

Uma solução de Business Intelligence para a área de recursos humanos da U.Porto

Solange Sampaio Perdigão

Mestrado Integrado em Engenharia de Redes e Sistemas
Informáticos

Departamento de Ciência de Computadores

2021

Orientador

Fernando Silva

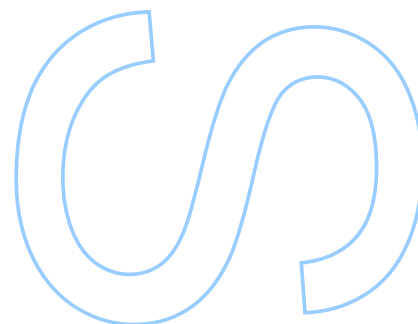
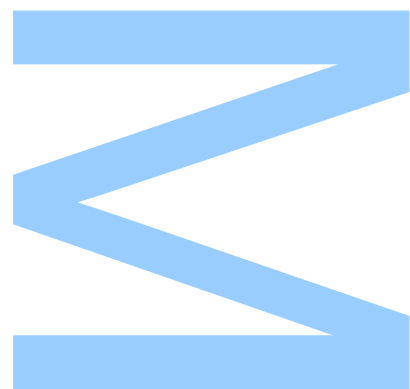
Professor Catedrático

Faculdade de Ciências da Universidade do Porto

Supervisor

Marco Nunes

UPdigital

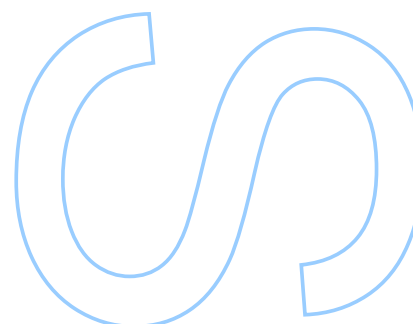
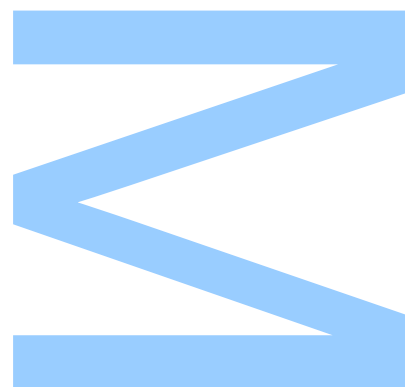




Todas as correções determinadas
pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, ____/____/____



Abstract

Nowadays data is fundamental for any organisation. Due to the huge amount of data produced there is a need for systems capable of extracting and analysing information in a timely manner as well as providing support in decision making. To fill this gap, Business Intelligence (BI) systems have emerged.

This dissertation project's goal is to develop a Business Intelligence solution for the University of Porto's Human Resources department. This solution is capable of extracting data from different sources within University of Porto. Furthermore, it can obtain quality data in a timely manner and provides support in decision making, fulfilling the organisation's requirements and needs.

This document describes the Business Intelligence solution developed under this project in different phases, namely: the requirements survey, development of a Data Warehouse, data extraction, processing and saving and finally dashboard designing to display information in a useful way.

In conclusion, the developed solution consists in a trusted and automatic system that fulfills the organisation needs - contributing to a more effective way of obtaining data and analysing information from multiple data sources.

Resumo

Nos dias de hoje a informação é considerada um recurso fundamental numa organização. Devido à enorme quantidade de dados produzidos pelas organizações, surge a necessidade da existência de sistemas capazes de extrair e analisar informação em tempo útil e darem suporte na tomada de decisões. Para dar resposta a esta necessidade, surgem os sistemas *Business Intelligence* (BI). Assim sendo, a utilização de sistemas BI nas organizações tem vindo a aumentar.

Este projeto de dissertação tem como objetivo o desenvolvimento de uma solução *Business Intelligence* para a área de recursos humanos da Universidade do Porto. Pretende-se uma solução capaz de extrair os dados das diferentes fontes de dados, que existem atualmente na organização. A solução desenvolvida permitirá obter informações de qualidade e em tempo útil, respondendo assim aos requisitos e necessidades da organização.

Assim, este documento descreve a solução *Business Intelligence* desenvolvida no âmbito deste projeto, desde o levantamento dos requisitos, definição da arquitetura, criação do *Data Warehouse*, desenvolvimento do processo de extração, transformação e carregamento dos dados e exploração da informação através da criação de vários *dashboards*.

Concluindo, a solução desenvolvida consiste num sistema confiável e automatizado que responde às necessidades definidas pela organização, contribuindo para uma forma mais eficaz de obter e analisar informação útil a partir dos dados provenientes das diversas fontes de dados.

Agradecimentos

O resultado desta dissertação não seria o mesmo sem o apoio diversas pessoas pelo que expresso os meus sinceros agradecimentos:

Primeiramente à minha família: avó, mãe e irmã, por todo o apoio, motivação, compreensão e por tudo aquilo que fizeram por mim, sem vocês isto não era possível.

Gostaria de agradecer ao Professor e orientador Fernando Silva, por esta oportunidade, apoio e orientação ao longo deste projeto.

Ao coordenador Marco Nunes, por todo o apoio, boa-disposição, atenção e ajuda disponibilizada no decorrer deste projeto.

À organização em estudo, aos seus colaboradores no geral e em particular à Ana Fernandes, pela simpatia, ajuda, paciência e esforço na realização deste projeto.

Um agradecimento especial aos meus amigos pela paciência, motivação e por estarem sempre presentes.

A todos, um sincero, obrigado!

Conteúdo

Abstract	i
Resumo	iii
Agradecimentos	v
Conteúdo	ix
Lista de Tabelas	xii
Lista de Figuras	xiv
Acrónimos	xv
1 Introdução	1
1.1 Enquadramento e Motivação	1
1.2 Objetivos	2
1.3 Contribuições	2
1.4 Estruturação da Dissertação	2
2 Enquadramento Teórico	5
2.1 Business Intelligence	5
2.1.1 Arquitetura	6
2.1.2 Desvantagens dos sistemas BI	6
2.1.3 Vantagens dos sistemas BI	7

2.2	Fontes de dados	7
2.3	<i>Extract, Transform and Load</i>	8
2.3.1	Fatores importantes na escolha da ferramenta ETL	9
2.3.2	Ferramentas ETL	10
2.3.3	Linguagens de <i>scripting</i>	11
2.4	<i>Data Warehouse</i>	11
2.4.1	Metodologia de Inmon	12
2.4.2	Metodologia de Kimball	14
2.4.3	Inmon vs Kimball	20
2.5	Modelos tabulares e multidimensionais	21
2.5.1	Modelo Multidimensional	21
2.5.2	Modelo Tabular	23
2.5.3	Modelo Multidimensional vs Modelo Tabular	24
2.6	Ferramentas BI	25
3	Estado da Arte	29
3.1	Uso de sistemas BI em organizações	29
3.2	Trabalho relacionado	30
3.2.1	Desenvolvimento e implementação de projetos de <i>Business Intelligence</i> nas áreas financeira e de logística	31
3.2.2	Solução Analítica de Suporte à Decisão no Âmbito da Gestão Acadêmica no Ensino Superior	32
3.2.3	Desenvolvimento de uma solução de <i>Business Intelligence</i> na área de restauração	33
4	Requisitos e Arquitetura	35
4.1	Levantamento dos Requisitos funcionais e Indicadores	35
4.2	Arquitetura	36
4.3	Solução atual	37

5	Desenho e Desenvolvimento	39
5.1	Modelação dimensional	39
5.2	Tabelas de facto	40
5.3	Tabelas de dimensão	43
5.4	Processo ETL	49
5.5	Cálculo dos indicadores	53
5.5.1	Importação do modelo	53
5.5.2	Medidas e colunas calculadas	54
5.6	<i>Dashboards</i>	57
5.6.1	Design dos <i>dashboards</i>	57
5.6.2	<i>Dashboards criados</i>	58
5.7	Validação e testes	76
6	Conclusões	77
6.1	Principais dificuldades	78
6.2	Trabalho Futuro	78
	Referências	79

Lista de Tabelas

2.1	Principais diferenças entre a metodologia de Inmon e Kimball	21
2.2	Principais recursos disponíveis no modelo multidimensional e tabular	25
5.1	Matriz de processos	39
5.2	Medidas e respetiva descrição da tabela <i>fact_trab_vinculos</i>	41
5.3	Medidas e respetiva descrição da tabela <i>fact_trab_entradas_saidas_e_encargos</i>	42
5.4	Atributos e respetiva descrição da tabela <i>dim_entidade</i>	44
5.5	Atributos e respetiva descrição da tabela <i>dim_data</i>	44
5.6	Atributos e respetiva descrição da tabela <i>dim_grupo_cargo_carreira_sioe</i>	45
5.7	Atributos e respetiva descrição da tabela <i>dim_modalidade_vinculacao</i>	45
5.8	Atributos e respetiva descrição da tabela <i>dim_nivel_antiguidade</i>	45
5.9	Atributos e respetiva descrição da tabela <i>dim_nivel_escolaridade</i>	46
5.10	Atributos e respetiva descrição da tabela <i>dim_pessoa</i>	47
5.11	Atributos e respetiva descrição da tabela <i>dim_pais</i>	48
5.12	Atributos e respetiva descrição da tabela <i>dim_periodo_normal_trabalho</i>	48
5.13	Atributos e respetiva descrição da tabela <i>dim_funcao_exercida</i>	48
5.14	Atributos e respetiva descrição da tabela <i>dim_regime_periodo_trabalho</i>	49
5.15	Tabela <i>dim_entidade</i> de exemplo	51
5.16	Tabela <i>aux_hie_entidade</i> de exemplo	51
5.17	Tabela com os tempos de execução do processo ETL para a tabela de factos Trabalhadores - Vínculos"	52

5.18	Tabela com os tempos de execução do processo ETL para a tabela de factos "Trabalhadores - Entradas, saídas e Encargos"	53
5.19	Medidas calculadas na tabela <i>fact_trab_percentagens</i>	56
5.20	Medidas calculadas na tabela <i>fact_trab_entradas_saidas_e_encargos</i>	56
5.21	Fórmula DAX para calcular a coluna idade	57

Lista de Figuras

2.1	Arquitetura BI	6
2.2	Arquitetura do DW de Inmon	13
2.3	Arquitetura do DW de Kimball	15
2.4	Tabela de factos	16
2.5	Tabelas de dimensão	16
2.6	<i>Star Schema</i>	18
2.7	<i>Snowflake Schema</i>	19
2.8	<i>Constellation Schema</i>	19
2.9	Exemplo de um cubo OLAP	22
2.10	Quadrante mágico de Gartner para plataformas BI	25
3.1	<i>Dashboard</i> : Visão 360°	31
3.2	Exemplo de <i>dashboard</i> desenvolvido no <i>Power BI</i>	32
3.3	Exemplo de <i>dashboard</i>	33
4.1	Arquitetura do sistema implementado	36
4.2	Arquitetura do sistema atual	37
5.1	Modelo desenvolvido para o <i>Data Mart</i> dos recursos humanos	40
5.2	Processo ETL	49
5.3	Processo ETL para a tabela <i>dim_pessoa</i>	50
5.4	Estrutura do <i>dashboard</i>	58

5.5	Menu inicial	59
5.6	Submenu da área de vínculos	59
5.7	Página - Resumo do Número de trabalhadores	60
5.8	Página - Evolução do número de trabalhadores	61
5.9	Trabalhadores segundo a Função Exercida e Nível de Escolaridade	62
5.10	Evolução das Percentagens segundo a Função Exercida e Habilitação	62
5.11	Trabalhadores segundo a modalidade de vinculação	63
5.12	Evolução do peso dos trabalhadores em funções públicas com contrato de trabalho ao abrigo do código de trabalho	64
5.13	Diversidade Trabalhadores - I	65
5.14	Diversidade Trabalhadores - II	66
5.15	Evolução da taxa de feminização	66
5.16	Docentes - I	67
5.17	Docentes - II	68
5.18	Evolução das percentagens - Docentes	68
5.19	Trabalhadores ETI	69
5.20	Evolução da percentagem de trabalhadores ETI por função exercida	70
5.21	Encargos com o pessoal	71
5.22	Trabalhadores admitidos e regressados	72
5.23	Trabalhadores saídas	73
5.24	Número de saídas por motivos de saída	74
5.25	Total de saídas de trabalhadores segundo o cargo/carreira, sexo e motivo de saída	74
5.26	Taxa de cobertura	75
5.27	Exemplo das duas formas utilizadas para colocar informação adicional no <i>dashboard</i>	76

Acrónimos

BI Business Intelligence

DW Data Warehouse

ETL Extract Transform Load

HOLAP Hybrid Online Analytical Processing

MOLAP Multidimensional Online Analytical Processing

OLAP Online Analytical Processing

OLTP Online Transactional Processing

ROLAP Relational Online Analytical Processing

SSAS SQL Server Analysis Services

SSIS SQL Server Integration Services

UP Universidade do Porto

Capítulo 1

Introdução

1.1 Enquadramento e Motivação

Atualmente, um dos principais desafios que as organizações enfrentam é obter a informação de forma rápida com o objetivo de otimizar a tomada de decisões. O conhecimento que é obtido através dos dados presentes nas organizações são recursos essenciais na tomada de decisões.

Cada vez mais as organizações sentem a necessidade de possuir informação com boa qualidade, fidedigna e no tempo certo, de forma a obterem uma vantagem competitiva. Desta forma, a utilização de sistemas de *Business Intelligence* nas organizações torna-se muito importante.

Os sistemas de *Business Intelligence* (BI) consistem num conjunto de técnicas, conceitos e ferramentas que permitem obter informação de forma fácil e eficaz [1]. Estes sistemas são capazes de extrair, transformar e armazenar dados. Isto permite uma maior qualidade da informação útil disponível para a tomada de decisões [1].

A necessidade de identificar problemas numa organização de ensino superior e melhorar os serviços apresentados pode-se tornar uma tarefa complexa. Desta forma, a implementação de uma solução BI nestas organizações traz uma grande vantagem na identificação e compreensão destes problemas.

Na Universidade do Porto (UP) são produzidos anualmente enormes volumes de dados a partir dos seus sistemas de gestão. Uma vez processados, estes dados tornam-se informação indispensável na monitorização do desempenho e no apoio à tomada de várias decisões. A recolha de indicadores eficientemente e ao longo do tempo requer uma estruturação adequada da informação. Deste modo, surge o projeto BI4UP, que procura o desenvolvimento da área de *business intelligence* da Universidade do Porto em diversos ramos, tais como: "Recursos Humanos", "Área académica", "Investigação", "Económico-financeira" e "Terceira missão". Assim sendo, o projeto descrito nesta dissertação foca-se na área de recursos humanos da UP e tem como finalidade a implementação de uma solução que vem dar resposta às necessidades da organização, ajudando na recolha, visualização e armazenamento dos dados de uma forma bem estruturada,

organizada e rápida.

1.2 Objetivos

Esta dissertação tem como objetivo o desenvolvimento de uma solução *Business Intelligence* para a área de recursos humanos da Universidade do Porto, de forma a facilitar a gestão desta organização.

Numa primeira fase, em colaboração com a organização, procurou-se identificar e analisar as diversas fontes de dados para os indicadores identificados. Em seguida, pretende-se um estudo da arquitetura do Data Warehouse (DW) a ser implementada, onde será integrado o *Data Mart* dos recursos humanos que se pretende desenvolver. Este *Data Mart* servirá de suporte à tomada de decisões. Pretende-se também, o desenho e desenvolvimento do modelo dimensional e de um processo (ETL). Este processo tem como função extrair, transformar e carregar os dados para o *Data Mart*. Por fim, pretende-se o desenvolvimento de vários *dashboards*.

1.3 Contribuições

As principais contribuições desta dissertação são:

- A construção de um *Data Mart* para área dos recursos humanos;
- Desenvolvimento de um processo ETL que permite extrair, transformar e carregar os dados para o *Data Mart* dos recursos humanos;
- Facilitar e tornar eficiente a obtenção de informação a partir dos dados produzidos pela organização, oferecendo aos utilizadores um sistema confiável e automatizado. Desta forma, são evitados os processos manuais que exigem um maior gasto de tempo e estão mais sujeitos a erros;
- Implementação de vários *dashboards* que permitem o acesso e análise dos indicadores identificados. Através destes *dashboards* o utilizador consegue obter um maior suporte na tomada de decisões.

1.4 Estruturação da Dissertação

Esta dissertação é composta por seis capítulos organizados da seguinte forma:

- **Capítulo 1:** É feita uma contextualização do tema do projeto, descrevendo a motivação que o levou a realizar, contributos e objetivos.

- **Capítulo 2:** Neste capítulo é apresentado um enquadramento teórico das metodologias utilizadas no desenvolvimento desta solução.
- **Capítulo 3:** É feita uma análise de trabalhos desenvolvidos anteriormente, que possuem semelhanças ao projeto desenvolvido nesta dissertação.
- **Capítulo 4:** Aqui é apresentado os requisitos definidos pela organização, a arquitetura definida para a solução e as tecnologias utilizadas. Além disso, é explicado brevemente a solução atual presente na organização.
- **Capítulo 5:** O quinto capítulo descreve todo o trabalho prático desenvolvido no projeto. Começando pela construção do modelo dimensional, passando pelo processo ETL desenvolvido. Posteriormente, é explicado todo o procedimento de modelação realizado nos dados e por fim, a construção de *dashboards*.
- **Capítulo 6:** Finalmente, neste capítulo é apresentado as conclusões do trabalho realizado, as principais dificuldades decorridas no desenvolvimento do projeto e por fim, o trabalho futuro.

Capítulo 2

Enquadramento Teórico

Este capítulo, apresenta alguns conceitos básicos relacionados com a área de *Business Intelligence* e ajuda a contextualizar e compreender os capítulos seguintes. Na secção 2.1, começa-se por analisar de uma forma geral a arquitetura dos sistemas *Business Intelligence*, bem como as suas vantagens e desvantagens. Nas secções seguintes, analisa-se em detalhe cada uma das componentes da arquitetura.

2.1 Business Intelligence

A primeira vez em que foi utilizada a expressão de *Business Intelligence* foi em 1865 por Richard Miller Devens [2] no seu livro *Cyclopædia of Commercial and Business Anecdotes*. Devens usou este conceito para descrever como o banqueiro Sir Henry Fumese usou *Business Intelligence* para ganhar vantagem sobre os seus concorrentes, através da recolha de informações sobre o mercado da sua época e o do seu negócio.

Em 1958, o investigador Hans Peter Luhn [3] publicou o artigo "*A Business Intelligence System*", onde definiu *Business Intelligence* como a capacidade de aprender interrelações com os factos existentes, de modo a tomar decisões para o objetivo pretendido.

Mais tarde, em 1989, Howard Dresner [4] descreveu o termo Business Intelligence (BI) como "conceitos e métodos para melhorar a tomada de decisões utilizando um sistema de suporte baseado em factos". Mas foi no final dos anos 90 que o termo foi amplamente adotado.

Em 2010, Turban, Sharda, Delen e King [1], definiram BI como um termo que combina várias arquiteturas, ferramentas analíticas, aplicações e metodologias. O principal objetivo do BI é permitir a manipulação dos dados interativamente, possibilitando que os gestores de negócios e analistas obtenham análises apropriadas e relevantes.

De uma forma geral, os sistemas BI oferecem um conjunto de tecnologias para transformação, armazenamento, acesso e análise dos dados para auxiliar os utilizadores na tomada de decisões [5].

A maioria das organizações possui grandes armazéns de dados, sendo que na maioria das vezes, nem sempre é fácil extrair informações úteis de forma fácil e em tempo útil. Desta forma, surgem os sistemas BI. Estes sistemas trazem um grande contributo para as organizações, oferecendo informação de qualidade e suporte na tomada de decisões, em contraste com as organizações que não possuem nenhum sistema BI.

2.1.1 Arquitetura

Os sistemas BI [6] são compostos por quatro componentes principais, como podemos observar na figura 2.1:

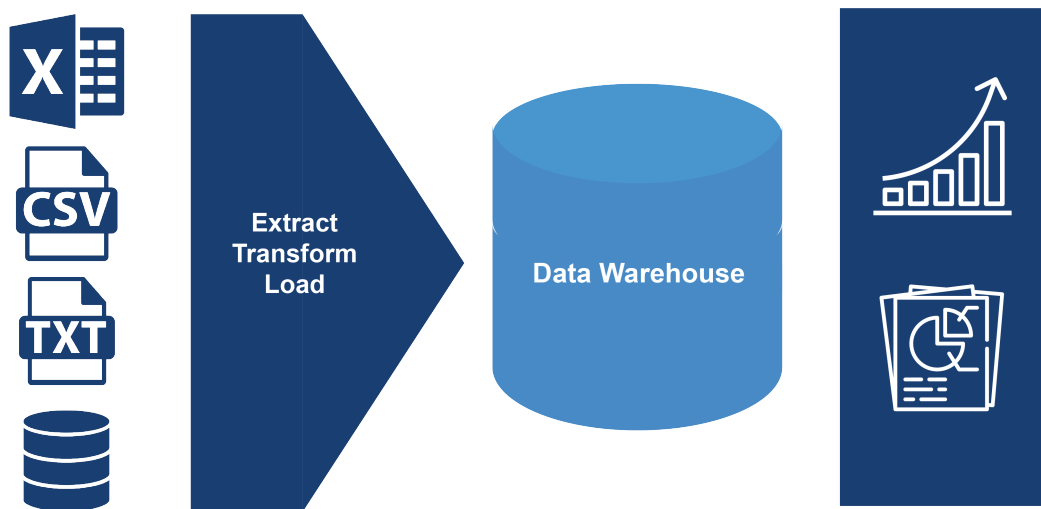


Figura 2.1: Arquitetura BI

- **Fontes de dados:** servem de suporte ao sistema, onde se encontram todas as origens dos dados;
- **Processo ETL:** é responsável por extrair, transformar e carregar os dados para o DW.
- **Data warehouse:** é utilizado para armazenar os dados.
- **Ferramentas de visualização de dados:** permitem analisar e visualizar os dados que estão no DW, através de análises OLAP, *Data Mining*, relatórios e *dashboards*.

Nas seguintes secções, estas componentes vão ser exploradas com mais detalhe.

2.1.2 Desvantagens dos sistemas BI

Segundo um estudo realizado por Silvana Quintanilha e Tricia Moraes [7], a principal desvantagem no uso de sistemas BI é o elevado custo do sistema. Muitas das vezes as organizações não têm capital suficiente para implementar uma solução deste tipo. Acresce que, por vezes é necessário

adaptar o sistema operativo e consequentemente, isso poderá trazer despesas adicionais ao nível de *hardware*, *software* e na contratação de especialistas na área. A sua manutenção também é considerada dispendiosa.

2.1.3 Vantagens dos sistemas BI

A utilização de um sistema de BI traz consigo diversos benefícios. Segundo Hugh Watson e Barbara Winxom [6] e tendo em conta o estudo de Quintanilha e Moraes [7], estas são algumas vantagens que a utilização de sistemas BI pode oferecer:

- Economiza tempo aos utilizadores devido à extração de informação de uma forma estruturada e eficiente;
- Mais informação e com uma melhor qualidade;
- Suporte na tomada de decisões;
- Melhoria dos processos de negócio;
- Maior suporte para a atingir os objetivos estratégicos do negócio;
- Análise dos dados através de *dashboards*, permitindo uma melhor perceção dos mesmos;
- Competitividade.

Tendo em consideração as desvantagens e vantagens referidas acima, pode-se concluir que a maior desvantagem da implementação de sistemas BI está relacionado com o seu custo de implementação e manutenção. Contudo, atualmente existem várias opções de ferramentas que não envolvem grandes investimentos, reduzindo o custo na implementação do sistema BI [7]. Um problema frequente, é as organizações não terem conhecimento técnico nem profissionais especializados para desenvolver estes sistemas [7], o que pode levar ao insucesso do sistema BI implementado.

