clientes em diferentes canais. Esta variável ajuda a distinguir clientes que preferem o comércio online daqueles que privilegiam a compra presencial, permitindo segmentações que refletem hábitos e preferências.

O **total de transações** reflete a frequência com que o cliente realiza compras, sendo um indicador direto do seu nível de engagement com a loja. Esta métrica é fundamental para identificar clientes mais ativos e avaliar padrões de fidelidade, ajudando a segmentar grupos com diferentes comportamentos de compra.

O **total de produtos comprados** traduz a quantidade efetiva de itens adquiridos, permitindo compreender a intensidade e diversidade do consumo de cada cliente. Esta variável complementa o total de transações, oferecendo uma visão mais detalhada sobre o volume de compras.

O **total gasto** representa o montante acumulado que um cliente despendeu na loja, sendo um indicador-chave do seu valor para o negócio. Esta métrica permite identificar clientes de alto valor e segmentar grupos segundo o seu potencial económico.

O **valor médio das transações** revela o ticket médio por compra, oferecendo insights sobre o comportamento de consumo em cada interação. Esta informação ajuda a diferenciar clientes que fazem poucas compras de valor elevado daqueles que realizam compras frequentes, mas de menor valor, enriquecendo a segmentação.

A feature **produtos únicos comprados** indica a diversidade de produtos adquiridos por cada cliente ao longo do tempo. Esta feature é útil para avaliar o grau de variedade no comportamento de compra, permitindo distinguir entre clientes com padrões de consumo especializados (compra repetida dos mesmos produtos) e clientes com preferências mais diversificadas.

A **média de dias entre compras** mede a frequência temporal do comportamento de compra do cliente. Esta métrica permite identificar clientes com hábitos regulares e outros com padrões mais esporádicos, sendo fundamental para a criação de perfis de cliente.

As features **dia e hora favoritos para comprar** ajudam a perceber quando os clientes tendem a efetuar compras, fornecendo insights sobre rotinas e preferências temporais.

A **média mensal de gasto** do cliente é um indicador útil para avaliar o seu comportamento de compra ao longo do tempo. Permite distinguir clientes com histórico relevante de clientes com pouca atividade.

O **desvio padrão mensal dos gastos** avalia a variabilidade no consumo mensal de cada cliente. Um valor alto indica flutuações significativas, enquanto um valor baixo sugere estabilidade.

A **tendência dos gastos** mede tendência de crescimento ou diminuição dos gastos ao longo do tempo. Esta feature ajuda a perceber se o cliente está a aumentar, manter ou reduzir o seu envolvimento com a loja.

Foi ainda determinada a **categoria preferida** do cliente. Esta feature apesar de não ser utilizada na criação de perfis de clientes, será útil para gerar recomendações baseadas em hábitos de compra.

7.4.3. Detecção e tratamento de outliers

Foi utilizado o algoritmo **Isolation Forest** para identificar outliers nos dados, com uma taxa de contaminação definida em 5%. Este método é eficaz na detecção de anomalias em conjuntos de dados multivariados, isolando pontos que se comportam de forma distinta dos restantes.

Após o ajuste do modelo, cada cliente foi classificado como outlier (1) ou inlier (0) através da coluna `Is_Outlier`. Esta informação pode ser utilizada para remover clientes com padrões de consumo atípicos, garantindo que a segmentação e outras análises se baseiam em dados representativos e consistentes.

7.4.4. Correlation analysis

Foi realizada uma análise de correlação entre as variáveis numéricas do conjunto de dados limpo. A matriz de correlação permite identificar relações lineares entre atributos, o que é útil para:

- Evitar redundância ao selecionar variáveis para clustering ou modelação;
- Compreender interdependências entre comportamentos dos clientes.

A utilização de um heatmap triangular com uma paleta personalizada facilita a interpretação visual das correlações mais fortes, tanto positivas como negativas.

7.4.5. Feature scaling

Para garantir que todas as variáveis numéricas relevantes têm escalas comparáveis, foi aplicada uma normalização utilizando o `StandardScaler`, que transforma os dados para que tenham média 0 e desvio padrão 1.

Foram excluídas da normalização todas as variáveis identificadoras, categóricas e temporais, que não beneficiariam deste tipo de transformação. Esta etapa é essencial para métodos de análise como clustering, onde diferenças de escala podem influenciar os resultados de forma indevida.

Chegamos a experimentar `cyclic scaling` com o atributo da hora e dia da semana da compra, mas este só piorou o `silhouette score` dos clusters por isso resolvemos não utilizá-lo.

7.4.6. Dimensionality reduction

Foi aplicada a técnica de **PCA** aos dados numéricos normalizados com o objetivo de reduzir a dimensionalidade do conjunto, preservando ao máximo a variância presente nos dados. Antes da transformação, os eventuais valores ausentes (não existentes no DW) seriam imputados com a média, e o índice foi definido como `Clienteld` para manter a referência aos clientes.

A análise da variância explicada por componente permitiu identificar que 4 componentes principais eram suficientes para reter a maior parte da informação relevante. Este número foi considerado o *k* ótimo, servindo de referência tanto para a redução dimensional como para orientar possíveis técnicas de clustering.

O resultado da transformação foi um novo conjunto de dados com componentes PC1 a PC4, que representam os dados originais num espaço mais compacto e eficiente para análise, mantendo a associação com cada cliente.

7.4.7. Determinação do número ótimo de clusters

7.4.7.1. Elbow method

Para identificar o número ótimo de clusters no conjunto de dados, foi utilizado o elbow method em conjunto com o algoritmo K-Means. Este método consiste em executar o algoritmo para diferentes valores de k (número de clusters). Através da análise visual do gráfico gerado, foi possível identificar o número de clusters que melhor representa a estrutura dos dados, evitando tanto o underfitting, com poucos clusters, como o overfitting, com excesso de clusters. Com este método obtivemos $k = 6$ com um score de 13604.

