

# **MESTRADO**

PEDRO CÔRTE-REAL MACHADO VENTURA

## **TRABALHO FINAL DE MESTRADO**

TRABALHO DE PROJETO

DASHBOARD DE TRÁFEGO NOS *WEBSITES* DE EMPRESAS EUROPEIAS

PEDRO CÔRTE-REAL MACHADO VENTURA

OUTUBRO DE 2019

# **MESTRADO**

PEDRO CÔRTE-REAL MACHADO VENTURA

## **TRABALHO FINAL DE MESTRADO**

TRABALHO DE PROJETO

DASHBOARD DE TRÁFEGO NOS *WEBSITES* DE EMPRESAS EUROPEIAS

PEDRO CÔRTE-REAL MACHADO VENTURA

**ORIENTAÇÃO:**

PROFESSOR DOUTOR JESUALDO FERNANDES

OUTUBRO DE 2019

*“You can’t manage what you don’t measure.”*

*Peter Drucker*



## GLOSSÁRIO

BI – Business Intelligence

CRM – Customer Relationship Management

CSV – Comma Separated Value

DW – Data Warehouse

ERP – Enterprise Resource Planning

ETL – Extract Transform Load

HTML - Hypertext Markup Language

IOT – Internet Of Things

JSON - JavaScript Object Notation

OLAP - Online Analytical Processing

PHP - Hypertext Preprocessor

SGI – Sistemas de Gestão de Informação

SSBI – Self Service Business Intelligence

SSD – Sistemas de Suporte à Decisão

TI – Tecnologias de Informação

URL – Uniform Resource Locator

## ABSTRACT

Business intelligence services offer many ways to process and analyze the richness of a data set in business today. In this project work, Microsoft Power BI and a *web* scraping tool were used to develop a business intelligence solution with a data set within the scope of *website* traffic.

To provide theoretical context on how to develop a business intelligence solution through a semi-structured data set, key principles of multidimensional modeling are introduced. The results of the development process are shown in the description and discussion of the results, examples of *dashboards* created for the solution. As this is a daily tool, technical performance and reading supporting documentation is important for the positive adoption of this tool by users. Functionality and performance aspects were analyzed and optimized based on literature review research, and the tool was put to work correctly.

**KEYWORDS:** Business Intelligence; *Web* scraping; *Dashboards*; Data Modelling.

## RESUMO

Os serviços de Business Intelligence oferecem várias formas de processar e analisar a riqueza de um conjunto de dados nas empresas nos dias de hoje. Neste trabalho de projeto, foi utilizado o Microsoft Power BI para desenvolver uma solução de business intelligence com um conjunto de dados no âmbito de tráfego de *websites*.

Para fornecer contexto teórico em como desenvolver uma solução de business intelligence através de um conjunto de dados semiestruturados, os princípios chave de modelação multidimensional são introduzidos. Os resultados do processo de desenvolvimento são evidenciados na parte da descrição e discussão dos resultados, de exemplos de *dashboards* criados para a solução. Sendo esta uma ferramenta com utilização diária, o desempenho técnico e a leitura de documentação de apoio é importante para uma adopção positiva desta ferramenta por parte dos utilizadores. Os aspectos de funcionalidade e desempenho foram analisados e otimizados com base na pesquisa da revisão de literatura, e a ferramenta foi colocada a funcionar corretamente.

**PALAVRAS-CHAVE:** Business Intelligence; *Web scraping*; *Dashboards*; Modelação de dados.

## TABELA DE CONTEÚDOS

Glossário.....	ii
Abstract.....	iii
Resumo .....	iv
Índice De Figuras.....	vii
Índice De Tabelas .....	ix
Agradecimentos .....	x
1. Capítulo 1 - Introdução .....	11
1.1. ENQUADRAMENTO E MOTIVAÇÃO .....	11
1.2. OBJETIVOS E ESTRUTURA .....	13
2. Capítulo 2 - Revisão de Literatura .....	13
2.1. BUSINESS INTELLIGENCE (BI) .....	13
2.1.1. COMPONENTES DE UM SISTEMA DE BI.....	16
2.2. MODELAÇÃO MULTIDIMENSIONAL.....	17
2.2.1. STAR SCHEMA .....	18
2.2.2. SNOWFLAKE SCHEMA .....	19
2.2.3. DIMENSÕES E FACTOS.....	19
2.2.3.1. DIMENSÕES.....	19
2.2.3.2. FACTOS.....	20
2.3. SSBI – SELF SERVICE BUSINESS INTELLIGENCE.....	20
2.3.1. MICROSOFT POWER BI .....	21
2.3.2. <i>DASHBOARDS</i> .....	23
2.3.3. RELATÓRIOS .....	23
2.4. <i>WEB SCRAPING</i> .....	24
2.4.1. O QUE É <i>WEB SCRAPING</i> .....	24



2.4.2.	COMPONENTES.....	25
3.	Capítulo 3 – Descrição do Trabalho de Projeto .....	25
3.1.	EXTRAÇÃO DE DADOS COM O DATA MINER.....	25
3.1.1.	FONTE DOS DADOS .....	27
3.1.2.	MÉTODO DE PESQUISA .....	28
3.1.2.1.	PESQUISA NO CRUNCHBASE.....	28
3.1.3.	EXPORTAÇÃO DE DADOS COM O DATA MINER .....	29
3.2.	FRAMEWORK DO PROJETO .....	31
3.3.	ESQUEMA EM ESTRELA NO POWER BI .....	32
3.3.1.	PREPARAÇÃO DOS DADOS.....	32
3.3.2.	CRIAR TABELAS DE FACTOS.....	33
3.3.3.	CRIAR TABELAS DE DIMENSÕES.....	34
3.3.4.	LIGAR DIMENSÕES E FACTOS NO POWER BI .....	36
3.4.	CRIAÇÃO DOS RELATÓRIOS E DASHBOARD .....	39
3.	Resultados.....	43
4.	Dificuldades e Limitações .....	44
5.	Conclusões.....	45
6.	Trabalhos Futuros .....	46
	Referências .....	48
	Anexo i .....	52
	Anexo ii .....	58

## ÍNDICE DE FIGURAS

Figura I - Relação do BI com outros sistemas de informação (Adaptado de Negash (2014)) .....	15
Figura II - Sede das empresas escolhidas .....	28
Figura III - Área das empresas escolhidas .....	29
Figura V - Seleção de colunas com o <i>Smart find</i> .....	30
Figura IV – Pré visualização de selecção das linhas com <i>Smart find</i> .....	30
Figura VI - Elemento Seletor de Navegação .....	31
Figura VII - Framework de elaboração do Projeto .....	31
Figura VIII – Ficheiro Excel de tráfego de websites de empresas Europeias na área da Indústria 4.0.....	32
Figura IX - Ambiente Power Query no Power BI.....	33
Figura X - Duplicar Tabela de Factos .....	34
Figura XI - Dimensão Empresa no menu Queries.....	35
Figura XII - Combinar Queries em ambiente Power BI.....	37
Figura XIII - Esquema em estrela do conjunto de dados extraídos.....	38
Figura XIV - Relatório de Visitas Mensais .....	40
Figura XV- Dashboard de Tráfego De Websites De Empresas Europeias .....	42
Figura XVI - Transformar Dados .....	52
Figura XVII - Obter Dados ou Get Data .....	52
Figura XVIII - Mudar tipo de dados ou Change Data Type para "Data" .....	53
Figura XIX - Mudar tipo de dados ou Change Data Type para "Número Inteiro" ..	53
Figura XX - Escolher colunas para a tabela de dimensões.....	54
Figura XXI - Coluna Index depois de adicionada .....	54
Figura XXII - Remover duplicados ou Remove Duplicates.....	55

Figura XXIII - Criar Index .....	55
Figura XXIV - Visualizações com o Power BI .....	55
Figura XXV – Gerir as relações no Power BI .....	56
Figura XXVI - Partilhar o dashboard .....	56
Figura XXVII - Partilhar o dashboard .....	56
Figura XXVIII- Relatório de Bounce Rate.....	58
Figura XXIX - Relatório Visit Duration (min).....	59
Figura XXX - Relatório de Pages Views / Visit.....	59

## ÍNDICE DE TABELAS

Tabela I - Tipos de seletores (Adaptado de <a href="https://data-miner.io/quick-guides/persona-business">https://data-miner.io/quick-guides/persona-business</a> ).....	27
---	----

## AGRADECIMENTOS

Em primeiro lugar gostaria de agradecer ao meu orientador Professor Doutor Jesualdo Fernandes, pela sua dedicação, entrega, profissionalismo e motivação transmitida ao longo deste percurso. Todo o empenho e os estímulos foram sem dúvida importantes para o desenvolvimento deste projeto.

Aos meus grandes amigos e parceiros desta jornada de fim de curso, João Sacramento e Filipe Oliveira. Um obrigado pelo incansável apoio, pela motivação, paciência e, sobretudo, dedicação nas horas mais difíceis. Um obrigado por todos os momentos de alegria, solidariedade e entreaajuda, e o apoio em todos os momentos e dificuldades durante este projeto.

A todos os meus familiares, sem exceção, que tiveram a sua cota parte de responsabilidade neste longo percurso. Deixo o meu último agradecimento, aos meus Pais e irmãs. Sem o apoio deles não teria tido motivação nem forças para acabar este projeto.

## 1. CAPÍTULO 1 - INTRODUÇÃO

Nesta primeira fase do documento é feito um enquadramento dos temas abordados no âmbito deste projeto. A motivação do estudo, a finalidade bem como os objetivos propostos à realização da mesma também são elaborados.

Na fase seguinte, é descrito o trabalho desenvolvido, sustentando o trabalho de investigação desenvolvido, bem como a descrição da estrutura do documento.

### 1.1. ENQUADRAMENTO E MOTIVAÇÃO

Nos dias de hoje a tecnologia e a informação estão cada vez mais presentes na vida das pessoas e das organizações o que, por sua vez, se pode traduzir em mais e em melhor conhecimento sobre o mundo que nos rodeia, quer a nível pessoal quer a nível organizacional (Shollo, 2012).

Com a quarta revolução indústria, ou a Indústria 4.0, o aparecimento de novas tecnologias como IoT (Internet of Things) e a sua utilização em paralelo com sistemas de Enterprise Resource Planning (ERP) levou a um rápido aumento dos dados armazenadas, culminando no termo *big data* (Khan et al., 2017). Quem toma a decisão procura uma forma de ver para além de todo o tipo de dados e figuras, para criar uma cadeia de informação eficaz e qualificada para utilizar na melhoria do processo de decisão (Loya & Carden, 2018).

Atualmente, ao nível da análise de dados que estas gerem ou que possuem, deparam-se com alguns desafios, como a produção de informação com melhor conteúdo que seja capaz de apoiar os processos organizacionais (Sivarajah et al., 2017). Por isso, qualquer organização ou empresa que se queira diferenciar da competição e alcançar um crescimento contínuo precisa de lidar com grandes volumes de informação e precisa de ter mão de obra qualificada (Castro, 2016). Mas segundo Ferrari & Russo (2016), esta crescente quantidade e complexidade dos dados não trazem valor por si só, visto que as pessoas têm uma capacidade de processamento de informações muito limitada, e por isso é preciso criar ferramentas que auxiliem a disponibilizar as informações relevantes para quem necessita delas para analisar e contruir os seus próprios painéis e ferramentas de

análise num ambiente de *business intelligence* sem necessidade de assistência especializada.

Desta forma, surge o conceito de *Business Intelligence*, sendo cada vez mais aceite e disseminado pela comunidade empresarial, que oferece uma opção viável e eficaz para ajudar a empresa a seguir os seus números de vendas, e utilizá-los em seu proveito para conseguir atingir um crescimento sustentável, oferecer serviços ou produtos personalizados e ficar à frente da concorrência (Malladi, 2013). Diante de um cenário de aumento da quantidade de dados com que nos encontramos atualmente, a tomada de decisão é otimizada uma vez que a informação apresentada por meio de *data warehouse* (DW) é organizada, *user-friendly*, ou seja, é um repositório central de dados estruturados e rápidos no sistema de pesquisa otimizado para análises (Adeoye, Raufu, & Omodara, 2011).

A implementação de um serviço de Business Intelligence (BI) não traz só vantagens apenas em questões de armazenamento e pesquisa rápida de grandes conjuntos de dados, mas também na procura inteligente de informações que permite a flexibilização de consultas e análises de informações consoante a necessidade do utilizador (Castro, 2016). Posto isto, é possível maximizar a utilidade de informação recolhida e o suporte aos processos de tomada de decisão nos negócios, através da adoção de um serviço de BI obtendo uma visão integrada do negócio que disponibiliza informação relevante para o decisor de forma rápida (Malladi, 2013).

Um destes serviços é o Microsoft Power BI, que até hoje é um dos serviços de BI mais presentes no mercado (Ferrari & Russo, 2016). A solução de business intelligence desenvolvida durante este projeto tese foi feita utilizando o Microsoft Power BI Desktop. Esta plataforma foi escolhida como ferramenta de desenvolvimento para este projeto, visto que é uma ferramenta de bastante fácil utilização e como uma interface de utilizador simples. Para além disso, foi uma ferramenta recomendada pelo orientador Prof. Dr. Jesualdo Fernandes e por terceiros que tinham experiência na área em questão.

## 1.2. OBJETIVOS E ESTRUTURA

Este projeto de final de mestrado tem como objetivo evidenciar a utilidade de ferramentas de extração de dados online e implementação e análise de dados com ferramentas de BI de forma a perceber como ajudam a ter *insights* dos dados e como ajudam na tomada de decisão baseada em dados .

Durante a fase empírica do projeto, um conjunto de dados foi extraído da Internet e foi desenvolvido um modelo para melhor se perceber as funções do Power BI.

Como resultado, foi criada de uma base de dados em Excel com dados de empresas retirados de um *website* (3.1.1) com recurso a uma ferramenta de *web scraping*, dados esses capazes de ser importados para o Microsoft Power BI. Posteriormente foi feita a criação de um *dashboard* e um relatório do Power BI Desktop com informações a retirar dos dados.

Neste projeto procura-se responder às seguintes questões:

- O que é o BI e *self-service* BI?
- O que é *web scraping*?
- Como utilizar o Power BI?
  - Como criar um *dashboard*?
  - Que tipo de informações podemos retirar dos dados?

Existem dois propósitos nesta pesquisa, sendo a primeira perceber o que é business intelligence e a segunda perceber como criar análises, dashboard e relatórios utilizando o Power BI de uma perspectiva de *end user*.

## 1. CAPÍTULO 2 - REVISÃO DE LITERATURA

### 2.1. BUSINESS INTELLIGENCE (BI)

À medida que a quantidade de informações disponíveis aumenta, com mudanças sucessivas e complexas no ambiente de negócios, as empresas são forçadas a tomar decisões mais rápidas devido às mudanças nas condições. Tudo o que está relacionado à



tomada de decisão precisa de quantidades consideráveis de dados que são a matéria prima da informação, de informações, que são o resultado do processamento de dados, e conhecimentos relevantes que são as informações trabalhadas. Visto que estes processos têm de ser eficientes e eficazes, precisam de suporte de computadores (Turban et al., 2008).