Desta forma, pode-se concluir que a implementação com sucesso de um sistema BI, trará vários benefícios à organização, fornecendo um maior suporte na tomada de decisões e informação de melhor qualidade à organização.

2.2 Fontes de dados

As fontes de dados contêm todos os dados necessários para o processo de extração, tratamento e carregamento para o DW. Estas podem ser internas ou externas. As fontes de dados internas são provenientes da organização, como dados da área financeira ou dos recursos humanos. As fontes de dados externas vêm de fora da organização, como relatórios do governo e notícias. A origem e

tipo destas fontes de dados podem ser diversas, desde ficheiros *Excel*, bases de dados, a sistemas transacionais (Online Transactional Processing (OLTP)).

2.3 *Extract, Transform and Load*

Os dados usados nos sistemas BI têm origem em diferentes fontes e na maioria das vezes, nem sempre estão prontos para serem usados diretamente no BI. Uma das operações mais importantes no BI e que consome mais tempo é a preparação dos dados para análise. Esta tarefa garante que os dados usados são consistentes e consequentemente, que os resultados obtidos com o sistema BI são corretos [8].

O processo ETL consiste então nestas três tarefas: [9]

- **Extração:** O objetivo do processo extração é obter os dados das diferentes fontes de dados e disponibilizá-los para o processamento. Existem duas formas de extração, completa e incremental. A primeira copia sempre os dados todos, enquanto a extração incremental só copia os dados que sofreram alterações desde a última extração realizada. A complexidade desta tarefa é, geralmente, determinada pela complexidade da fonte de dados.
- **Transformação:** Consiste no processo de aplicar várias regras e funções aos dados extraídos até conseguir um formato adequado. Existe uma variedade de transformações que são possíveis fazer, tais como:
 - Limpar os dados corrigindo erros ortográficos, conflitos de domínio (como um nome de um país não corresponder ao código correto) e dados em falta;
 - Eliminar dados que não são úteis para o *Data Warehouse*;
 - Combinar várias fontes de dados.

É este processo que garante a qualidade dos dados antes destes serem carregados no DW.

- **Carregamento:** O processo de carregamento é responsável por carregar os dados extraídos e transformados no *data warehouse*. Existem dois métodos de carregamentos: *refreshing* e *updating*. O primeiro é usado quando o DW é criado, enquanto o segundo é usado para manter o DW atualizado.

Após a extração dos dados das fontes de dados, são armazenados numa área chamada *staging area*. Esta área pode ser física como, por exemplo, uma base de dados relacional ou podemos guardar os dados diretamente na memória [10].

Uma das vantagens da *staging area* localizar-se num sistema físico é, caso ocorra um problema na transformação dos dados ou uma falha no carregamento, não será necessário extrair ou transformar novamente esses dados. Outra vantagem é permitir a verificação da integridade dos dados. Como os dados originalmente extraídos estão guardados, é possível fazer uma comparação

entre os dados transformados e os originais. Desta forma, verifica-se se existe algum problema no processo ETL [10].

Em contrapartida, se optarmos por não guardar os dados extraídos antes de realizar as transformações, não conseguimos ter as vantagens referidas anteriormente. No entanto, se o nosso conjunto de dados não for muito grande e as transformações que necessitamos fazer não forem demasiado complexas, podemos optar por esta estratégia, tornando o processo mais rápido e simples [10].

2.3.1 Fatores importantes na escolha da ferramenta ETL

Um dos processos considerado vital na construção de um *Data Warehouse* é o ETL [8]. Portanto, antes da escolha da ferramenta utilizada neste processo, é fundamental ter em consideração alguns fatores. Segundo Goldfedder, no seu livro "Building a Data Integration Team" [11] são necessárias ter em consideração as seguintes características:

- **Extração:**

- **Estabelecer conexões** – Garantir que a ferramenta ETL consegue estabelecer ligação com todas as nossas fontes de dados;
- **Guardar dados temporariamente** – Deve conseguir manter seguros os dados armazenados na *staging area*, pois pode conter dados sensíveis. Para isso podem ser usadas técnicas de encriptação ou uma *firewall*;
- **Validação dos dados extraídos** – Garantir que haja uma correspondência correta entre os dados que extraímos e os dados que armazenamos.

- **Transformação**

- **Transformação dos dados** – A ferramenta ETL deve ser capaz de fazer todas as transformações necessárias;
- **Validação dos dados** – Garantir que as transformações feitas foram corretamente aplicadas.

- **Carregamento**

- **Estabelecer conexões** – Garantir que o carregamento dos dados não demora muito tempo;
- **Determinação dos tipos de carregamento** – Deve permitir escolher o tipo de carregamento;
- ***Rolling Back Errors*** – Deve ajudar a encontrar erros que possam ter ocorrido durante o carregamento.

2.3.2 Ferramentas ETL

Atualmente, existe uma grande variedade de ferramentas ETL no mercado[12], tais como: Talend[13], Pentaho Data Integration (PDI)[14], Informatica[15] e Microsoft SQL Server Integration Services(SSIS)[16]. A seguir, vamos analisar as vantagens e desvantagens de cada uma destas ferramentas.

Talend

O Talend é uma ferramenta *open source* que possui vários serviços que permitem integrar, gerir e armazenar os dados em cloud [13]. É uma ferramenta que consegue lidar com cenários de ETL complexos.

Vantagens: É uma ferramenta *open source* que é adequada para uma equipa de desenvolvimento com grande experiência em Java [11].

Desvantagens: Para uma equipa de desenvolvimento com pouca experiência em Java pode tornar o desenvolvimento muito mais difícil [11].

Pentaho Data Integration (PDI)

O Pentaho é uma outra ferramenta *open source* que fornece recursos para extrair, transformar e carregar os dados de diferentes fontes, utilizando ambientes de processamento em *cloud* [14].

Vantagens: O Pentaho possui compatibilidade e flexibilidade com vários sistemas operativos e uma comunidade forte. Relativamente às transformações nos dados, tem várias funcionalidades que permitem aplicar qualquer transformação. Por fim, a plataforma tem uma *interface* intuitiva [17].

Desvantagens: A instalação consome bastante tempo. [18].

Microsoft SQL Server Integration Services (SSIS)

O SSIS é uma componente do Microsoft SQL Server que serve para executar diversas tarefas de migração de dados. SSIS é uma ferramenta rápida e flexível usada para extração, transformação e carregamento dos dados de diversas fontes de dados [19].

Vantagens: Se a base da solução forem produtos da Microsoft, o SSIS é a ferramenta mais adequada para o processo de ETL. Além disso, tem um excelente suporte [19].

Desvantagens: Por vezes podem aparecer problemas se o ambiente de utilização não for Windows e pode trazer problemas ao tentar integrar com outros produtos que não são da Microsoft. Por estas razões, só se deve optar por esta ferramenta se for integrada com outros

produtos da Microsoft [11].

Informatica

Informatica tem vários produtos focados na integração dos dados. Tal como as outras ferramentas referidas anteriormente, facilita todo o processo ETL [15].

Vantagens: Oferece suporte a todos os aspetos de migração/integração[11].

Desvantagens: Devido à quantidade de funcionalidades que esta ferramenta possui, pode tornar-se complexa a sua aprendizagem. Assim sendo, esta ferramenta não é adequada para um volume de dados pequeno, porque são normalmente realizadas tarefas mais simples, não sendo necessária uma ferramenta tão robusta [11].

2.3.3 Linguagens de *scripting*

Com o aumento da popularidade das linguagens de *script* como Python [20] e R [21], muitas das organizações optam por utilizar estas linguagens para o processo ETL, ao invés de ferramentas ETL. Desta forma, não é necessário um investimento financeiro e ganha-se em flexibilidade.

Vantagens: Por vezes as ferramentas ETL têm algumas limitações relativamente às transformações efetuadas nos dados, no entanto, as linguagens de *script* são bastante versáteis e possuem várias bibliotecas e estruturas de dados que facilitam estas transformações [11]. Além disso, a sua utilização não possui nenhum custo associado.

Desvantagens: Para pessoas que não possuem experiência em programação e no desenvolvimento de processo ETL, pode tornar-se uma tarefa mais complexa a utilização de linguagens de *scripting*. Enquanto as ferramentas ETL possuem *interfaces user-friendly*, aqui isso não existe. Relativamente à manutenção e escalabilidade, também poderá ser mais fácil ao utilizar ferramentas ETL [11].

2.4 *Data Warehouse*

Conforme descrito na secção 2.1, o *Data Warehouse* é um dos principais componentes num sistema *Business Intelligence*.

Um *Data Warehouse*, em português, armazém de dados, é um repositório central de informação que guarda dados históricos e atuais, obtidos de diferentes fontes de dados. O DW serve de suporte para a tomada de decisões com o objetivo de tornar mais fácil e rápido o acesso à informação. Este sistema é usado para análise de um grande volume de dados e para a criação de relatórios/*dashboards*, facilitando a tomada de decisões [22]

Segundo Kimball e Ross [10] um *Data Warehouse* deve cumprir os seguintes requisitos:

- Facilitar o acesso à informação;
- Tornar a informação da organização consistente;
- Deve ser uma fonte de informação adaptável e resiliente;
- Deve manter a informação segura;
- Deve ser a base para a tomada de decisões.

Várias metodologias podem ser usadas na criação de um *data warehouse*. As duas metodologias mais usadas são a Ralph Kimball e William Bill Inmon. Nas subsecções [2.4.1 e 2.4.2] seguintes, vamos analisá-las em detalhe.

2.4.1 Metodologia de Inmon

Esta abordagem foi fundada na década de 90 para responder aos problemas das organizações e facilitar os seus sistemas de decisão. A metodologia Inmon é baseada no modelo entidade-relação [23]. Esta abordagem é também chamada *top-down*.

Inmon definiu um *data warehouse* como uma coleção de dados que [24]:

- **São integrados:** O DW deve ser a única e abrangente fonte de informação sobre um dado negócio. Portanto, com um DW não há necessidade de aceder a várias fontes de dados para responder às questões existentes.
- **São organizados por assuntos:** Os dados devem ser apresentados por assuntos, com base nas necessidades dos utilizadores.
- **Variam no tempo:** o DW contém informações históricas sobre o negócio permitindo análises da evolução histórica.
- **Não são voláteis:** Os dados que chegam ao DW não são apagados. Portanto, os dados são sempre adicionados regularmente para fornecer históricos corretos e confiáveis.

Neste modelo começa-se por analisar os dados provenientes das diversas fontes de dados da organização. Esses dados são tratados e carregados para uma *staging area* e só depois para o *data warehouse*, através do conjunto de processos ETL [24]. Após os dados estarem completamente integrados no DW são gerados vários *data marts* (subconjunto lógico do *data warehouse*) consoante as necessidades da organização, específicos para cada departamento. Como a única fonte de dados é o DW, isto garante que a integridade e a consistência sejam mantidas [23]. Por fim é realizada a exploração e análise dos dados, através de relatórios, *dashboards*, entre outros.

Esta abordagem é recomendada se os requisitos de análise não estiverem definidos e as fontes de origem forem estáveis [23]. Para organizações de grandes dimensões, que têm a necessidade de obter novas fontes de informação constantemente, esta arquitetura pode não ser a melhor solução, pois exige um enorme gasto de tempo [23].

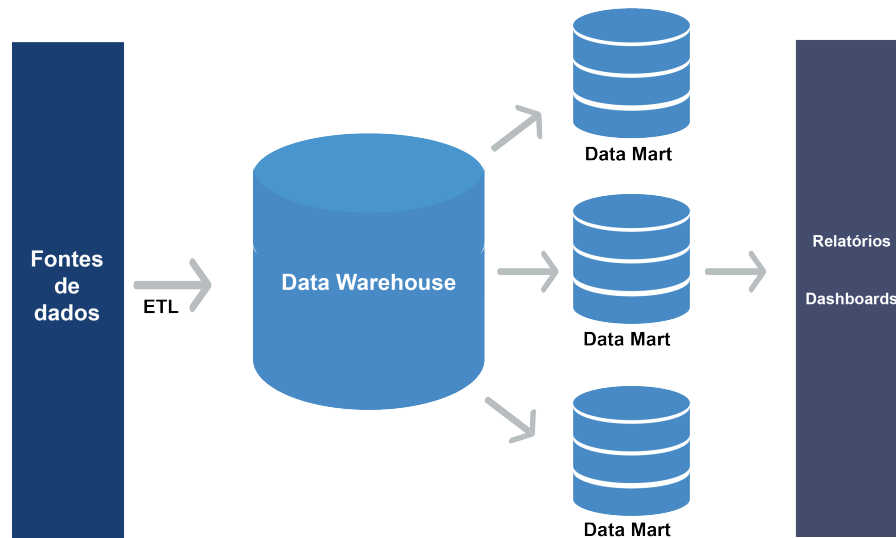


Figura 2.2: Arquitetura do DW de Inmon

Modelo normalizado

Inmon utiliza o modelo normalizado para construção do DW. Neste modelo, os dados no DW são guardados conforme as regras de normalização. As tabelas são agrupadas por áreas temáticas que refletem as categorias gerais dos dados, como, por exemplo, produto, cliente, finanças, entre outros [23]. A estrutura normalizada divide os dados em entidades, o que cria várias tabelas na base de dados relacional. O grande objetivo da normalização é reduzir a redundância nos dados. No entanto, isto provoca um enorme número de tabelas [25].

As regras de normalização são divididas na seguinte forma [25]:

- **Primeira Forma Normal (1NF)** - Remove atributos multi-valor, isto é, todos os atributos de relações são atômicos;
- **Segunda Forma Normal (2NF)** - Remove dependências parciais, ou seja, nenhum atributo não-chave depende funcionalmente de uma chave parcial;
- **Terceira Forma Normal (3NF)** - Remove dependências transitivas, isto significa que nenhum atributo não-chave pode depender transitivamente da chave primária.

Vantagens

A metodologia de Inmon têm as seguintes vantagens [26]:

- A redundância nos dados é muito baixa. Portanto, há menos possibilidade de irregularidades na atualização dos dados.
- O DW é a única fonte unificada, onde todos os dados estão integrados.

- Grande flexibilidade, pois é fácil de atualizar o DW caso haja alguma mudança nos requisitos.

Desvantagens

Contudo, também apresenta algumas desvantagens [26]:

- A complexidade aumenta à medida que várias tabelas são adicionadas ao modelo com o tempo.
- As consultas são muito lentas devido à estrutura do modelo.

2.4.2 Metodologia de Kimball

A metodologia de Kimball, chamada também de *bottom-up* foi fundada nos anos 90, oferecendo uma nova visão da arquitetura de um *data warehouse*. Esta abordagem tem por base o modelo dimensional.

Tal como no modelo de Inmon, as fontes de dados são identificadas e é feito todo o processo de ETL. Após estas duas etapas, são criados os *data marts* necessários para cada área da organização e só de seguida, são integrados no *data warehouse* [10].

Esta abordagem tem por base o modelo dimensional que é desnormalizado por natureza. Este modelo divide os dados em duas tabelas. Tabelas de facto que contêm os dados transacionais e tabelas de dimensão, que contêm as informações que suportam os factos.

Segundo Kimball, um *data warehouse* deve: [10]

- **Facilitar o acesso às informações da organização:** O conteúdo do *data warehouse* deve ser compreensível e as ferramentas devem ser fáceis de utilizar.
- **Apresentar as informações de forma consistente:** As informações devem ser credíveis. Não pode haver diferentes valores para a mesma informação.
- **Ser adaptável e flexível a mudanças:** Deve ser projetado para mudanças, tal como: novas necessidades dos utilizadores, condições do negócio, informações ou tecnologias podem ser alteradas ao longo do tempo.
- **Ser seguro de maneira a proteger as informações da organização:** Muitas das informações que o DW armazena podem ser dados pessoais, desta forma, é necessário haver um controlo sobre o acesso a estas informações.
- **Ajudar na tomada de decisões.**
- **Ser utilizado pelos utilizadores da organização.**

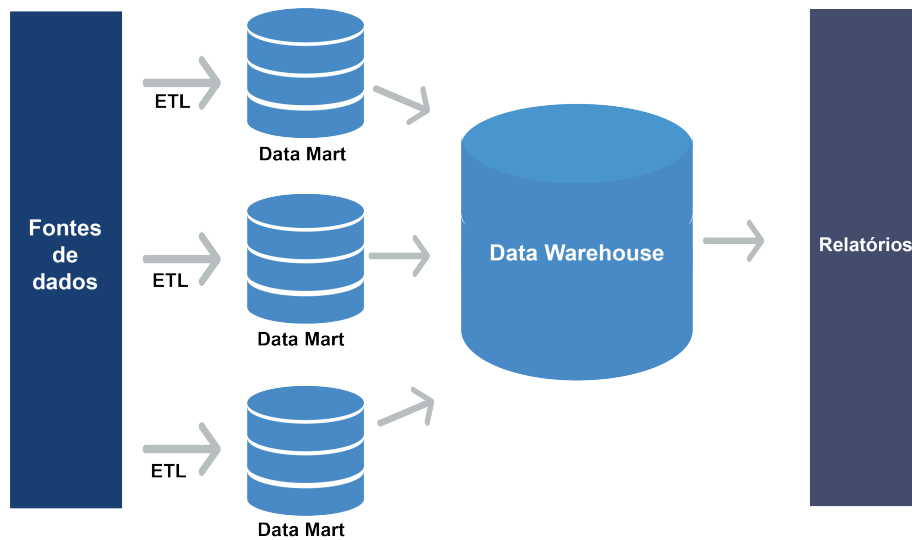


Figura 2.3: Arquitetura do DW de Kimball

A abordagem *bottom-up* é recomendada se os requisitos forem conhecidos e bem definidos. Devido à construção incremental, oferece alto desempenho nas consultas [26]. Contudo, como nesta abordagem os *data marts* são desenvolvidos independentemente entre si, podem conter dados redundantes, desta forma é necessário garantir que todos os *data marts* tenham dimensões que possam ser partilhadas entre tabelas de facto de vários *data marts*, garantindo a integridade destes.

Modelo dimensional

De acordo com Kimball e Ross [10] a construção de um *data warehouse* tem por base o modelo dimensional. É necessário garantir a simplicidade no modelo, para que a compreensão do modelo seja fácil por parte dos utilizadores. Desta forma, é possível obter informações de forma rápida. Um modelo dimensional é constituído por uma tabela de factos e conjunto de tabelas de dimensões, podendo assemelhar-se a uma estrela [10]. Por esse motivo é que o modelo de Kimball é também conhecido como modelo estrela.

Tabela de factos

A tabela de factos é o centro do modelo e armazena grandes quantidades de informação. Uma tabela de factos normalmente contém dois tipos de colunas: colunas que têm os factos e colunas que são uma chave estrangeira (FK) que permite o relacionamento com as tabelas de dimensão. A chave primária da tabela de factos, é normalmente composta por todas as chaves estrangeiras [22].

Existem 3 tipos de factos:

- **Aditivo:** fatos que podem ser somados em todas as dimensões;

- **Semi-aditivos:** factos que podem ser adicionados em algumas dimensões;
- **Não-aditivos:** factos que não podem ser somados.

Fact_Vendas	
PK, FK1	Id_Data
PK, FK2	Id_Loja
PK, FK3	Id_Produto
Facto	Unid_Vendas

Figura 2.4: Tabela de factos

Dimensões

As dimensões fornecem o contexto do facto, por exemplo, para a tabela de factos representada na figura 2.4 temos o facto "unidades vendidas". Esse valor por si só não fornece muita informação, contudo através das dimensões data, loja e produto conseguimos saber quantas unidades foram vendidas de um determinado produto, para um dado período de tempo específico, numa dada loja. É através das dimensões que conseguimos obter o contexto dos factos.

Dim_Data		Dim_Loja		Dim_Produto	
PK	Id_Data	PK	Id_Loja	PK	Id_Produto
Atributo	Data	Atributo	Nome_Loja	Atributo	Nome_Produto
Atributo	Dia	Atributo	Morada	Atributo	Cod_Produto
Atributo	Mês				
Atributo	Ano				

Figura 2.5: Tabelas de dimensão

As tabelas de dimensões são constituídas por atributos e são estes que permitem detalhar os factos. Nestas tabelas a chave primária é gerada automaticamente e serve para fazer a ligação a uma ou várias tabelas de facto [22]. Em comparação as tabelas de facto, as dimensões possuem muitos menos registos.

As dimensões podem ter diferentes tipos [27]:

- **Dimensões de mudança lenta** – Os atributos da dimensão podem mudar ao longo do tempo, mas não muito frequentemente, como, por exemplo, nome ou morada de uma pessoa. Para estas situações, existem duas formas de alterar o registo. Na primeira forma, o valor antigo é substituído pelo novo, o que implica que o histórico não é mantido. Na segunda forma, é acrescentando um novo registo e assim o histórico é mantido;

- **Dimensões conformadas** – As dimensões conformadas são partilhadas por diferentes tabelas de facto.
- **Dimensões degeneradas** – Quando o atributo da dimensão é armazenado na tabela de factos e não numa tabela separada.
- **Dimensões lixo** – Esta dimensão é útil quando existem várias dimensões com poucos atributos e poucos valores. Desta forma, evita-se um grande número de chaves estrangeiras na tabela de factos.
- **Dimensões *Role-Playing*** – É quando uma dimensão é referenciada várias vezes numa tabela de factos. Por exemplo, se tivermos uma tabela de factos que tenha a data em que foi realizada a encomenda e data de envio da mesma, a dimensão data será referenciada duas vezes.

Para se obter um bom modelo dimensional é necessário ter em consideração as seguintes etapas [10]:

Identificar o processo de negócio

Um DW pode ser usado em diversas áreas como marketing, vendas, recursos humanos, entre outros. Posto isto, a primeira coisa a ter em atenção quando se vai construir o modelo dimensional é a área que o DW vai abranger. Para além disto, também é importante perceber as necessidades da organização e a qualidade dos dados disponíveis para este processo. A escolha do processo de negócio é uma etapa bastante importante porque define o objetivo do projeto e permite a construção da matriz de processos. É nesta matriz que são identificados os processos principais da organização e permite que a granularidade, dimensões e factos sejam declarados. Cada processo de negócio corresponde a uma linha na matriz.

Identificar a granularidade

A granularidade descreve o nível de detalhe dos dados no DW. Depois de escolher o processo de negócio, é necessário identificar o nível mais baixo de informação, designado como grão atómico. Segundo Kimball e Ross, deve-se ter um grão o mais atómico possível, para garantir que não ocorram problemas na agregação. Se uma tabela contém dados relativos a vendas diárias, então a granularidade vai ser diária.

Identificar dimensões

A próxima etapa é a escolha das dimensões. Utiliza-se a matriz de processos para escolher as dimensões que descrevem melhor os factos.

Identificar factos

Depois da escolha das dimensões, deve-se escolher as medidas que vão constituir a tabela de factos, de maneira que possam responder às necessidades da organização. Por exemplo, se um utilizador quiser saber quantos produtos específicos são vendidos diariamente em lojas diferentes, o facto aqui será a soma das vendas por produto, tempo e local.

Construção do modelo

Por fim, após estarem definidas as dimensões e os factos, faz-se o desenho do modelo. Existem várias técnicas de modelação dimensional, mas as mais utilizadas são as seguintes:

- O **Star Schema**, em português, modelo em estrela, é o modelo com a estrutura mais simples. No modelo em estrela, a tabela de factos encontra-se no centro, envolvida pelas dimensões através das chaves estrangeiras. Neste modelo, cada dimensão é representada por uma única tabela e não existe ligação entre duas tabelas de dimensão. A maioria das ferramentas BI suportam este modelo [10]. Na figura seguinte, temos um exemplo de um modelo em estrela.