7.4.7.2. Silhouette method

Para complementar a determinação do número ótimo de clusters, foi aplicado o silhouette method, que avalia a qualidade da segmentação resultante para diferentes valores de k . Ao calcular a média dos coeficientes para diferentes números de clusters, é possível identificar o valor de k que maximiza esta média, indicando a segmentação mais coesa e bem definida. Obtivemos $k = 3$ com este método e um score de 0.20.

Tendo em consideração que obtivemos dois valores possíveis para k e que os respetivos scores de avaliação não foram particularmente satisfatórios, optámos por definir o número de clusters como 4. Esta decisão baseou-se numa análise visual dos agrupamentos, na qual considerámos que esta configuração proporcionava uma segmentação mais coerente e adequada para futuras recomendações.

7.4.8. Clustering com K-Means

Após definir o número ótimo de clusters, foi aplicado o algoritmo K-Means com 4 clusters ao conjunto de dados numéricos imputados. O modelo foi configurado para múltiplas reinicializações ($n_init = 10$) e um limite de 100 iterações para assegurar a convergência.

Posteriormente, foi calculada a frequência de elementos em cada cluster para ordenar os clusters por tamanho. Com base nesta ordenação, os rótulos originais foram reatribuídos para que o cluster mais populoso tenha o rótulo 0, o segundo mais populoso o rótulo 1, e assim sucessivamente. Esta reclassificação dos rótulos permite uma interpretação mais intuitiva dos clusters, facilitando a análise e comunicação dos resultados.

Vale a pena referir que também testámos a utilização do algoritmo DBSCAN para o agrupamento. No entanto, revelou-se pouco funcional no nosso contexto, uma vez que, independentemente dos valores atribuídos aos parâmetros, os resultados obtidos não foram satisfatórios. Em todos os cenários testados, verificaram-se apenas duas possibilidades: ou o algoritmo criava mais de 10 clusters, sendo que um deles concentrava cerca de 98% dos clientes enquanto os restantes tinham até apenas um cliente cada; ou então formava apenas dois clusters, dos quais um incluía 1740 registos e o outro apenas 18. Este comportamento indicou uma segmentação inadequada para os nossos objetivos, pelo que optámos por não prosseguir com o DBSCAN.

7.5. Validação do Desempenho do Modelo

De forma a podermos avaliar o desempenho do modelo desenvolvido, foi criada uma visualização dos clusters de clientes para análise. O gráfico permite observar a distribuição espacial e a separação entre os clusters, ajudando a validar visualmente a segmentação realizada.

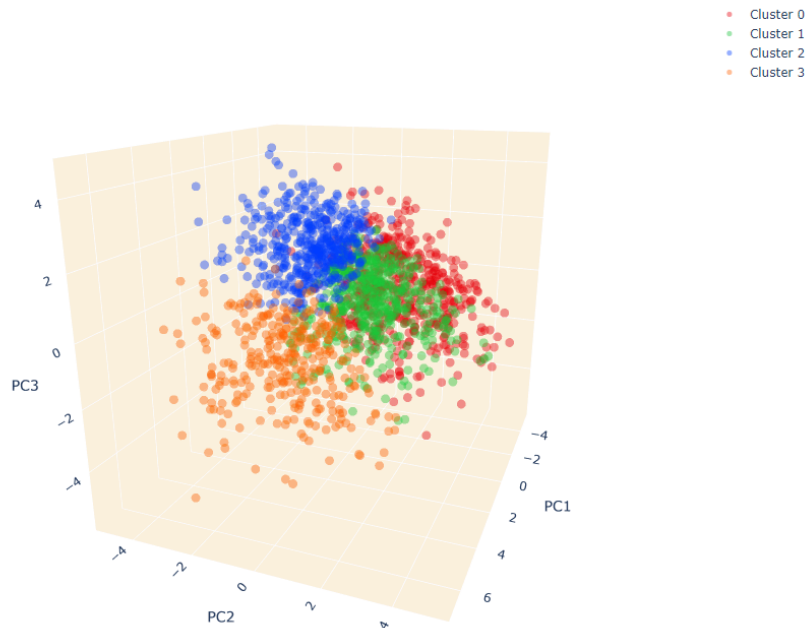


Figura 41: Visualização 3D dos clusters dos clientes

Com esta visualização consegue-se entender que há alguma sobreposição entre clusters, mas mesmo assim ainda conseguimos definir pontos mais individuais de cada cluster em concreto.

Para validar a qualidade da segmentação obtida com o algoritmo K-Means, foram calculadas três métricas distintas que avaliam diferentes aspetos da formação dos clusters:

- **Silhouette Score:** Mede a coesão e separação dos clusters, com valores entre -1 e 1 . Valores próximos de 1 indicam clusters bem definidos, enquanto valores negativos indicam sobreposição ou má separação.
- **Calinski-Harabasz Score:** Avalia a razão entre a dispersão intercluster e intracluster. Valores mais elevados indicam uma melhor separação e compactação dos clusters.
- **Davies-Bouldin Score:** Mede a similaridade entre clusters, considerando a distância entre clusters e a dispersão dentro dos clusters. Valores mais baixos indicam clusters mais distintos e compactos.

Os resultados obtidos fornecem uma avaliação quantitativa da qualidade da segmentação, permitindo confirmar a adequação do número de clusters escolhido e a robustez da análise.

Metрика	Valor
<i>Silhouette Score</i>	0.2457
<i>Calinski-Harabasz Score</i>	632.54
<i>Davies-Bouldin Score</i>	1.2024

Estes foram os resultados que obtivemos para o caso dos 4 clusters, configuração que decidimos adotar após análise conjunta dos resultados do método do cotovelo (Elbow Method), do Silhouette Score e da visualização tridimensional dos dados.

7.6. Avaliação dos Resultados

Depois da escolha final de utilizarmos 4 clusters a partir da reflexão de todas as métricas que analisamos procedemos à avaliação dos resultados, ou seja, verificar se a distribuição entre os 4 clusters estaria equilibrada, analisar cada perfil de forma a facilitar futuras recomendações e ver as diferenças dos valores de cada atributo em diferentes clusters.