Atualmente, as empresas precisam de um método de tomada de decisão eficaz, e para preencher essa lacuna, começaram a ser criados sistemas para toda a empresa que contém visualizações adicionais e capacidades de medidas de alerta e de desempenho. Este tipo de produtos, constituem o termo de Business Intelligence (BI) (Loya & Carden, 2018).

O termo de *Business Intelligence* (BI) foi proposto em 1996 por Howard Dresner, que pertencia ao Gartner Group. Segundo Dresner, BI é um "termo abrangente que inclui as aplicações, as infraestruturas, ferramentas e as melhores práticas que permitem o acesso e análise de informação para melhorar e otimizar decisões e o desempenho" (Gartner, 2019).

Existem outras definições de BI, como a da Microsoft que define BI como um meio de simplificação da descoberta de informação e análises, que fornece capacidades de acesso, de compreensão, de análise e de colaboração de informação para ajudar os tomadores de decisões, em todas áreas e níveis da organização em qualquer altura e em qualquer lugar (Ferrari & Russo, 2016).

Segundo Santos & Ramos (2009), os sistemas de BI fizeram uma série de contribuições significativas para as organizações. Eles contribuem para aumentar a inteligência coletiva, a capacidade de aprendizagem e a criatividade das organizações, promovendo o desenvolvimento de novos pensamentos, produtos ou serviços que permitem que a organização se adapte dinamicamente.

Business Intelligence é a forma como as organizações utilizam tecnologias modernas para recolher, gerir e analisar dados de negócio e informação. Os objetivos passam por acumular e recolher conhecimento de negócio e intuições, para tomar melhores decisões, tornar as operações mais eficientes e eficazes, para melhorar os processos de negócio, para ganhar vantagem competitiva no mercado (Loshin, 2012).

Por outro lado, temos Negash (2015), onde refere que BI implica tomada de decisões on-line ou decisões instantâneas. Segundo Negash, os sistemas de BI têm como foco a diminuição da janela temporal entre a recolha dos dados e a disponibilização das informações. A expansão da internet, e dos DW, a evolução de técnicas de limpeza de dados, e o aumento das capacidades de hardware e software, todas juntas, permitem a criação de sistemas de BI. A Figura I abaixo, representa alguns dos sistemas de informação que são utilizados pelo BI:

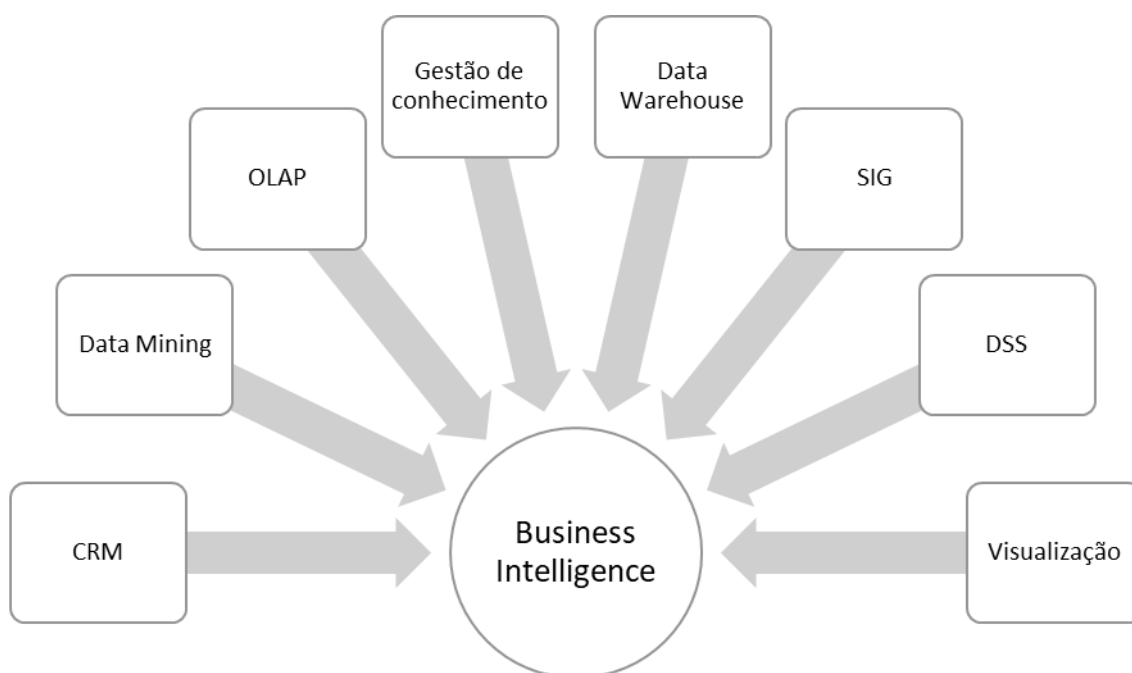


Figura I - Relação do BI com outros sistemas de informação (Adaptado de Negash (2014))

Pelo que se pode observar na Figura I, BI é um conceito versátil e amplo suportados por um conjunto de sistemas e tecnologias, de entre os quais se destacam data mining, CRM, OLAP, Gestão de conhecimento, SIG, Data Warehouse, Visualizações, SSD, para cumprir com as exigências e necessidades de negócio tornando-o mais eficiente e útil. Depois de implementados, estes sistemas têm como objetivo criar previsões com base em dados históricos, o desempenho passado e atual da empresa como possibilidade de criação de cenários que mostrem o impacto da alteração de algumas variáveis. Permite o acesso ad hoc aos dados para responder a questões não previstas, e não específicas o que permite obter um conhecimento estratégico da organização (Negash, 2015).

Numa organização, o BI pode ser utilizado em várias áreas como:

1. Entender o processo de operação e a situação, ou seja, pode ajudar uma organização a perceber as suas condições de funcionamento bem como as suas forças motoras, podendo até revelar tendências futuras (Santos & Ramos, 2009);
2. Medir o desempenho, ou seja, pode ser utilizado para medir e monitorizar desempenhos de funcionários e empresa, interligando os dois (Bordeleau, et al., 2018);
3. Melhorar relações, ou seja, BI pode ser integrado com outros sistemas de uma organização. Por exemplo, integrando uma ferramenta de BI com CRM (Customer Relationship Management) ajudar a tornar os dados do CRM mais eficientes, ao permitir a partilha de dados, melhorar a comunicação entre operações, e melhorar relações com clientes e fornecedores, etc. (Baars & Kemper, 2008)<sup>1</sup>;
4. Criação de oportunidades, ou seja, as empresas fazem dinheiro através de informação que é útil, e o BI organiza e gere os dados que pode levar a organização a encontrar informações valiosas de forma eficiente (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Tendo em conta as definições supramencionadas, é possível afirmar que uma das principais vantagens do BI é a capacidade de criação de mecanismos de análise com precisão de informação histórica, como instrumento de apoio para atingir o sucesso e objetivos empresariais.

### 2.1.1. COMPONENTES DE UM SISTEMA DE BI

Segundo Turban et al. (2008), um sistema de BI é formado por quatro componentes:

1. *Data Warehouse*, que é considerado um conjunto de dados históricos, preparados e prontos a prestar suporte à tomada de decisão, não sendo mais que a união de todos os *data marts* (Kimball et al., 2010);

---

<sup>1</sup> Esta pesquisa estuda a integração entre CRM e BI de uma organização proeminente orientada para o cliente chamada *Skånetafiken*. Nesta pesquisa, uma estrutura funcional de integração de CRM e BI é derivada com sucesso. Eles concluem que o uso de CRM e BI ajuda na integração de pessoas, processos, tecnologia e dados, levando a melhorias no gerenciamento de relacionamento com clientes (Shinde & Sunjita, 2018).

2. *Data Marts*; são feitos para entregar objetivos de negócio aos departamentos numa organização. Podem ser descritos com um subconjunto de um DW, sendo que este é um algerado de todos os *data marts*, cada um representado um processo de nesgocio na organização, por meio de um *star schema*;
3. *Business Analytics*, compreende um conjunto de ferramentas, como técnicas de análise avançada, processamento analítico, queries, relatórios, data mining e *web mining*, que ajudam o utilizador a transformar dados do DW em informações concretas com valor (McAfee et al., 2012);
4. *Business Performance Management*, que monitoriza e compara o desempenho da organização com os objetivos definidos e metas a ser atingidas, como por exemplo os *balanced scorecards* (Mircea, 2012);
5. Interface do utilizador, que diz respeito à forma como a informação gerada é disponibilizada.

## 2.2. MODELAÇÃO MULTIDIMENSIONAL

Kimball (2010), defende que a modelação multidimensional é a técnica mais adequada para ambientes de DW, visto que foi desenvolvida para a construção dos mesmos colmatando limitações de métodos tradicionais que utilizam a normalização dos dados para garantir a sua consistência, e redução do espaço de armazenamento e de redundâncias.

Utiliza-se esta técnica quando queremos produzir estruturas de dados de fácil compreensão para o utilizador e que tenha capacidade de otimização de desempenho de processamento de queries no sistema, em vez de atualizações (Moody & Kortink, 2003), o que é o caso deste projeto, pretendendo mostrar como transformar e utilizar um conjunto de dados numa ferramenta de BI.

Esta técnica, tem como objetivo a remoção de dados redundantes, fazendo uma pré seleção de dados e dividindo-os em entidades distintas, em que cada uma delas é representada por uma tabela (Höpken, Keil, & Lexhagen, 2015).

As bases de dados são constituídas por várias tabelas, e essas tabelas podem ser de factos ou de dimensões e relacionam-se através de um modelo de dados e apresentam-se em forma de esquema (Kazi, Kazi, & Radulovic, 2012). Os esquemas mais utilizados

na modelação dimensional são o esquema em estrela (star schema) e o floco de neve (snowflake).

### 2.2.1. STAR SCHEMA

Para se conseguir analisar algo numa solução de BI em dados extraídos de diferentes fontes e que estão armazenados em várias tabelas que de algum modo estão relacionadas entre si, é necessário um modelo de dados. Um modelo de dados de dados traz benefícios de facilidade de uso e também de desempenho (Ari & Tolvanen, 2019). Este esquema compreende uma única tabela de factos que está rodeada por várias tabelas de dimensão (Moody & Kortink, 2003)

Depois de extraídos os dados, são carregados em tabelas. Este conjunto desconectado de tabelas pode ser formatado para um modelo de dados ao criarmos relações entre eles. Segundo Ferrari & Russo (2016), um modelo de dados é um conjunto de tabelas que estão interligadas entre si com relações através de chaves identificadoras: colunas partilhadas entre cada conjunto de tabelas.

À medida que vão sendo adicionadas tabelas ao modelo de dados ele torna-se cada vez mais complexo, e é necessário realizar uma técnica de modelagem de dados que é *star schema*. Ao desenhar um modelo de dados de acordo com o *star schema*, existem duas categorias onde devem ser enquadradas as tabelas no modelo de dados: as tabelas de dimensões e as tabelas de factos (Ari & Tolvanen, 2019).

As tabelas dimensionais contêm dados com atributos, são informativas, como fornecedores ou artigos. As tabelas de factos contêm métricas sobre eventos relacionados a dimensões, como as ordens de compra (Höpken et al., 2015).

O *star schema* tem este nome devido à forma que o modelo de dados forma quando a tabela de factos é colocada no meio e as suas respetivas dimensões à sua volta. Ao organizar os dados desta forma é possível prevenir ambiguidade nas relações do modelo ao mesmo tempo que se minimiza as repetições desnecessárias do mesmo valor numa coluna ou diferentes linhas, o que torna o modelo eficiente (Ari & Tolvanen, 2019).

### 2.2.2. SNOWFLAKE SCHEMA

Um modelo que também pode ser adotado é o esquema em floco de neve ou *snowflake schema* que surge como uma variação do esquema referido no ponto 2.2.1.. Relativamente à estrutura este esquema não difere muito do esquema em estrela, no entanto as tabelas de dimensão podem ser organizadas por hierarquias e estão todas normalizadas (Elmasri & Navathe, 1989).

Os esquemas em floco de neve são esquemas em estrela com as dimensões normalizadas, e têm este nome devido a aparência que assumem quando construídos, mas ao invés de ter uma estrutura regular como o *star schema*. Neste caso as ramificações do floco de neve podem crescer para qualquer direção arbitrariamente (Moody & Kortink, 2003). Isto é, ao invés de ter uma estrutura regular como o *star schema*, as várias tabelas de dimensões do esquema podem ter várias tabelas associadas a si, atingindo um nível de estrutura muito mais complexo. A maior vantagem deste tipo de esquemas, é que em vez de parecer um conjunto de dados não estruturados mostra explicitamente a estrutura de cada dimensão (Turban et al., 2008).

### 2.2.3. DIMENSÕES E FACTOS

#### 2.2.3.1. DIMENSÕES

Quando falamos de bases de dados multidimensionais, dimensões são um conceito importante de reter. As dimensões são o que permite a análise de diferentes perspetivas da informação e são utilizadas para *slice*, ou triagem, de dados e agrupá-los de acordo com o nível de detalhe desejado (Jensen, Pedersen, & Thomsen, 2010).

As dimensões são constituídas por tabelas e têm capacidade para integrar vários atributos ou colunas que, habitualmente, incluem descrições que vão permitir a contextualização e enquadramento das métricas utilizadas. Cada tabela das dimensões vai ter uma chave primária única que é idêntica a uma das partes da chave composta da tabela de factos que a ela está associada, ou seja, a chave estrangeira da tabela de factos (Santos & Ramos, 2009).

### 2.2.3.2. FACTOS

As tabelas de factos compõem um dos principais componentes do modelo multidimensional, ao permitir armazenar os temas objetos de análise. Os factos são objetos ou instâncias que representam determinado assunto ou tema da análise pretendida, isto é, é um assunto relevante para a empresa que deve ser analisado para se obter um melhor entendimento desse assunto (Jensen et al., 2010; Santos & Ramos, 2009).

As tabelas de dimensões e de factos relacionam-se uma vez que, na maioria dos modelos multidimensionais, os factos encontram-se definidos pelas combinações das dimensões. A tabela de factos vai conter uma chave primária, ou *surrogate key*<sup>2</sup>, que por sua vez, é composta por duas ou mais chaves estrangeiras ou *foreign keys*, e é sempre definida por uma relação de um para muitos com as tabelas de dimensão (Jensen et al., 2010; Kimball et al., 2010).

### 2.3. SSBI – SELF SERVICE BUSINESS INTELLIGENCE

Em 2009, a Microsoft introduz um produto, o Microsoft Power BI, alinhado com a ideia de Self Service Business Intelligence (SSBI). O objetivo dos SSBI é fornecer aos utilizadores a capacidade de criar e recolher relatórios customizados dentro de uma arquitetura aprovada e suportada pela organização (Castro, 2016).