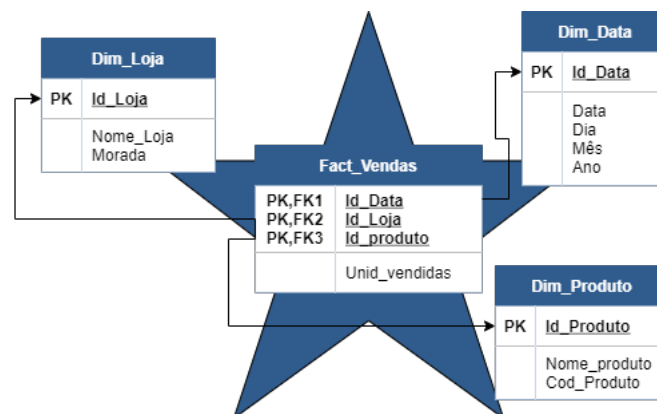
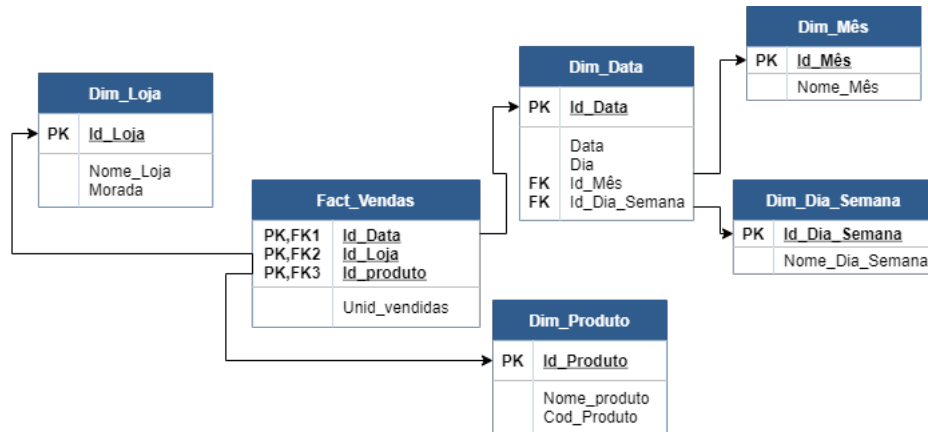
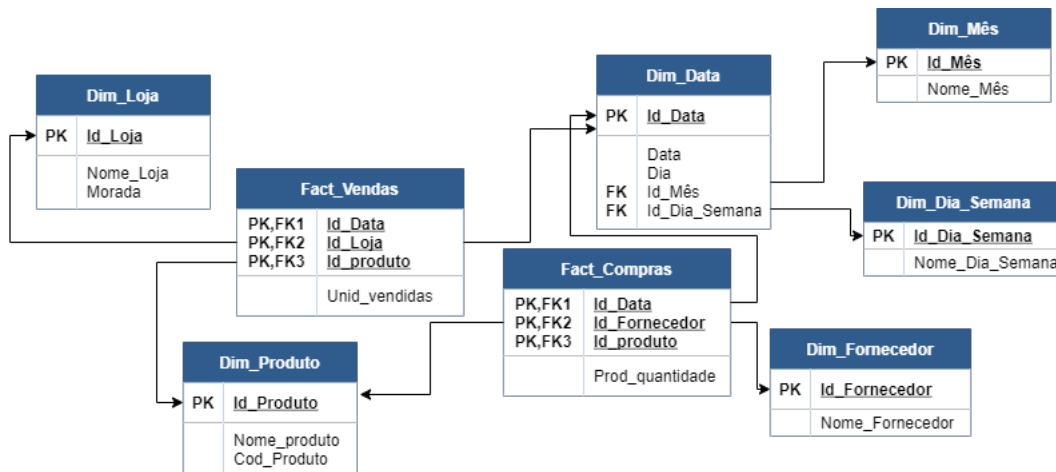


Figura 2.6: *Star Schema*

- **Snowflake Schema**, em português, modelo do floco de neve, é uma extensão do modelo em estrela. Neste modelo as tabelas de dimensão relacionam-se com a tabela de factos, mas existem algumas dimensões que apenas se relacionam com outras dimensões. Isto ocorre, porque ao contrário do modelo em estrela, as dimensões estão totalmente normalizadas, ocupando assim menos espaço, sendo esta a principal vantagem deste modelo [28]. No entanto, como existem mais tabelas, o desempenho nas consultas é reduzido.

Figura 2.7: *Snowflake Schema*

- **Constellation Schema**, em português, modelo em constelação, é constituído por mais que uma tabela de factos ligadas a várias tabelas de dimensão partilhadas. É um modelo mais complexo que o *star schema* e *snowflake schema*. Este modelo é mais usado para implementação de *data warehouses*, enquanto o *star schema* e *snowflake schema* são geralmente implementados no início da construção do DW [29].

Figura 2.8: *Constellation Schema*

Benefícios da modelação dimensional

A modelação dimensional traz consigo os seguintes benefícios [10]:

- **Compreensibilidade** – O modelo dimensional é intuitivo. As informações são agrupadas em dimensões, tornando-as mais fáceis de ler e interpretar. Devido à simplicidade do modelo, navegar na base de dados é muito mais eficiente;
- **Desempenho da consulta** – Os modelos dimensionais mais desnormalizados são otimizados para consultas, enquanto os modelos dimensionais normalizados tentam eliminar

redundâncias nos dados e são otimizados para o carregamento e atualização;

- **Flexível** – O design do modelo dimensional pode ser facilmente alterado se necessário, para integrar novos requisitos.

Vantagens e Desvantagens

A metodologia de Kimball possui as seguintes vantagens [26]:

- A construção é rápida e simples. Os custos iniciais são baixos;
- Não é necessária uma grande equipa para o desenvolvimento e gestão do DW;
- O esquema em estrela é facilmente compreendido pelos utilizadores.

No entanto, também apresenta algumas desvantagens [26]:

- Dados redundantes podem causar anomalias na atualização dos dados ao longo do tempo;
- Novos requerimentos têm impacto no modelo;
- Mudanças frequentes nas fontes de dados afetam o modelo.

2.4.3 Inmon vs Kimball

As principais diferenças entre a metodologia de Inmon e de Kimball estão resumidas na seguinte tabela [26].

	Inmon	Kimball
Objetivo	Fornece uma solução técnica útil para os profissionais de IT	Fornece uma solução que facilita as tarefas dos utilizadores finais
Arquitetura	DW de nível atômico alimenta os <i>data marts</i>	O conjunto de <i>data marts</i> constitui o DW
Complexidade	Complexo	Simples
Custos	Os custos iniciais são altos, mas depois os custos de desenvolvimento são baixos	Os custos iniciais são baixos, mas depois os custos de desenvolvimento são altos
Tempo	O tempo de desenvolvimento é alto	O tempo de desenvolvimento é baixo
Complexidade do ETL	As regras de transformação são simples se o modelo de dados for semelhante ao modelo das fontes de dados	As transformações são complexas
Flexibilidade	Se houver alterações nas fontes de dados é necessário alterar o modelo Se houver novos requisitos, é necessário alterar o modelo se os dados não existirem no DW	Mudanças frequentes na origem dos dados, afetam o modelo do DW Novos requisitos têm impacto no modelo
Performance das Queries	A consulta é muito lenta por causa da estrutura do modelo	O modelo é projetado para ser eficiente

Tabela 2.1: Principais diferenças entre a metodologia de Inmon e Kimball

2.5 Modelos tabulares e multidimensionais

Após os dados estarem integrados no DW, um dos objetivos é analisar e obter informação a partir dos dados. Desta forma, os modelos tabulares e multidimensionais permitem realizar análises de uma forma eficiente e rápida. Em seguida, analisaremos cada um destes modelos.

2.5.1 Modelo Multidimensional

De uma forma geral, a modelação multidimensional cria cubos constituídos por medidas e dimensões, com base nos dados que estão presentes em bases de dados relacionais. Estes modelos são utilizados pelas ferramentas Online Analytical Processing (OLAP) [30].

Foi por volta da década de 90 que as ferramentas de OLAP apareceram. Estas ferramentas têm como principal objetivo a análise dos dados. Executam análises multidimensionais nos dados e têm a capacidade de realizar cálculos complexos rapidamente, análises de tendências

e modelação dos dados [31]. Desta forma, os utilizadores têm uma maior facilidade na análise desses dados, de modo a tomar decisões mais bem informadas, com precisão e rapidez [31].

Cubo OLAP

O cubo OLAP é uma estrutura que armazena os dados do negócio num formato multidimensional, tornando-os mais fácil de analisar. No cubo, cada célula representa medidas ou valores baseado num conjunto de dimensões.

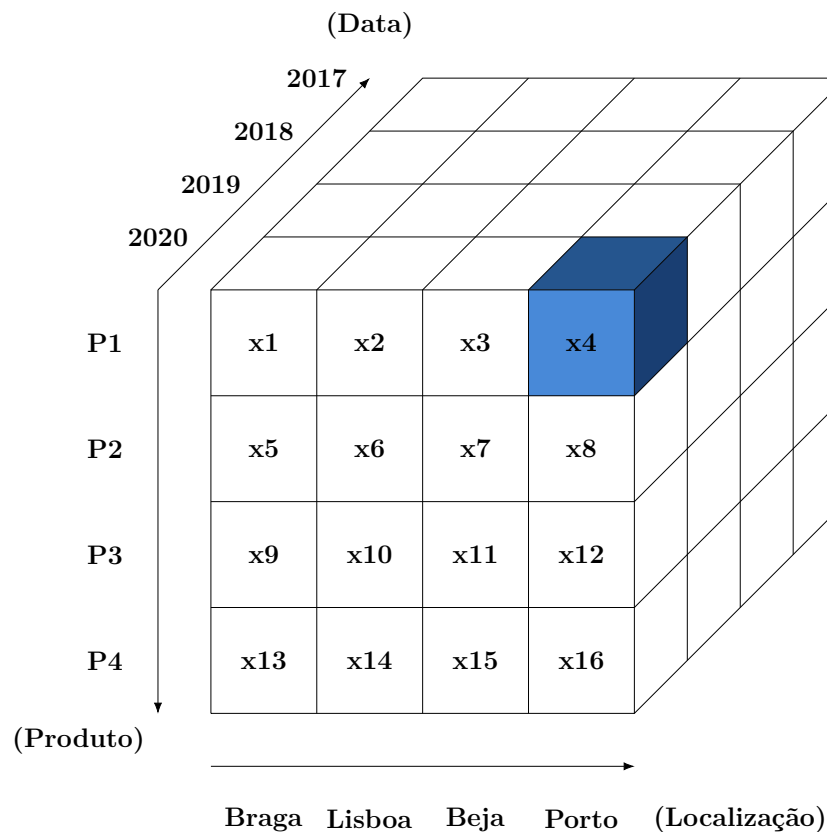


Figura 2.9: Exemplo de um cubo OLAP

A figura 2.9 representa um cubo composto por três dimensões: produto, localização e data; E uma medida: total de vendas [32]. Por exemplo, a célula em azul(x4) representa as vendas do produto P1 no Porto em 2020. As dimensões agrupam os dados do negócio relacionados e podem formar uma hierarquia, isto é, a dimensão data tem uma hierarquia de dia-mês-ano, sendo o dia o nível mais baixo de granularidade e ano o mais alto [32].

Os cubos OLAP permitem quatro tipos de operações: [33]

Slice e Dice

A operação *Slice* permite seleccionar dados de uma única dimensão do cubo OLAP. Enquanto a operação *Dice* isola um subcubo de duas ou mais dimensões.

Drill-down

Examina os dados com maior detalhe através de dois métodos - descendo na hierarquia ou adicionando uma nova dimensão.

Roll up

É o oposto da operação *drill-down*, combina vários dados de uma ou mais dimensões, reduzindo o número de dimensões.

Pivot

Esta operação permite visualizar o cubo atual numa nova perspectiva.

Sistemas OLAP

Existem três tipos de arquitetura nos sistemas OLAP: Multidimensional Online Analytical Processing (MOLAP), Relational Online Analytical Processing (ROLAP) e Hybrid Online Analytical Processing (HOLAP).

MOLAP

No MOLAP (*Multidimensional OLAP*), os dados são armazenados de forma multidimensional. Esta versão do OLAP foi desenvolvida para velocidade. O MOLAP armazena os dados em estruturas lógicas construídas exclusivamente para acelerar a recuperação. Esta é a forma mais tradicional da análise OLAP [34].

ROLAP

ROLAP (*Relational OLAP*), esta análise opera diretamente sobre os dados em tabelas relacionais, sem reorganizar primeiro os dados no cubo [33]. O ROLAP é uma melhor opção para quando se está trabalhar com grandes volumes de dados e o desempenho não é muito importante [34].

HOLAP

HOLAP (*Hybrid OLAP*), tenta combinar as vantagens dos sistemas MOLAP e ROLAP numa única arquitetura. A estrutura do cubo está pronta para lidar com um elevado número de dimensões. Este sistema oferece um desempenho e atualizações rápidas para análises e criação de relatórios [34].

2.5.2 Modelo Tabular

Os modelos tabulares são baseados no armazenamento dos dados em memória. Utilizam algoritmos de compressão, processamento *multithread* de consultas e o mecanismo de análise *xVelocity* que oferece acesso rápido aos dados do modelo tabular, através de relatórios de aplicações como o Power BI e Excel [35].

No entanto, apesar do modo *"in-memory"* ser o mais comum quando se opta por utilizar os modelos tabulares, existe o modo *"DirectQuery"* utilizado para quando os modelos são muito grandes e não cabem em memória [35].

2.5.3 Modelo Multidimensional vs Modelo Tabular

Os modelos tabulares e multidimensionais possuem algumas características importantes a ter em consideração ao decidir o tipo de modelo que se adequa melhor aos nossos dados.

Compressão

Tanto o modelo tabular, como o modelo multidimensional comprimem os dados. Como existem vários fatores que influenciam a compactação dos dados, não se sabe com precisão a percentagem de compressão, no entanto, nos modelos multidimensionais o tamanho será cerca de um terço do tamanho dos dados originais. Já nos modelos tabulares, por vezes consegue-se obter uma melhor compactação, cerca de um décimo do tamanho [36].

Linguagem

Os modelos tabulares oferecem suporte para cálculos/consultas realizadas na linguagem DAX e MDX, no entanto, a linguagem mais utilizada nestes modelos é DAX, por ser fácil de aprender e apresentar algumas semelhanças com as funções utilizadas no Excel. Nos modelos multidimensionais, é normalmente utilizada a linguagem MDX, apesar de também suportar DAX [36].

Relatórios

Ambos os modelos possuem algumas funcionalidades diferentes. A tabela a seguir resume os recursos disponíveis para cada um respetivamente [36].

	Multidimensional	Tabular
Ações	Sim	Não
Agregações	Sim	Não
Colunas calculadas	Não	Sim
Rollups personalizados	Sim	Não
Drillthrough	Sim	Sim
Hierarquias	Sim	Sim
Relacionamentos muito para muitos	Sim	Não
Hierarquias pai-filho	Sim	Sim

Tabela 2.2: Principais recursos disponíveis no modelo multidimensional e tabular

2.6 Ferramentas BI

Existe uma enorme variedade de ferramentas para o desenvolvimento de soluções BI. Segundo o quadrante mágico de Gartner [37], estas são as plataformas líderes: Power BI [38], Tableau [39] e QlikView [40].



Figura 2.10: Quadrante mágico de Gartner para plataformas BI

Power BI

O Power BI, pertencente à Microsoft, é uma das ferramentas de *Business Intelligence* mais utilizadas. Esta plataforma permite aos utilizadores criar diversos gráficos, indicadores e relatórios através das várias fontes de dados [38]. Além disso, é fácil de utilizar.

De seguida, temos algumas das principais características do Power BI [41]:

- **Conectividade da fonte de dados:** Extrai os dados de centenas fontes de dados desde Excel, JSON, base de dados, entre outros.
- **Preparação dos dados:** Permite manipular e aplicar transformações aos dados, através do Power Query;
- **Visualização de dados:** Oferece uma ampla variedade de tipos de gráficos;
- **Relatórios:** Possui modelos predefinidos de relatórios para várias áreas de negócio. No entanto, também oferece a possibilidade de criar os relatórios de uma forma fácil, bastando arrastar e soltar a visualização pretendida;
- **Compatibilidade móvel:** Todos os relatórios são otimizados para dispositivos móveis e podem ser publicados na cloud ou localmente.

Limitações

- Não suporta consultas SQL;
- É difícil de trabalhar com grandes conjuntos de dados no Power BI;
- Não permite aos utilizadores criar relatórios programados e visualizações personalizadas pelo utilizador.

Tableau

O Tableau é uma ferramenta de visualização e análise de dados que combina informações de uma ampla variedade de fontes de dados e permite a exploração de dados de uma forma intuitiva [41].

Principais características [41]:

- **Visualizações de arrastar e soltar:** Permite criar relatórios rapidamente devido a esta facilidade;
- **Preparação dos dados:** Utiliza *machine learning* para limpar e transformar os dados, reduzindo o trabalho de processamento;

- **Integração das fontes de dados:** Extrai os dados de uma variedade de fontes de dados;
- **Acesso em tempo real:** Atualiza automaticamente as informações as aplicações Web e bases de dados na cloud;
- **Visualizações recomendadas:** Recomenda tipos de visualizações tendo em conta as variáveis selecionadas para análise;
- **Aplicação móvel:** Oferece uma versão móvel com várias funcionalidades da versão desktop.

Limitações

- Fraco suporte;
- Problemas de segurança;
- Suporte de TI para tarefas mais complexas;

QlikView

É uma ferramenta de BI para integração e visualização dos dados. Suporta várias fontes de dados e fornece resultados rapidamente através de um recurso de processamento de dados em memória, permitindo que os dados sejam processados até 10% do seu tamanho original [41].

Principais características [41]:

- **Plataforma *memory*:** As consultas e pesquisas são processadas de uma forma mais rápida e eficiente;
- **Configuração rápida:** Permite uma implementação rápida porque os dados não necessitam ser armazenados num cubo;
- **Manipulação de dados robustos:** Permite a manipulação fácil dos dados na plataforma, através de relatórios, visualizações e filtros.

Limitações

- Não suporta processamento analítico online (OLAP);
- Carece de funcionalidades na criação de relatórios;
- Não permite o agendamento de relatórios em horários específicos.

Capítulo 3

Estado da Arte

No âmbito deste projeto foi feita uma investigação para perceber como o uso de *Business Intelligence* afeta as organizações. Além disso, foi realizada uma pesquisa, para encontrar projetos desenvolvidos semelhantes ao nosso.

3.1 Uso de sistemas BI em organizações

Na área de recursos humanos a implementação de soluções de *Business Intelligence* tem sido cada vez mais comum, devido à existência de uma maior competitividade no mundo de negócios. Os sistemas *Business Intelligence*, em comparação com as soluções tradicionais existentes nas organizações, trazem vantagens na criação de estratégias e fornecem suporte para a tomada de decisões.

Existem alguns estudos que procuram avaliar o impacto que estes sistemas têm nas organizações.

Um estudo realizado por Otmane Azeroual e Horst Theel [42] em 2018, faz a análise dos efeitos da utilização de sistemas BI na gestão e tomada de decisões por empresas *start-up*. Com o aumento do volume de dados produzidos pelas organizações, obter informações importantes manualmente torna-se mais complexo. Portanto, a adoção de soluções BI tem aumentado ao longo dos últimos anos [42]. Assim sendo, foram identificados várias vantagens com a utilização dos sistemas BI, tais como:

- Rapidez na obtenção de informação;
- Apresentação da informação de forma adequada (gráficos, textos, imagens);
- Informação consistente;
- Processamento e avaliação de informações por métodos poderosos.

No entanto, também foram relatadas algumas desvantagens:

- Problemas de desenvolvimento (dificuldades no uso das ferramentas);
- Problemas de manutenção.

Os resultados indicam que o uso de sistemas BI tendem a ter um impacto positivo nas organizações, havendo uma baixa probabilidade de trazerem desvantagens às organizações. Para isso, é necessário ter em conta os requisitos técnicos antes da implementação de um sistema BI [42].

Um outro estudo realizado pela Dresner Advisory Associates [43], mostra que 32% das equipas de desenvolvimento e investigação usam regularmente quatro ou mais ferramentas BI. Os departamentos de vendas, operacionais e finanças são os principais motivadores do uso dos sistemas BI, procurando novas formas de reduzir os custos e perceber com uma maior precisão o impacto que as decisões operacionais podem ter na organização [43].

Neste mesmo estudo concluiu-se também que a criação de relatórios, *dashboards*, integração de dados e implementação de *data warehouses* são das principais tecnologias/iniciativas estratégicas adotadas pelas organizações para BI [43]. Por fim, segundo este estudo, a priorização da utilização de BI em 2019 – 2020 é de 21% na área de recursos humanos, encontrando-se em terceiro lugar dos departamentos que mais priorizaram a importância do BI nesse ano [43].

Finalmente, um estudo realizado por Ivo Dias e Maria José Sousa [44], analisou a utilização de ferramentas BI para auxiliar na tomada de decisões sobre a gestão de recursos humanos nas organizações portuguesas. A evolução da tecnologia da informação tem contribuído para um aumento do BI aplicado à gestão de RH [44].

Neste contexto, foi realizado um estudo para verificar se o BI está relacionado positivamente aos RH e se ajuda na previsão de decisões. Estas hipóteses foram validadas através de uma amostra de 43 gestores de recursos humanos que forneceram informações e exemplos do que contribui para um melhor processo de tomada de decisão com base nas ferramentas BI [44].

Desta forma, os resultados obtidos confirmam que o BI é fundamental em qualquer organização, sendo que na gestão de RH é importante disponibilizar de forma eficiente e consistente a informação, para os restantes departamentos, portanto o uso de BI é uma mais-valia nestas tarefas [44].

3.2 Trabalho relacionado

Nesta secção serão apresentados os diferentes projetos implementados em várias áreas que recorrem à utilização de *Business Intelligence*. O primeiro projeto [45] diz respeito a uma solução desenvolvida em 2014, para a área financeira e logística de duas organizações diferentes. O segundo projeto [46] foi desenvolvido em 2016 e descreve a implementação de uma solução BI no âmbito da gestão académica do Ensino Superior. Por fim, o último trabalho [47] aqui descrito, descreve uma solução BI desenvolvida para a área de restauração, em 2017. Os dois primeiros

projetos foram desenvolvidos na Faculdade de Engenharia da Universidade do Porto e o último na Faculdade de Ciências da Universidade de Lisboa.

Em seguida, encontram-se descritos cada um destes projetos.

3.2.1 Desenvolvimento e implementação de projetos de *Business Intelligence* nas áreas financeira e de logística

Atualmente, as enormes quantidades de dados produzidos pelas organizações exigem sistemas capazes de extrair informação orientada e de qualidade, para apoiar as tomadas de decisão. Neste contexto, foi desenvolvido um sistema que pretende responder às necessidades de controlo e tomada de decisão na área financeira e logística [45]. A solução desenvolvida implicou a criação do *Data Mart*, processo ETL e criação de relatórios. Para tal, foi utilizado a solução oferecida pela Oracle, denominada por OBIEE (*Oracle Business Intelligence Enterprise Edition*) [48], que inclui diversas ferramentas que permitem a criação e gestão de repositórios e criação de relatórios.

Relativamente aos relatórios desenvolvidos, foi desenvolvido um *dashboard* para cada área e criado diferentes separadores que permitem perspetivas diferentes dos dados: Visão Mensal, Visão Acumulada, Visão Anual e Visão 360°. Além destes separadores, existe a possibilidade de aplicar filtros aos dados e fazer *drill-down*. Na figura abaixo, encontra-se representado um exemplo de um *dashboard* desenvolvido nesta solução [45].