7.6.1. Distribuição de clientes por cluster

Analisando o gráfico apresentado abaixo, verificamos que os clusters estão, de forma geral, equilibrados, com exceção do Cluster 3, que apresenta uma menor percentagem de clientes em comparação com os restantes.

Concluimos, portanto, que a escolha de uma segmentação com 4 clusters, em vez de 3, permite que os clientes do Cluster 3 sejam representados de forma distinta, ao invés de serem diluídos entre os outros clusters. Apesar da sua menor representatividade, o segmento possui características suficientemente específicas para justificar a sua independência.

É importante destacar que, ao aumentar o número de clusters, tenderiam a surgir grupos com percentagens muito reduzidas, o que poderia prejudicar a eficácia de futuras estratégias de recomendação. Isto porque, ao criar clusters excessivamente segmentados, correríamos o risco de gerar recomendações demasiado específicas para um número muito pequeno de clientes. Esses microgrupos, muitas vezes, poderiam ser integrados em clusters semelhantes, otimizando os recursos e permitindo recomendações mais abrangentes e eficazes.

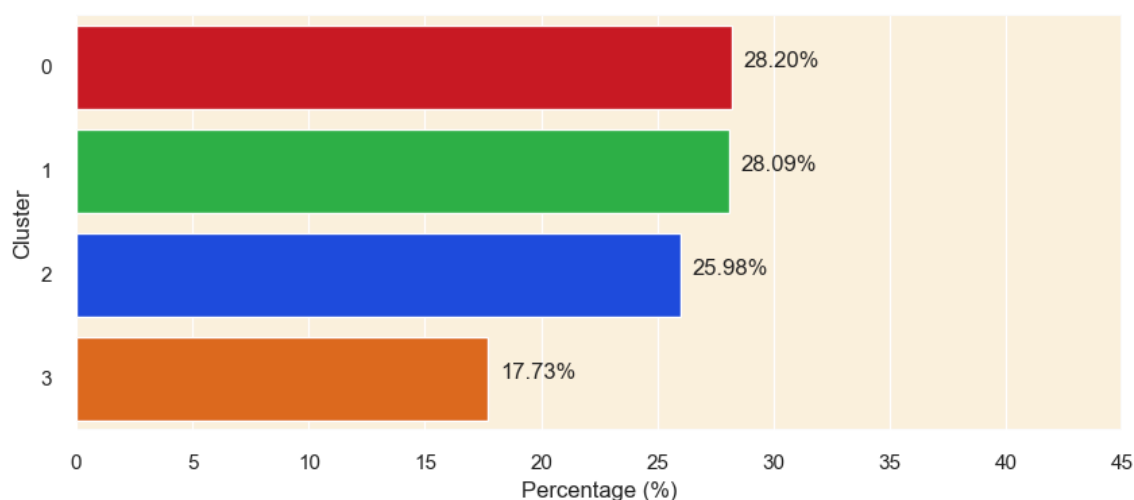


Figura 42: Distribuição dos clientes por Cluster

7.6.2. Análise do perfil dos clientes de cada Cluster

Após a análise da distribuição por clusters, passamos a examinar o perfil de cada grupo com base em gráficos que mostram, de forma visual, como cada atributo se comporta em relação à média geral. A partir da distância em cada gráfico, conseguimos perceber se determinado atributo tende a apresentar valores mais altos ou mais baixos naquele cluster específico.

Por exemplo, valores elevados no atributo `Day_Of_Week` indicam que os consumidores daquele cluster tendem a realizar compras mais próximo do fim de semana. Já valores mais baixos sugerem preferência por compras no início da semana.

Com essa lógica em mente, iniciamos a análise dos quatro clusters, focando apenas nas tendências apresentadas nos gráficos — sem considerar, por enquanto, a distribuição numérica específica de cada atributo.

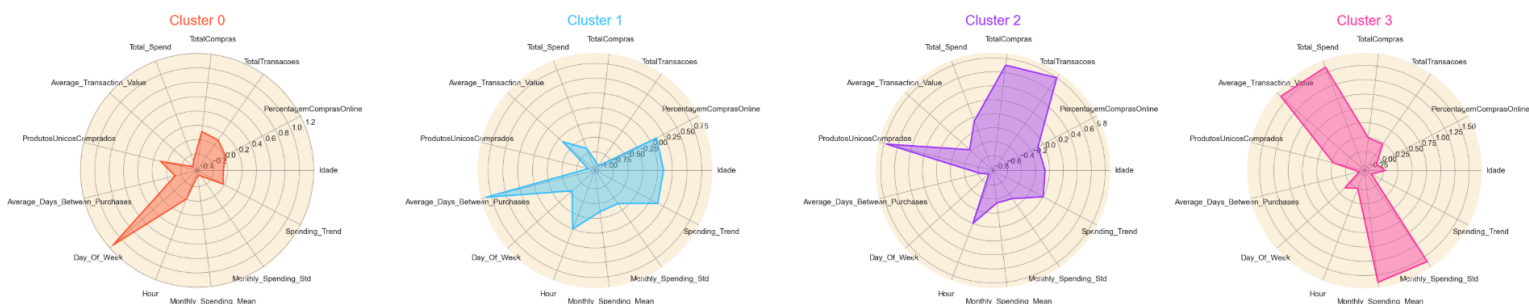


Figura 43: Gráfico do Perfil de cada cliente por Cluster

7.6.2.1. Cluster 0 (Cliente de compras pontuais e baixo envolvimento)

O Cluster 0 corresponde aos clientes com baixo nível de engagement e gastos reduzidos. Estes clientes apresentam valores baixos em variáveis como total gasto, número total de compras e transações, bem como valor médio por transação. Compram poucos produtos diferentes e têm uma média mensal de gastos bastante baixa, tal como o valor médio por transação. No entanto, nota-se um pico no atributo de compras acerca do dia da semana, o que sugere que concentram as suas compras em dias específicos - provavelmente ao fim de semana. Este grupo tende a fazer compras pontuais de baixo custo e não demonstra um comportamento de consumo consistente ao longo do tempo. São os clientes que se costuma chamar como os “Compradores pontuais de fim de semana”.