SSBI também pode ser definido como sendo uma abordagem analítica que permite aos utilizadores terem acesso a dados da empresa sem estarem a trabalhar com um perfil corporativo (Wigmore & Rouse, 2014). Para além de se tornarem mais independentes na organização, vão permitir que as equipas de IT e de BI estejam mais isentas da responsabilidade de criar a maioria dos relatórios, permitindo que estas possam ajudar onde é mais necessário, de forma a atingir os objetivos (Castro, 2016).

Visto que a atenção do SSBI está virada para os utilizadores, e tendo em conta que podem ou não ter conhecimentos técnicos, é preciso que a interface do utilizador seja simples, intuitiva, com navegação e *dashboards* simples.

De acordo com (Imhoff & White, 2011), SSBI têm quatro objetivos principais:

---

<sup>2</sup> Neste projeto as chaves primárias são *surrogate keys*. Uma *surrogate key* é qualquer coluna ou colunas que podem ser declaradas como chave primária de uma tabela (Han & Kamber, 2006)

1. Simplicidade na interpretação e melhoria dos resultados de BI, ou seja, o SSBI deve ter um ambiente que seja simples para descobrir, aceder e partilhar informações e análises. Os utilizadores devem ser capazes de realizar e personalizar os seus relatórios e *dashboards*;
2. Simplicidade na utilização das ferramentas de BI, ou seja, as ferramentas a ser utilizadas devem ser simples, tendo como objetivo torná-las mais diretas, o que permite acesso mais rápido a informação;
3. Rapidez na implementação e gestão de DW, ou seja, os SSBI tendem a aplicar mecanismos alternativos para redução de custos e suportar processamento de dados, recorrendo a metodologias Agile, SaaS onde o utilizador não precisa de manter os servidores físicos ou em cloud pagando uma assinatura para aceder a uma aplicação já desenvolvida na *web*, ou em cloud onde um utilizador pode personalizar e gerir uma aplicação hospedada num servidor remoto, que visam um ambiente de SSBI com bom desempenho e escalável.
4. Simplicidade no acesso a fontes de dados, ou seja, visto que nem todos os dados precisam de estar armazenados em DW, existe a possibilidade de aceder a dados externos (meteorológicos, geográficos...) sem assistência de IT. Torna-se essencial dar uma visão mais ampla e completa do panorama corporativo, juntando dados não estruturados e dados estruturados criando uma infraestrutura que permita o livre trânsito de dados de todas as fontes necessárias.

### 2.3.1. MICROSOFT POWER BI

Uma das principais características do BI é fornecer ferramentas user-friendly que tornem os dados verdadeiramente disponíveis para o utilizador final. Querendo trabalhar em conjunto com a finalidade de atingir objetivos comuns da empresa, a solução para partilhar informações num ambiente corporativo, passa pela utilização da gestão do conhecimento, da gestão do conteúdo e de plataformas organizacionais (Ferrari & Russo, 2016).



A Microsoft desenvolveu o Microsoft Power BI, que é uma aplicação que contém várias ferramentas de análise de dados referentes a um determinado negócio. O Power BI consegue conjugar e armazenar vários conjuntos de dados de uma organização, na cloud ou localmente. Permite ainda a ligação a bases de dados que estejam no Microsoft SQL Server, ou outras fontes de dados. É através de partilha de informação e utilização de ferramentas de análise de BI que uma empresa retirará vantagens competitivas da sua estrutura, isto com decisões suportadas em factos e em análises rigorosas da empresa e do seu desempenho.

De acordo com Microsoft (2019), Power BI é uma solução de SSBI, que significa que é possível construir e ativar rapidamente uma solução. Power BI é composto por três características principais e serviços:

1. Características de Excel: Power Query, Power Pivot, Power View e Power Map, todas estas características são utilizadas para fornecer análise de dados aos utilizadores;
2. Power BI para Office 365: aqui os utilizadores podem partilhar relatórios, realizar queries online e melhorar a capacidade de aceder a dados e relatórios;
3. Infraestrutura de TI para Power BI, cria uma forma simples de administração e gestão de dados.

Existem outras aplicações de *business intelligence*, mas esta foi escolhida pelas seguintes características:

1. É uma aplicação que está localizada na *cloud*, por isso permite todos os utilizadores possam aceder aos relatórios de todos os tipos de dispositivos;
2. Permite recolher dados localizados na *cloud* ou localmente e fornece acesso rápido e fácil aos dados;
3. É possível ver todos os dados numa só tela e estes são sempre atualizados independentemente do sítio onde estamos a vê-los;
4. Possui ferramentas de análise intuitivas, onde é possível explorar os dados subjacentes, facilitando a localização exata das respostas necessárias,
5. É possível combinar dados de várias localizações, arquivos e serviços da *web* com ferramentas visuais do Power BI;
6. Permite que tomada de decisão seja um processo mais informado;

7. Permite a integração de dados de outras aplicações como Microsoft Office Excel.

A Microsoft define os utilizadores alvo desta ferramenta em três categorias: criadores de relatórios (beneficiam das capacidades analíticas e ferramentas de pesquisa), como os analistas de dados ou consultores, administrador de dados (gerem a conexão de dados e estrutura) como um cientista de dados ou profissional de TI, e utilizadores de relatórios.

## 2.4. DASHBOARDS

Os objetivos dos *dashboards* em BI, é fornecer aos membros de equipa e a utilizadores individuais visibilidade de informação em tempo real. De acordo com Microsoft (2019), um *dashboard* é uma página única, apelidada de tela, que conta uma história através de visualizações gráficas. Visto que está limitada a uma página, um *dashboard* vai conter apenas os elementos mais importantes daquela história (Dedić & Stanier, 2016).

As visualizações de um *dashboard* são provenientes de relatórios e cada relatório é baseado num conjunto de dados, e no caso deste projeto, num conjunto de dados com informações de tráfego de *webistes*. Um *dashboard* pode ser considerado como uma porta de entrada para os relatórios e conjuntos de dados subjacentes. Constituem uma boa forma de monitorizar o negócio e procurar por respostas e ver as métricas mais importantes na generalidade (Eckerson, 2010).

### 2.4.1. RELATÓRIOS

Um relatório de Power BI é uma exibição de várias de várias perspetivas de um conjunto de dados, com recursos visuais e que representam diferentes descobertas e detalhes desse mesmo conjunto de dados. Um relatório tanto pode ter apenas um visual ou várias páginas com visuais dos dados (Dedić & Stanier, 2016).

Os designers dos relatórios criam os visuais num relatório que representam uma parte pequena da informação, não sendo estáticos, ou seja, vão atualizando à medida que os dados mudam por trás.

## 2.5. WEB SCRAPING

### 2.5.1. O QUE É WEB SCRAPING

Embora *web scraping* não seja um termo recente, os últimos anos esta prática tem sido apelidado de *screen scraping* ou *web harvesting* (Mitchell, 2015).

Em teoria, *web scraping* é o processo de extrair informações e dados de um *website*, transformando as informações presentes na página num conjunto de dados estruturados para serem analisados. Isto é mais provável ser cumprido ao escrever um programa automatizado que executa *queries* no servidor *web*, e depois analisa os dados. (Vargiu & Urru, 2012).

Na prática, *web scraping* inclui uma variedade de técnicas de programação e tecnologias. Por exemplo, copiar a lista de contactos de um diretório de internet é um exemplo de *web scraping*, mas esta tarefa executada manualmente funciona para pequenos conjuntos de dados e consome muito tempo útil e então, a automação vem ajudar no aumento da eficiência na recolha de dados em grandes conjuntos de dados (Boeing & Waddell, 2017).

Os dados normalmente são transferidos pela internet através de um conjunto de dados estruturados e formal que é facilmente processado por um computador. No entanto, a internet está cheia de dados não estruturados e semiestruturados que nunca estão disponíveis para serem facilmente interpretados por uma máquina (Haddaway, 2016). *Web scraping* preenche essa lacuna e fornece um leque vasto de dados ao investigador ao extraírem conjuntos de dados estruturados de conteúdo que é criado para humanos (Mitchell, 2015).

Um *web scraper* entra nas páginas *web*, procura elementos de dados específicos na página (podem ser formatos em HTML, PHP, etc.), extrai-os, transforma-os se necessário, e por fim, guarda os dados num conjunto de dados estruturados (Haddaway, 2016).

Para este projeto, foi utilizado o conteúdo dos dados extraídos para análise posterior depois de terem sido limpos e organizados, ou seja, utilizou-se um *web scraper* direcionado para extrair informação textual a partir de páginas *web*. Um *web scraper* automatiza um processo moroso de recolha dados de muitas páginas *web* e juntar

conjuntos de dados estruturados a partir de texto aparentemente sem nexos e não estruturado, que pode estar espalhado por milhares de páginas (Hirschey, 2014).

### 2.5.2. COMPONENTES

Uma ferramenta de *web scraping* é um processo Extract-Transform-Load (ETL). Os *web scrapers* procuram *websites*, extraem dados a partir deles e transformam-nos num formato estruturado, utilizável e pronto para ser carregado num ficheiro ou numa base de dados (Matei, Russell, & Bertino, 2015). É composto pelos seguintes componentes:

1. *Web crawling*: Navega pelo *website* alvo fazendo pedidos http para URL's seguindo um determinado padrão ou outra lógica de paginação. Faz o download dos objetos requeridos como conteúdo HTML e passa esses dados para um extrator (Mitchell, 2015);
2. *Data Parsing e Extração*: O conteúdo HTML é processado utilizando um parser (divide os dados em elementos menores para facilitar a tradução noutra idioma) que extrai os dados requeridos de cada página utilizando diferentes técnicas (Hirschey, 2014);
3. *Limpeza dos dados e transformação*: Converte os dados analisados num formato mais adequado a ser guardado como CSV, JSON ou numa base de dados (Mitchell, 2015);
4. *Armazenamento*: Lê os registos e guarda dos dados num formato CSV, JSON, XML, XLSX ou carrega os dados numa base de dados relacional ou não relacional consoante a estrutura dos dados (Mitchell, 2015).

## 3. CAPÍTULO 3 – DESCRIÇÃO DO TRABALHO DE PROJETO

### 7.1. EXTRAÇÃO DE DADOS COM O DATA MINER

No presente trabalho foi utilizada uma extensão de data mining e *web scraping* para o Google Chrome, denominada Data Miner. Esta extensão é um software que assiste na extração de dados que conseguimos ver no browser de navegação e permite guardar os dados extraídos num formato compatível com Excel (.xlsx, e .csv), ou seja, vem ajudar

na transformação de dados em formato HTML para o formato em tabela (Software Innovation Lab LLC, 2019).

Existem outras aplicações para fazer data scraping, mas esta foi escolhida pelas seguintes características:

1. Simplicidade de utilização;
2. Software com licenciamento gratuito;
3. Permite a extração gratuita de 500 páginas por mês;
4. Permite alterar a velocidade de extração dos dados;
5. Permite a utilização de mais de 50000 receitas criadas por outros utilizadores, ou seja, contém uma componente colaborativa;
6. Não utilizam dados que o utilizador extrai e não vende dados a utilizadores;
7. Permite o scraping automático de páginas através da criação de receitas customizadas (Software Innovation Lab LLC, 2019).

Como referido anteriormente, o Data Miner assiste na extração de dados de páginas *web* para um formato compatível com Excel. Esta extração é feita através da extensão do browser quando estamos numa determinada página *web* e é possível extrair os dados ao correr uma receita.

Uma receita consiste numa lista de instruções que o Data Miner vai ler para prosseguir com a extração dos dados da página. As receitas utilizadas tanto podem ter sido criadas por outros utilizadores como podem ter sido criadas pelo próprio utilizador.

Quando estamos a criar uma receita nova, podemos utilizar a ferramenta *smart find* que permite extrair os dados pretendidos sem ser preciso escrever linhas de código. Primeiramente temos de determinar se o que pretendemos é uma página de lista ou uma página de detalhes. Uma página de lista tem linhas e parece-se como uma página de resultados, enquanto uma página de detalhes não tem linhas e parece-se com uma página de perfil ou uma página de um produto. Para este projeto foi selecionada uma página de lista.

Posteriormente, começamos a capturar os dados utilizando seletores, que são pedaços de HTML retirados diretamente da página e que informam o Data Miner onde procurar os dados a extrair (Software Innovation Lab LLC, 2019). Na Tabela I abaixo encontram-se alguns seletores:

<b>Tipo de seletor</b>	<b>Exemplo</b>	<b>Significado</b>
<i>Id</i>	<i>#name</i>	<i>Ids</i> são o tipo mais específico de seletor e, na maioria dos casos, selecionam apenas um item. São bons para páginas de detalhes e maus para páginas de listas.
<i>class</i>	<i>.name</i>	As classes são o melhor tipo de seletor, são gerais o suficiente para capturar o elemento numa página de detalhe ou lista, mas podem selecionar demasiados itens.
Elemento <i>HTML</i>	<i>div</i>	Estes são <i>tags HTML</i> utilizados para estruturar a página. São utilizados com frequência e geralmente fornecem demasiados resultados.
<i>Link</i>	<a>	Estes são <i>tags HTML</i> utilizados para links. São empregues quando é necessário capturar <i>URLs</i> .

Tabela I - Tipos de seletores (Adaptado de <https://data-miner.io/quick-guides/persona-business>)

### 7.1.1. FONTE DOS DADOS

Para este projeto, utilizou-se o Data Miner para a extração de dados em várias páginas de lista do *website* Crunchbase Inc. Esta é uma plataforma líder para profissionais, investidores e market researchers descobrirem empresas inovadoras, para se interligarem com as pessoas por detrás das empresas, perseguir novas oportunidades de negócio e para informar melhor as suas decisões de negócio.

Foi escolhido o Crunchbase Inc. como fonte de dados visto que é a fonte de dados de empresa com melhor qualidade. Isto é, o Crunchbase Inc. foi escolhido devido às seguintes características:

1. Possui a maior rede de investidores: mais de 3700 empresas de investimento que enviam atualizações mensais de portfolio para o *website*. Este

relacionamento vem garantir que o mesmo tenha acesso em primeira mão às informações das empresas, permitindo angariar dados de origem fidedigna;

2. Possui contribuidores ativos da comunidade empresarial: A sua comunidade de executivos, empreendedores e investidores contribui ativamente para as páginas de perfil da empresa. Isto vem permitir que a sua base de dados esteja em constante crescimento e melhoria;
3. Possui uma Inteligência Artificial (IA) e um enriquecimento de dados auxiliado por aprendizagem de máquinas (machine learning): Esta IA e os algoritmos de machine learning validam a precisão dos dados inseridos na base de dados e examinam anomalias e alertam sobre possíveis conflitos de dados;
4. Possui uma equipa de pessoas especializadas: Os analistas da empresa fornecem validação e curadoria manual dos dados inseridos no *website*. Estes também analisam interligações importantes nos dados para desenvolver algoritmos e fornecer informações valiosas (Crunchbase Inc., 2019).

