Universidade do Minho
Escola de Engenharia

Luciana Gomes Machado

Data Science – Metodologias e Ferramentas

Dissertação do Mestrado Integrado em Engenharia e Gestão de Sistemas de Informação (MiEGSI)

Trabalho efetuado sob a orientação do
Professor Doutor Carlos Filipe Portela
e do
Professor Doutor Jorge Oliveira e Sá

Agosto de 2022

RESUMO

Esta pré-dissertação enquadra-se num projeto de dissertação do mestrado integrado em Engenharia e Gestão de Sistemas de Informação na Universidade do Minho, sendo o tema da dissertação "*Data Science* – Metodologias e Ferramentas". Este tema é fruto de uma relação entre o professor Filipe Portela, o *Intelligent Data Systems Group – Algoritmi Research Centre* e a Escola de Engenharia da Universidade do Minho. Várias são as metodologias e ferramentas de *Data Science* que as entidades nomeadas anteriormente tem contacto diariamente, no entanto o apoio à decisão em tempo-real com base em dados gerados no momento é visto por estas entidades e por muitas organizações como um fator decisivo para o sucesso na tomada de uma decisão. Devido à complexidade, quantidade e diversidade dos dados existentes atualmente, têm surgido um conjunto de metodologias e ferramentas *Data Science, Big Data, Machine-Learning, Data Mining,* ou *Business Intelligence* que ajudam na implementação das soluções. Este tema surge, fundamentalmente, com o propósito de levantar o maior número de metodologias e ferramentas, de modo a compará-las, relacioná-las e proporcionar aos interessados aquela(s) que mais abrangente é e que mais se destaca no mercado. O trabalho desenvolvido seguirá a metodologia Benchmarking de modo a fazer a comparação das metodologias e ferramentas. Ainda no presente documento são identificados os objetivos, as motivações e o enquadramento do tema. Após isto, são definidas as metodologias a serem abordadas no desenvolvimento da dissertação, bem como todo o planeamento do projeto e os possíveis riscos, de modo a precaver falhas.

ABSTRACT

This pre-dissertation is part of a dissertation project of the integrated master's degree in Engineering and Management of Information Systems at the University of Minho, being the theme of the dissertation "Data Science – Methodologies and Tools". This theme is the result of a relationship between Professor Filipe Portela, the Intelligent Data Systems Group – Algoritmi Research Center and the School of Engineering of the University of Minho. There are several Data Science methodologies and tools that the aforementioned entities have daily contact with, however real-time decision support based on data generated at the moment is seen by these entities and by many organizations as a decisive factor for the success in making a decision. Due to the complexity, quantity and diversity of data currently available, a set of Data Science, Big Data, Machine-Learning, Data Mining, or Business Intelligence methodologies and tools have emerged that help in the implementation of solutions. This theme arises, fundamentally, with the purpose of raising the largest number of methodologies and tools, in order to compare them, relate them and provide interested parties with the one(s) that are more comprehensive and that stand out in the market. The work developed will follow the Benchmarking methodology in order to compare methodologies and tools. Also in this document, the objectives, motivations and framework of the theme are identified. After that, the methodologies to be addressed in the development of the dissertation are defined, as well as all the project planning and possible risks, in order to prevent failures.

Índice

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

1. INTRODUÇÃO

Na introdução está presente toda a contextualização, enquadramento e motivação para a realização da dissertação, bem como os seus objetivos e resultados esperados, e ainda a estrutura do presente documento.

1.1. Contextualização

Atualmente vivemos num mundo que está em mudança constante, resultante principalmente da evolução tecnológica. Diariamente a população mundial é bombardeada com tecnologia, desde o momento em que acorda com o despertador até ao que adormece a ver uma série. Mas para que isto não aconteça, ou seja, para que ninguém adormeça a ver uma série, é que a informação se torna importante. Só com a informação certa é que é possível analisar os gostos pessoais e aconselhar conteúdos interessantes a cada um.

Como disse Ana Caroline da Rosa e Calos de Oliveira no artigo "O ERP como ferramenta de Gestão de *Stock* nas franquias de *fast-food* na cidade de Itaquaquecetuba", a "Informação é um conjunto de dados que possui valor"(Caroline et al., 2020). Através desta é possível verificar diversos detalhes que podem ser usados pela sociedade em geral para retirar conclusões em prol dos seus objetivos.

As responsáveis pela obtenção, proteção, gestão, armazenamento e uso das informações são as Tecnologias de Informação.

O uso massivo das Tecnologias de informação veio fomentar a produção de dados, estes são produzidos e armazenados de modo a, posteriormente, serem usados na obtenção de informações.