7.6.2.2. Cluster 1 (Cliente moderado e consistente)

O Cluster 1 representa um perfil de cliente mais moderado e consistente em relação ao anterior. A maioria dos valores deste grupo são medianos, o que indica um comportamento equilibrado. Apresentam um intervalo elevado entre compras, ou seja, compram com pouca frequência, mas de forma regular. Têm um padrão mensal de gastos relativamente estável, sem grandes variações, o que sugere um comportamento previsível e planeado, além do facto de terem um valor baixo nos produtos únicos, podendo indicar a compra de um item repetido. A idade média dos clientes deste cluster é superior à dos outros grupos, apontando para um perfil possivelmente mais conservador e metódico nas decisões de compra, provavelmente planeando-as com antecedência.

7.6.2.3. Cluster 2 (Comprador frequente e diversificado)

O Cluster 2 agrupa os compradores mais frequentes e diversificados. Este grupo destaca-se pelo número elevado de compras e transações, assim como pela grande variedade de produtos adquiridos. O total gasto também é elevado, embora o valor médio por transação seja relativamente baixo. Os clientes deste cluster têm uma elevada percentagem de compras realizadas online, o que indica uma forte preferência pelo comércio eletrónico (apesar de uma percentagem mais baixa que no Cluster 1, devido ao número de transações deste cluster, podemos afirmar que estes clientes têm um maior número de transações online). O seu comportamento sugere um padrão de consumo regular, com alguma diversidade, o que os torna altamente ativos em termos de volume de interações comerciais e dos melhores candidatos a recomendações.

7.6.2.4. Cluster 3 (Cliente dispendioso e volátil)

O Cluster 3 identifica os clientes dispendiosos (são o grupo com maior gasto total) e com comportamento volátil. Estes clientes realizam poucas compras, mas gastam muito em cada transação, apresentando o valor médio por transação mais elevado de todos os grupos. Além disso, registam

uma grande variação nos gastos mensais, alternando entre períodos de grande consumo e outros de inatividade. Apesar de comprarem poucos produtos e realizarem poucas transações, o seu impacto financeiro é elevado. Este perfil pode corresponder a consumidores que fazem compras esporádicas mais impulsivas/imprevisíveis ou de produtos de maior valor, como artigos de tecnologia luxuosos.

7.6.3. Análise dos diferentes atributos em cada cluster

Após analisarmos os clusters, procedemos a uma análise mais aprofundada dos valores dos atributos em cada grupo. Para isso, recorremos à visualização por histogramas, que nos permitiu observar a distribuição de cada variável dentro de cada cluster. Esta abordagem possibilitou uma caracterização mais precisa dos perfis de clientes, revelando padrões de comportamento, preferências de compra, variações no gasto e frequência de transações ao longo do tempo. Com estas análises as nossas recomendações futuras terão um melhor fundamento.



Figura 44: Distribuição dos Valores de cada atributo por Cluster

7.6.3.1. Idade

No que diz respeito à idade dos clientes, observam-se diferenças marcantes entre os clusters. O Cluster 0 agrupa maioritariamente indivíduos entre os 25 e os 45 anos, refletindo um público relativamente jovem. Já o Cluster 1 distingue-se por uma faixa etária mais elevada, com muitos clientes acima dos 50 anos, indicando um perfil sénior. O Cluster 2 concentra-se na faixa dos 35 aos 50 anos, uma população adulta ativa, enquanto o Cluster 3 apresenta uma distribuição mais heterogénea, embora com uma ligeira predominância nas faixas médias, revelando uma mistura de consumidores jovens e de meia-idade.

7.6.3.2. Percentagem de Compras Online

A percentagem de compras realizadas online também varia de forma expressiva entre os clusters. O Cluster 2 evidencia uma forte inclinação para o comércio digital, com uma elevada percentagem de transações online. O Cluster 0 apresenta uma percentagem mais equilibrada, revelando hábitos mistos entre compras físicas e digitais. Por contraste, o Cluster 1 mostra uma menor adesão às compras online, o que está em linha com o perfil mais sénior deste grupo, apesar de no cluster ter valor elevado, devido à existência de alguns outliers que só fizeram compras online. O Cluster 3 apresenta uma ampla variabilidade, mas tende a incluir consumidores com comportamentos digitais diversificados.

7.6.3.3. Total de Transações

Analisando o total de transações realizadas, o Cluster 2 destaca-se com um número bastante elevado, refletindo um padrão de consumo frequente. O Cluster 0 também apresenta um volume considerável, embora inferior ao do Cluster 2. O Cluster 1 regista o menor número de transações, condizente com o perfil de menor atividade de compra. Já o Cluster 3 apresenta um número moderado de transações, embora os seus valores de compra sejam, em média, mais elevados.

7.6.3.4. Total de Compras

No total de compras, a tendência repete-se: o Cluster 2 lidera em volume, seguido de perto pelo Cluster 0. O Cluster 1 volta a registar os valores mais baixos, confirmando um padrão de consumo mais esporádico. O Cluster 3 situa-se entre o Cluster 0 e 1, com um número intermédio de compras, mas, como veremos adiante, com maior peso financeiro por transação.

7.6.3.5. Total Gasto

Quando se observa o total gasto pelos clientes, o Cluster 3 destaca-se claramente com os valores mais elevados, resultado do alto valor médio por transação. O Cluster 2 também regista um gasto total significativo, compatível com o elevado número de transações. O Cluster 0 mostra um gasto total moderado, enquanto o Cluster 1 apresenta o menor nível de despesa, o que está alinhado com a sua menor frequência de compras e transações.

7.6.3.6. Valor Médio por Transação

O valor médio por transação permite perceber diferenças no tipo de consumo. O Cluster 3 lidera com uma média bastante elevada, indicando compras de maior valor unitário. O Cluster 2 e o Cluster 1 apresentam valores intermédios, embora no caso do Cluster 1 esse valor reflita compras pouco frequentes. O Cluster 0, por outro lado, revela um valor médio por transação mais baixo, o que indica hábitos de consumo mais regulares mas com menor peso por compra.