## 7.1.2. MÉTODO DE PESQUISA

### 7.1.2.1. PESQUISA NO CRUNCHBASE

Primeiramente foram definidos no *website* Crunchbase os critérios de pesquisa das empresas a ser incluídas na base de dados:

1. Foram escolhidas empresas sediadas na União Europeia (UE);

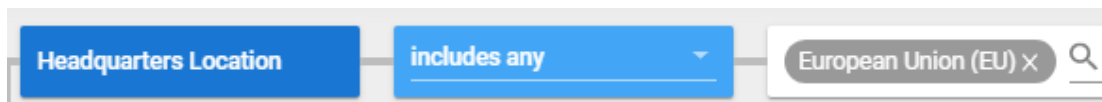


Figura II - Sede das empresas escolhidas

- Foram escolhidas empresas que estejam ligadas à Indústria 4.0, ou seja, que se incluísse nas suas características as seguintes categorias: Big Data, Analytics, Data Mining, Machine learning, E-commerce, Robotics, Artificial Intelligence (IA), Business Intelligence, Information Technology, Cloud Computing e Internet of Things (IOT), como mostra a Figura III abaixo.

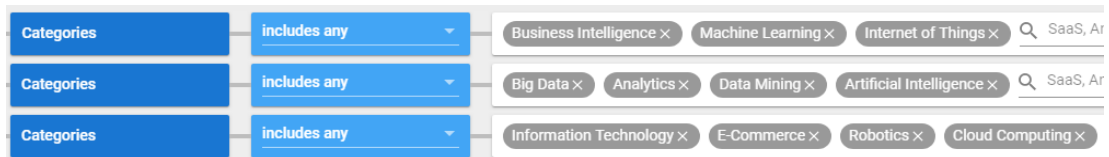


Figura III - Área das empresas escolhidas

Depois de terem sido definidos os critérios de pesquisa, foi gerado um conjunto de resultados em formato de lista que continha as seguintes colunas: *Organization Name, Categories, Headquarters Location, City, Country, State, Description, CB Rank (Company), Operating Status, Company Type, Number of Articles, Founded Date, Number of Founders, Number of Employees, Number of Investors, Monthly Visits, Page Views / Visit, Bounce Rate, Visit duration, Number of Competitors, Total Products Active, Number of Funding Rounds e Total Funding Amount.*

Depois de definidos os critérios de pesquisa e ter sido gerada a página de resultados, foi posta em prática a extensão Data Miner para ir buscar os dados à página *web* e exportá-los para o formato Excel (.xlsx).

### 7.1.3. EXTRACT-TRANSFORM-LOAD COM O DATA MINER

Para conseguirmos exportar os dados com o Data Miner, primeiramente escolhemos o tipo de página que queremos analisar, ou seja, página de lista ou página de detalhes. Neste caso, o tipo escolhido é uma página de lista.

Depois de escolhido o tipo de página, vamos utilizar a ferramenta *smart find* para seleccionar as linhas que contêm os dados que queremos exportar, e para tal, vamos utilizar o tipo de seletor class denominado “*.component--grid-row*”. Esta ferramenta mostra a verde aquilo que estamos a seleccionar, como na Figura IV abaixo:







<input type="checkbox"/>	Organization Name	Categories	Headquarters Location
<input type="checkbox"/>	1.  Cambridge Analytica	Analytics, Data Mining, Information Technolo...	London, England, United Kingdom
<input type="checkbox"/>	2.  Onfido	Artificial Intelligence, Identity Management, L...	London, England, United Kingdom
<input type="checkbox"/>	3.  Darktrace	Artificial Intelligence, Cyber Security, Informa...	Cambridge, Cambridgeshire, United Kingdom
<input type="checkbox"/>	4.  StethoMe	Apps, Artificial Intelligence, Home Health Car...	Poznan, Wielkopolskie, Poland

Figura V – Pré visualização de selecção das linhas com *Smart find*

Posteriormente vamos especificar as colunas pretendidas para o ficheiro output, mais uma vez utilizando a ferramenta *smart find*. Aqui é permitido escolher o nome que queremos dar à coluna bem como o tipo de dados que queremos exportar dessa coluna (texto, HTML, URL, image URL ou other attribute). Neste ponto, o seletor utilizado foi “.column-id-” seguido do nome da coluna que o identifica em formato HTML, por exemplo, no caso da coluna “Organization Name” o seu elemento seletor é “.column-id-identifier” e identifica a coluna com cor roxa:

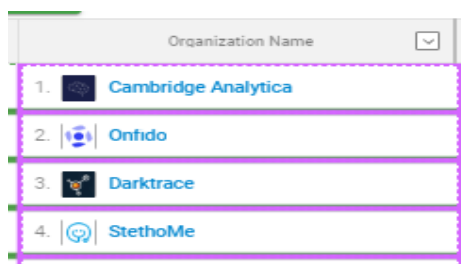


Figura IV - Seleção de colunas com o *Smart find*

Depois de definidas as linhas e as colunas necessárias para o ficheiro output, vamos indicar ao Data Miner que existem outras páginas que contêm informação a ser exportada. Para tal, vamos à página de navegação da extensão indicar que existe um elemento seletor de navegação, que neste caso será a classe “.page-button-next” como mostra a Figura VI. Isto indica à extensão, que após ter feito o scraping numa determinada página, ele deverá passar a página seguinte e iniciar novamente o processo de scraping.

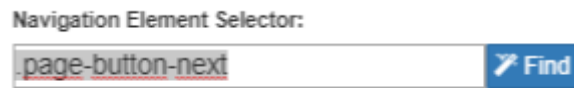


Figura VI - Elemento Seletor de Navegação

Depois de definidos os elementos seletores das linhas, das colunas e de navegação o Data Miner inicia o processo de scraping da página *web* e exporta para um ficheiro output em formato *.xlsx*. Neste caso, o ficheiro criado contém um universo de 5604 empresas sediadas na União Europeia ligadas à área da Indústria 4.0.

## 7.2. FRAMEWORK DO PROJETO

A seguinte lista de atividades mostra como foi construído e partilhada uma solução BI com ferramentas de Power BI:

1. Exportação dos dados com *web scraping*;
2. Importação dos dados para Power BI;
3. Transformação dos dados;
4. Enriquecimento dos dados;
5. Criação de relatórios e *dashboard*.

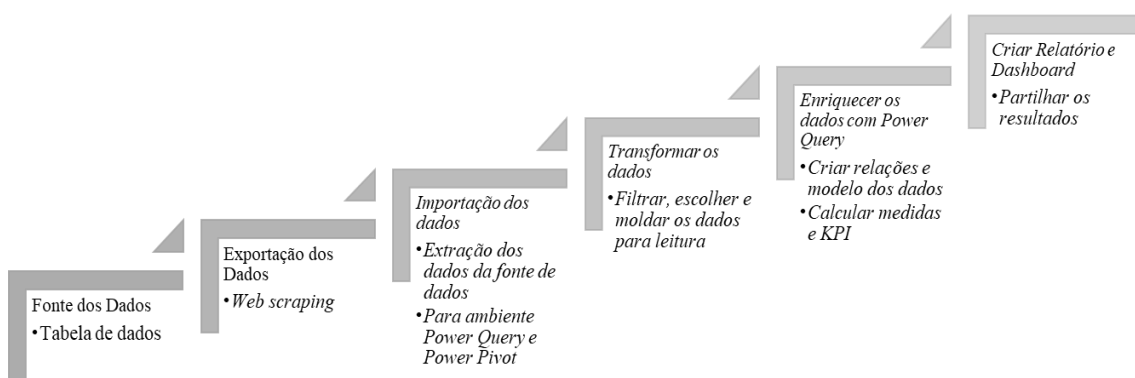


Figura VII - Framework de elaboração do Projeto

### 7.3. ESQUEMA EM ESTRELA NO POWER BI

Muitas vezes uma análise começa com um conjunto de dados em bruto que numa tabela que contem todas as colunas pertinentes ao estudo que queremos fazer. Como foi referido, as dimensões e os factos são duas partes essenciais da construção de um modelo de dados e as relações entre si no *star schema* é a melhor opção para fazer a modelagem.

Na Figura VIII podemos ver como a base de dados era depois de extraída pelo Data Miner. Como podemos observar, todos estes dados estão desformatados e necessitam de algum tratamento.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Organiza	FoundingD	Company	Number c	Number c	Number c	Number c	Patents (	Tradema	CB Rank	CB Rank	Operatin	Number c	Country	State	City	Artificial I
2	Nordic E	01/11/2016	Profit	5	B	11-50	Pequena			83908	10001-10	Active	3	Denmark	Hovedsta	Copenh	Artificial Inte
3	Dafcode	01/01/2011	Profit	1	B	101-250	Média			104818	100001-5	Active	40	Poland	Mazowie	Warsaw	Bi
4	Korber AI	01/01/1995	Profit	1	B	10001+	Grande	139	39	100887	100001-5	Active	12721	Germany	Hamburg	Hamburg	
5	Build Up	01/01/2014	Profit	1	B	1-10	Micro			128536	100001-5	Active	#N/A	Portugal	Lisboa	Lisbon	
6	Nina Cap	01/01/2019	Profit	1	B	1-10	Micro			94608	10001-10	Active		Spain	Catalonia	Barcelon	Artificial Inte
7	MyBucks	01/01/2011	Profit	1	B	1001-500	Grande			106791	100001-5	Active		Luxembc	Luxembc	Luxembc	Artificial Inte
8	Sequitur	01/07/2018	Profit	1	B	1-10	Micro			128910	100001-5	Active		Italy	Lombard	Milan	Artificial Inte
9	Sharpe C	01/01/2015		3	B	1-10	Micro			151348	100001-5	Closed		United Ki	England	London	
10	KraussM	01/01/1900	Profit	1	B	5001-100	Grande			155543	100001-5	Active	96	Germany	Bayern	Munich	

Figura VIII – Ficheiro Excel de tráfego de websites de empresas Europeias na área da Indústria 4.0

#### 7.3.1. PREPARAÇÃO DOS DADOS

A Figura VIII, é uma amostra das informações recolhidas pela ferramenta Data Miner numa só tabela. Ao carregar o nosso conjunto de dados num ambiente de Power Query, é possível começar a preparar os dados. No caso deste projeto, o conjunto de dados está contido numa tabela de Excel, por isso ao importar no Power BI escolhemos a opção *Get Data* → *Excel File* (Figura XVI, Anexo I).



Normalmente, a tabela de factos é contruída com base na tabela pré-existente, por isso a tabela será duplicada.

### 7.3.3. CRIAR TABELAS DE DIMENSÕES

Como referido no ponto 2.2.2.1, uma tabela de dimensões contém informação descritiva, que é ou pode ser utilizada para cortar e seccionar dados, eixo dos gráficos, como a coluna “Date” e “Organization Name”. Estes campos podem ser tipo texto, ou datas, ou mesmo um número.

Neste projeto foram identificadas as seguintes dimensões:

1. Área
2. Date
3. Empresa
4. Funding
5. Localização
6. Redes Sociais

Na secção seguinte será explicada a construção da dimensão “Empresa”, seguido de expressões em linguagem DAX utilizada pelo Power BI. Para construir a dimensão Empresa, temos de criar uma cópia da tabela de factos, ou neste caso, a tabela original, e podemos fazê-lo duplicando, como na Figura X abaixo:

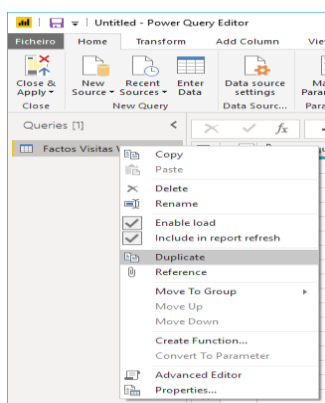


Figura X - Duplicar  
Tabela de Factos

Após esta mudança, no menu de queries passamos a ter duas entradas de dados: “Factos Visitas Website” e “Empresa”, depois de renomeado (Figura XI).

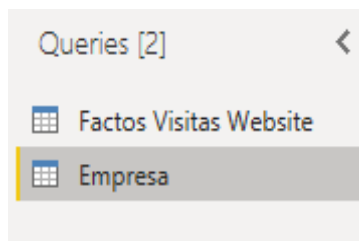


Figura XI - Dimensão  
Empresa no menu Queries

Esta nova dimensão só terá um campo descritivo, por isso é possível remover todas as outras colunas da tabela de Factos Visitas Website, mas queremos ter a certeza de que a nova dimensão “Area” não tem nenhum elemento duplicado, por isso clicamos em “Choose Columns” e selecionamos só aquela que queremos (Figura XX, Anexo I), com recurso à seguinte expressão DAX:

```
= Table.SelectColumns("#Changed Type1", {"Organization Name", "Company Type",  
"Number of Founders", "Number of Founders categorizado", "Number of Employees", "Number  
of Employees categorizado", "Patents Granted", "Trademarks Registered", "CB Rank  
(Company)", "CB Rank (Company) categorized", "Operating Status", "Number of Employees2"})
```

De seguida, selecionamos as colunas da nova dimensão e vamos remover todos os elementos duplicados, em “Remove Duplicates” (ver Figura XXII, Anexo I), de forma a ter uma lista de áreas únicas no nosso conjunto de dados, o utilizando a expressão DAX em baixo:

```
= Table.Distinct("#Removed Other Columns")
```

Estes passos são aplicados para todas as tabelas dimensões supramencionadas, até termos uma tabela única para cada dimensão bem como uma tabela única de factos.

#### 7.3.4. LIGAR DIMENSÕES E FACTOS NO POWER BI

Depois de serem construídas as diversas tabelas do conjunto de dados, é necessário criar as relações entre elas, e estas relações só podem ser criadas com base num único campo em cada tabela. Mantendo o exemplo da dimensão Empresa acima, temos 8 campos da tabela para serem ligados (“Organization Name”, “Company Type”, “Number of Founders”, “Number of Employees”, “Patents Granted”, “Trademarks registered”, “CB Rank”, “Operating Status”), sendo necessário criar uma chave identificadora de cada um destes campos.

Para tal, existem dois métodos para realizar esta tarefa: a primeira é criar uma coluna de chaves para a tabela de dimensão, recorrendo à opção *Index Column*, e o outro método seria criar um campo de valores concatenados, mas que não será utilizado visto que algumas dimensões possuem texto longo e isso causa problemas de eficiência, e então a abordagem de *Index* é mais adequada neste projeto.