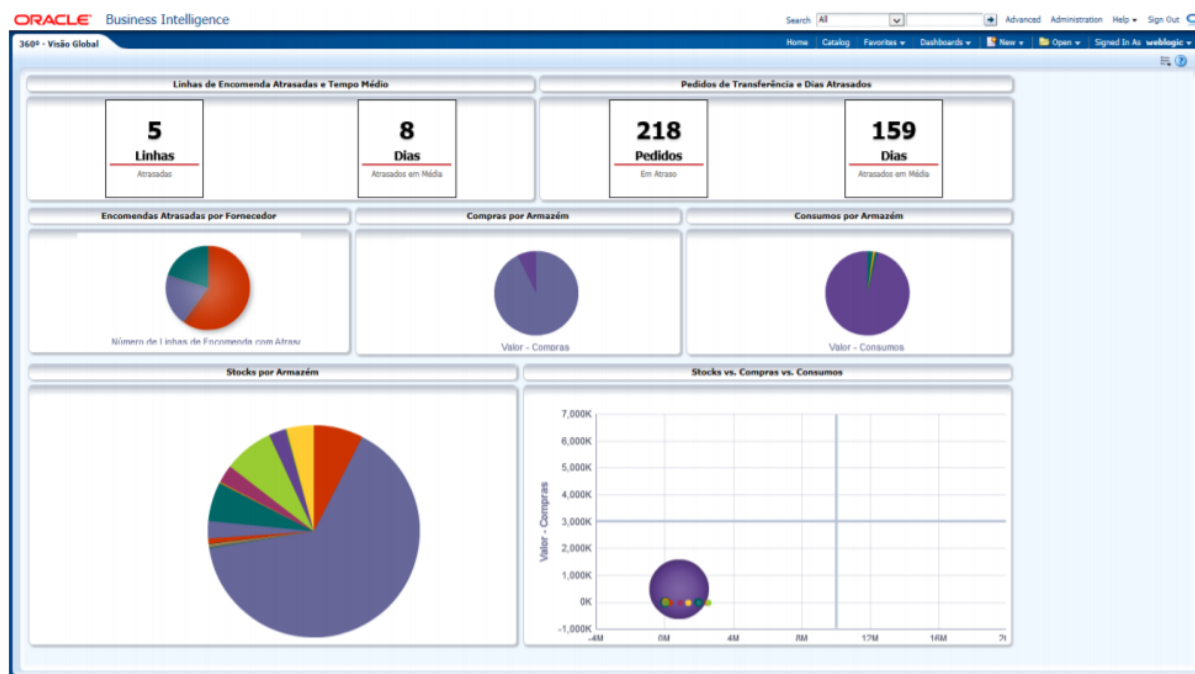


Figura 3.1: *Dashboard: Visão 360°*

A solução desenvolvida para ambas as áreas, permitiu às organizações obter uma visão global dos dados, permitindo obter informação em tempo útil sobre o seu negócio, estando assim melhor

capacitada para tomar decisões [45].

3.2.2 Solução Analítica de Suporte à Decisão no Âmbito da Gestão Académica no Ensino Superior

Neste projeto [46] foi desenvolvida uma solução que pretende responder às necessidades de controlo e tomada de decisão da gestão académica, através da análise da procura e eficácia dos cursos superiores, receita, assiduidade e planeamento. Esta solução BI, implicou a criação de um repositório de dados, processo ETL e criação de relatórios. Para o desenvolvimento desta solução foram utilizadas ferramentas da Microsoft, tais como: *Microsoft SQL Server* para a *staging area*, *Microsoft Analysis Server* para a construção de cubos multidimensionais e *Power BI* para a construção de relatórios. Além do *Power BI*, foi utilizada a ferramenta *Sharepoint* para construção de relatórios. O repositório de dados encontra-se montado na base de dados MySQL. Na figura seguinte, encontra-se um exemplo de um *dashboard* desenvolvido no *Power BI* nesta solução [46].

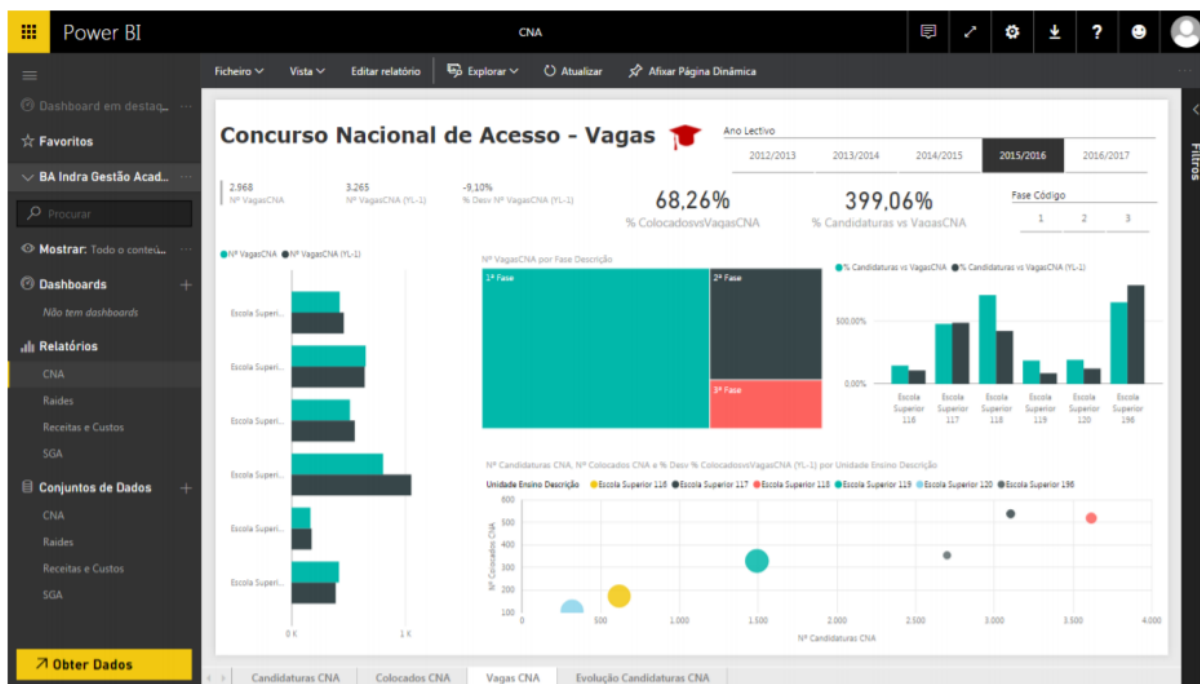


Figura 3.2: Exemplo de *dashboard* desenvolvido no *Power BI*

A solução desenvolvida permitiu à organização a que se destina, a obtenção de informação rapidamente e uma visão global dos dados, que até então não era possível, devido à dificuldade no acesso e interpretação dos dados.

3.2.3 Desenvolvimento de uma solução de *Business Intelligence* na área de restauração

A solução desenvolvida neste projeto [47] veio combater a necessidade de integrar um volume considerável de informação relevante para o negócio. A criação de novos processos de tratamento e análise de dados, permitiu obter informação de forma mais rápida e dar suporte na tomada de decisões. Esta solução abrange diversas áreas do negócio, como: vendas, contabilidade, recursos humanos e áreas de compras.

As ferramentas utilizadas na implementação desta solução são da *Microsoft*. Desta forma, para a implementação do armazém de dados foi utilizado *SQL Server Database*, para o processo ETL o *SQL Server Integration Services*, para a criação dos modelos multidimensionais o *SQL Server Analysis Services* e por último, o *Power BI* para a criação de relatórios.

Na construção dos *dashboards* foi pretendido a disponibilização da informação de forma visual, com um acesso simples, permitindo uma interpretação rápida e objetiva dos indicadores representados [47]. Um outro requisito, foi uma navegação fácil, intuitiva e interativa [47]. Na figura seguinte está representado um exemplo de um *dashboard* desenvolvido neste projeto.



Figura 3.3: Exemplo de *dashboard*

Esta solução permitiu ao cliente o acesso à informação através de uma *interface* única, garantindo a capacidade de realizar análises dinâmicas e flexíveis sobre os dados, evitando a utilização de processo manuais [47]. A solução garante ainda que os dados sejam atualizados automaticamente.

Com estes exemplos foi possível analisar as várias técnicas e ferramentas utilizadas no

desenvolvimento de sistemas de *business intelligence*. Apesar de serem projetos com objetivos e dados diferentes, apresentam bastantes semelhanças ao projeto que será desenvolvido, desde metodologias utilizadas para a implementação do armazém de dados, desenvolvimento do processo ETL e criação de relatórios.

Capítulo 4

Requisitos e Arquitetura

Como foi anteriormente referido no capítulo 1, pretende-se realizar a construção de um *data mart* para a área de recursos humanos. A primeira etapa reside no levantamento dos requisitos, indicadores e na definição da arquitetura a seguir no desenvolvimento da solução.

4.1 Levantamento dos Requisitos funcionais e Indicadores

Através da análise de vários relatórios existentes e de diversas reuniões, foi desenvolvido em conjunto com os recursos humanos, um conjunto de requisitos funcionais e indicadores que cobrem as necessidades dessa área e tentam atingir os objetivos de análise pretendidos. No capítulo 5 estes indicadores vão ser descritos com mais detalhe.

No que diz respeito aos requisitos funcionais, a solução desenvolvida deve cumprir os seguintes requisitos:

- Análise histórica dos dados;
- Filtros;
- Capacidade de *Drill-Down* e *Roll-Up*;
- Interface simples e *user-friendly*;
- Criação e interatividade com os relatórios;
- Exportação de relatórios;
- Integração com os outros *Data Marts* existentes para as outras áreas.
- Escalabilidade.

4.2 Arquitetura

Na figura 4.1 está representada a arquitetura definida para este projeto.

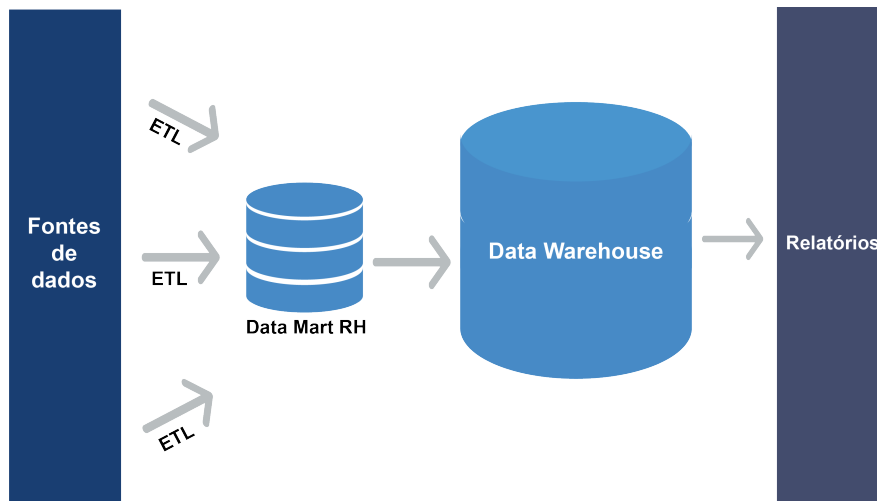


Figura 4.1: Arquitetura do sistema implementado

Fontes de dados

Tal como já foi referido no capítulo 2, as fontes de dados são a primeira componente na arquitetura de um DW, pois contêm todos os dados que precisamos para os próximos passos no desenvolvimento. A solução deve integrar informação do sistema de informação que dá suporte à gestão de recursos humanos na Universidade do Porto, o GRH, implementado numa base de dados Oracle.

ETL

Relativamente a todo o processo ETL, não foi utilizada uma ferramenta ETL comercial, tal como algumas das ferramentas referidas na subsecção 2.3.2. Optou-se por utilizar a linguagem Python, por ser uma linguagem simples, versátil, de fácil aprendizagem e adequada para a análise de dados [49].

Data Warehouse

A arquitetura adotada para o desenvolvimento deste projeto foi a de Kimball, ou seja, a abordagem *bottom-up*. Foi desenvolvido um *data mart*, relativamente aos recursos humanos que integrará o DW com outros *data marts* pertencentes ao projeto BI4UP. Para a implementação destes *data marts* foi utilizado o sistema de base de dados PostgreSQL, que tem compatibilidade com Python,

grande suporte e não requer nenhum custo adicional à solução. Outra vantagem do PostgreSQL é ser fácil de interoperar com bases de dados Oracle.

Relatórios

A criação de relatórios é um dos principais objetivos nas soluções BI, pois permitem o acesso e visualização dos dados de uma forma fácil e intuitiva. Desta forma, foi utilizada a ferramenta Power BI para a exploração da informação. Como referido anteriormente (2.6), o Power BI é a ferramenta líder para o desenvolvimento de soluções BI, oferecendo aos utilizadores diversas vantagens.

4.3 Solução atual

A Universidade do Porto teve há alguns anos uma primeira abordagem ao BI na área de recursos humanos. Esta não respondeu às necessidades atualmente existentes, assim como apresenta problemas ao nível de eficiência e usabilidade da ferramenta utilizada na construção de relatórios e *dashboards*. Além disso, os utilizadores relatam que o processo de carregamento dos dados para o *Data Warehouse* é muito demorado.

A arquitetura da solução BI atualmente implementada, encontra-se representada na figura 4.2.

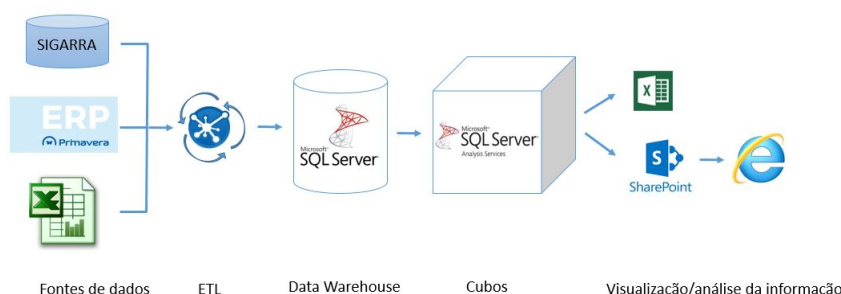


Figura 4.2: Arquitetura do sistema atual

Nesta solução, o processo de extrair, transformar e carregar os dados é efetuado através da ferramenta Pentaho. Relativamente ao *Data Warehouse* e aos cubos multidimensionais, são utilizadas ferramentas da Microsoft. O DW encontra-se construído na base de dados SQL Server e os cubos multidimensionais são implementados através da ferramenta SQL Server Analysis Services (SSAS). Por último, para a construção de relatórios e *dashboards* são utilizadas as ferramentas Sharepoint e o Excel.

A solução que irá ser desenvolvida neste projeto, deve responder às necessidades dos utilizadores e combater os problemas relatados.

Capítulo 5

Desenho e Desenvolvimento

Neste capítulo é descrito o desenvolvimento do sistema BI deste projeto. Inicia-se o capítulo com a descrição detalhada do modelo dimensional produzido. Seguidamente, é apresentado todo o processo ETL desenvolvido para integrar os dados no DW. Posteriormente, é descrito o processo de cálculo dos indicadores, e por fim são descritos os *dashboards* desenvolvidos.

5.1 Modelação dimensional

O desenvolvimento de um DW implica a construção de um modelo dimensional, de forma a garantir um melhor desempenho na obtenção de informação. Para isso é necessário ter em consideração os requisitos da organização para a construção do modelo, pois este servirá de suporte para a tomada de decisões da organização.

A primeira etapa é identificar os diferentes processos de negócio existentes e construir a matriz de processos. A matriz de processos é uma ferramenta importante no desenvolvimento do DW, pois permite uma visão geral sobre todas as dimensões e processos [10].

	Nível de antiguidade	Pessoa	Data	Tipo de função	Nível de escolaridade	Período normal de trabalho	Modalidade vinculação	Entidade	Grupo/cargo/carreira SIOE	Regime do período de trabalho
Trabalhadores - Vínculos	x	x	x		x	x	x	x	x	x
Trabalhadores - Percentagens		x	x					x		
Trabalhadores - Entrada, Saídas e Encargos		x	x	x				x	x	x

Tabela 5.1: Matriz de processos

Após definida a matriz de processos, representada na tabela 5.1, o próximo passo é identificar a granularidade dos dados, ou seja, o nível de detalhe pretendido. Neste projeto estabeleceu-se a granularidade seguinte: pessoa / em cada entidade / em cada trimestre. Por fim, fez-se a identificação das tabelas de facto e das dimensões, descritas detalhadamente nas secções 5.2 e 5.3.

A figura seguinte representa o modelo dimensional desenhado para o *Data Mart* dos recursos humanos que será integrado no DW, com os restantes *Data Marts* desenvolvidos no âmbito do projeto BI4UP.

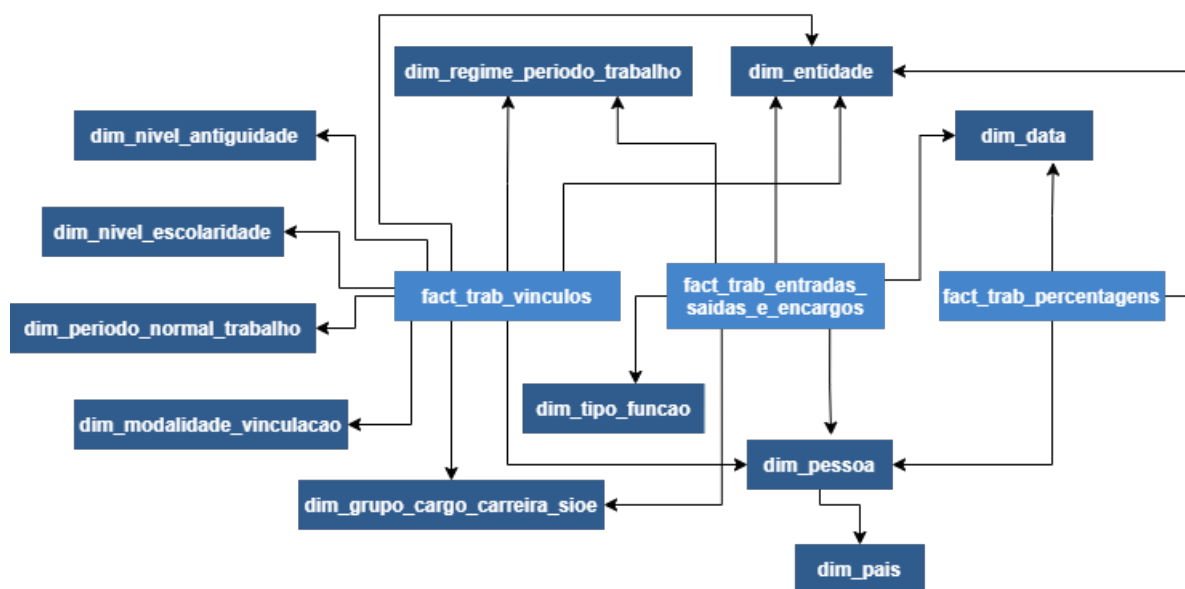


Figura 5.1: Modelo desenvolvido para o *Data Mart* dos recursos humanos

5.2 Tabelas de facto

Como podemos observar na Figura 5.1, este modelo é constituído por três tabelas de facto representadas em detalhe em seguida.

Trabalhadores - Vínculos

A tabela de factos *fact_trab_vinculos* contém informações relativamente aos trabalhadores da organização e é a tabela que possui um maior número de dimensões. Está relacionada com as dimensões, data, entidade, nível de escolaridade, pessoa, modalidade de vinculação, período normal de trabalho, regime do período de trabalho, nível de antiguidade e grupo/cargo/carreira SIOE.

Na tabela seguinte estão descritos em detalhe os atributos pertencentes a esta tabela:

Medidas	Descrição
id_data	ID da dimensão data
id_entidade	ID da dimensão entidade
id_grupo_cargo_carreira_sioe	ID da dimensão grupo/cargo/carreira SIOE
id_modalidade_vinculacao	ID da dimensão modalidade de vinculação
id_nivel_antiguidade	ID da dimensão nível de antiguidade
id_nivel_escolaridade	ID da dimensão nível de escolaridade
id_pessoa	ID da dimensão pessoa
id_periodo_normal_trabalho	ID da dimensão período normal de trabalho
id_regime_periodo_trabalho	ID da dimensão regime do período de trabalho
trabalhadores	Número mecanográfico de um trabalhador
vinculos	Toma sempre o valor um
trabalhadores_eti	Valor ETI de um trabalhador
ndoc_ninv	Número mecanográfico de um trabalhador - não docente não investigador
investigadores	Número mecanográfico de um trabalhador - investigador
investigadores_eti	Valor ETI de um trabalhador - investigador
docentes	Número mecanográfico de um trabalhador - docente
docentes_eti	Valor ETI de um trabalhador - docente
docentes_doutorados_eti	Valor ETI de um trabalhador - docente doutorado
docentes_carreira	Número mecanográfico de um trabalhador - docente de carreira
docentes_carreira_eti	Valor ETI de um trabalhador - docente de carreira
docentes_agregados	Número mecanográfico de um trabalhador - docente agregado
docentes_agregados_eti	Valor ETI de um trabalhador - docente agregado
docentes_exclusividade	Número mecanográfico de um trabalhador - docente em regime de exclusividade
docentes_exclusividade_eti	Valor ETI de um trabalhador - docente em regime de exclusividade
docentes_catedraticos_carreira	Número mecanográfico de um trabalhador docente catedrático de carreira
docentes_catedraticos_carreira_eti	Valor ETI de um trabalhador - docente catedrático de carreira
docentes_associados_carreira	Número mecanográfico de um trabalhador docente associado de carreira
docentes_associados_carreira_eti	Valor ETI de um trabalhador - docente associado de carreira
investigadores_carreira_eti	Valor ETI de um trabalhador - investigador de carreira
investigadores_doutorados_eti	Valor ETI de um trabalhador - investigador doutorado
ndoc_niv_carreira_eti	Valor ETI de um trabalhador - não docente não investigador de carreira
trabalhadores_femininos	Número mecanográfico de um trabalhador feminino
trabalhadores_deficiencia	Número mecanográfico de um trabalhador portador de deficiência

Tabela 5.2: Medidas e respetiva descrição da tabela *fact_trab_vinculos*

Trabalhadores - Entradas, saídas e Encargos

A tabela de factos *fact_trab_entradas_saidas_e_encargos* contém informações relativas à entrada e saída dos trabalhadores e encargos. Relativamente ao relacionamento com as tabelas de dimensão, possui ligação com as dimensões data, pessoa, grupo/cargo/carreira SIOE, entidade, função exercida e regime do período de trabalho.

A tabela seguinte contém os atributos que compõem esta tabela de factos.

Medidas	Descrição
id_data	ID da dimensão data
id_entidade	ID da dimensão entidade
id_grupo_cargo_carreira_sioe	ID da dimensão grupo/cargo/carreira SIOE
id_pessoa	ID da dimensão pessoa
id_funcao_exercida	ID da dimensão função exercida
id_regime_periodo_trabalho	ID da dimensão regime do período de trabalho
enc_remuneracao_base	Valor dos encargos da remuneração base
enc_outro_enc_pessoal	Valor de outros encargos com o pessoal
enc_supl_remuneratorios	Valor dos encargos de suplementos remuneratórios
enc_prest_sociais	Valor dos encargos com prestações sociais
admi_regress	Valor binário: 1 = trabalhador admitido e regressado, 0: caso contrário
saidos	Valor binário: 1 = saída do trabalhador, 0: caso contrário
saidos_lim_idade	Valor binário: 1 = saída por limite de idade, 0: caso contrário
saidos_aposentacao	Valor binário: 1 = saída por aposentação, 0: caso contrário
saidos_caducidade	Valor binário: 1 = saída por caducidade, 0: caso contrário
saidos_denuncia	Valor binário: 1 = saída por denúncia, 0: caso contrário
saidos_mobilidade	Valor binário: 1 = saída por mobilidade, 0: caso contrário
saidos_revogacao	Valor binário: 1 = saída por revogação, 0: caso contrário
saidos_cedencia	Valor binário: 1 = saída por cedência, 0: caso contrário
saidos_conc_sem_sucess_per_exp	Valor binário: 1 = saída por conclusão sem sucesso o período de experiência, 0: caso contrário
saidos_outras_situacoes	Valor binário: 1 = saída por outras situações, 0: caso contrário

Tabela 5.3: Medidas e respetiva descrição da tabela *fact_trab_entradas_saidas_e_encargos*

Trabalhadores - Percentagens

Esta tabela possui informação relativa a percentagens relacionadas com os trabalhadores e é composta pelas medidas da tabela *fact_trab_vinculos* (5.3) com exceção das seguintes: *docentes_agregados*, *docentes_agregados_eti*, *docentes_exclusividade*, *investigadores_carreira_eti*, *ndoc_ninv_carreira_eti*, *trabalhadores_carreira_eti*, *trabalhadores_deficiencia* e *vinculos*. Relativamente aos relacionamentos com as dimensões, apenas possui relação com a *dim_pessoa*, *dim_entidade* e *dim_data*.