7.6.3.7. Produtos Únicos Comprados

No que toca à diversidade de produtos adquiridos, o Cluster 2 evidencia uma boa variedade, sugerindo um perfil de consumo mais diversificado. O Cluster 0 e o Cluster 3 seguem uma tendência semelhante,

embora com ligeiras diferenças no número médio de produtos únicos. Já o Cluster 1 volta a evidenciar um comportamento mais restrito, adquirindo uma menor diversidade de produtos.

7.6.3.8. Dias Médios entre Compras

Este indicador permite avaliar a frequência de compra. O Cluster 2 apresenta os intervalos mais curtos entre compras, o que confirma o seu padrão de consumo intenso. O Cluster 3 também mostra uma frequência relativamente alta, mas com alguns valores elevados. Por outro lado, o Cluster 1 apresenta os maiores intervalos, reforçando o seu perfil de consumidor esporádico. O Cluster 0 encontra-se numa posição intermédia, com uma frequência moderada.

7.6.3.9. Dia da Semana

Relativamente ao dia da semana em que ocorrem as compras, o Cluster 0 concentra a sua atividade a partir de sexta, sugerindo um perfil de consumidor que realiza compras no final de semana. O Cluster 1 prefere os dias úteis, especialmente segunda e quarta-feira. O Cluster 2 apresenta uma distribuição mais homogênea pelos dias da semana (principalmente no início), enquanto o Cluster 3 exibe alguma variabilidade, mas com ainda alguma incidência ao fim de semana.

7.6.3.10. Média de Gastos Mensais

Em relação à despesa média mensal, o Cluster 3 destaca-se com os valores mais altos, seguido do Cluster 2. Estes dois grupos revelam perfis de consumo mais intensos e consistentes. O Cluster 0 apresenta valores moderados, e o Cluster 1 novamente regista os valores mais baixos, condizentes com a sua menor atividade global.

7.6.3.11. Desvio Padrão dos Gastos Mensais

O desvio padrão da despesa mensal permite avaliar a consistência do comportamento de compra. O Cluster 3 apresenta elevada variabilidade, sugerindo meses com gastos muito distintos. O Cluster 2 também mostra alguma flutuação, enquanto o Cluster 0 e, sobretudo, o Cluster 1 apresentam padrões mais estáveis e controlados ao longo do tempo.

7.6.3.12. Tendência de Gastos

Por fim, a tendência de gasto ao longo do tempo mostra uma trajetória crescente nos Clusters 2 e 3, o que pode indicar um aumento do engagement destes clientes ou mudanças no seu padrão de consumo. O Cluster 0 mantém uma tendência relativamente estável, enquanto o Cluster 1 apresenta pouca variação ao longo do tempo, mantendo o seu perfil conservador.

8. Personalização de Ofertas de Produtos e Serviços

8.1. Definição do Problema e Compreensão dos Elementos de Análise Envolvidos

No contexto do negócio em questão, torna-se fundamental realizar uma análise aprofundada dos hábitos de compra dos clientes. Compreender padrões de consumo, preferências individuais e comportamentos recorrentes permite à loja alinhar melhor a sua oferta com as expectativas e necessidades do público-alvo. Através desta análise, é possível implementar sistemas de recomendação personalizados que sugiram produtos ajustados aos interesses específicos de cada cliente ou segmentos de clientes. É ainda relevante perceber quais são as *trends* de vendas. Desta forma é possível recomendar produtos populares a clientes novos na loja que não têm um histórico de compras na loja.

8.2. Seleção e Preparação dos Dados

Apontando para a Secção 7.2, é importante destacar que os sistemas de recomendação representam uma extensão natural do trabalho realizado durante a segmentação de clientes. Os dados utilizados foram previamente preparados no âmbito do processo de ETL. Posteriormente, foi realizada uma etapa de feature engineering, na qual foram geradas novas variáveis a partir dos dados existentes no data warehouse, enriquecendo assim o conjunto de dados para suportar os modelos de recomendação.

8.3. Identificação e Fundamentação da Técnica de Análise

O modelo de recomendação deve considerar dois tipos distintos de clientes:

- **Cliente novo:** trata-se de um cliente que ainda não efetuou qualquer compra na loja. Por esse motivo, não existem dados que permitam inferir os seus hábitos de consumo, nem é possível determinar a que segmento de clientes pertence.
- **Cliente existente:** este cliente já realizou compras previamente na loja, o que permite reunir informação sobre os seus padrões de consumo e classificá-lo num determinado segmento de clientes.

8.3.1. Estratégia para clientes novos

No que diz respeito aos clientes novos, a ausência de dados históricos impede a personalização das recomendações com base em preferências individuais. Nestes casos, a abordagem mais eficaz consiste

em sugerir os produtos mais vendidos. Esta estratégia assenta na técnica de **demographic filtering**, recorrendo ao volume de vendas para identificar os produtos com maior probabilidade de atrair o interesse de novos clientes.

8.3.2. Estratégia para clientes existentes

Para os clientes existentes, que já realizaram compras na loja, é possível tirar partido dos dados históricos para oferecer recomendações mais personalizadas e eficazes. Esta abordagem permite não só aumentar a relevância das recomendações, mas também reforçar a relação com o cliente, incentivando a repetição de compras e promovendo a fidelização. Para este tipo de clientes iremos utilizar três abordagens diferentes:

- **Content-based filtering**: gerar recomendações com base no histórico de compras de um dado cliente
- **Collaborative filtering**: gerar recomendações com base nas características de um perfil de cliente
- **Recomendação híbrida**: combinar os dois métodos mencionados anteriormente através da média ponderada

8.4. Construção do Modelo de Análise

8.4.1. Demographic filtering

Produtos com elevada procura tendem a ser apelativos para um público mais vasto, incluindo aqueles que ainda não interagiram com a loja. Para garantir que a trend representa hábitos de compra mais atuais, vamos incluir apenas compras realizadas até um mês antes da compra mais recente. Visto que a compra mais antiga foi feita em 2010, considerou-se relevante excluir produtos que tiveram altos volumes de vendas no passado. Por exemplo, um smartphone que em 2010 teve um número recorde de vendas, já não terá a mesma relevância em 2025 visto que a tecnologia evoluiu e é de esperar que o seu sucessor tenha um menor volume de vendas.