Primeiramente, depois de removidos os elementos duplicados, é adicionada uma coluna no menu *Add Column* → *Index Column* → *From 1*, que irá criar uma nova coluna com o índice a começar no número 1 (Figura XXIII, Anexo I):

```
= Table.AddIndexColumn("#Removed Duplicates", "Index", 1, 1)
```

Esta nova coluna será a nossa *Surrogate Key* ou chave estrangeira denominada de *SK\_Empresa*. Ou seja, é um identificador exclusivo de um objeto ou uma entidade e não é derivada de outros dados na base de dados e pode ou não ser utilizada como chave primária:

```
= Table.RenameColumns("#Added Index", {"Index", "SK_Empresa"})
```

O próximo passo é juntar a “SK\_Empresa” na tabela Factos *Visitas Website*, utilizando a opção *Merge*. É possível observar na Figura XII, como se fazem as ligações

entre queries, selecionando um campo de valores de uma tabela que esteja presente na outra tabela de forma a identificar valores cruzando dados.

= `Table.NestedJoin("#Changed Type1", {"Organization Name"}, Empresa, {"Organization Name"}, "Empresa", JoinKind.LeftOuter)`

## Merge

Select a table and matching columns to create a merged table.

Factos Visitas Website

Organization Name	Company Type	Number of Founders	Number of Founders categorizado	Number of Em
Nordic Eye Venture Capital	Profit	5	B	11-50
Nordic Eye Venture Capital	Profit	5	B	11-50
Nordic Eye Venture Capital	Profit	5	B	11-50
Nordic Eye Venture Capital	Profit	5	B	11-50
Nordic Eye Venture Capital	Profit	5	B	11-50

Empresa

Organization Name	Company Type	Number of Founders	Number of Founders categorizado	Number of Em
Nordic Eye Venture Capital	Profit	5	B	11-50
Daftcode	Profit	null	B	101-250
Korber AG	Profit	null	B	10001+
Build Up Labs	Profit	1	B	1-10
Nina Capital	Profit	1	B	1-10

Join Kind

Left Outer (all from first, matching from second)

Use fuzzy matching to perform the merge

> Fuzzy merge options

• Estimating matches based on data previews

OK

Cancel

Figura XII - Combinar Queries em ambiente Power BI

Esta opção serve para combinar queries em linhas correspondentes de outra tabela. O resultado de Merge é uma única query com critérios de associação ou correspondência entre as duas consultas, o número de linhas depender dos critérios de correspondência entre as colunas e o número de colunas irá depender também, das colunas selecionadas no conjunto de resultados. Após termos utilizado a ferramenta *Merge*, podemos ir buscar a chave identificadora “SK\_Empresa” e ligá-la à tabela de factos como mostra a Figura XII e a expressão DAX, em baixo:



```

= Table.ExpandTableColumn(#"Merged Queries", "Empresa", {"SK_Empresa"}, {"Empresa.SK_Empresa"})

```

Depois de concluídos estes passos para todas as dimensões identificadas, é possível apagar as colunas correspondentes da tabela Factos Visitas Website e carregar as tabelas para o Power BI e criar as relações entre elas com base nas *surrogate keys* ou chave primária de cada uma das tabelas dimansionais, o que nos leva a criar o *star schema* dos dados, com relações de um para muitos na tabela de Factos Visitas Website, como mostra a Figura XIII:

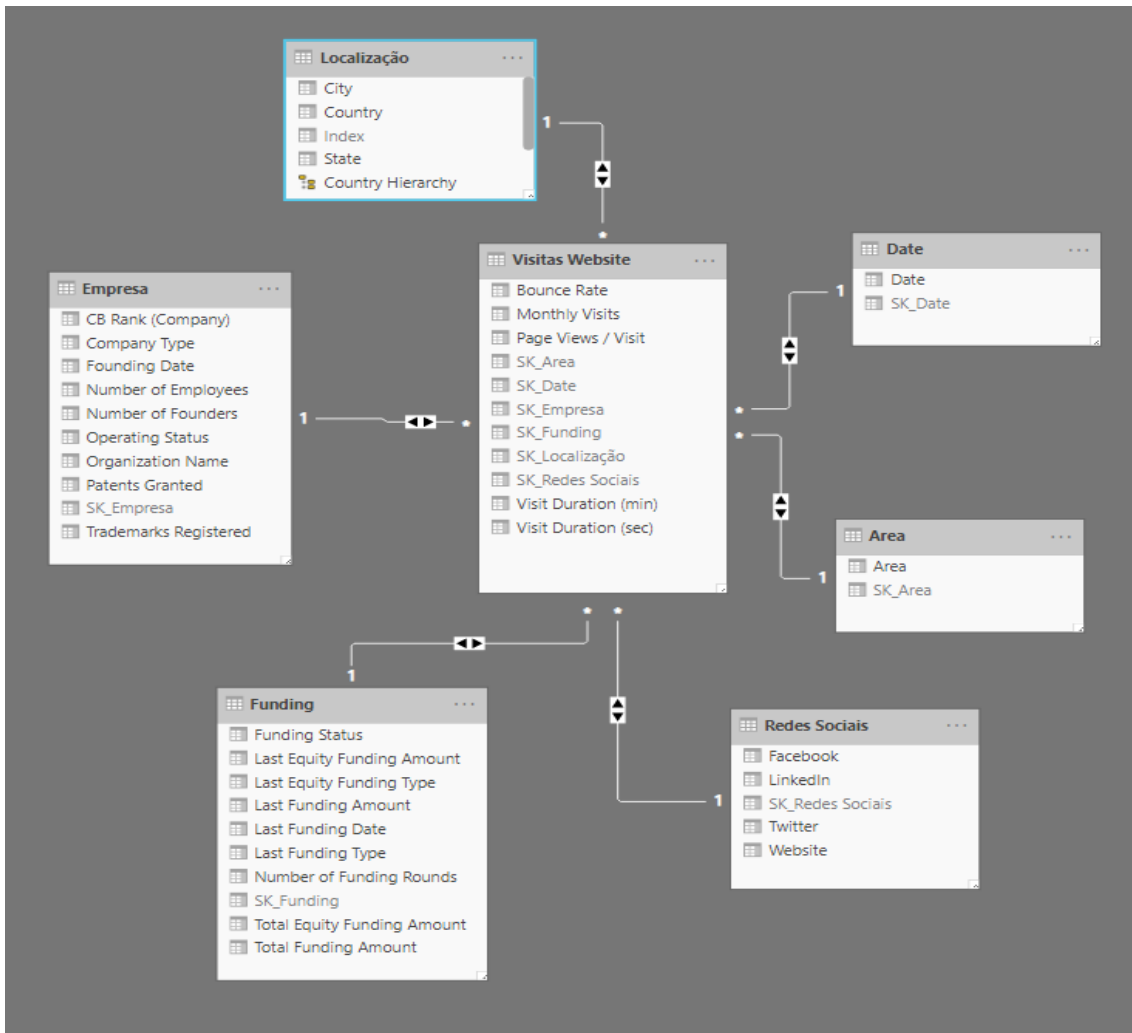


Figura XIII - Esquema em estrela do conjunto de dados extraídos

Depois de termos criado o modelo de dados a ser utilizado, é possível começar a realizar consultas como gerir as relações (ver Figura XXV, Anexo I) entre eles e posteriormente criar os relatórios e o *dashboard* de visitas a *websites*.

#### 7.4. CRIAÇÃO DOS RELATÓRIOS E DASHBOARD

Como referido no ponto 2.3.2, um dashboard é uma página única, ou uma tela que recorre a visualizações para evidenciar determinados aspetos de determinados assuntos. Visto que isto é um tipo de informação relativamente limitado, contém somente os elementos mais importantes de um assunto.

As visualizações que conseguimos ver no dashboard são chamadas de tiles ou azulejos, e são afixados no dashboard pelos utilizadores (Ver Figura XXIV, Anexo I). Na maioria dos casos, ao selecionarmos um azulejo, esse mesmo encaminha-nos para a página de relatório onde ele foi criado (Microsoft, 2019).

Estas visualizações que preenchem o dashboard são provenientes dos relatórios e cada relatório é feito com base num conjunto de dados representando informações que foram descobertas nos dados. Por isso podemos ver os *dashboards* como sendo uma forma de entrada para os relatórios e conjuntos de dados subjacentes, e ao clicarmos numa visualização de um dashboard, somos transportados para o relatório ou conjunto de dados que levou a sua criação (Dedić & Stanier, 2016).

A base de dados criada para este projeto, é relativa a dados de tráfego de *websites* em empresas da área da Indústria 4.0 sediadas na União Europeia face ao ano 2019. Posto isto, é possível criar um dashboard que represente a situação do tráfego aos *websites* na U.E que seja suportado por relatórios que contenham visualizações ou representações visuais da média de visitas mensais aos *websites*, do número médio de páginas visitadas por visita e do bounce rate médio (percentagem de pessoas que após visitar uma página muda de *website*) bem como do tempo médio de duração da visita ao *website* em segundos e minutos. Ao todo foram elaborados quatro relatórios (ver Anexo II), um para cada variável da tabela de factos.

Um relatório de Power BI permite-nos, com recurso a vários tipos de visuais que representam diferentes descobertas e detalhes de informação, ter uma vista com várias

perspetivas de um conjunto de dados. Um relatório pode ter mais do que uma representação visual, dependendo do tipo de análise pretendida (Ferrari & Russo, 2016).

No presente projeto, visto que a variável a estudar é “*Monthly Visits*”, queremos ver esta variável explicada por um elemento que a resume num ponto principal, sendo esse a média de visitas mensais nos últimos meses.

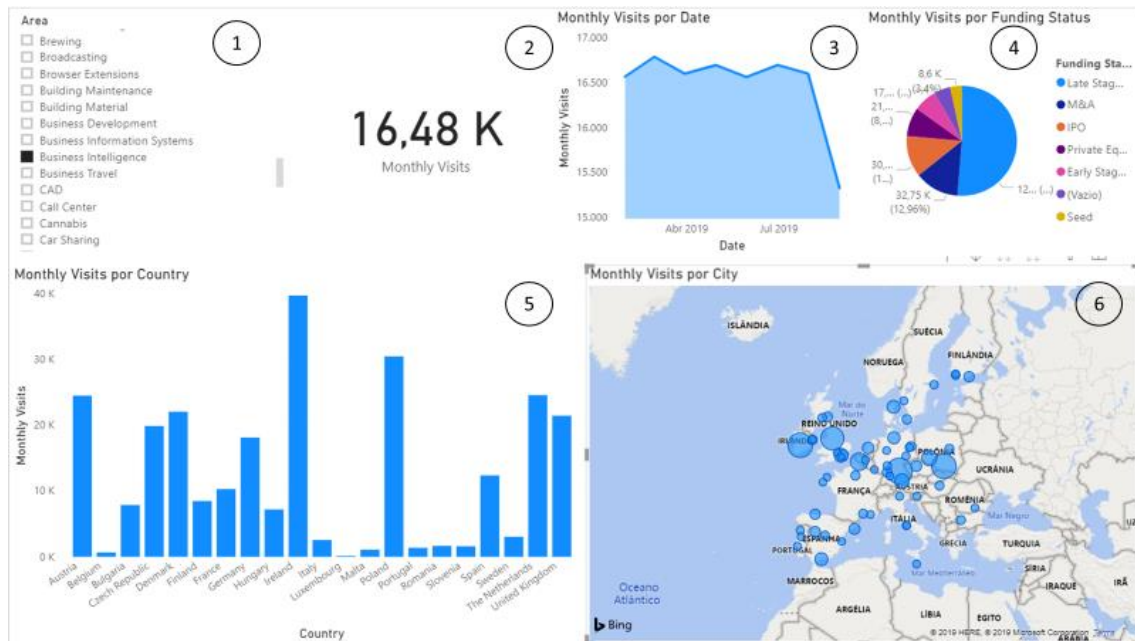


Figura XIV - Relatório de Visitas Mensais

A Figura XIV acima, representa um relatório elaborado com os dados extraídos de tráfego de *websites*, constituído por várias visualizações com recurso a várias dimensões:

- No ponto 1 é apresentada uma visualização do tipo *slicer* ou segmentação de dados que aplica um filtro que influencia as visualizações de toda a página, mediante a variável escolhida que, neste caso é “*Business Intelligence*”;
- No ponto 2 é apresentada uma visualização de tipo *card* que apresenta a média de visitas mensais, apresentando valores de 16480 visitas, considerando a segmentação de dados escolhida anteriormente. Sendo este o elemento escolhido para resumir a variável num ponto principal para inserir no *dashboard*;
- No ponto 3 é possível observar um *area chart* que mostra os dados em forma de gráfico dados quantitativos, neste caso comparando o número de visitas mensais com a dimensão *Date*. Este tipo de gráfico realça a importância da alteração do

número de visitas mensais ao longo do tempo e ao mesmo tempo chama a atenção para uma tendência de queda no valor total;

- No ponto 4 é possível observar um *pie chart* que mostram os dados em proporção de um todo, em forma de percentagem. Neste caso estamos a comparar o número de visitas mensais com o estatuto de financiamento das empresas. Podemos então observar que neste momento as empresas classificadas como *Late Stage Venture*, representam mais de 50 % do número de visitas mensais;
- No ponto 5, é apresentado um *clustered column chart* que ajuda a representar o número médio de visitas mensais por país;
- Por fim, no ponto 6, os dados das visitas mensais e os dados hierarquizados da dimensão “Localização”, são comparados, como mostra a Figura XV abaixo. Para a criação deste tipo de gráficos é necessário ter atenção à atribuição de uma categoria (*Country, State, City*).

Depois de terem sido desenhados e construídos os relatórios de dados de tráfego de *websites*, é possível fazer o upload para a versão online, editar o relatório e escolher quais os azulejos que pretendemos para criar o dashboard selecionando cinco elementos representativos, um de cada uma das variáveis da tabela de factos, presentes no Anexo II.

A Figura XV, representa o dashboard criado, com a variável bounce rate com os pontos 2, 3 e 5 do relatório da mesma (Figura XXVII, Anexo II); com os pontos 2,5 e 6 dos relatórios das variáveis monthly visits e visit duration (Figura XXVIII, Anexo II) e por fim com os pontos 2,3 e 6 do relatório da variável pages views / visit (Figura XXIX, Anexo II). O dashboard criado apresenta-se como uma solução intuitiva e interativa, o que vai de encontro aos objetivos e resultados esperados.