Em comparação com as outras tabelas de facto implementadas, esta tabela possui algumas diferenças em relação ao modo de implementação. Enquanto as tabelas *fact_trab_vinculos* e *fact_trab_entradas_saidas_e_encargos* estão implementadas no DW, a tabela *fact_trab_percentagens* não está. O motivo para esta diferença, deve-se ao facto desta tabela possuir maioritariamente as mesmas medidas que a tabela *fact_trab_vinculos* no entanto, não possui as mesmas dimensões. Assim sendo, optou-se por duplicar a tabela *fact_trab_vinculos* no Power BI, alterou-se o nome e mantiveram-se apenas os relacionamentos com as dimensões pretendidas. Deste modo, evitou-se a criação de uma tabela no DW que possui quase os mesmos dados que a tabela *fact_trab_vinculos*.

De modo geral, estas três tabelas de facto são compostas por várias medidas que serão utilizadas como indicadores simples, bem como, medidas utilizadas para o cálculo de indicadores. Mais à frente, analisaremos em detalhe estes indicadores. A chave primária de cada uma das tabelas é composta por todas as chaves estrangeiras que fazem a relação com as diversas dimensões.

5.3 Tabelas de dimensão

Para além das três tabelas de facto, o modelo contém onze tabelas de dimensão, que veremos em detalhe de seguida.

Dimensão Entidade

A *dim_entidade* armazena na coluna *hierarquia_entidade*, a hierarquia completa de cada entidade associada a cada pessoa. Esta dimensão trata-se de uma hierarquia de níveis variáveis, no entanto, o Power BI não suporta diretamente este tipo de hierarquias [50]. Desta forma, optou-se por realizar uma implementação diferente, criando uma tabela auxiliar para esta dimensão, que permite representá-la, de uma forma correta no Power BI. Mais à frente será explicado detalhadamente o processo de desenvolvimento desta tabela auxiliar.

Atributos	Descrição
id_entidade	ID que identifica cada registo
hierarquia_entidade	Descrição da hierarquia da entidade

Tabela 5.4: Atributos e respetiva descrição da tabela *dim_entidade*

Dimensão Data

A *dim_data* armazena informação sobre as datas em que um determinado facto ocorre.

Atributos	Descrição
id_data	ID que identifica cada registo
data	Data no formato (yyyy-mm-dd)
mes	Mês no formato (mm)
mes_extenso	Descrição do mês
trimestre	Trimestre
trimestre_extenso	Descrição do trimestre
ano	Ano no formato (yyyy)
num_dia_semana	Número do dia da semana
dia_semana	Descrição do dia da semana
dia_mes	Número do dia do mês
num_semana_mes	Número da semana do mês
num_semana_ano	Número da semana do ano
indc_fim_de_semana	Indicador binário se é fim de semana
indc_ultimo_dia_mes	Indicador binário se é o último dia do mês

Tabela 5.5: Atributos e respetiva descrição da tabela *dim_data*

A *dim_data* contém todas as datas do calendário até ao final do ano corrente. Estes dados são gerados através de um *script* que está incluído no processo de ETL.

Dimensão Grupo/Cargo/Carreira SIOE

A *dim_grupo_cargo_carreira_sioe* guarda informação dos tipos de grupo, cargos e carreira SIOE associados a cada pessoa.

Atributos	Descrição
id_grupo_cargo_carreira_sioe	ID que identifica cada registo
grupo	Descrição do grupo
cargo	Descrição do cargo
categoria_profissional	Descrição da categoria profissional

Tabela 5.6: Atributos e respetiva descrição da tabela *dim_grupo_cargo_carreira_sioe*

Esta tabela apresenta uma estrutura hierárquica composta por três níveis: Grupo, Cargo e Categoria Profissional.

Dimensão Modalidade de vinculação

A *dim_modalidade_vinculacao* guarda informação sobre os tipos de modalidade de vinculação associados a cada pessoa.

Atributos	Descrição
id_modalidade_vinculacao	ID que identifica cada registo
modalidade_vinculacao	Descrição da modalidade de vinculação
regime	Descrição do regime
relacao_juridica_emprego	Descrição da relação jurídica de emprego

Tabela 5.7: Atributos e respetiva descrição da tabela *dim_modalidade_vinculacao*

Tal como a dimensão Grupo/Cargo/Carreira SIOE, esta dimensão tem uma estrutura hierárquica com três níveis.

Dimensão Nível de antiguidade

A *dim_nivel_antiguidade* armazena o nível de antiguidade, correspondente aos anos de trabalho associado a cada pessoa.

Atributos	Descrição
id_nivel_antiguidade	ID que identifica cada registo
nivel_antiguidade	Descrição do nível de antiguidade
numero_anos	Número de anos

Tabela 5.8: Atributos e respetiva descrição da tabela *dim_nivel_antiguidade*

Dimensão Nível de escolaridade

A *dim_nivel_escolaridade* guarda informação relativa ao nível de escolaridade de cada pessoa.

Atributos	Descrição
id_nivel_escolaridade	ID que identifica cada registo
codigo_raides	Código RAIDES do nível de escolaridade
nome_raides	Nome RAIDES do nível de escolaridade
nome	Nível de escolaridade
name_raides	Nome em inglês do nível de escolaridade
nivel_escolaridade_order	Ordem do nível de escolaridade

Tabela 5.9: Atributos e respetiva descrição da tabela *dim_nivel_escolaridade*

Dimensão Pessoa

A *dim_pessoa* armazena todas as informações relativas às pessoas. Esta dimensão para além da relação com as tabelas de facto, possui uma ligação à dimensão *dim_pais* através da chave *id_pais*. Esta é a dimensão com maior número de atributos e registos.

Atributos	Descrição
id_pessoa	ID que identifica cada registo
id_pessoa_so	ID da tabela pes_pessoa da gestão académica
nic	Número de identificação pessoal
nif	Número fiscal
num_passaporte	Número de passaporte
aut_residencia	Número de autorização de residência
nome_pessoa	Nome da pessoa
nome proprio_pessoa	Nome próprio da pessoa
apelido_pessoa	Apelido da pessoa
nome_abreviado_pessoa	Nome abreviado da pessoa
sexo	Sexo da pessoa (F ou M)
desc_sexo	Descrição do sexo
indc_sexo_masculino	Indicador sexo masculino
indc_sexo_feminino	Indicador sexo feminino
data_nascimento	Data de nascimento da pessoa
num_mecanografico	Número mecanográfico
estud_num_unico_inst	Número único de estudante
id_pais_nacionalidade	ID da nacionalidade
id_pais_nacionalidade_outra	ID de outra nacionalidade
validade_inicio_registo	Validade de início do registo
validade_fim_registo	Validade de fim do registo

Tabela 5.10: Atributos e respetiva descrição da tabela *dim_pessoa*

Dimensão País

A *dim_pais* guarda todos os dados relativos a países. Ao contrário das outras dimensões, não possui nenhum relacionamento com as tabelas de facto, apenas tem uma relação com a *dim_pessoa*.

Atributos	Descrição
id_pais	ID que identifica cada registo
codigo_iso_num	Código numérico do país
codigo_iso_alpha_2	Código ISO alpha-2 do país
nome_pais	Nome do país
name_pais	Nome em inglês do país
nacionalidade	Nacionalidade do país
nationality	Nacionalidade do país em inglês

Tabela 5.11: Atributos e respetiva descrição da tabela *dim_pais*

Dimensão Período normal de trabalho

A *dim_periodo_normal_trabalho* guarda informação relativa ao período normal de trabalho de cada pessoa, isto é, o número de horas semanais que os trabalhadores fazem. Além disso, guarda o intervalo a que cada valor corresponde e a respetiva ordem para que mais tarde, seja possível fazer a representação pela ordem pretendida.

Atributos	Descrição
id_periodo_normal_trabalho	ID que identifica cada registo
periodo_normal_trabalho	Descrição do período normal de trabalho
periodo_normal_trabalho_range	Descrição do período normal de trabalho em intervalos
periodo_normal_trabalho_order	Ordem do período normal de trabalho

Tabela 5.12: Atributos e respetiva descrição da tabela *dim_periodo_normal_trabalho*

Dimensão Função exercida

A *dim_funcao_exercida* guarda as diferentes funções exercidas existentes. Esta dimensão apenas se relaciona com a *fact_trab_entradas_saidas_e_encargos*.

Atributos	Descrição
id_funcao_exercida	ID que identifica cada registo
funcao_exercida	Descrição da função exercida

Tabela 5.13: Atributos e respetiva descrição da tabela *dim_funcao_exercida*

Dimensão Regime do período de trabalho

A *dim_regime_periodo_trabalho* guarda informação sobre o regime do período de trabalho de cada pessoa. Esta tabela apenas possui três registos, sendo a dimensão de menor tamanho.

Atributos	Descrição
id_regime_periodo_trabalho	ID que identifica cada registo
regime_periodo_trabalho	Descrição do regime do período de trabalho

Tabela 5.14: Atributos e respetiva descrição da tabela *dim_regime_periodo_trabalho*

Para todas as tabelas de dimensão, foram utilizadas chaves substitutas como chaves primárias, em vez de chaves compostas. Estas chaves são um número inteiro gerado de forma sequencial, a cada registo novo inserido. Assim sendo, isto permite que não haja duplicação de chaves ao inserir novos registos.

Nesta mesma fase, foram criadas as respetivas tabelas no servidor PostgreSQL, para onde os dados vão ser carregados.

Fazendo uma análise dos tipos de cada dimensão, temos como dimensões conformadas a *dim_grupo_cargo_carreira_sioe*, *dim_regime_do_periodo_de_trabalho*, *dim_data*, *dim_pessoa* e *dim_entidade*, pois são partilhadas entre as tabelas de facto. Além disso, a *dim_pessoa* é considerada uma dimensão de mudança lenta, em que neste caso, optou-se por acrescentar uma nova linha, sempre que ocorre uma mudança, mantendo desta forma o histórico.

5.4 Processo ETL

Concluído o modelo dimensional, procedeu-se o processo ETL. Este processo, tal como referido anteriormente, foi desenvolvido com a linguagem Python. A figura 5.2 ilustra as etapas seguidas no processo ETL

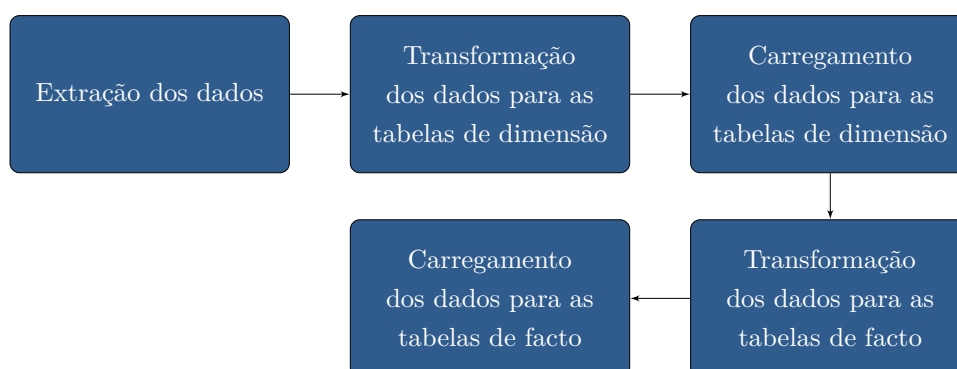


Figura 5.2: Processo ETL

A primeira etapa do processo ETL, corresponde à extração dos dados. Começou-se por criar a conexão com a fonte de dados da qual os dados vão ser extraídos. Para isso foi utilizado o módulo *cx_Oracle* do Python que permite o acesso a bases de dados Oracle [51]. Utilizou-se ainda a biblioteca *pandas* para armazenar os dados extraídos em *dataframes*. Esta biblioteca foi

desenhada para ser flexível, fácil de usar e bastante poderosa para manipulação de dados [52].

Após a extração dos dados, a próxima etapa no ETL passa pela transformação dos dados para as respectivas tabelas de dimensão. Para as dimensões *dim_grupo_cargo_carreira_sioe*, *dim_regime_periodo_trabalho*, *dim_modalidade_vinculacao*, *dim_periodo_normal_trabalho*, *dim_nivel_antiguidade*, *dim_pais*, *dim_nivel_escolaridade* e *dim_funcao_exercida* não foi necessário transformações avançadas. Apenas se teve em atenção a possibilidade da existência de valores em falta e nesses casos optou-se por substituir o valor por “Sem informação”. Além disso, foi necessário alterar o nome das colunas nos *dataframes* de forma a que seja compatível com as colunas das tabelas para onde os dados são carregados.

Relativamente à tabela *dim_pessoa*, foi necessário implementar histórico para esta dimensão. Sempre que são introduzidos novos registos, é necessário verificar algumas condições. Num primeiro carregamento todos os registos colocados possuem o atributo *validade_fim_registo* a NULL e o atributo *validade_inicio_registo* preenchido com a data a que os dados extraídos correspondem. Sempre que ocorre um novo carregamento, é necessário verificar se na tabela *dim_pessoa* já existe algum registo para aquela pessoa, se não existir, é adicionado o registo com os valores referidos anteriormente nos atributos *validade_inicio_registo* e *validade_fim_registo*. Caso contrário, é preciso verificar se ocorreu alguma modificação no registo dessa pessoa, como a alteração do nome ou da nacionalidade, sendo essas as mais comuns. Se isto ocorrer, é adicionado um novo registo para essa pessoa e o atributo *validade_fim_registo* do registo que já estava anteriormente na tabela é alterado. Deixa de ser NULL e passa a ter o valor da data para qual os dados foram extraídos. Se não houver nenhuma alteração, não é adicionado nenhum novo registo, nem é feito qualquer alteração nos registos que já estão na tabela *dim_pessoa*.

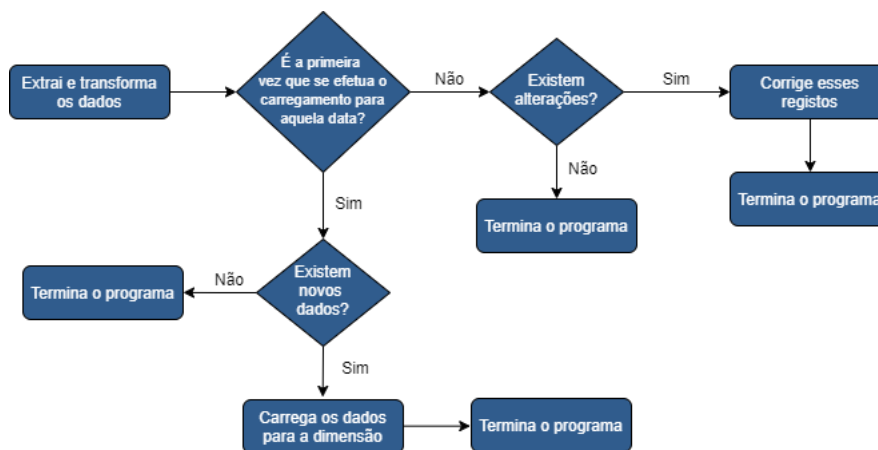


Figura 5.3: Processo ETL para a tabela *dim_pessoa*

Além da existência de histórico nesta dimensão, foi necessário fazer uma verificação de erros. Por vezes podem ocorrer erros em registos na fonte de dados, como um nome mal escrito. Quando esse tipo de erros ocorre, o atributo nesse registo é corrigido diretamente na fonte de dados. Posto isto, é essencial que essa mesma correção ocorra na tabela *dim_pessoa*, de forma a garantir a integridade dos dados.

Em relação à *dim_entidade* não foi preciso grandes transformações nos dados. No entanto, como referido anteriormente, foi necessário criar uma tabela auxiliar para a representação das entidades. Na tabela *dim_entidade* é guardada a hierarquia completa da entidade no atributo *hierarquia_entidade*, mas como a hierarquia tem níveis de profundidade variáveis, o Power BI tem algumas limitações a representar essas hierarquias. A solução encontrada passou pela criação de uma tabela auxiliar, *aux_hie_entidade*, que tem por base a tabela *dim_entidade*.

Para implementação da tabela *aux_hie_entidade*, começou-se por obter todos os dados que estão na tabela *dim_entidade* e utilizou-se a função *split*, que dividiu a coluna *hierarquia_entidade* em diferentes colunas, utilizando a *string* ' - ' como referência e a função *shape* para obter o número de colunas, ou seja, o número de níveis máximo na hierarquia. Posto isto, foi possível saber o número de colunas necessárias criar para tabela *aux_hie_entidade*.

Após a tabela criada, são carregados os dados que estão na *dim_entidade*, mas com as respetivas alterações, como demonstrado no seguinte exemplo:

id_entidade	hierarquia_entidade
1	UP - FCUP - Dep.CC
2	UP - FEUP
3	UP - FLUP - Dep.Sociol.
4	UP - FAUP - Administrativos - Gestão Financeira

Tabela 5.15: Tabela *dim_entidade* de exemplo

id_aux_hie_entidade	hie_level_0	hie_level_1	hie_level_2	hie_level_3
1	UP	FCUP	Dep.CC	Dep.CC
2	UP	FEUP	FEUP	FEUP
3	UP	FLUP	Dep.Sociol.	Dep.Sociol.
4	UP	FAUP	Administrativos	Gestão Financeira

Tabela 5.16: Tabela *aux_hie_entidade* de exemplo

Sempre que são adicionados novos registos na tabela *dim_entidade*, a tabela *aux_hie_entidade* é apagada e é criada novamente. O *id_entidade* e o *id_aux_hie_entidade* vão ser iguais. Desta forma, é possível criar uma relação entre o *id_entidade* das tabelas de factos e o *id_aux_hie_entidade* da tabela auxiliar, para a representação correta das hierarquias.

Quanto ao processo de carregamento dos dados para as tabelas de dimensão, é necessário garantir que não existem linhas duplicadas, porque isso levaria a problemas ao criar as relações entre as tabelas de dimensão e as tabelas de facto. Desta forma, sempre que é pretendido realizar um carregamento, é feita uma comparação dos dados que já estão nas tabelas de dimensão com os dados que pretendemos carregar de modo a carregar apenas os dados que ainda não estão lá.

É importante referir que todos os dados carregados nas dimensões são extraídos da fonte de dados (GRH), à exceção da *dim_data*, em que os dados são gerados através de um *script*.

Acresce que, existem algumas dimensões partilhadas com outros *Data Marts*, como é o caso da *dim_pessoa*, *dim_data*, *dim_nivel_escolaridade* e *dim_pais*.

Após concluída a implementação de todo o processo ETL para as tabelas de dimensão, prosseguiu-se para o ETL das tabelas de facto. Como estas tabelas estão dependentes das tabelas de dimensão, devido aos relacionamentos através das chaves estrangeiras, foi necessário proceder ao processo de ETL das tabelas de dimensão em primeira lugar.

As transformações dos dados para as tabelas de facto são mais complexas, comparativamente às que são efetuadas para as tabelas de dimensão. Ao contrário das tabelas de dimensão, os atributos das tabelas de facto são todos numéricos, portanto os dados extraídos da origem necessitam de tratamento, não podendo ser carregados diretamente nas tabelas de facto, como é o caso de algumas dimensões. Além disso, as tabelas de facto são constituídas por chaves estrangeiras, que fazem o relacionamento com as tabelas de dimensão, portanto, é também necessário efetuar o respetivo relacionamento. Assim, após os dados passarem pelo processo de transformações, é feita a união com as tabelas de dimensão, de modo a obter as chaves estrangeiras. Finalmente, os dados são carregados para as tabelas de facto, ficando assim concluído o processo ETL. Neste projeto, foi definido o carregamento dos últimos cinco anos para o DW.

Relativamente aos tempos de execução do ETL, o maior gasto de tempo é na extração dos dados, demorando em média cerca de 22s para a tabela de factos *fact_trab_vinculos* e 219s para a tabela de factos *fact_trab_entradas_saidas_e_encargos*. Em relação ao tempo de transformações e carregamento, para a tabela de factos *fact_trab_vinculos* demora cerca de 6s e para a tabela de factos *fact_trab_entradas_saidas_e_encargos* cerca de 9s. Na tabela seguinte, está representado o tempo de extração, o número de registos extraídos e tempo de total nos últimos quatro trimestres do ano de 2020, para cada uma das tabelas.

Data	Tempo de extração	Nº de registos extraídos	Tempo total
1º Trim 2020	27.67s	4251	33.58s
2º Trim 2020	26.64s	4496	37.80s
3º Trim 2020	22.53s	4499	27.85s
4º Trim 2020	23.11s	4798	27.80s

Tabela 5.17: Tabela com os tempos de execução do processo ETL para a tabela de factos Trabalhadores - Vínculos"

Data	Tempo de extração	Nº de registos extraídos	Tempo total
1º Trim 2020	197.39s	56806	206.25s
2º Trim 2020	338.34s	114425	347.69s
3º Trim 2020	151.18s	49506	163.04s
4º Trim 2020	220.83s	72220	227.26s

Tabela 5.18: Tabela com os tempos de execução do processo ETL para a tabela de factos "Trabalhadores - Entradas, saídas e Encargos"

Pretende-se que a execução do processo ETL seja trimestral.

Após concluído o processo ETL, em que os dados são extraídos, transformados e carregados para o DW, a próxima fase diz respeito ao acesso aos dados através de ferramentas BI. Nesta fase os dados serão explorados, de forma a obter informação através de relatórios e *dashboards*.

5.5 Cálculo dos indicadores

Após os dados estarem integrados no DW e antes da criação dos *dashboards* é necessário fazer uma análise dos dados. Algumas das medidas presentes nas tabelas de facto são utilizadas como indicadores simples, porém existem outros indicadores, como percentagens que não são obtidos diretamente, necessitando de serem calculados através de outras medidas. Desta forma, antes de se avançar para a criação dos *dashboards* é necessário criar estes indicadores previamente definidos.

Optou-se pela utilização do Power BI nesta etapa. O Power BI utiliza a mesma tecnologia utilizada para a construção de modelos tabulares no SQL Server Analysis Services (SSAS) [53]. Estes modelos permitem realizar análises eficientemente. Assim sendo, foi utilizado o Power BI, em alternativa à ferramenta SSAS, que implicaria um custo maior no desenvolvimento desta solução.

5.5.1 Importação do modelo

O primeiro passo é a importação do modelo para o Power BI, seleccionando a opção "obter dados" de uma base dados PostgreSQL. O Power BI oferece duas opções para o modo de importação dos dados. A primeira é "Importar" ou "*in-memory*", como referido na subsecção 2.5.2. Este modo é utilizado quando os modelos não são muito grandes, ficando os dados armazenados em memória, ao contrário do "*DirectQuery*".

Para esta solução, foi utilizado o modo "Importar", dado que o nosso modelo não é grande, para além de que esta opção permite o acesso a uma quantidade maior de funcionalidades no Power BI [54].

5.5.2 Medidas e colunas calculadas

As medidas e colunas calculadas são implementadas nas tabelas através da linguagem DAX. Todas as medidas e colunas foram criadas nas tabelas de facto. As tabelas seguintes mostram em detalhe as medidas e colunas criadas no modelo, começando pelas que foram criadas na tabela *fact_trab_vinculos*. Posteriormente na tabela *fact_trab_entradas_saidas_e_encargos* e por fim na tabela *fact_trab_percentagens*.