8.4.2. Content-based filtering

Uma abordagem comum para clientes já existentes é a aplicação de algoritmos de **content-based filtering**, que analisam as características dos produtos adquiridos anteriormente e os comparam com os de outros artigos disponíveis na loja. Esta técnica permite gerar recomendações alinhadas com os interesses e preferências demonstrados pelo cliente ao longo do tempo, oferecendo sugestões com base em características como categoria, marca, preço, ou funcionalidades.

Ao centrar-se no perfil individual do utilizador, o content-based filtering proporciona uma experiência de compra personalizada e consistente, reforçando o envolvimento do cliente com a plataforma e aumentando a probabilidade de conversão.

Passando para a implementação da estratégia, o primeiro passo é obter o histórico de compras do cliente. De seguida, as características mais relevantes dos produtos — como a categoria e a marca — são convertidas numa representação binária, o que permite uma comparação mais eficaz entre os diferentes artigos. Com base nesta representação, é elaborado um perfil do cliente que sintetiza as suas preferências médias, considerando os atributos dos produtos anteriormente adquiridos. Posteriormente, calcula-se a similaridade entre este perfil e todos os produtos disponíveis, através de uma métrica que quantifica o grau de afinidade entre eles. Por fim, são filtrados e ordenados os produtos que o cliente ainda não adquiriu, privilegiando-se aqueles que apresentam maior correspondência com o seu perfil, os quais são então recomendados.

8.4.3. Collaborative filtering

Outra abordagem consiste na utilização de algoritmos de **collaborative filtering**, que comparam os hábitos de compra de um cliente com os de outros clientes com perfis semelhantes. Esta técnica permite identificar padrões de comportamento partilhados dentro de um determinado segmento e, com base nesses padrões, gerar recomendações mais relevantes.

Ao alinhar as sugestões com as preferências de um grupo de utilizadores com características semelhantes, é possível oferecer produtos que, embora ainda não tenham sido adquiridos pelo cliente em questão, apresentam uma elevada probabilidade de correspondência com os seus interesses. Esta abordagem contribui para uma experiência de recomendação mais robusta, mesmo na ausência de um histórico de compras extenso.

Na implementação desta estratégia, os clientes são previamente agrupados em clusters com base em características comuns. Para cada cluster, é criada uma matriz que relaciona clientes e produtos, refletindo a frequência de compra dos artigos pelos membros desse grupo. Seguidamente, calcula-se a popularidade de cada produto dentro de cada cluster, considerando métricas como o número de clientes que adquiriram o produto, a quantidade total vendida e o volume de receita gerada, assim como a taxa de penetração, que representa a percentagem de clientes do cluster que compraram aquele produto. Os produtos são depois ordenados segundo estes indicadores, permitindo identificar os mais relevantes para cada grupo. Por fim, as recomendações para um cliente específico são feitas com base no seu cluster, excluindo-se os produtos que já comprou, e sugerindo os artigos mais populares e relevantes dentro do seu grupo.

8.4.4. Recomendação híbrida

Adicionalmente, é possível **combinar o content-based filtering com o collaborative filtering**, recorrendo a modelos híbridos que exploram tanto os comportamentos individuais como os de clientes com perfis semelhantes. Esta abordagem permite aumentar a precisão das recomendações, promovendo uma experiência de compra mais personalizada e contribuindo para a fidelização do cliente e o aumento das vendas.

Ao conjugar estas duas técnicas, deixamos de depender exclusivamente dos hábitos de compra do utilizador, o que facilita a introdução de produtos que não fazem parte do seu histórico, mas que apresentam elevada probabilidade de interesse com base em padrões partilhados por outros clientes. Desta forma, não só se alarga a diversidade das recomendações, como também se abre espaço à descoberta de novos produtos, incentivando a exploração e o cruzamento entre diferentes perfis de consumo.

Para gerar as recomendações usou-se a **média ponderada**. Após obtidas as recomendações do collaborative filtering e do content-based filtering, estas foram combinadas atribuindo-se pesos distintos a cada método, com maior importância dada ao content-based filtering (por exemplo, 60%) e um peso menor ao collaborative filtering (cerca de 40%). Esta ponderação permite privilegiar as preferências individuais do cliente, ao mesmo tempo que incorpora tendências e padrões observados em clientes semelhantes, resultando numa lista final de produtos mais equilibrada, precisa e diversificada.

8.5. Validação do Desempenho do Modelo

8.5.1. Demographic filtering

Tendo em conta que, no período de tempo considerado, a loja registou um total de 630 vendas, e que o produto mais vendido representa apenas cerca de 4% do volume de negócios, verifica-se que os produtos com maior número de vendas não são necessariamente os mais populares ou relevantes em termos de interesse geral dos clientes. Isto indica uma grande dispersão nas preferências de compra, o que pode dificultar a identificação de padrões claros de popularidade.

Desta forma, um sistema de recomendação baseado apenas nos produtos mais vendidos pode não ser o mais eficaz, especialmente no caso de clientes novos, para os quais não existe histórico de compras. Este tipo de abordagem pressupõe a existência de produtos significativamente populares e uma distribuição de vendas mais concentrada, o que não se verifica neste caso. Além disso, este método não considera fatores importantes como a sazonalidade ou a complementaridade entre produtos.

8.5.2. Content-based filtering

Para validar as recomendações geradas com base no histórico de compras de um determinado cliente, foi realizada uma análise da distribuição percentual das categorias dos produtos adquiridos por esse cliente. Este procedimento permite avaliar a relevância das recomendações, comparando-as com os padrões de compra previamente observados.