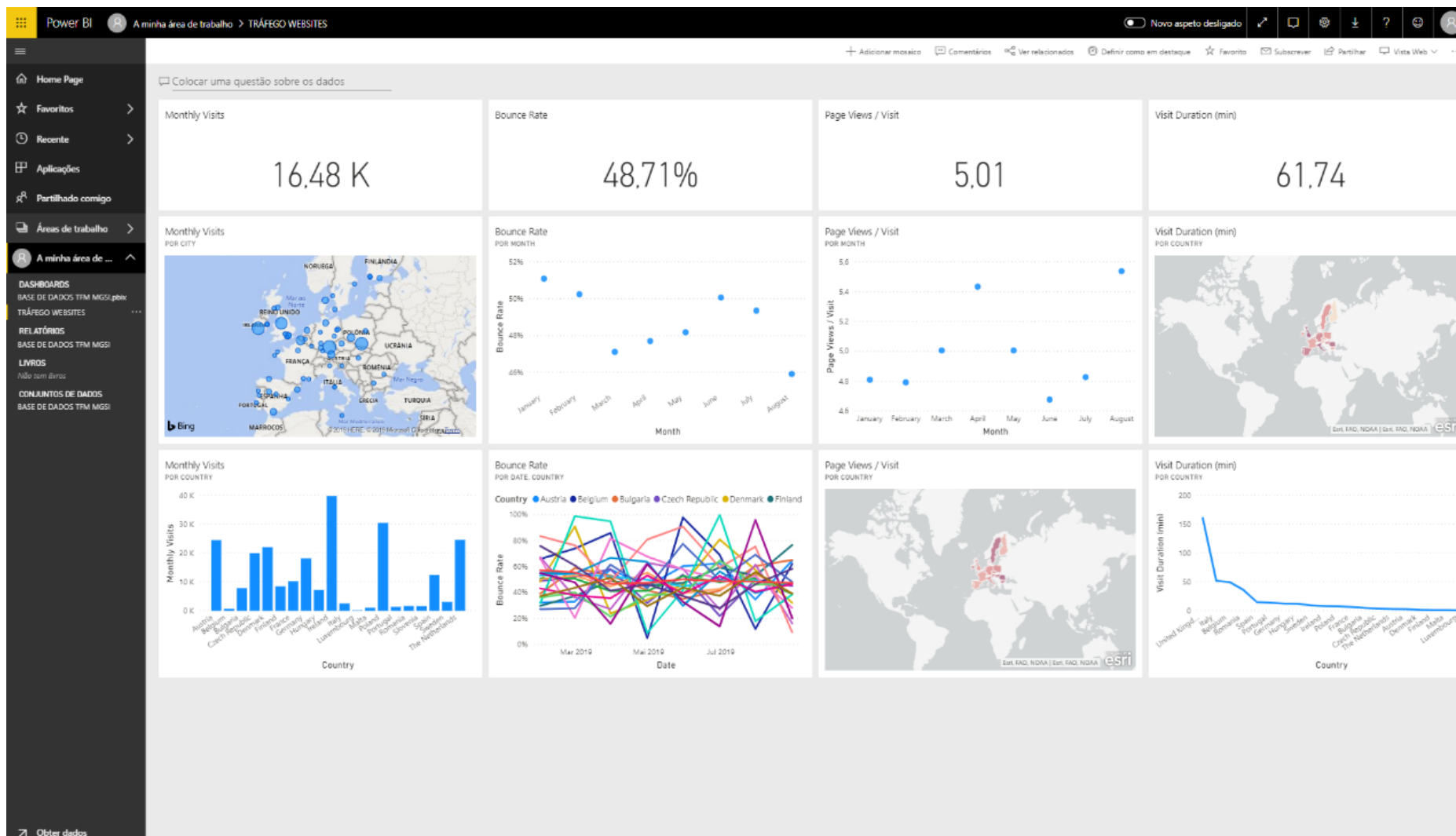


Figura XV- Dashboard de Tráfego De Websites De Empresas Europeias

## 2. RESULTADOS

O objetivo deste trabalho de projeto foi desenvolver e criar uma solução de business intelligence para resolver questões de âmbito analítico, e neste caso sobre dados de tráfego de *websites*. Para guiar o processo empírico, a secção teórica deste projeto foi construída de forma a tentar responder às seguintes questões: 1. “O que é o BI e o self-service BI?”, 2. “O que é *web scraping* e como utilizar?”, 3. “Como utilizar o Power BI?”, 4. “Como criar um dashboard?”, 5. “Que tipo de informações podemos retirar dos dados?”.

A literatura revista tem como propósito estabelecer uma ligação desde a extração da fonte de dados, à ligação à fonte de dados e às especificidades de construir um modelo de dados. Neste sentido, este projeto revelou responder a todas estas questões tentando ser conciso. No entanto, no final, a literatura escrita ao responder às questões colocadas, vem suportar a secção empírica deste projeto, incluindo a extração de dados, modelação de dados e ferramentas de análise utilizadas, e todos os elementos introduzidos nela.

A solução e o método desenvolvido neste projeto são uma prova de conceito, ou seja, seguindo uma abordagem de business intelligence e utilizando ferramentas da área, é possível descobrir detalhes e padrões de forma mais fácil. Apesar de ser uma ferramenta utilizada para desenvolver projetos de larga escala é também possível aplicar esta ferramenta como alternativa a ferramentas de análise tradicionais como Excel.

Os resultados deste projeto mostram que é possível implementar uma solução de business intelligence, que pode trazer benefícios e vantagens com uma melhor visualização dos dados, de forma mais intuitiva e analítica que permite uma rápida compreensão dos dados e permite perceber detalhes que dantes passariam despercebidos.

Relativamente aos dados obtidos na construção do dashboard, este apresenta-se como uma solução elegante e simplista de resumir os dados e mostrar alguns principais alvos de estudo, tornado fácil e intuitiva a interpretação dos dados. Neste caso o dashboard foi contruído utilizando a aplicação do Power BI online, e utilizando um slicer no valor “Business Intelligence” na dimensão “Area”. Por isso, ao observar o dashboard podemos constatar que a situação geral das empresas cuja área de atuação é “Business Intelligence”, tem valores de média mensal de visitas aos seus *websites* de 16480 visitas por mês sendo que a Irlanda, Holanda, Áustria e Reino Unido se destacam entre os

demais, que têm uma percentagem média de bounce rate de 48,71% onde Janeiro e Fevereiro são os meses em que destaca uma maior percentagem de bounce rate, que o número médio de páginas visitadas é 5 mas onde as empresas registam maior número de páginas visitadas por visita é na Roménia e em média visitam 20 páginas, e que o tempo em média despendido é cerca de 61,74 minutos por visita no ultimo ano, sendo este número inflacionado pelo Reino Unido que apresenta valores muito elevados de tempo de visita.

Portanto, os resultados mostram que mesmo tendo pouca experiência em termos de recolha e modelação de dados, é possível demonstrar como relatórios e *dashboards* podem ser construídos a partir de um conjunto de dados a partida não estruturado, aplicando ferramentas e conceitos de BI e de sistemas de informação.

### 3. DIFICULDADES E LIMITAÇÕES

Os sistemas de BI apresentam-se como sistemas que facilitam o processo de tomada de decisão, caracterizando-se como um processo de recolha, processamento e análise dos dados.

Uma das maiores dificuldades encontradas neste projeto, revelou ser a falta de documentação científica sobre a construção e o design de relatórios e *dashboards*. Deste modo foi necessário recorrer a tutoriais e documentação da própria aplicação Power BI de forma a melhor entender como realizar a integração dos dados com a ferramenta e também para fazer as várias ligações entre as tabelas de dimensões e a tabela de factos como explicado no ponto 3.3.4.

Foram sentidas dificuldades também na recolha dos dados com a ferramenta de *web scraping*, no sentido em que foi necessário compreender um pouco de estruturas de HTML para conseguir detetar qual a classe de dados a selecionar para exportação.

O tratamento dos dados no Power BI, também acabou por ser um fator que acabou por atrasar o projeto, sendo que foi necessário fazer uma seleção dos dados a utilizar no modelo, bem como o tratamento posterior destes valores a atribuir a cada um o formato correto de dados de forma a garantir uma maior fiabilidade nos resultados obtidos.

#### 4. CONCLUSÕES

O projeto aqui desenvolvido e apresentado, apesar de se revelar funcional e de todos os objetivos terem sido atingidos, apresenta alguns aspetos que carecem de evolução e investigação futura e necessita de melhorias e implementações futuras, para ultrapassar algumas das limitações pré-existentes. São então referidas algumas implementações futuras que visam dar continuidade ao trabalho aqui iniciado: seria pertinente rever o modelo de dados incluindo novas medidas (factos) que permitam novas análises; seria também relevante definir KPI's e implementar o projeto a nível live data; seria importante também alargar a base de dados incluindo novos dados que possam explicar de melhor forma as variáveis escolhidas; seria relevante também utilizar novas ferramentas de extração de dados online de forma a garantir maior diversidade e conhecimento tácito nesta área.

O termo de BI bem como os conceitos que o complementam revelaram ser bastante importantes. Mostrou-se que ao utilizar tecnologias de informação e seus conceitos para construir e entregar informação útil e valiosa aos tomadores de decisão, é essencial tanto para realizar estudos e projetos como também na esfera empresarial. As ferramentas de BI, em específico o Power BI, mostraram que é possível recolher dados estruturados que são mais fáceis de pesquisar e semiestruturados que contêm a informação necessária para análise e tomada de decisão.

Foi dada uma especial atenção ao SSBI por ser uma ferramenta que se tem diferenciado dentro da área. O facto de esta conseguir transmitir aos utilizadores parte de capacidade de BI, permite uma utilização recorrente desta ferramenta, o que torna os utilizadores mais autónomos e mais curiosos. A ferramenta mostrou ser fácil de utilização e intuitiva, mas ao mesmo tempo são uma ferramenta abrangente, onde a informação está centralizada, atualizada e corretamente processada. Existem vários tipos de utilizadores destas ferramentas, uns com mais aptidões analíticas que outros e para responder às necessidades desses utilizadores mais avançados, estes sistemas fornecem ferramentas capazes de análises bem mais complexas que aquelas realizadas neste projeto. A interação que estes têm entre si também constitui parte do sistema de SSBI.

Este projeto demonstra que, é possível apresentar conceitos de TI aos utilizadores não experientes, como *star-schema*, *surrogate keys* e *dimensions*, e ao conseguir aproximar



os utilizadores de sistemas de SSBI, é possível transmitir uma visão diferente do que é uma ferramenta de BI tradicional, que ficava à responsabilidade das áreas das TI. Um sistema de SSBI bem construído e mantido, consegue dar aos seus utilizadores a capacidade realizar perguntas e obter respostas, de forma imediata e interativa, iterativamente até se atingir o resultado pretendido. Ao conseguir transmitir aos utilizadores aquilo que necessitam para realizar as suas funções, consegue angariar e cativar mais utilizadores e aumentar o seu valor dentro da organização como sendo uma ferramenta valiosa para o crescimento de negócio.

A transmissão desta capacidade de escolha de funções e interações a ter com a ferramenta, permite que os utilizadores tenham acesso a várias funcionalidades consoante o seu propósito, podendo ser utilizadores finais que pretendem utilizar estas ferramentas no âmbito de atividade do dia-a-dia de forma autónoma ou utilizadores com características analíticas e técnicas mais avançadas que pretendem descobrir novas oportunidades.

O objetivo deste projeto foi alcançado, visto que a informação recolhida para análise demonstrou qual a direção a seguir mediante a problemática apresentada, evidenciando oportunidades e possíveis dificuldades que possam surgir e o modo como podem ser ultrapassados. Foi demonstrado como se pode criar um conjunto de dados, modelá-lo e analisá-lo utilizando técnicas e ferramentas de BI.

## 5. TRABALHOS FUTUROS

O projeto apresenta alguns aspetos que carecem de evolução e investigação futura e necessita de melhorias e implementações futuras, para ultrapassar algumas das limitações pré-existentes. São então referidas algumas implementações futuras que visam dar continuidade ao trabalho aqui iniciado:

1. Seria pertinente testar no âmbito deste projeto, diferentes aplicações que tenham funcionalidades similares, de forma a perceber quais as melhores aplicações a utilizar mediante um determinado caso prático;
2. Seria pertinente também perceber que aplicações e funcionalidades os utilizadores realmente procuram e se de facto as ferramentas de SSBI ajudam a mitigar essas falhas;

3. Seria também pertinente estudar outras ferramentas que permitam a extração de dados *online*, descobrir ou apresentar novos métodos eficazes de recolha de informação.

## REFERÊNCIAS

- Adeoye, T., Raufu, O., & Omodara, O. (2011). Design of Data Warehouse and Business Intelligence System A case study of a Retail Industry Thesis submitted for completion of Master of Science (60 credits), (June).
- Ari, H., & Tolvanen, A. (2019). Developing an efficient business intelligence solution for day-to-day supply chain management: Case Pulp and Paper Industry Workshop.
- Baars, H., & Kemper, H. G. (2008). Management support with structured and unstructured data - An integrated business intelligence framework. *Information Systems Management*, 25(2), 132–148. <https://doi.org/10.1080/10580530801941058>
- Boeing, G., & Waddell, P. (2017). New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings. *Journal of Planning Education and Research*, 37(4), 457–476. <https://doi.org/10.1177/0739456X16664789>
- Bordeleau, Fanny-Eve, S. U., Mosconi, Elaine, S. U., & Santa-Eulalia, De, S. U. (2018). Business Intelligence in Industry 4.0: State of the art and research opportunities. In *Proceedings of the 51st Hawaii International Conference on System Sciences* (Vol. 9, pp. 3944–3953). Retrieved from <http://hdl.handle.net/10125/50383>
- Castro, J. M. L. T. de. (2016). Tendências de Business Intelligence. Retrieved from <https://run.unl.pt/bitstream/10362/19434/1/TGI0065.pdf>
- Crunchbase Inc. (2019). The Crunchbase Data Difference. Retrieved August 2, 2019, from <https://about.crunchbase.com/products/the-crunchbase-difference/>
- Dedić, N., & Stanier, C. (2016). Measuring the success of changes to existing business intelligence solutions to improve business intelligence reporting. *Lecture Notes in Business Information Processing*, 268(0), 225–236. [https://doi.org/10.1007/978-3-319-49944-4\\_17](https://doi.org/10.1007/978-3-319-49944-4_17)
- Eckerson, W. W. (2010). *Performance Dashboards: Measuring, Monitoring, and Managing Your Business*. Wiley. Retrieved from <https://books.google.pt/books?id=daiXfV1jcakC>

- Elmasri, R., & Navathe, S. B. (1989). *Fundamentals of Database Systems*. Redwood City, CA, USA: Benjamin-Cummings Publishing Co., Inc.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3), 37. <https://doi.org/10.1609/aimag.v17i3.1230>
- Ferrari, A., & Russo, M. (2016). *Introducing Microsoft Power BI*. (R. Caperton, O. P. Russel, Diane, & O. P. Russel, Bob, Eds.). Redmond: Microsoft Press.
- Gartner. (2019). Analytics and Business Intelligence (ABI). Retrieved October 12, 2019, from <https://www.gartner.com/en/information-technology/glossary/business-intelligence-bi>
- Haddaway, N. R. (2016). The use of *web-scraping* software in searching for grey literature. *Grey Journal*, 11(February), 186–190.
- Han, J., & Kamber, M. (2006). *Data Mining: Concepts and Techniques*. *Soft Computing* (Vol. 54). <https://doi.org/10.1007/978-3-642-19721-5>
- Hirschey, J. K. (2014). Article 16 8-1-2014 Symbiotic Relationships: Pragmatic Acceptance of Data Scraping, 29 Berkeley Tech. *Berkeley Technology Law Journal*, 29. <https://doi.org/10.15779/Z38B39B>
- Höpken, W., Fuchs, M., Keil, D., & Lexhagen, M. (2015). Business intelligence for cross-process knowledge extraction at tourism destinations. *Information Technology and Tourism*, 15(2), 101–130. <https://doi.org/10.1007/s40558-015-0023-2>
- Imhoff, C., & White, C. (2011). Self-Service Business intelligence TDWI best practice report. *TDWI - Transforming Data With Inteligence*, 1–38. Retrieved from <http://triangleinformationmanagement.com/wp-content/uploads/2014/02/Self-Service-Business-Intelligence-empowering-users-to-generate-insights.pdf>
- Jensen, C. S., Pedersen, T. B., & Thomsen, C. (2010). *Multidimensional Databases and Data Warehousing*. *Synthesis Lectures on Data Management* (Vol. 2). <https://doi.org/10.2200/s00299ed1v01y201009dtm009>
- Kazi, L., Kazi, Z., & Radulovic, B. (2012). Data warehouse based evaluation of students' achievements in information systems education. *MIPRO 2012 - 35th International*

*Convention on Information and Communication Technology, Electronics and Microelectronics - Proceedings*, 1377–1382.