Como foi referido anteriormente, a tabela *fact_trab_percentagens* utiliza as medidas presentes na tabela *fact_trab_vinculos*. Portanto, algumas das medidas existentes na tabela *fact_trab_vinculos* servirão apenas como medidas auxiliares no cálculo de outras medidas na tabela *fact_trab_percentagens*.

As seguintes medidas necessitaram de transformações para a sua representação correta na ferramenta de visualização, como, por exemplo, evitar o aparecimento de espaços em branco na representação do indicador, substituindo por 0:

docentes, investigadores, ndoc_ninv, trabalhadores, trabalhadores_deficiencia, docentes_carreira, docentes_associados_carreira, docentes_catedraticos_carreira e *trabalhadores_femininos*.

Para cada uma destas medidas foram usadas as seguintes fórmulas DAX, respetivamente:

Fórmula DAX	Medidas
<pre>VAR aux = DISTINCTCOUNTNOBLANK('public fact_trab_vinculos'[medida]) RETURN IF(ISBLANK(aux),0,aux)</pre>	docentes, investigadores, ndoc_ninv, trabalhadores, trabalhadores_deficiencia
<pre>VAR aux = DISTINCTCOUNTNOBLANK('public fact_trab_percentagens'[medida]) RETURN IF(ISBLANK(aux),0,aux)</pre>	docentes, investigadores, ndoc_ninv, trabalhadores, docentes_associados_carreira, docentes_carreira, docentes_catedraticos_carreira, trabalhadores_femininos

Consequentemente, deu origem às seguintes medidas, respetivamente:

- N° de Docentes
- N° de Investigadores
- N° de Não Docentes/Não Investigadores

- N° de Trabalhadores
- N° de Trabalhadores portadores de deficiência
- N° de Docentes Associados Carreira
- N° de Docentes Carreira
- N° de Docentes Catedráticos Carreira
- N° de Trabalhadores Femininos

Na tabela seguinte estão representadas as medidas criadas na tabela *fact_trab_percentagens*, que dependem de outras para serem calculadas.

Medida	Fórmula DAX
N° Não Docente/Não Investigador com habilitação superior	<code>CALCULATE([N° de Não Docentes/Não Investigadores], FILTER(RELATEDTABLE('public dim_nivel_escolaridade'), 'public dim_nivel_escolaridade'[nome]="Bacharelato" 'public dim_nivel_escolaridade'[nome]= "Licenciatura" 'public dim_nivel_escolaridade'[nome]="Mestrado" 'public dim_nivel_escolaridade'[nome]="Doutoramento"))</code>
Percentagem de docentes associados e catedráticos de carreira	<code>DIVIDE((([N° de Docentes Associados Carreira] + [N° de Docentes Catedráticos Carreira]), [N° de Docentes Carreira])</code>
Percentagem de docentes associados e catedráticos de carreira ETI	<code>DIVIDE((SUM('public fact_trab_percentagens' [docentes_associados_carreira_eti]) + SUM('public fact_trab_percentagens'[docentes_catedraticos_carreira_eti])), SUM('public fact_trab_percentagens'[docentes_carreira_eti]))</code>
Percentagem de docentes doutorados (ETI)	<code>DIVIDE(SUM('public fact_trab_percentagens'[docentes_doutorados_eti]), SUM('public fact_trab_percentagens'[docentes_eti]))</code>
Percentagem de docentes e investigadores doutorados (ETI)	<code>DIVIDE((SUM('public fact_trab_percentagens' [docentes_doutorados_eti]) + SUM('public fact_trab_percentagens'[investigadores_doutorados_eti])), (SUM('public fact_trab_percentagens'[docentes_eti]) + SUM('public fact_trab_percentagens'[investigadores_eti]))))</code>
Percentagem de docentes em regime de exclusividade (ETI)	<code>DIVIDE(SUM('public fact_trab_percentagens'[docentes_exclusividade_eti]), SUM('public fact_trab_percentagens'[docentes_eti]))</code>
Percentagem de investigadores doutorados (ETI)	<code>DIVIDE(SUM('public fact_trab_percentagens' [investigadores_doutorados_eti]), SUM('public fact_trab_percentagens'[investigadores_eti]))</code>

Percentagem de Não docentes/Não Investigadores com habilitação superior	<code>DIVIDE</code> ([Nº Não Docente(Não Investigador com habilitação superior], [Nº de Não Docentes/Não Investigadores])*100
Percentagem de pessoal docente e investigador	<code>DIVIDE</code> ('public fact_trab_percentagens'[Nº de Docentes] + 'public fact_trab_percentagens'[Nº de Investigadores], 'public fact_trab_percentagens'[Nº de Trabalhadores])
Percentagem de trabalhadores ETI por função exercida (Docentes)	<code>DIVIDE</code> (<code>SUM</code> ('public fact_trab_percentagens'[docentes_eti]), <code>SUM</code> ('public fact_trab_percentagens'[trabalhadores_eti]))
Percentagem de trabalhadores ETI por função exercida (Investigadores)	<code>DIVIDE</code> (<code>SUM</code> ('public fact_trab_percentagens'[investigadores_eti]), <code>SUM</code> ('public fact_trab_percentagens'[trabalhadores_eti]))
Percentagem de trabalhadores ETI por função exercida (Não docentes/Não investigadores)	<code>DIVIDE</code> (<code>SUM</code> ('public fact_trab_percentagens'[ndoc_ninv_eti]), <code>SUM</code> ('public fact_trab_percentagens'[trabalhadores_eti]))
Taxa de feminização	<code>DIVIDE</code> (<code>DISTINCTCOUNT</code> ('public fact_trab_percentagens'[trabalhadores_femininos]), <code>DISTINCTCOUNT</code> ('public fact_trab_percentagens'[trabalhadores]))

Tabela 5.19: Medidas calculadas na tabela *fact_trab_percentagens*

Relativamente às medidas criadas na tabela *fact_trab_entradas_saidas_e_encargos*, estas encontram-se representadas na tabela seguinte:

Medida	Fórmula DAX
Nº Admitidos e Regressados	<code>IF</code> (<code>SUM</code> ('public fact_trab_entradas_saidas_e_encargos'[admi_regress]) < 0,0, <code>SUM</code> ('public fact_trab_entradas_saidas_e_encargos'[admi_regress]))
Nº Saidas	<code>IF</code> (<code>SUM</code> ('public fact_trab_entradas_saidas_e_encargos'[saidos]) < 0,0, <code>SUM</code> ('public fact_trab_entradas_saidas_e_encargos'[saidos]))
Taxa de Cobertura	<code>DIVIDE</code> ([Nº Admitidos e Regressados], [Nº Saidas])*100
Total de encargos	<code>SUM</code> ('public fact_trab_entradas_saidas_e_encargos'[enc_outros_enc_pessoal]) + <code>SUM</code> ('public fact_trab_entradas_saidas_e_encargos'[enc_prest_sociais]) + <code>SUM</code> ('public fact_trab_entradas_saidas_e_encargos'[enc_remuneracao_base]) + <code>SUM</code> ('public fact_trab_entradas_saidas_e_encargos'[enc_supl_remuneratorios])

Tabela 5.20: Medidas calculadas na tabela *fact_trab_entradas_saidas_e_encargos*

Em todas as tabelas de facto para além das medidas calculadas, foi calculada a coluna idade.

Coluna	Fórmula DAX
Idade	<pre>VAR data_ref = RELATED('public dim_data'[data]) VAR data_nascimento = RELATED('public dim_pessoa'[data_nascimento]) RETURN INT(YEARFRAC(data_nascimento, data_ref))</pre>

Tabela 5.21: Fórmula DAX para calcular a coluna idade

As medidas calculadas em cada uma destas tabelas, representam os restantes indicadores que não são obtidos diretamente.

De uma forma geral, foram criadas trinta medidas das quais, cinco medidas na tabela de factos *fact_trab_vinculos*, vinte e uma medidas na tabela *fact_trab_percentagens* e quatro medidas na tabela *fact_trab_entradas_saidas_e_encargos*. É importante referir que estas medidas calculadas estão ao nível do *dashboard* e não implementadas na base de dados. Desta forma, só serão utilizadas nos *dashboards* desenvolvidos.

5.6 Dashboards

Após os dados estarem integrados no *Data Warehouse* e os indicadores já calculados, a última etapa desta solução consiste no desenvolvimento de *dashboards*. Os *dashboards* permitem aceder e visualizar a informação de uma forma fácil, rápida e intuitiva, sendo essenciais na análise da informação e no suporte na tomada de decisões.

A forma como os *dashboards* estão construídos e os elementos que os compõem, como, por exemplo, gráficos, mapas, tabelas, entre outros elementos visuais, influenciam a interpretação e análise dos dados por parte dos utilizadores, portanto é necessário algum cuidado no seu desenvolvimento.

5.6.1 Design dos *dashboards*

Na construção dos *dashboards* foi necessário ter em consideração alguns detalhes importantes. Entre eles:

- *Interface* simples;
- Pouca variação de cores, utilizar sempre o mesmo esquema de cores;
- Todos os *dashboards* identificadas com o logótipo da Universidade do Porto e o nome do projeto (BI4UP);
- Utilizar a mesma estrutura em todas os *dashboards*;

- Possibilitar a definição de filtros.

Todos os *dashboards* desenvolvidos seguiram a estrutura apresentada a seguir:

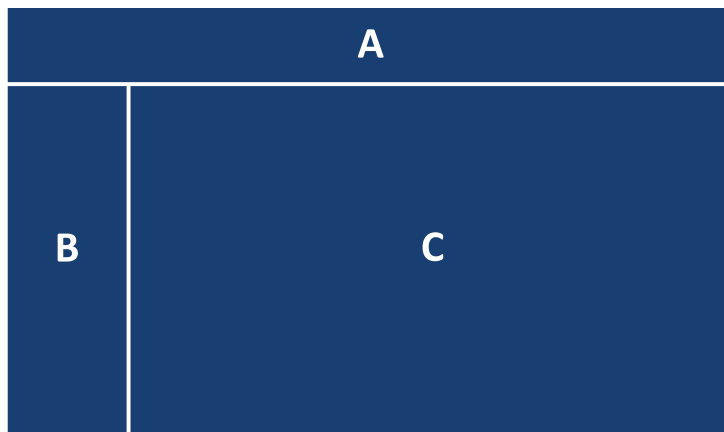


Figura 5.4: Estrutura do *dashboard*

A: Na parte superior do *dashboard* encontra-se um botão que permite voltar ao menu/submenu inicial, o título, o logótipo da Universidade do Porto e nome do projeto (BI4UP). Com exceção do título, os outros três atributos são estáticos.

B: No lado esquerdo do *dashboard*, estão localizados os diversos filtros que permitem uma maior flexibilidade do *dashboard*.

C: Todas as visualizações de dados e gráficos encontram-se representados nesta área.

Assim sendo, neste projeto foram desenvolvidos vários *dashboards* para a área de recursos humanos, com base nos requerimentos definidos e nos indicadores identificados.

5.6.2 *Dashboards criados*

Antes de avançar diretamente para os *dashboards* que contêm os indicadores, consideramos importante a existência de uma página (figuras 5.5 e 5.6) que dá a conhecer ao utilizador as três áreas (Vínculos, Percentagens e Entradas, Saídas e Encargos) de informação dos RH que pode encontrar ao longo dos *dashboards*, permitindo navegar para um submenu correspondente a cada área.



Figura 5.5: Menu inicial



Figura 5.6: Submenu da área de vínculos

No total, foram desenvolvidos vinte *dashboards* para a área de recursos humanos. Em seguida, será analisado cada um destes *dashboards*.

Resumo do Número de trabalhadores

Tal como o título indica, neste *dashboard* (figura 5.7) é apresentado um resumo geral sobre o número de trabalhadores. Aqui, o utilizador consegue obter o número de trabalhadores nas diferentes entidades constitutivas e nos diferentes grupos/carreiras e tem a possibilidade de fazer *drill-down* e *roll-up* nos gráficos. Além disso, é possível ver a percentagem de trabalhadores femininos e masculinos, através do gráfico circular. Por fim, é possível saber o número total de vínculos.



Figura 5.7: Página - Resumo do Número de trabalhadores

Para além dos gráficos, este *dashboard* possui um botão que permite ao utilizador navegar até outro *dashboard* (figura 5.8), onde estão presentes dois gráficos que mostram a evolução do número de trabalhadores por função exercida e sexo.

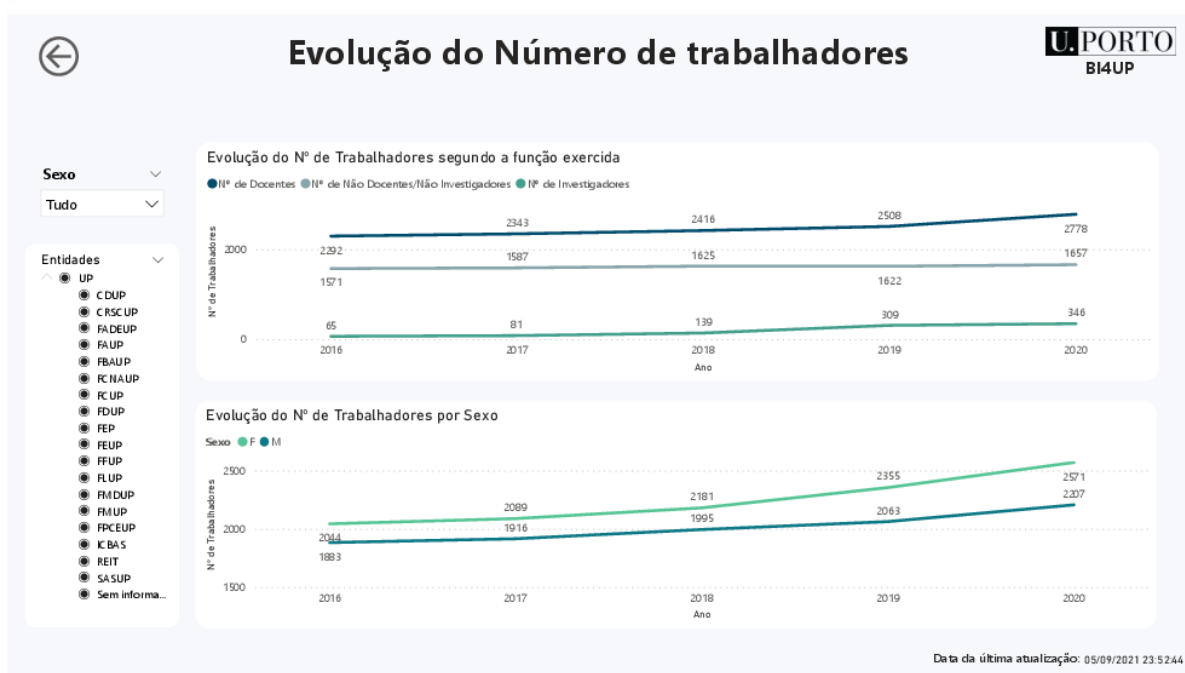


Figura 5.8: Página - Evolução do número de trabalhadores

Trabalhadores segundo a Função Exercida e Nível de Escolaridade

Neste *dashboard* (figura 5.9) encontra-se representado o número de trabalhadores por nível de escolaridade, ordenado do nível mais baixo para o mais alto. No gráfico de barras à direita encontra-se o número de trabalhadores segundo a função exercida e sexo. Por fim, o gráfico na parte de baixo do *dashboard* mostra a percentagem de trabalhadores segundo a função exercida e o nível de escolaridade.

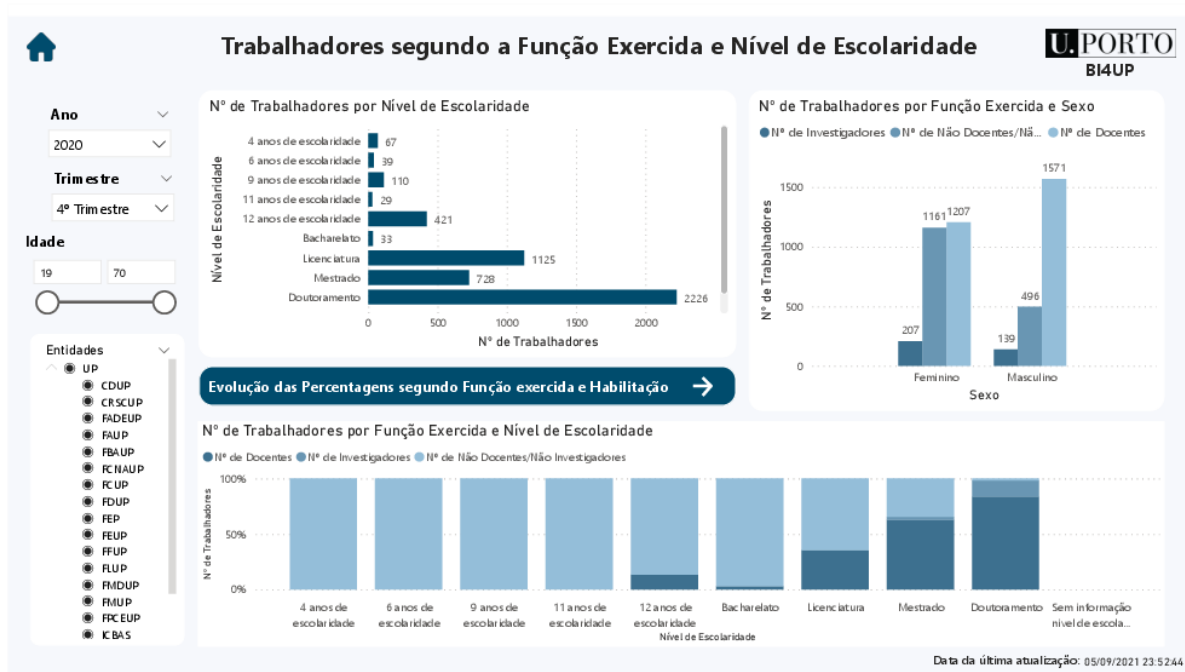


Figura 5.9: Trabalhadores segundo a Função Exercida e Nível de Escolaridade

Do mesmo modo que o primeiro *dashboard*, existe um botão que remete o utilizador para um *dashboard* (figura 5.10) com três gráficos evolutivos.

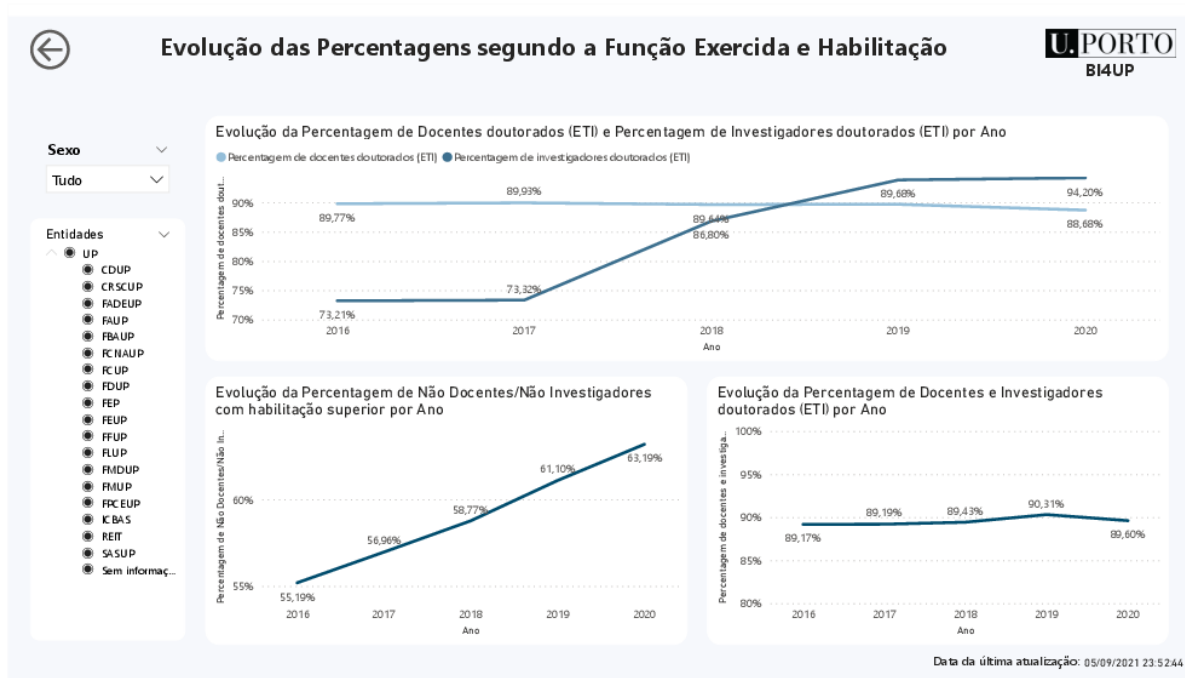


Figura 5.10: Evolução das Percentagens segundo a Função Exercida e Habilitação

Trabalhadores segundo a Modalidade de Vinculação

Neste *dashboard* (figura 5.11) apresenta-se uma tabela com número de trabalhadores em regime privado e público, e um gráfico de barras que mostra a percentagem de trabalhadores em regime privado e público nas diferentes entidades constitutivas. Em ambas as visualizações é possível fazer *drill-down* e *roll-up*. Por último, este *dashboard* contém três gráficos circulares que dão a informação do número de trabalhadores por função exercida nos dois diferentes regimes.

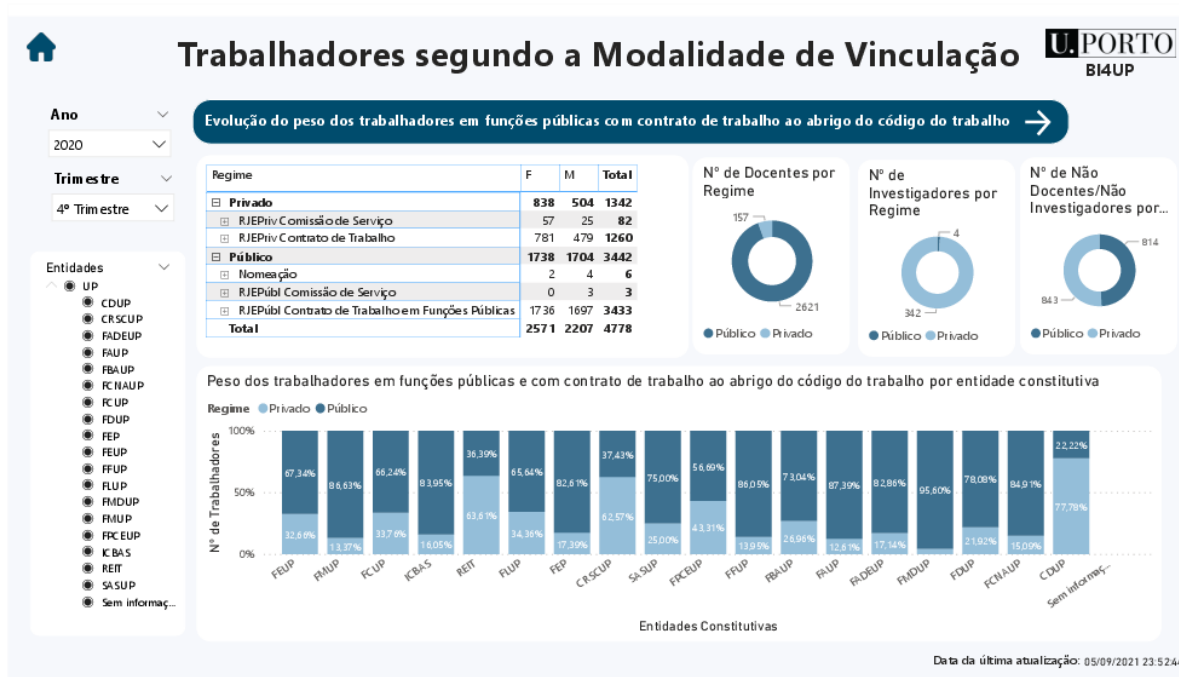


Figura 5.11: Trabalhadores segundo a modalidade de vinculação

Além disso, existe um botão que leva o utilizador para um *dashboard* (figura 5.12) com três gráficos que mostram a evolução do número de trabalhadores por regime privado e público nas três funções exercidas.