Verificou-se que o cliente em questão apresenta uma forte preferência por monitores e periféricos, categorias que representam uma parte significativa do seu histórico de compras. As recomendações geradas, por sua vez, incluem predominantemente produtos dessas mesmas categorias, o que indica que são coerentes com os interesses do cliente e, portanto, plausivelmente válidas e realistas.

8.5.3. Collaborative filtering

Para validar o desempenho de um sistema de recomendação baseado em perfis de cliente, optou-se por analisar os tipos de produtos recomendados para cada perfil. Verificou-se que os produtos recomendados, de forma geral, caem nas categorias de impressoras e componentes.

Como consequência, há uma limitação na diversidade das recomendações, o que pode levar à saturação de certas categorias e à falta de promoção de outras potencialmente relevantes. Este comportamento pode comprometer a personalização esperada por parte dos utilizadores, afetando negativamente a sua experiência e diminuindo o potencial de descoberta de novos produtos. Além disso, pode limitar as oportunidades comerciais da plataforma, ao não explorar de forma equilibrada o portfólio completo de produtos disponíveis.

8.6. Avaliação dos Resultados

8.6.1. Demographic filtering

Para gerar as recomendações do demographic filtering, começamos por filtrar os produtos que foram vendidos até um mês antes da compra mais recente. Feita essa filtragem conseguimos ordenar os produtos vendidos por número de vendas e gerar o top 10 produtos mais vendidos.

Produtoid	Nº vendas	Nome	Marca	Categoria
117	25	Placa Gráfica NVIDIA GTX 1660 Super	NVIDIA	Componentes
102	19	Placa Gráfica Gigabyte RTX 3060 Ti	Gigabyte	Componentes
74	19	Impressora Epson EcoTank L3150	Epson	Impressoras
69	16	Impressora HP DeskJet 2720	HP	Impressoras
104	15	Impressora Brother HL-L2350DW	Brother	Impressoras
15	14	Smartphone Xiaomi Redmi Note 9	Xiaomi	Smartphones
2	14	Logitech Rato MX Anywhere	Logitech	Periféricos
134	12	Headset SteelSeries Arctis 1	SteelSeries	Audio
119	12	Impressora HP LaserJet Pro M15w	HP	Impressoras
48	12	Monitor 24 BenQ	BenQ	Monitores

Como mencionado anteriormente, o volume de vendas dos produtos recomendados durante o período considerado é relativamente baixo, logo, estes não parecem ser efetivamente populares e do interesse geral. Desta forma, este tipo de recomendação poderá não ser a mais indicada para clientes para os quais não existem dados.

8.6.2. Content-based filtering

Para este tipo de recomendação tivemos em conta os dados relativos ao cliente com ID 100.

Produtoid	Nome
5	Monitor Dell 24"
113	Monitor 24 Polegadas ASUS ROG
128	Monitor 32 HP Omen
173	Monitor 27 ASUS ProArt
21	Rato SteelSeries Rival 3
51	Rato SteelSeries Aerox 3
81	Rato ASUS ROG Chakram
151	Rato SteelSeries Prime

171	Rato SteelSeries Rival 5
18	Monitor 32 LG"

As recomendações geradas, apesar de estarem alinhadas com o histórico de compras do cliente em questão, apresentam uma limitação que pode impactar negativamente a experiência de compra na loja. Uma vez que o sistema se baseia exclusivamente em produtos anteriormente adquiridos pelo cliente, não promove a descoberta de novos produtos ou categorias, restringindo a diversidade das sugestões apresentadas.

Este tipo de abordagem tende a reforçar padrões de consumo já existentes, o que pode resultar numa experiência repetitiva e pouco inspiradora para o utilizador.

8.6.3. Collaborative filtering

Para este tipo de recomendação mantivemos os dados relativos ao cliente com ID 100.

Obtivemos as seguintes recomendações para o cluster onde o cliente em questão se insere:

Produto Id	Nome	Nº Compras	Nº Clientes	Qtd Total
74	Impressora Epson EcoTank L3150	233	184	466
104	Impressora Brother HL-L2350DW	231	184	515
72	Placa Gráfica ASUS RTX 3070	227	182	502
102	Placa Gráfica Gigabyte RTX 3060 Ti	231	180	487
87	Placa Gráfica MSI RTX 3080	229	173	489
89	Impressora Canon PIXMA G6020	218	173	475
139	Impressora Epson WorkForce WF-110	216	168	472
67	Placa Gráfica NVIDIA RTX 3060	226	166	520
69	Impressora HP DeskJet 2720	212	166	471
137	Placa Gráfica ASUS TUF RTX 3090	206	161	462

Comparando a tabela de produtos recomendados com os dados reais de compras efetuadas por este cliente, observa-se que a maioria das sugestões incide sobre impressoras e componentes. Estes tipos de artigos representam uma fatia significativa do volume total de compras do cliente em análise, o que indica que as recomendações para o perfil onde o cliente se insere, estão alinhadas com os seus interesses e padrões de consumo. Assim, estas sugestões podem ser consideradas relevantes, realistas e coerentes com o comportamento de compra previamente registado.

No entanto, para evitar redundância e melhorar a utilidade das recomendações, optou-se por remover da lista os produtos que o cliente já adquiriu. Esta abordagem tem como objetivo promover a descoberta de novos artigos.

Com base nos dados anteriores, gerou-se as seguintes recomendações para o cliente em questão, removendo das recomendações para o cluster os produtos já comprados pelo cliente:

Produto Id	Nome
104	Impressora Brother HL-L2350DW
102	Placa Gráfica Gigabyte RTX 3060 Ti
87	Placa Gráfica MSI RTX 3080
139	Impressora Epson WorkForce WF-110
67	Placa Gráfica NVIDIA RTX 3060
69	Impressora HP DeskJet 2720
117	Placa Gráfica NVIDIA GTX 1660 Super
65	Smartphone Samsung Galaxy Note 20
60	Smartphone Google Pixel 5
35	Smartphone iPhone 12

Os resultados mantêm-se alinhados com os padrões de consumo tanto do perfil de cliente típico como do cliente específico em análise, com a diferença de que passam agora a incluir apenas produtos que ainda não foram adquiridos por este. Esta abordagem permite preservar a relevância das recomendações, ao mesmo tempo que introduz variedade e promove a descoberta de novos artigos potencialmente interessantes.