- Khan, M., Wu, X., Xu, X., & Dou, W. (2017). Big data challenges and opportunities in the hype of Industry 4.0. In *2017 IEEE International Conference on Communications (ICC)* (pp. 1–6). IEEE. <https://doi.org/10.1109/ICC.2017.7996801>
- Kimball, R., Ross, M., Becker, B., Mundy, J., & Thornthwaite, W. (2010). *The Kimball Group Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence*. Wiley. Retrieved from <https://books.google.pt/books?id=tCr5sMmkiPAC>
- Loshin, D. (2012). *Business Intelligence: The Savvy Manager's Guide*. Elsevier Science. Retrieved from <https://books.google.pt/books?id=L7SLNIS1ao8C>
- Loya, T., & Carden, G. (2018). *Business intelligence and analytics. Higher Education Strategy and Planning*. <https://doi.org/10.4324/9781315206455-12>
- Malladi, S. (2013). Adoption of Business Intelligence & Analytics in Organizations – An Empirical Study of Antecedents. *AMCIS 2013 Proceedings, 2016*, 1–11. Retrieved from <http://aisel.aisnet.org/amcis2013/BusinessIntelligence/GeneralPresentations/3>
- Matei, S. A., Russell, M. G., & Bertino, E. (2015). *Transparency in Social Media: Tools, Methods and Algorithms for Mediating Online Interactions*. Springer International Publishing. Retrieved from [https://books.google.pt/books?id=fx4\\_CgAAQBAJ](https://books.google.pt/books?id=fx4_CgAAQBAJ)
- McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., & Barton, D. (2012). Big data: the management revolution. *Harvard Business Review*, *90*(10), 61–67. <https://doi.org/00475394>
- Microsoft. (2019). Microsoft Power BI. Retrieved October 12, 2019, from <https://docs.microsoft.com/en-us/power-bi/guided-learning/index>
- Mircea, M. (2012). *Business Intelligence: Solution for Business Development*. IntechOpen. Retrieved from <https://books.google.pt/books?id=hPeZDwAAQBAJ>
- Mitchell, R. (2015). *Web Scraping with Python*. (S. St. Laurent & A. MacDonald, Eds.). O'Reilly Media, Inc.
- Moody, D. L., & Kortink, M. a R. (2003). From ER Models to Dimensional Models:

Bridging the Gap between OLTP and OLAP Design, Part I. *Business Intelligence Journal*, (June), 7–24.

Negash, S. (2015). *Business Intelligence.*, (January 2003).

Santos, M., & Ramos, I. (2009). *Business Intelligence: Tecnologias da Informação na Gestão de Conhecimento. FCA - Editora de Informática*, (August), 25. Retrieved from [http://repositorium.sdum.uminho.pt/bitstream/1822/6198/1/Resumo\\_Livro\\_BI\\_MYS\\_IR.pdf](http://repositorium.sdum.uminho.pt/bitstream/1822/6198/1/Resumo_Livro_BI_MYS_IR.pdf)

Shinde, S. R., & Sunjita. (2018). *Integration between Customer Relationship Management and Business Intelligence.*

Shollo, A. (2012). *The role of Business Intelligence in Organizational Memory support.* Retrieved from <http://repositorium.sdum.uminho.pt/bitstream/1822/26319/1/thesis.pdf>

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, 70, 263–286. <https://doi.org/10.1016/j.jbusres.2016.08.001>

Software Innovation Lab LLC. (2019). How It works - Data Miner. Retrieved July 31, 2019, from <https://data-miner.io/how-it-works>

Turban, E., Sharda, R., Aronson, J. E., & King, D. (2008). *Business Intelligence: A Managerial Approach.* Pearson Prentice Hall. Retrieved from <https://books.google.pt/books?id=NWPuAAAAMAAJ>

Vargiu, E., & Urru, M. (2012). Exploiting *web* scraping in a collaborative filtering- based approach to *web* advertising. *Artificial Intelligence Research*, 2(1), 44–54. <https://doi.org/10.5430/air.v2n1p44>

Wigmore, I., & Rouse, M. (2014). What is semi-structured data? - Definition from WhatIs.com. Retrieved September 13, 2017, from <http://whatis.techtarget.com/definition/semi-structured-data>

## ANEXO I

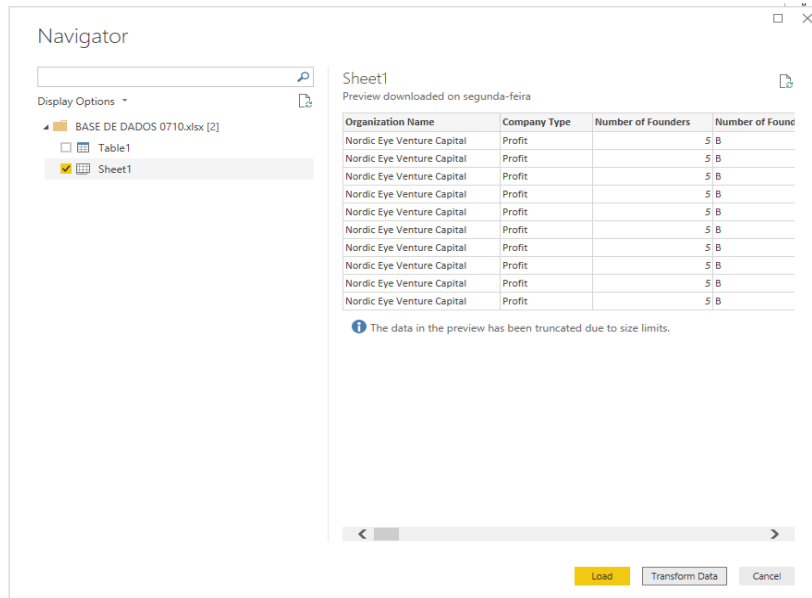


Figura XVI - Transformar Dados

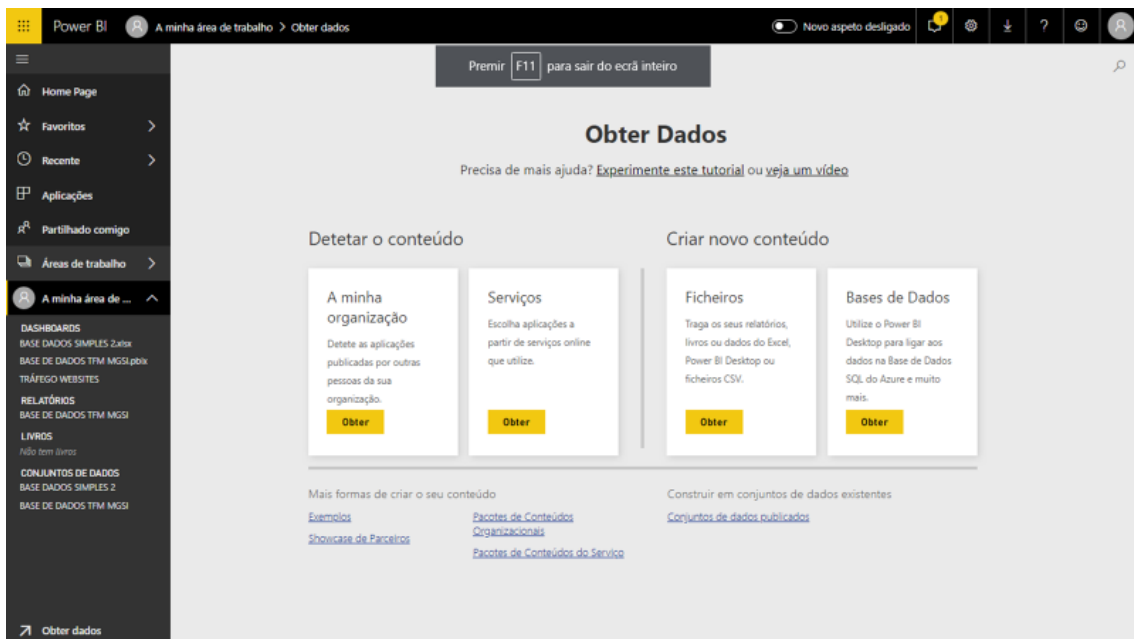


Figura XVII - Obter Dados ou Get Data

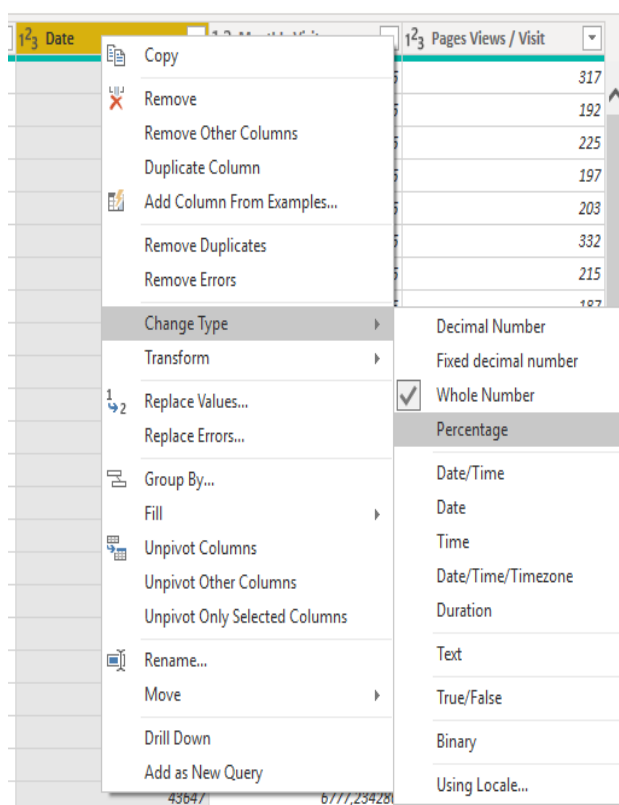
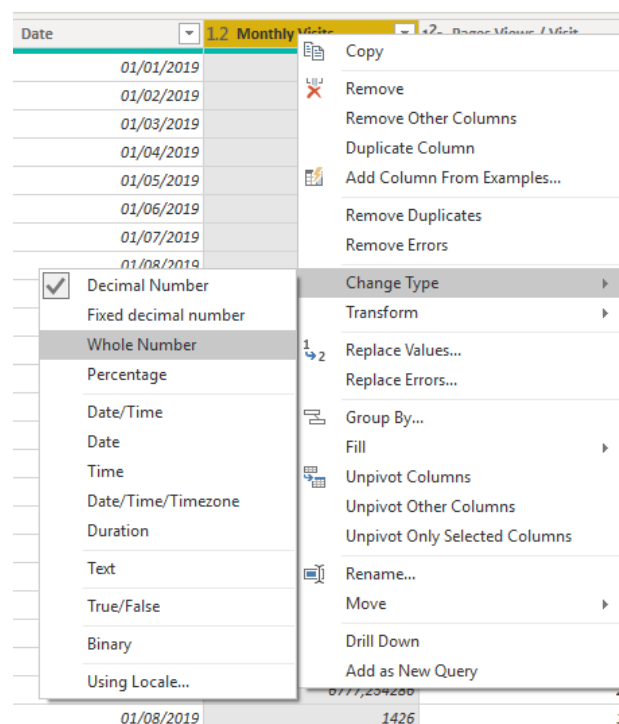


Figura XVIII - Mudar tipo de dados ou Change Data Type para "Data"





+

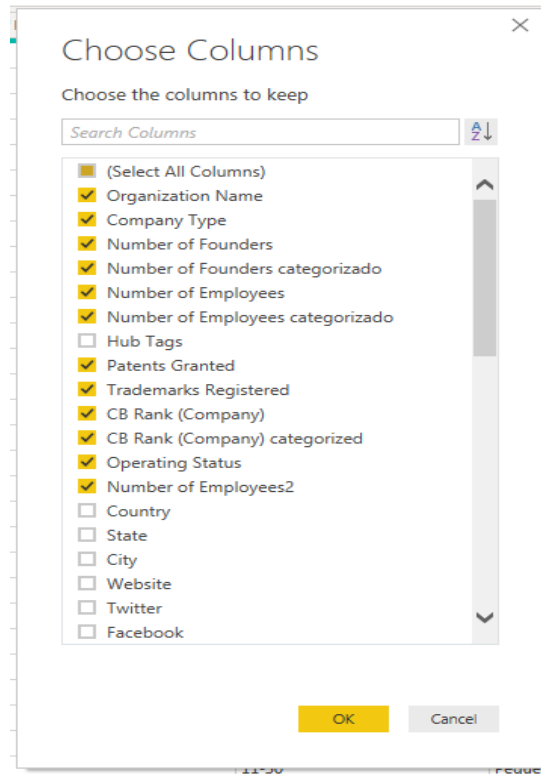


Figura XX - Escolher colunas para a tabela de dimensões

Operating Status	Number of Employees2	Index
Active		742
Active		752
Active		756
Active		783
Active		798
Active		799
Active		800
Active		807
Active		809