Data Science está cada vez mais presente no quotidiano da população mundial, dessa forma, esta ciência está constantemente em alteração e evolução. Obter informações através da extração de grandes quantidades de dados tornou-se uma necessidade, no entanto a compreensão e análise dos mesmos é algo complexo. Para além de complexa, a existência de grandes quantidades de dados embarga, também, mais responsabilidades para as organizações e individuais. Estes não têm capacidade para lidar com os sete V's, ou seja, o

volume, velocidade, variedade, valor, veracidade, volatilidade e a visualização dos dados que estão sob. Desse modo, têm de recorrer a ferramentas que lhes forneçam ajuda suficiente para retirar conclusões, mas ainda assim advém um problema. E é aqui que surge esta dissertação, a presente pesquisa tem como problema o seguinte. Acontece que existem variadas ferramentas, porém muitas organizações de desenvolvimento de soluções de dados não têm noção de qual delas usar, acabando por escolher uma sem qualquer critério ou, tendo um critério, recorrem ao uso de duas ou mais para conseguir dar vazão a tarefas diferentes. Para contornar isto é necessário o levantamento de todas metodologias de *Data Science*, de forma a criar diretrizes e fazer uma junção perfeita com a ferramenta que melhor se encaixa. Para além disso, é de salientar a importância que uma ferramenta que se conseguiria adequar a várias metodologias teria.

1.2. Enquadramento e Motivação

O apoio à decisão em tempo-real com base em dados gerados no momento é visto pelas organizações como um fator decisivo para o sucesso na tomada de uma decisão, até porque quanto mais rápido esse processo é feito, mais as organizações conseguem prever e antecipar futuros problemas. No entanto, as organizações têm dificuldade em fazer uma análise desses dados em tempo real, devido à complexidade, quantidade e diversidade dos mesmos. Para dar resposta a esta problemática têm surgido um conjunto de metodologias e ferramentas de *Data Science, Big Data, Machine-Learning, Data Mining, Business Intelligence*, entre outras, que ajudam na implementação das soluções. Este tema tem como principal objetivo e propósito o levantamento das metodologias e de ferramentas existentes e a realização de uma análise comparativa entre elas, de modo a encontrar a ferramenta que melhor se adequa a cada metodologia e, também, a ferramenta mais completa para todos os utilizadores. Se não for possível encontrar uma que satisfaça todos os utilizadores, ou seja, que cumpra o propósito mencionado anteriormente, o objetivo passa por idealizar uma ferramenta que retire o melhor de cada uma analisada previamente, criando a ferramenta perfeita.

Metodologia segundo a IBM no artigo "Metodologia de Base para Ciência de Dados" é "uma estratégia geral que orienta os processos e atividades dentro de um determinado

domínio"(Rollins, 2015), e ferramentas, segundo o dicionário Priberam são "utensílios usados para atingir um fim", no caso do *Data Science*, o meu conhecimento diz-me que ferramenta é o instrumento de trabalho ou objeto intermediário usado para recolher, analisar, visualizar dados, entre outros.

A motivação para a realização desta dissertação surge do interesse pela área de *Data Science*, pelo gosto do *benchmarking* e, também, pelo desconhecimento de programação, o que fez com que ingressar num tema teórico fosse a melhor solução. Esta preferência agregada à dificuldade de várias organizações encontrarem a ferramenta indicada para o seu verdadeiro sucesso e, atualmente, a importância grandiosa dos dados contribuíram como fator motivacional para a realização desta dissertação.

1.2.1. Data Science

Data Science segundo Atilas Ferreira de Paiva no artigo "Aplicação de *data science* como ferramenta de apoio a tomada de decisão orientada por dados" resume-se à "ciência que procura extrair informações, conhecimentos e *insights* a partir das mais diversas fontes de dados, através da recolha, do processamento e da análise dos dados" (Paiva & Branco, 2018). Esta ciência é aplicada maioritariamente a negócios, de forma a captar novos clientes, a descobrir fraudes ou para encontrar as forças e as ameaças de qualquer organização.

As principais disciplinas do *Data Science* segundo o artigo "Aplicação do *Big Data* e *Data Science* no âmbito jurídico" escrito por Nicoly dos santos são "a estatística ou matemática aplicada, juntamente com a ciência e também a área dos negócios. Contendo estas três áreas, é possível analisar e gerar informações com base no apoio na tomada de decisões" (Santos, 2022).