No entanto, apesar de favorecer a exposição a novos produtos, esta estratégia reduz ligeiramente a influência direta do histórico de compras do cliente. Ao excluir itens já adquiridos, corre-se o risco de desvalorizar padrões de preferência mais consistentes, o que pode afetar negativamente a personalização em determinados casos.

8.6.4. Recomendação híbrida

Tendo em conta os pesos mencionados anteriormente, foram geradas as seguintes recomendações para o cliente com ID 100.

Produto Id	Nome	Pontuação
5	Monitor Dell 24"	0.60
113	Monitor 24 Polegadas ASUS ROG	0.54
128	Monitor 32 HP Omen"	0.48
173	Monitor 27 ASUS ProArt"	0.42
104	Impressora Brother HL-L2350DW	0.40
102	Placa Gráfica Gigabyte RTX 3060 Ti	0.36
21	Rato SteelSeries Rival 3	0.36
87	Placa Gráfica MSI RTX 3080	0.32
51	Rato SteelSeries Aerox 3	0.30
139	Impressora Epson WorkForce WF-110	0.28

A recomendação híbrida com maior peso no content-based filtering revelou-se eficaz na personalização das sugestões, alinhando-as com os interesses individuais dos clientes. Esta abordagem garantiu relevância ao basear-se nas características dos produtos já adquiridos, mantendo consistência com o perfil de consumo.

A componente de collaborative filtering, embora com menor influência, introduziu alguma diversidade, permitindo sugerir produtos ainda não explorados mas relevantes com base em padrões de clientes semelhantes.

Apesar da qualidade das recomendações, a forte dependência do histórico individual pode limitar a descoberta de novos produtos. Assim, embora os resultados sejam positivos, especialmente para clientes com um histórico consistente, seria vantajoso ajustar dinamicamente o equilíbrio entre os dois métodos consoante o perfil do utilizador. Por exemplo, no caso de um cliente com um histórico de compras mais curto ou pouco representativo, faz sentido atribuir maior peso às preferências de clientes semelhantes, aproveitando a componente de collaborative filtering para enriquecer as recomendações.

9. Conclusões e Trabalho Futuro

O desenvolvimento deste sistema de suporte à decisão representou uma transformação significativa para a *Perifericum*, convertendo dados operacionais dispersos em ativos estratégicos valiosos. A implementação de um armazém de dados centralizado, complementado por pipelines ETL automatizadas e ferramentas avançadas de visualização e análise preditiva, dotou a organização de capacidades analíticas robustas e orientadas ao negócio.

Os resultados alcançados demonstram o cumprimento integral dos objetivos inicialmente estabelecidos. A caracterização detalhada dos perfis de clientes, através de técnicas de clustering, permitiu uma segmentação refinada da base de clientes, possibilitando a personalização eficaz de ofertas comerciais. Os dashboards interativos desenvolvidos permitiram o acesso à informação crítica, facilitando a tomada de decisão em diferentes níveis organizacionais. Esta nova infraestrutura analítica permitiu transformar dados em conhecimento, o qual fundamenta decisões mais estratégicas e centradas no cliente.

A arquitetura implementada provou ser não apenas tecnicamente sólida, mas fundamentalmente alinhada com a visão estratégica da empresa, proporcionando uma base escalável para o seu posicionamento competitivo nos ambientes físico e digital. Contudo, a jornada de transformação digital não termina com este projeto.

Para evolução futura, identifica-se um conjunto de oportunidades estratégicas. A implementação de integração de dados em tempo real representaria um avanço significativo, permitindo responder com maior agilidade às dinâmicas do mercado e às mudanças comportamentais dos consumidores. A expansão do sistema para incorporar *Data Marts* adicionais em áreas como Logística, Gestão de Inventário e Marketing potencializaria análises interdepartamentais mais abrangentes, catalisando sinergias operacionais. Adicionalmente, o refinamento dos algoritmos de recomendação, explorando técnicas avançadas de filtragem colaborativa e híbrida, elevaria a precisão e relevância das sugestões personalizadas, aprofundando o relacionamento com os clientes.

Em suma, o sistema desenvolvido não só modernizou fundamentalmente as capacidades analíticas da loja, como estabeleceu alicerces sólidos para inovações futuras. A transformação iniciada posiciona a organização na vanguarda da tomada de decisão baseada em dados, permitindo-lhe antecipar tendências, responder proativamente às necessidades dos clientes e capitalizar oportunidades emergentes num mercado cada vez mais competitivo e digitalizado.

Referências

- Golfarelli, M. (2009,). *From User Requirements to Conceptual Design in Data Warehouse Design: A Survey*. https://www.researchgate.net/publication/228618702_From_User_Requirements_to_Conceptual_Design_in_Data_Warehouse_Design-a_Survey
- Golfarelli, M., Maio, D., & Rizzi, S. (1998,). *The Dimensional Fact Model: A Conceptual Model for Data Warehouses*. https://www.researchgate.net/publication/220095043_The_Dimensional_Fact_Model_A_Conceptual_Model_for_Data_Warehouses
- McKinsey & Company. (2021,). *The Value of Getting Personalization Right—or Wrong—Is Multiplying*. <https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/the-value-of-getting-personalization-right-or-wrong-is-multiplying>
- Oliveira, B., Oliveira, Ó., & Belo, O. (2021,). *Using BPMN for ETL Conceptual Modelling: A Case Study*. https://www.researchgate.net/publication/353396624_Using_BPMN_for_ETL_Conceptual_Modelling_A_Case_Study

Lista de Siglas e Acrónimos

<i>DW</i>	<i>Data Warehouse</i>
<i>UX</i>	<i>User Experience</i>
<i>ETL</i>	Extração, Transformação e Carregamento
<i>DFM</i>	Dimensional Fact Model
<i>SCD</i>	Slowly Changing Dimension