Figura XXI - Coluna Index depois de adicionada

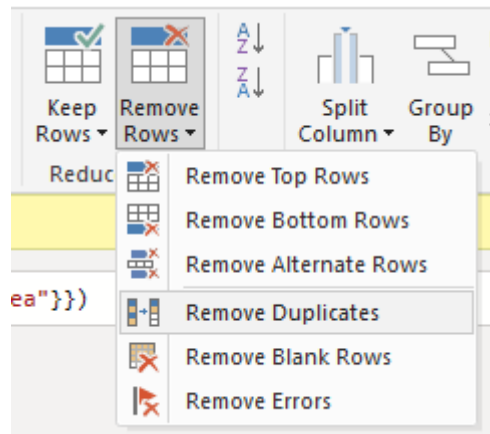


Figura XXII - Remover duplicados ou Remove Duplicates

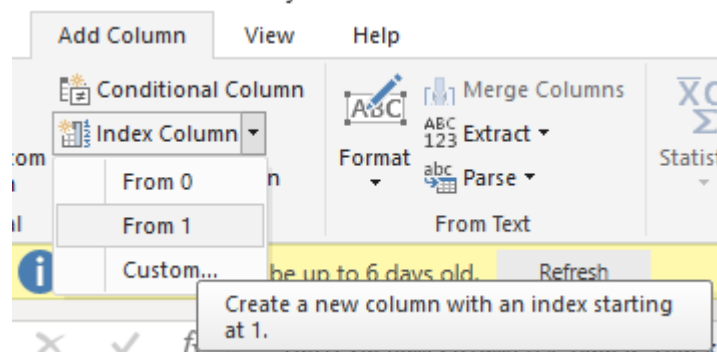


Figura XXIII - Criar Index

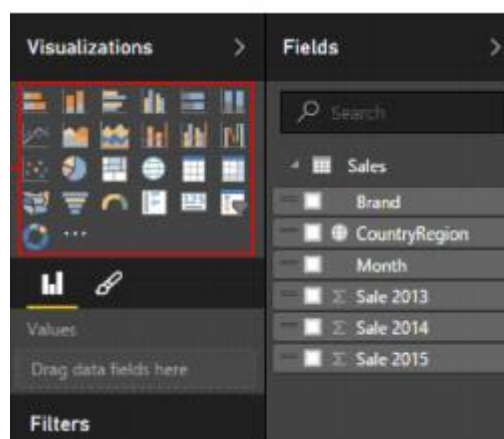


Figura XXIV - Visualizações com o

## Manage relationships

Active	From: Table (Column)	To: Table (Column)
<input checked="" type="checkbox"/>	Visitas Website (Area.1.SK_Area)	Area (SK_Area)
<input checked="" type="checkbox"/>	Visitas Website (Date.1.Index)	Date (Index)
<input checked="" type="checkbox"/>	Visitas Website (Empresa.Index)	Empresa (Index)
<input checked="" type="checkbox"/>	Visitas Website (Funding.Index)	Funding (Index)
<input checked="" type="checkbox"/>	Visitas Website (Localização.Index)	Localização (Index)
<input checked="" type="checkbox"/>	Visitas Website (Redes Sociais.Index)	Redes Sociais (Index)

New... Autodetect... Edit... Delete

Close

Figura XXV – Gerir as relações no Power BI



## ANEXO II

### Bounce Rate

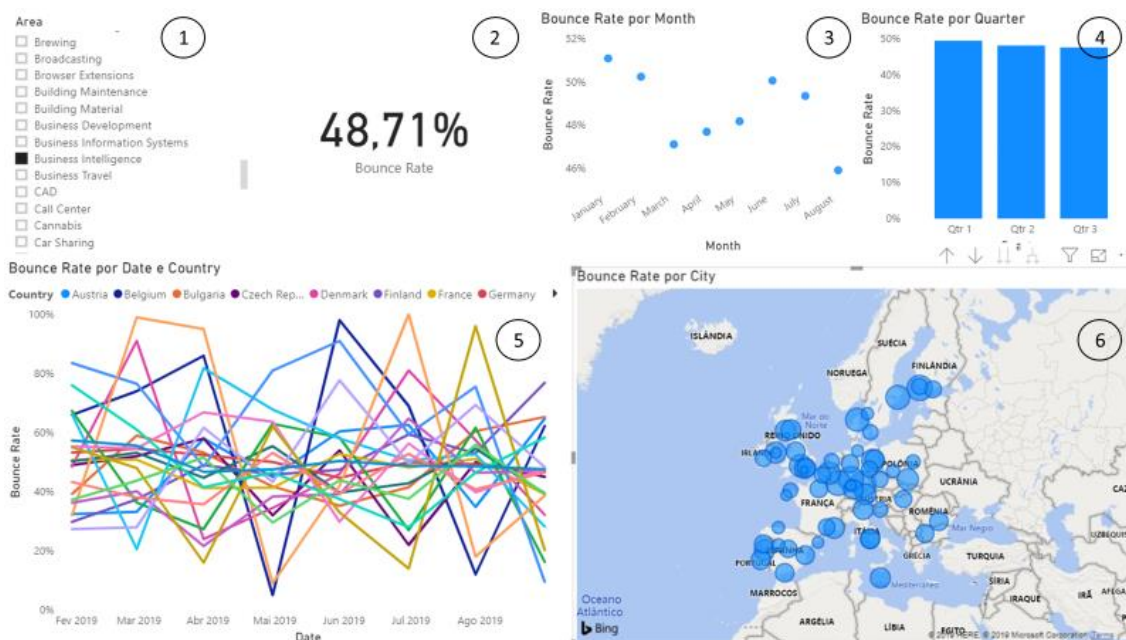


Figura XXVIII- Relatório de Bounce Rate

No ponto 1 é apresentada uma visualização do tipo *slicer* ou segmentação de dados que aplica um filtro que influencia as visualizações de toda a página, mediante a variável escolhida que, neste caso é “Business Intelligence”;

- No ponto 2 é apresentada uma visualização de tipo *card* que apresenta a média de *bounce rate*, apresentando valores de 48,71%, considerando a segmentação de dados escolhida anteriormente. Sendo este o elemento que resume a variável num ponto principal para inserir também no *dashboard*;
- No ponto 3 é possível observar um *scatter chart* ou um gráfico de dispersão que mostra os dados em forma de intersecção de dados, neste caso comparando a percentagem de *bounce rate* com a dimensão *Date*. Neste caso é possível constatar que as empresas na “Area” de “Business Intelligence”, têm uma crescente percentagem de *bounce rate*, exceto em Agosto onde encontraram um súbito declínio;
- No ponto 4, é apresentado um *clustered column chart* que ajuda a representar a percentagem média de *bounce rate* por trimestre, onde podemos observar um declínio desta variável ao longo do tempo;

- No ponto 5, encontramos um gráfico de linhas que é um serie de pontos de dados representados por pontos e ligados com linhas. Neste caso, o gráfico mostra a variação da percentagem de *bounce rate* ao longo do tempo e segmentada por país o que nos permite ter uma visão espacial e temporal desta variável;
- Por fim, no ponto 6, os dados das visitas mensais e os dados hierarquizados da dimensão “Localização”, são comparados. Para a criação deste tipo de gráficos é necessário ter atenção à atribuição de uma categoria (*Country, State, City*). Neste relatório, esta visualização mostra a distribuição por país das maiores percentagens de *bounce rate*.

### Visit Duration (min)

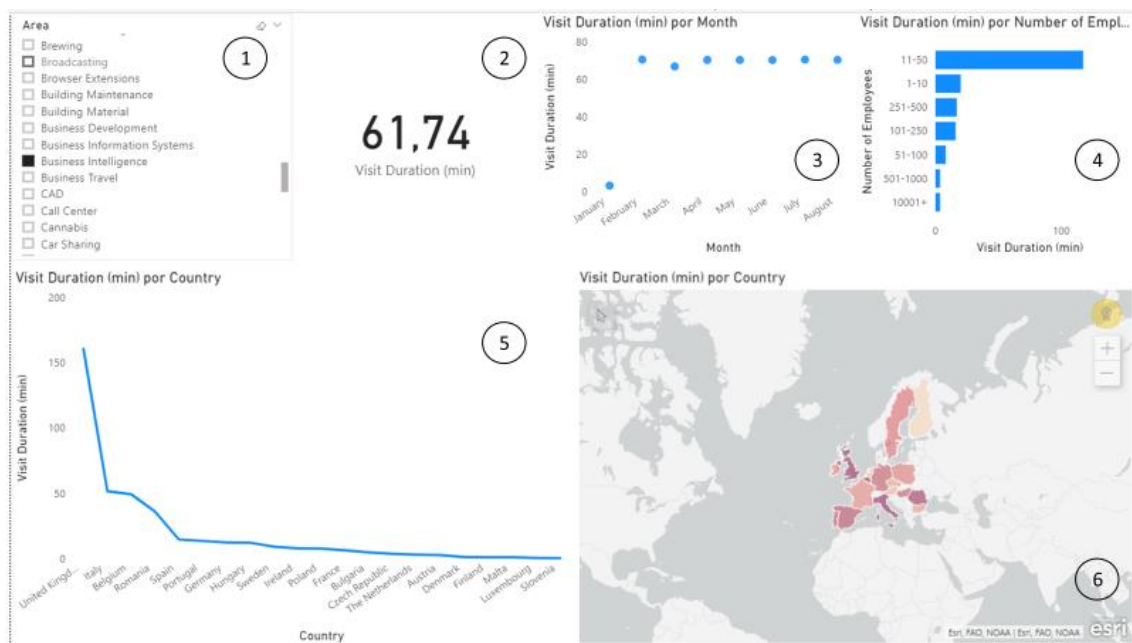


Figura XXIX - Relatório Visit Duration (min)

- No ponto 1 é apresentada uma visualização do tipo *slicer* ou segmentação de dados que aplica um filtro que influencia as visualizações de toda a página, mediante a variável escolhida que, neste caso é “Business Intelligence”;
- No ponto 2 é apresentada uma visualização de tipo *card* que apresenta a média de *visit duration (min)*, apresentando valores de 61,74 min, considerando a

segmentação de dados escolhida anteriormente. Sendo este o elemento que resume a variável num ponto principal para inserir também no *dashboard*;

- No ponto 3 é possível observar um *scatter chart* ou um gráfico de dispersão que mostra os dados em forma de intersecção de dados, neste caso comparando a o número médio de minutos de visita com a dimensão hierarquizada *Date*. Neste caso é possível constatar que as empresas na “Area” de “Business Intelligence”, têm valores médio de duração de visita bastante similares ao longo dos meses;
- No ponto 4, é apresentado um *clustered column chart* que ajuda a representar o valor médio de tempo de visita por classificação do número de empregados da empresa, onde podemos observar que os utilizadores passam mais tempo nos *websites* de empresas que têm entre 11 e 50 empregados;
- No ponto 5, encontramos um gráfico de linhas que é um serie de pontos de dados representados por pontos e ligados com linhas. Neste caso, o gráfico mostra a variação do valor médio do tempo de visita ao *website* ao longo do tempo e segmentada por país o que nos permite ter uma visão espacial e temporal desta variável, podendo ser possível observar rapidamente quais os países quem têm valores de tempo de visita mais elevados;
- Por fim, no ponto 6, os dados do tempo médio de visita e os dados hierarquizados da dimensão “Localização”, são comparados. Para a criação deste tipo de gráficos é necessário ter atenção à atribuição de uma categoria (*Country, State, City*). Neste relatório, esta visualização mostra a distribuição por país, recorrendo a um mapa de calor, dos valores médios por país.

## Pages Views / Visit

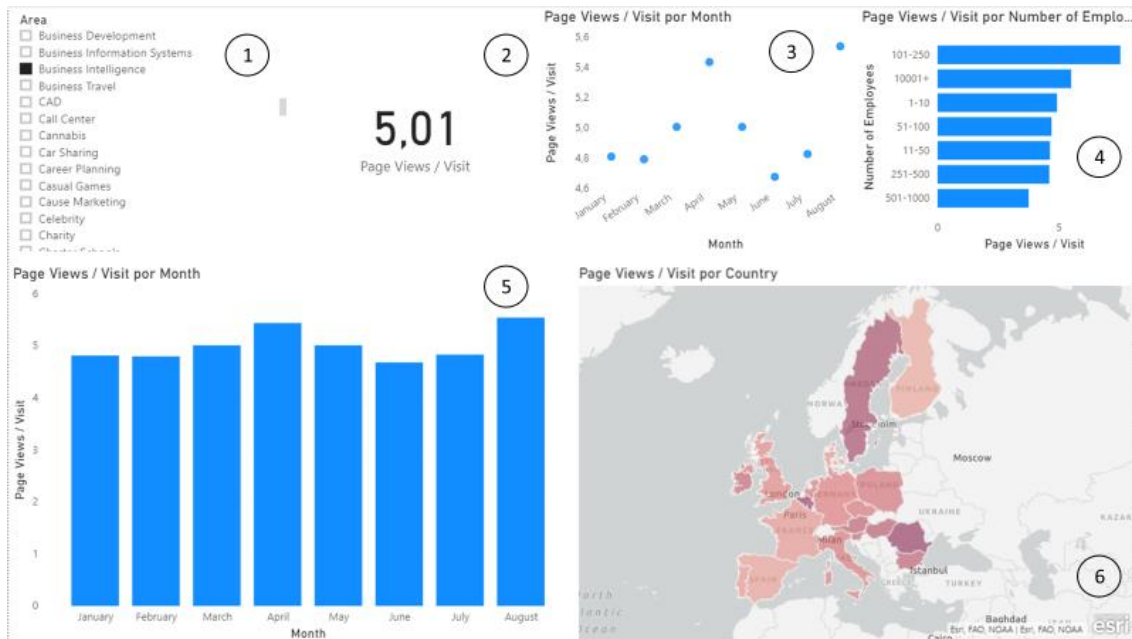


Figura XXX - Relatório de Pages Views / Visit

- No ponto 1 é apresentada uma visualização do tipo *slicer* ou segmentação de dados que aplica um filtro que influencia as visualizações de toda a página, mediante a variável escolhida que, neste caso é “Business Intelligence”;
- No ponto 2 é apresentada uma visualização de tipo *card* que apresenta a média de *pages views / visit*, apresentando valores de 5,01 páginas visitadas em média por mês, considerando a segmentação de dados escolhida anteriormente. Sendo este o elemento que resume a variável num ponto principal para inserir também no *dashboard*;
- No ponto 3 é possível observar um *scatter chart* ou um gráfico de dispersão que mostra os dados em forma de intersecção de dados, neste caso comparando a páginas visitadas em média com a dimensão hierarquizada *Date*. Neste caso é possível constatar que as empresas na “Area” de “Business Intelligence”, registraram um pico de páginas visitadas em média por mês em Abril;
- No ponto 4, é apresentado um *clustered column chart* que ajuda a representar as páginas visitadas em média por classificação do número de empregados da empresa, onde podemos observar que os utilizadores visitam mais páginas de *websites* de empresas que têm entre 101 e 250 empregados;



- No ponto 5, encontramos um gráfico de linhas ou um *line chart* que é uma série de pontos de dados representados por pontos e ligados com linhas. Neste caso, o gráfico mostra a variação do número de páginas visitadas em média por mês e segmentada por país o que nos permite ter uma visão espacial e temporal desta variável, podendo ser possível observar rapidamente quais os países quem têm valores de tempo de visita mais elevados;
- Por fim, no ponto 6, os dados do tempo médio de visita e os dados hierarquizados da dimensão “Localização”, são comparados. Para a criação deste tipo de gráficos é necessário ter atenção à atribuição de uma categoria (*Country, State, City*). Neste relatório, esta visualização mostra a distribuição por país, recorrendo a um mapa de calor, dos valores médios por país.