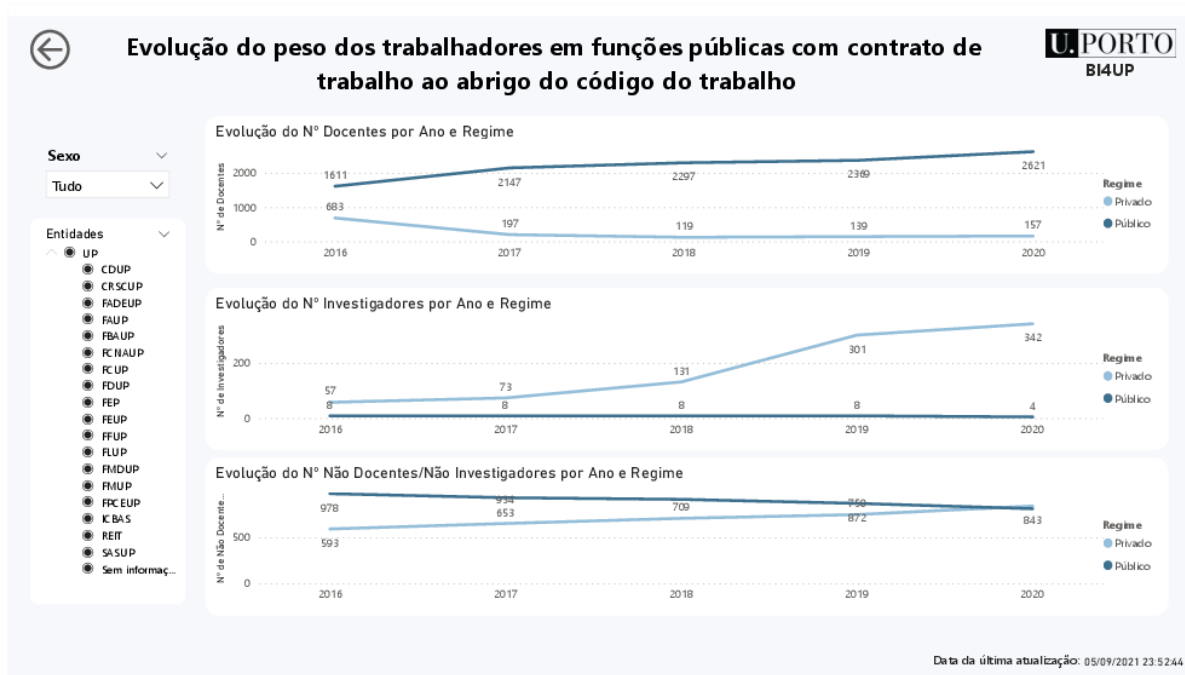


Figura 5.12: Evolução do peso dos trabalhadores em funções públicas com contrato de trabalho ao abrigo do código de trabalho

Diversidade Trabalhadores I e II

Nestes dois *dashboards* estão representadas várias visualizações relacionadas com a diversidade dos trabalhadores. No primeiro (figura 5.13), o utilizador pode observar a variação de idades dos trabalhadores e a diversidade de nacionalidades, além da percentagem de trabalhadores femininos e masculinos a tempo parcial ou integral e do número de trabalhadores portadores de deficiência.



Figura 5.13: Diversidade Trabalhadores - I

Relativamente ao segundo *dashboard* sobre a diversidade dos trabalhadores (figura 5.14), encontra-se representado o número de trabalhadores por nível de antiguidade, o número de trabalhadores por modalidade de vinculação e a percentagem de trabalhadores femininos e masculinos por período normal de trabalho. Por fim, dispõe de um botão que encaminha o utilizador para um *dashboard* (figura 5.15) que contém a evolução da taxa de feminização.

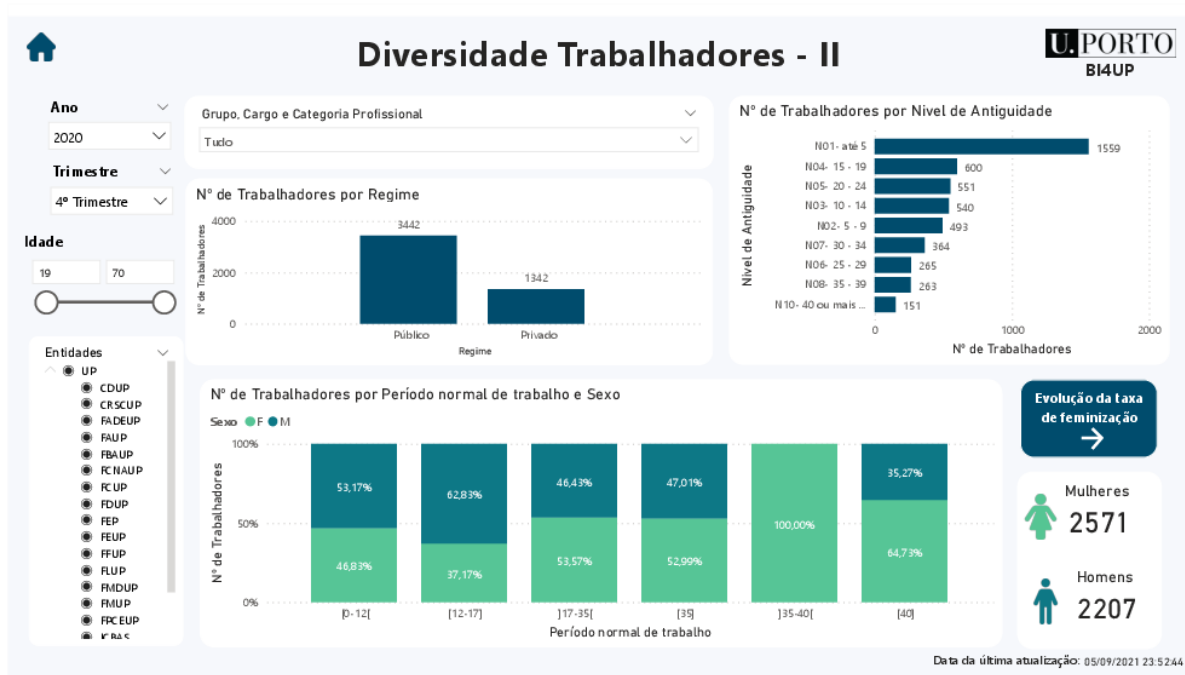


Figura 5.14: Diversidade Trabalhadores - II



Figura 5.15: Evolução da taxa de feminização

Em ambas os *dashboards* (figuras 5.13 e 5.14) é apresentado ao utilizador o número de trabalhadores femininos e masculinos. Estes valores são alterados, consoante as interações que o utilizador tem com os restantes gráficos. Por exemplo, ao interagir com a visualização do

mapa, poderá filtrar apenas os trabalhadores com nacionalidade espanhola, e assim, poderá saber quantos trabalhadores femininos e masculinos com nacionalidade espanhola existem na organização.

Docentes - I

Neste *dashboard* (figura 5.16) estão representados vários indicadores relacionados com os docentes. Nesta página, o utilizador consegue obter a informação de quantos docentes existem em cada entidade constitutiva, o número de docentes(ETI), o número de docentes de carreira e o número de docentes de carreira(ETI). Além disso, ainda consegue obter a percentagem de docentes femininos e masculinos.

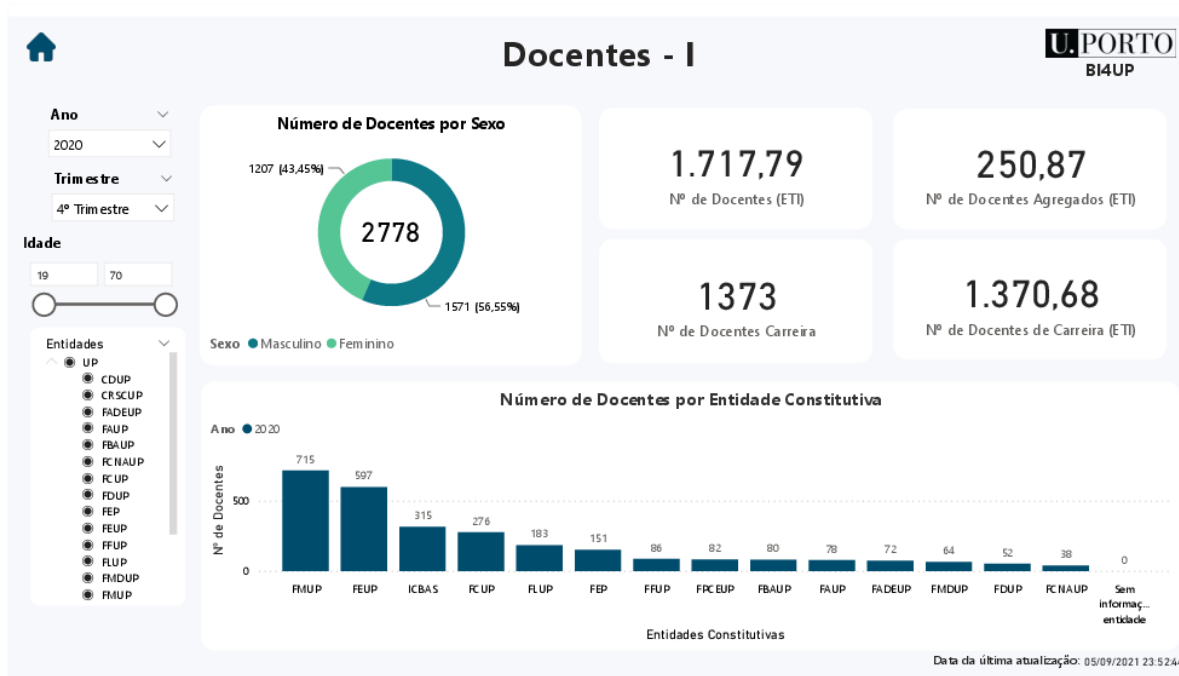


Figura 5.16: Docentes - I

Docentes - II

Neste *dashboard* (figura 5.17) é apresentado o número de docentes por nível de antiguidade e variação de idades dos docentes.

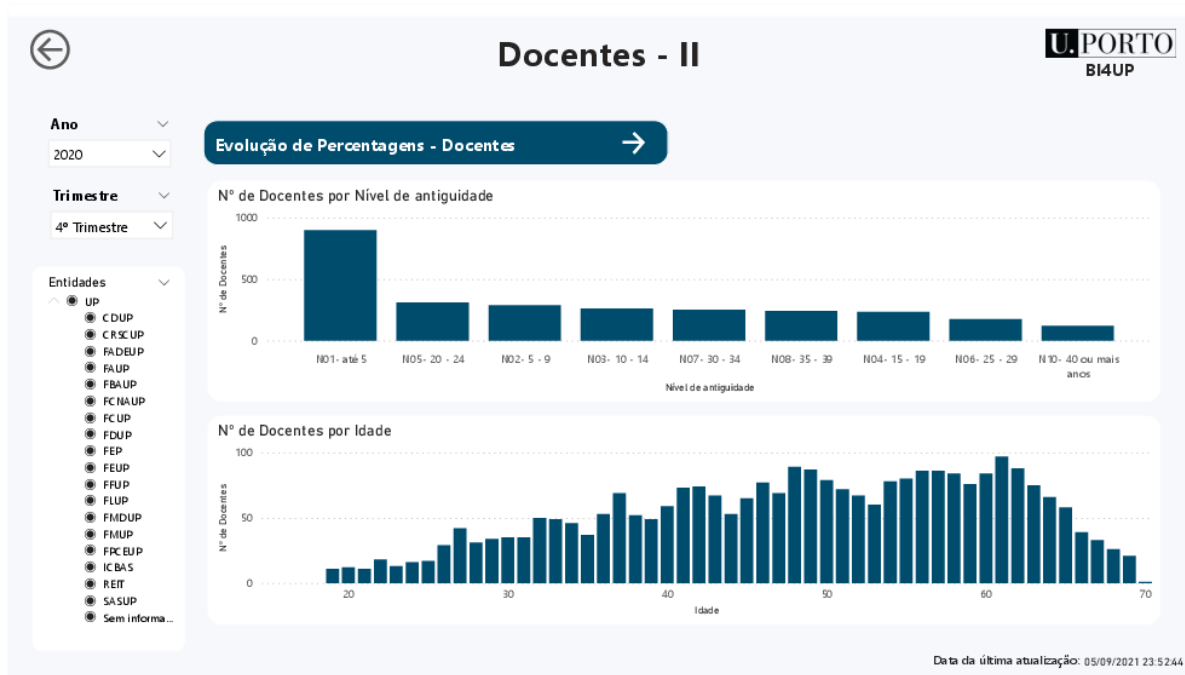


Figura 5.17: Docentes - II

Neste *dashboard* (figura 5.18) existe um botão que encaminha o utilizador para um *dashboard* que contém três gráficos evolutivos de indicadores relacionados com os docentes.

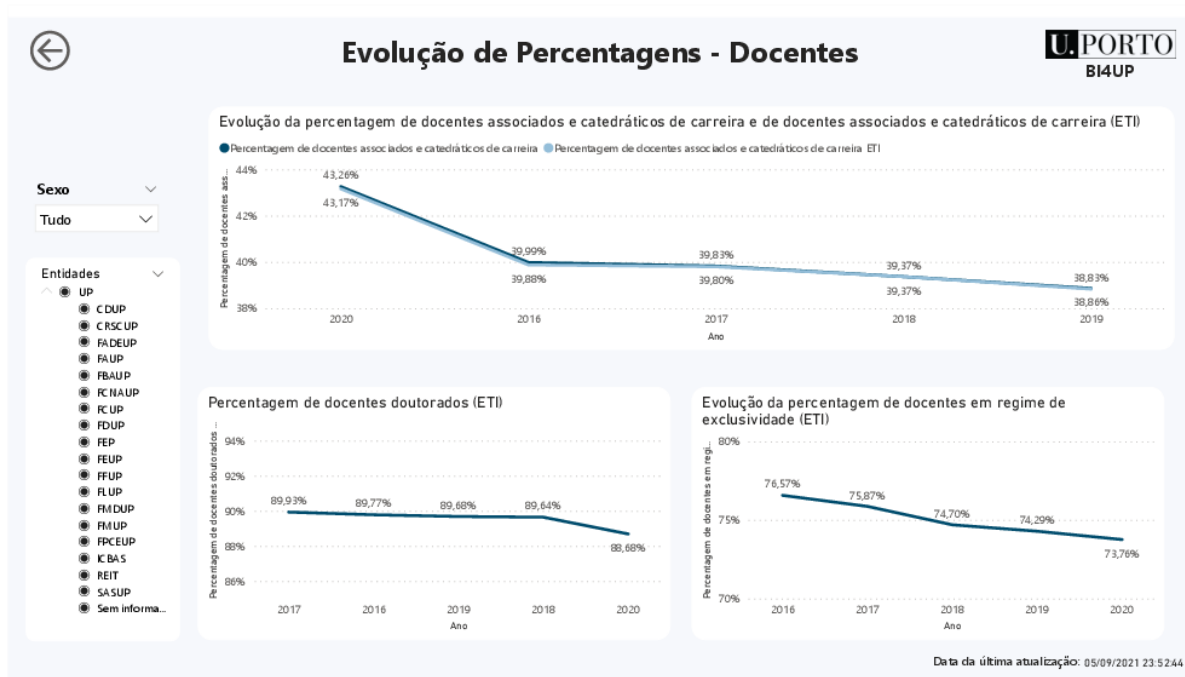


Figura 5.18: Evolução das percentagens - Docentes

Trabalhadores ETI

Neste *dashboard* (figura 5.19) são apresentados vários valores relacionados com o ETI dos trabalhadores. Encontra-se representado o número total de trabalhadores, o número total de trabalhadores ETI e o número ETI por regime, que permite ao utilizador fazer uma comparação de forma rápida. Representado num gráfico circular, encontram-se os valores ETI por função exercida.

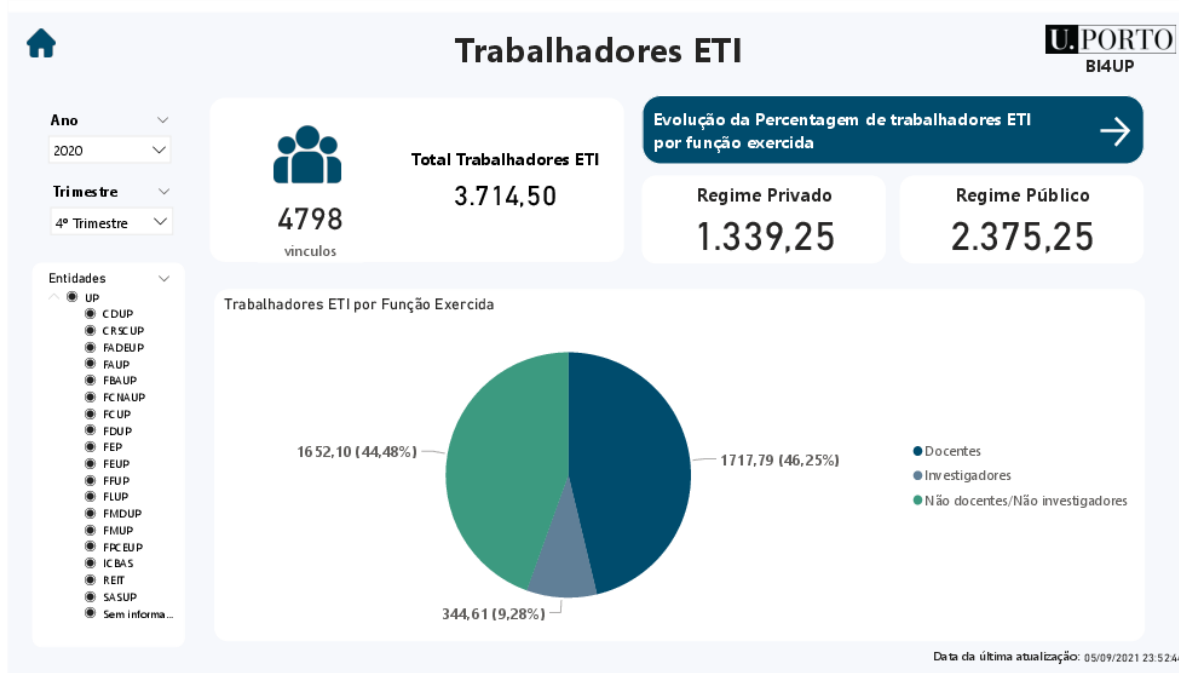


Figura 5.19: Trabalhadores ETI

Além destes gráficos, existe um botão para um novo *dashboard* (figura 5.20) que dá a conhecer ao utilizador a evolução dos valores ETI por função exercida nos últimos anos.



Figura 5.21: Encargos com o pessoal

Trabalhadores admitidos e regressados

Neste *dashboard* (figura 5.22) são apresentados vários gráficos que mostram o número de admitidos e regressados pelas diferentes dimensões. No primeiro gráfico está representado o número de admitidos e regressados por função exercida, logo a seguir encontram-se um gráfico que mostra a variação de idades dos trabalhadores admitidos e regressados e ao lado, um gráfico que mostra a percentagem de admitidos e regressados a tempo integral e parcial. Por fim, na parte de baixo do *dashboard* está representado um gráfico que contém o número de admitidos e regressados segundo o cargo/carreira e sexo.

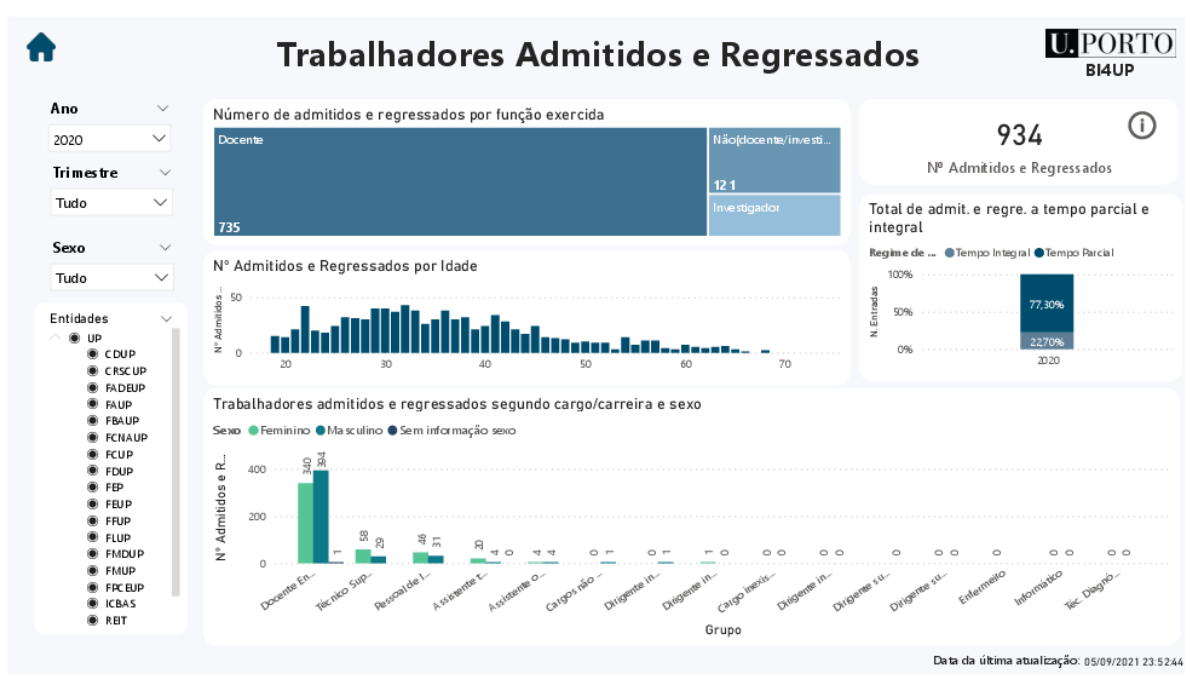


Figura 5.22: Trabalhadores admitidos e regressados

Trabalhadores Saídas

Os gráficos presentes neste *dashboard* (figura 5.23) são os mesmos que estão presentes no *dashboard* dos trabalhadores admitidos e regressados, mas relativos aos trabalhadores que saíram, com exceção do gráfico que mostra a percentagem de admitidos e regressados a tempo integral e parcial.

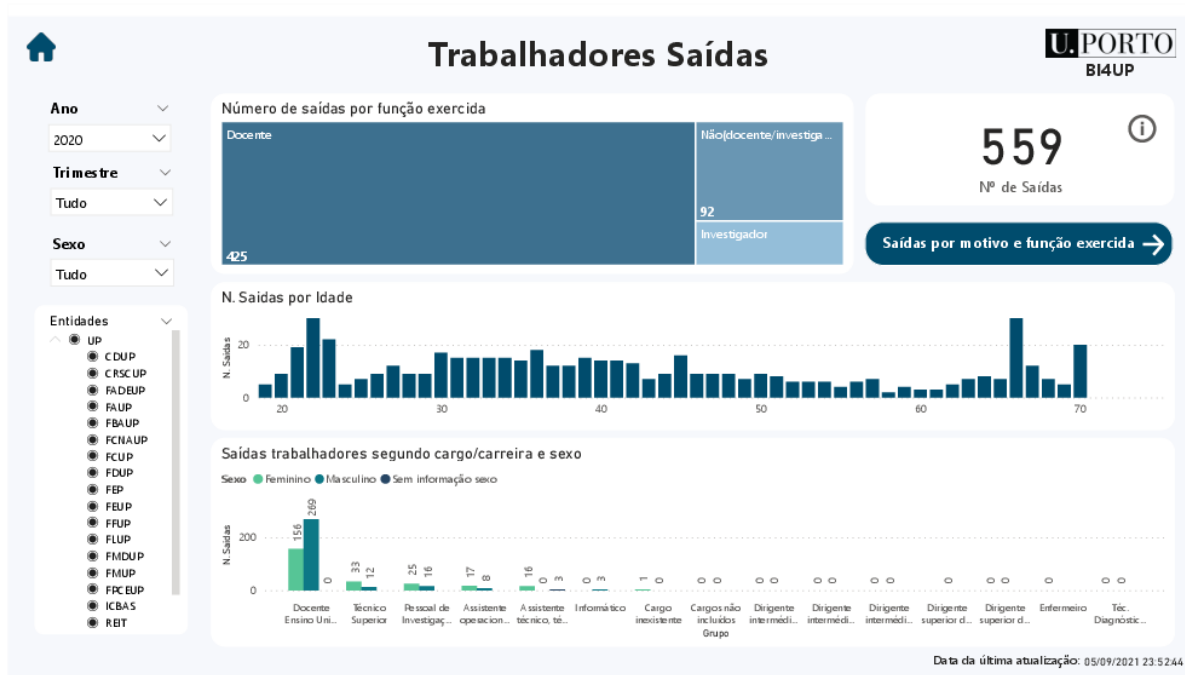


Figura 5.23: Trabalhadores saídas

No entanto, ao contrário do *dashboard* (figura 5.22), existe um botão que encaminha o utilizador para um novo *dashboard* (figura 5.24) que mostra a evolução do número de saídas ao longo dos anos e número de trabalhadores que saíram por função exercida e motivo de saída. Neste *dashboard* também existe um botão, que encaminha o utilizador para um *dashboard* (figura 5.25) que mostra o número de trabalhadores que saíram segundo o motivo de saída, cargo/carreira e sexo.

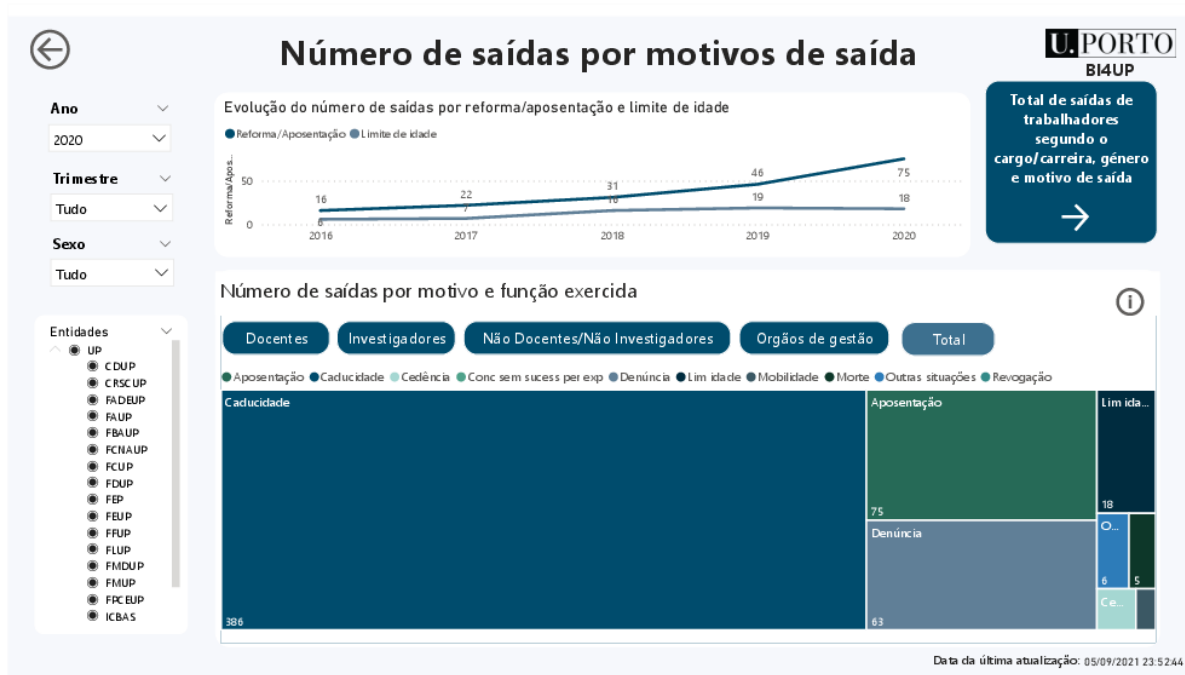


Figura 5.24: Número de saídas por motivos de saída



Figura 5.25: Total de saídas de trabalhadores segundo o cargo/carreira, sexo e motivo de saída

Taxa de cobertura

Neste *dashboard* (figura 5.26) encontram-se dois gráficos. O primeiro mostra o número de saídas e o número de admitidos e regressados, mais a respetiva taxa de cobertura segundo as diferentes entidades constitutivas. No segundo gráfico está representado a taxa de cobertura segundo a função exercida.

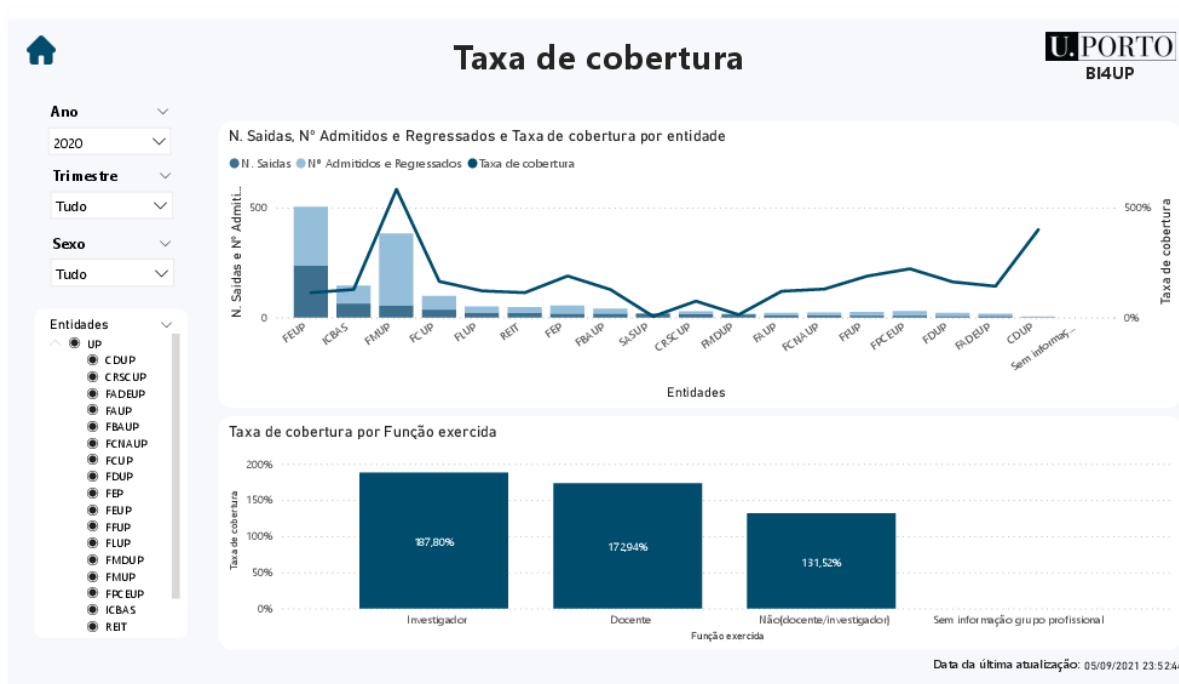
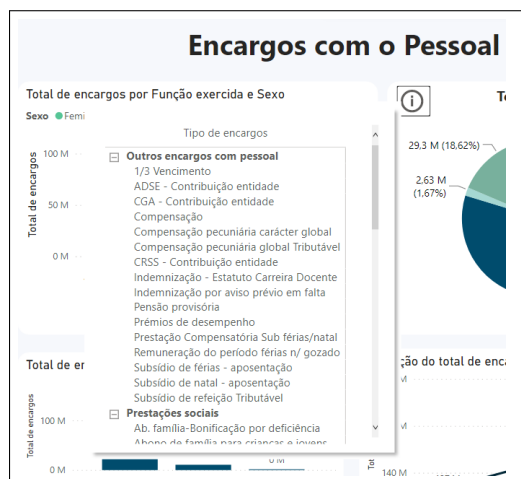
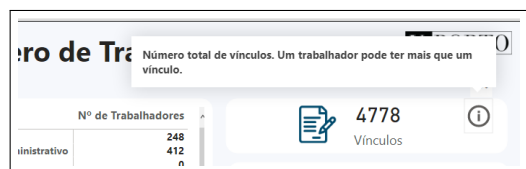


Figura 5.26: Taxa de cobertura

Analisando de uma forma geral todos os *dashboards* desenvolvidos, no lado esquerdo de todos eles estão presentes vários filtros que oferecem a possibilidade ao utilizador de filtrar por sexo, entidade constitutiva ou até mesmo num intervalo de idades. Além disso, também é possível alterar o ano e o trimestre a que os dados correspondem. Em todos os *dashboards* existe um ícone no canto superior esquerdo, que permite navegar até ao submenu correspondente a cada área.

Por fim, em algumas visualizações estão presentes *tooltips* e *pop-ups* que contêm uma explicação adicional sobre a visualização. Foi então colocado um ícone de um *i* informativo em algumas visualizações, que é ativado ao passar o rato por cima no caso das *tooltips*. No caso dos *pop-ups*, ao clicar, é aberta uma janela flutuante (figura 5.27) que contém informação em detalhe sobre o que está representado.

(a) Exemplo de um *pop-up*(b) Exemplo de uma *tooltip*Figura 5.27: Exemplo das duas formas utilizadas para colocar informação adicional no *dashboard*

Os *dashboards* produzidos são constituídos por tabelas e gráficos que permitem explorar e extrair informação de forma autónoma e servem de suporte na tomada de decisões, representando um grande benefício para a organização. Além disso, os utilizadores também têm a possibilidade de criar *dashboards* personalizados com mais, ou, menos informação.

5.7 Validação e testes

Após os dados estarem integrados no *Data Warehouse* e os *dashboards* criados, foram realizados testes ao processo ETL desenvolvido e à integridade dos dados.

Para os testes do processo ETL, foi realizado um carregamento inicial dos últimos cinco anos para o DW, de forma a avaliar o seu comportamento. Este carregamento demorou sensivelmente 1 hora e 30 minutos, carregando cerca de 180 mil factos, não ocorrendo nenhum tipo de problema no processo ETL. Após efetuado este primeiro carregamento, o carregamento é feito a cada trimestre, demorando cerca de 5 minutos.

Relativamente aos testes de integridade da informação obtida, foram efetuados através da comparação da informação obtida nos *dashboards* desenvolvidos, com relatórios já devidamente validados. Foram detetadas algumas diferenças relativas ao número de trabalhadores femininos e masculinos, apesar de o número total de trabalhadores ser o mesmo. Embora as diferenças não sejam muito significativas, foi realizada uma análise detalhada, de forma a identificar e a resolver o problema.

Capítulo 6

Conclusões

Este trabalho teve como principal objetivo o desenvolvimento de um sistema *Business Intelligence* que suportasse a gestão dos recursos humanos da Universidade do Porto, permitindo à organização em causa e, em particular aos utilizadores finais, uma visão global da informação existente. Contribuindo para a obtenção de informação de uma forma rápida e bem estruturada, melhorando o processo de gestão e tomada de decisões.

Os objetivos propostos para a concretização deste projeto começaram pela análise detalhada das necessidades da organização e dos requisitos pretendidos, de forma a identificar a metodologia a seguir no desenvolvimento da solução, bem como a arquitetura a implementar. Foi desenvolvido um repositório de dados consolidado e fiável, que permitiu a integração das fontes de dados, servindo como suporte para todas as tomadas de decisão na área dos recursos humanos. Este armazém de dados é constituído por outros *Data Marts*, além do de recursos humanos pertencentes ao projeto BI4UP. Outro objetivo proposto e o mais complexo, foi o desenvolvimento do processo ETL. Este processo alimenta o armazém de dados, portanto se ocorrer algum problema nesta etapa e os dados não forem corretamente tratados antes de serem carregados, pode-se obter incoerências e erros na informação obtida a partir do armazém de dados. Por fim, o último objetivo foi a construção de vários *dashboards*, que permitem o acesso e a visualização da informação de uma forma fácil e rápida. Nesta fase, foram também realizados testes ao processo ETL desenvolvido, de forma a verificar a existência de erros. Além disso, foi feita uma verificação dos dados obtidos, de modo a validar os dados. Esta verificação foi realizada através da comparação com relatórios anteriormente desenvolvidos pelos utilizadores finais, garantindo a qualidade da informação obtida e validando a solução implementada.

Atualmente, a solução BI implementada na organização já não corresponde às necessidades dos utilizadores, o que levou à realização deste projeto.

Com a implementação desta nova solução BI, a organização conseguirá obter um sistema que permite a extração, tratamento e armazenamento dos dados num único local. Este sistema, permite um rápido acesso à informação e possibilita a organização de obter informação em tempo útil, auxiliando a tomada de decisões.

Por fim, o desenvolvimento deste projeto permitiu-me adquirir várias competências na área de *Business Intelligence*, desde do levantamento dos requisitos, desenho da arquitetura, desenvolvimento do processo ETL e criação dos *dashboards*. Além disso, permitiu também explorar e desenvolver competências na utilização da ferramenta Power BI. Deste modo, desenvolver este projeto foi uma experiência enriquecedora tanto a nível profissional como pessoal.

6.1 Principais dificuldades

No desenvolvimento deste projeto de dissertação foram encontradas algumas dificuldades que provocaram um aumento no tempo de implementação da solução, mas que foram ultrapassadas.

A etapa que implicou um maior gasto de tempo foi o desenvolvimento de todo o processo de ETL, devido ao facto deste ter sido implementado todo de raiz. Por vezes, ocorriam erros no código devido a trocas do nome de variáveis ou valores inesperados e nem sempre era fácil de perceber onde se encontrava o erro, o que provocava um grande gasto de tempo.

Outra dificuldade encontrada, foi no desenvolvimento dos *dashboards*, dado que não existia experiência com a ferramenta Power BI, provocando um gasto de tempo na sua aprendizagem. Além disso, foi necessário perceber a parte técnica na área dos recursos humanos, de forma a que os *dashboards* criados sejam os mais adequados.

6.2 Trabalho Futuro

O trabalho desenvolvido cumpre os objetivos definidos para este projeto de dissertação, no entanto, ainda existe algum trabalho a ser feito, de modo a que esta solução possa ser disponibilizada aos utilizadores, bem como superar algumas limitações atualmente existentes. Assim sendo, destacam-se os seguintes pontos a ter em consideração na continuidade deste projeto:

- Implementar funções de segurança no Power BI para que os utilizadores tenham acesso apenas aos seus dados;
- Definição e implementação de novos indicadores na área de recursos humanos;
- Apesar de já terem sido realizados testes para verificar a integridade dos dados obtidos, são necessários mais testes ao sistema e testes de usabilidade para entrada em produção;
- Criação de novas tabelas de facto e dimensões, para integrar novos processos de negócio, como, por exemplo, a área financeira;

Referências

- [1] Efraim Turban. *Business intelligence : a managerial approach*. Prentice Hall, Boston, 2011. ISBN: 978-0136100669.
- [2] Richard Miller Devens. *texts Cyclopædia of commercial and business anecdotes*; New York, London, D. Appleton and company, 1865.
- [3] H. P. Luhn. A business intelligence system. *IBM Journal of Research and Development*, 2: 314–319, 1958.
- [4] Power D. J. [A brief history of decision support systems](#). *DSSResources.COM*, 2007.
- [5] Barb Wixom and Hugh Watson. [The bi-based organization](#). *IJBIR*, 1:13–28, 01 2010. doi:10.4018/jbir.2010071702.
- [6] Hugh Watson and Barb Wixom. [The current state of business intelligence](#). *Computer*, 40:96 – 99, 10 2007. doi:10.1109/MC.2007.331.
- [7] Trícia Karla Moraes, Silvana Quintanilha, and Heitor Quintella. Vantagens e desvantagens do "business intelligence", como forma de projetar a inteligência nos negócios de pequenas e médias empresas. *Revista Carioca De Produção*, 09 2009.
- [8] Nur Alia Hamizah Mohamad Rodzi, Mohd Shahizan Othman, and Lizawati Mi Yusuf. [Significance of data integration and etl in business intelligence framework for higher education](#). In *2015 International Conference on Science in Information Technology (ICSITech)*, pages 181–186, 2015. doi:10.1109/ICSITech.2015.7407800.
- [9] K. Sun and Y. Lan. [Setl: A scalable and high performance etl system](#). In *2012 3rd International Conference on System Science, Engineering Design and Manufacturing Informatization*, volume 1, pages 6–9, 2012. doi:10.1109/ICSSEM.2012.6340727.
- [10] Margy Ross Ralph Kimball. *The Data Warehouse Toolkit*. Wiley, 2002.
- [11] Jarrett Goldfedder. [Building a Data Integration Team: Skills, Requirements, and Solutions for Designing Integrations](#). Apress, 01 2020. ISBN: 978-1-4842-5652-7. doi:10.1007/978-1-4842-5653-4.
- [12] Ehtisham Zaidi, Sharat Menon, Robert Thanaraj, Eric Thoo, and Nina Showell. [Gartner magic quadrant for data integration tools](#). *Gartner*, 2021.

-
- [13] Talend. [Talend - a cloud data integration leader](#). *Talend Real-Time Open Source Data Integration Software*, 2021.
- [14] Pentaho. [Pentaho data integration](#). *Pentaho Documentation*, Nov 2016.
- [15] Informatica. [Informatica](#). *Informatica.com*, 2021.
- [16] Grace Gu. [Sql server integration services - sql server integration services \(ssis\)](#), Jul 2018.
- [17] Carlos Gameiro. [Implementation of business intelligence tools using open source approach](#). In *Proceedings of the 2011 Workshop on Open Source and Design of Communication*, OSDOC '11, page 27–32, New York, NY, USA, 2011. Association for Computing Machinery. ISBN: 9781450308731. doi:10.1145/2016716.2016723.
- [18] Nuno Leite, Isabel Pedrosa, and Jorge Bernardino. [Open source business intelligence on a sme: A case study using pentaho](#). In *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–7, 2019. doi:10.23919/CISTI.2019.8760740.
- [19] Mr Mali and Mr Sachinbojewar. [A survey of etl tools](#). *International Journal of Computer Techniques*, 2, 2015.
- [20] Python. [Python](#). *Python.org*, Ago 2021.
- [21] R. [R](#). *R-project.org*, 2021.
- [22] José Oliveira e Miguel Baptista Carlos Sezões. *Business Intelligence*. SPI – Sociedade Portuguesa de Inovação, 2006.
- [23] N. Jukic. [Modeling strategies and alternatives for data warehousing projects](#), 2006.
- [24] William H. Inmon. *Building the Data Warehouse*. John Wiley; Sons, Inc., USA, 2002. ISBN: 0-471-08130-2.
- [25] Yasmine Nagm Aldeen. [Data warehouse- dimensional model vs normalized model](#), Jul 2020.
- [26] L. Yessad and A. Labiod. [Comparative study of data warehouses modeling approaches: Inmon, kimball and data vault](#). In *2016 International Conference on System Reliability and Science (ICSRS)*, pages 95–99, 2016. doi:10.1109/ICSRS.2016.7815845.
- [27] Myers Peter. [Compreender o que é um esquema de estrela e qual a importância para o power bi - power bi](#), Set 2019.
- [28] Mark Levene and George Loizou. [Why is the snowflake schema a good data warehouse design?](#) *Information Systems*, 28(3):225–240, 2003. ISSN: 0306-4379. doi:https://doi.org/10.1016/S0306-4379(02)00021-2.
- [29] Harco Leslie Hendrie Spits Warners and Richard Randriatoamanana. [Datawarehouser: A data warehouse artist who have ability to understand data warehouse schema pictures](#). In *2016 IEEE Region 10 Conference (TENCON)*, pages 2205–2208, 2016. doi:10.1109/TENCON.2016.7848419.

- [30] Souriau poke. [Choosing a tabular or multidimensional modeling experience in sql server 2012 analysis services microsoft business intelligence technical article authors hitachi consulting](#), 2012.
- [31] OLAP. [What is the definition of olap?](#) *OLAP.com*, 2020.
- [32] Anindya Datta and Helen Thomas. [The cube data model: a conceptual model and algebra for on-line analytical processing in data warehouses](#). *Decision Support Systems*, 27(3): 289–301, Dec 1999. doi:10.1016/s0167-9236(99)00052-4.
- [33] IBM Cloud Education. [What is olap?](#), Jun 2020.
- [34] S. Scheps. *Business Intelligence For Dummies*. Wiley, 2011. ISBN: 9781118051412.
- [35] Minewiskan. [Tabular modeling overview - analysis services](#). *Microsoft.com*, Abr 2020.
- [36] Minewiskan. [Comparing analysis services tabular and multidimensional models](#). *Microsoft.com*, Dez 2020.
- [37] Arun Ulag. [Microsoft named a leader in the 2021 gartner magic quadrant for analytics and bi platforms](#). *Microsoft.com*, 2021.
- [38] Michele Hart. [O que é power bi? - power bi](#). *Microsoft.com*, Set 2020.
- [39] Tableau. [Tableau: business intelligence and analytics software](#). *Tableau*, 2021.
- [40] Qlikview. [Qlikview – powerful interactive analytics and dashboards](#). *Qlik*, 2014.
- [41] Selecthub. [Business intelligence software tools comparison](#). *Selecthub.com*, 2021.
- [42] Otmane Azeroual and Horst Theel. [The effects of using business intelligence systems on an excellence management and decision-making process by start-up companies: A case study](#). *INTERNATIONAL JOURNAL OF MANAGEMENT SCIENCE AND BUSINESS ADMINISTRATION*, 4(3):30–40, 2018. doi:10.18775/ijmsba.1849-5664-5419.2014.43.1004.
- [43] Dresner Advisory Services. [Dresner advisory services - products - 2020 wisdom of crowds business intelligence market study report](#), 2020.
- [44] Maria José Sousa and Ivo Dias. [Business intelligence for human capital management](#). *International Journal of Business Intelligence Research*, 11(1):38–49, Jan 2020. doi:10.4018/ijbir.2020010103.
- [45] Carlos Duarte Valente de Barros. [Desenvolvimento e implementação de projetos de business intelligence nas áreas financeira e de logística](#). Master’s thesis, Faculdade de Engenharia da Universidade do Porto, 2014.
- [46] Sérgio Manuel Soares Esteves. [Solução analítica de suporte à decisão no âmbito da gestão académica no ensino superior](#). Master’s thesis, Faculdade de Engenharia da Universidade do Porto, 2016.

-
- [47] Pedro Miguel Martins Assis. [Desenvolvimento de uma solução de business intelligence](#). Master's thesis, Faculdade de Ciências da Universidade de Lisboa, 2017.
 - [48] Oracle. [Oracle business intelligence enterprise edition](#). *Oracle.com*, 2020.
 - [49] Jake VanderPlas. *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly Media, Inc., 1st edition, 2016. ISBN: 1491912057.
 - [50] Marco Russo and Alberto Ferrari. *DAX Patterns: Second Edition*. Sqlbi, 2020. ISBN: 9781735365206.
 - [51] Oracle. [cx_oracle - python interface for oracle database](#). *Github.io*, 2021.
 - [52] Pandas. [pandas - python data analysis library](#), 2021.
 - [53] Philip Seamark and Thomas Martens. *Pro DAX with Power BI*. Apress, 2019. doi:10.1007/978-1-4842-4897-3.
 - [54] Mitchell Pearson, Brian Knight, Devin Knight, and Manuel Quintana. *Connecting to Data*, chapter 17, pages 204–207. Apress, Berkeley, CA, 2020. ISBN: 978-1-4842-6008-1. doi:10.1007/978-1-4842-6008-1_17.