1.2.2. Principais tecnologias aplicadas no *Data Science:*

Em conformidade com o que foi escrito na secção anterior, existem diferentes tipos de data science, nomeadamente:

a) Data Analytics

Data Analytics é o processo de analisar conjuntos de dados de modo a encontrar tendências e tirar conclusões através das mesmas. Cada vez surgem mais ferramentas e

softwares para auxiliar este processo. As metodologias e técnicas de *Data Analytics* são cada vez mais usadas pelas organizações e por pesquisadores de forma a tomarem melhores as suas decisões e de modo a verificarem modelos e teorias científicas, respetivamente.

b) Data Mining

De acordo com o artigo "Descoberta do Conhecimento em Banco de Dados: Mineração de dados Educacionais" escrito por Rosana Massahud, Simão Assaid e Cristhian de Carvalho, *Data Mining* é o "processo de extrair e explorar grandes quantidades de dados de modo a encontrar padrões consistentes" (Assaid et al., n.d.).

O *Data Mining* surgiu no início dos anos 80 quando os profissionais das organizações começaram a ter uma maior preocupação com a quantidade de dados informáticos inúteis dentro da empresa. Nesta época, *Data Mining* consistia basicamente na extração de informação de bases de dados enormes da forma mais automática possível. Hoje em dia, esta procura essencialmente encontrar padrões.

c) Business Intelligence

Segundo Anuj Tripathi, Teena Bagga e Rashmi Aggarwal no artigo "*Strategic impact of business intelligence: A review of literature", Business Intelligence* é "processo que recolhe, organiza, analisa, partilha e monitoriza os dados que dão suporte à gestão de negócio, juntamente com um conjunto de técnicas que auxiliam a transformação de dados brutos em informações relevantes e úteis para a organização" (Tripathi et al., 2020).

Apesar do *Business Intelligence* ter surgido recentemente, este descende dos primórdios da sociedade. O mundo dos negócios já tem alguns anos e há quem diga que se iniciou na pré-história, quando ainda só existia o setor primário e as trocas que se faziam eram entre sal e batatas. Mais tarde, com o aparecimento de outros setores, os negócios começaram a crescer, e os dados começaram também a aumentar. Estes inicialmente eram mantidos em papel, mas ao longo do tempo, com o avançar tecnológico, começaram a ser armazenados em bases de dados. Assim as organizações conseguiam reservar, agrupar, analisar e tratar múltiplos dados históricos, de modo a extrair informações e tomar melhores decisões. Assim sendo, a conversão dos dados para informação de modo a dar suporte às decisões tomadas pela organização é nada mais nada menos do que *Business Intelligence*.

d) Data Warehousing

Data Warehouse é o processo de armazenamento de grandes quantidades de dados não voláteis e informações relativas às atividades de uma dada organização.

De acordo com o livro "*The data warehouse toolkit: the complete guide to dimensional modeling*", escrito por Margy Ross e Ralph Kimball, o *Data Warehousing* "surgiu nos anos 1960 quando General Mills e Dartmouth College, num projeto de pesquisa comum, desenvolveram os termos "dimensões" e "factos". Na década de 1970, Bill Inmon começou a definir e a discutir sobre Data Warehouse. No final da década de 1980 dois investigadores, Barry Devlin e Paul Murphy, introduziram o "*Business Data Warehouse*" que tinha como propósito fornecer um modelo de arquitetura para o fluxo de dados de sistemas operativos para suporte à decisão" (Kimball & Ross, 2002).

e) Big Data

Em conformidade com as palavras usadas por Oussous Ahmed no artigo "Tecnologias *de big data*: uma pesquisa", para o *Journal of King Saud University-Computer and Information Sciences, Big Data* é o "processo que estuda o tratamento, análise e obtenção de conjuntos de dados em quantidades superiores ao normal e que, desse modo, não é possível analisá-los pelos meios ditos normais" (Oussous et al., 2018).

Consoante as palavras de Trevor Barnes no artigo "*Big data, little history",* o *Big Data* "iniciou-se em 1663 por Jonh Graunt, quando este decidiu obter informações de grandes quantidades de dados através de diferentes fontes para estudar a pandemia existente na Europa, a peste negra. Em 1890, durante a realização dos Censos nos Estados Unidos da América, foram usados os primeiros equipamentos para processar dados, este processo demorava cerca de seis semanas. Já no século XX, surgiram os primeiros sistemas de armazenamento de informações e, durante a Segunda Guerra Mundial, foi produzida a primeira máquina digital de processamento de dados, nomeada de Colossus, esta foi desenvolvida pelos britânicos de modo a decifrar os códigos nazis. Em 1989, surgiu o *World Wide Web*, conhecido como WWW, este foi criado por Tim Berners-Lee com o propósito de facilitar a troca de informações entre pessoas, tendo, posteriormente, revolucionado a forma como os dados são gerados e quantidade de dados criados. O termo *Big Data* começou a ser utilizado em 1997, sendo que só foi oficialmente nomeado em 2005" (Barnes, 2013).

f) Machine Learning

De acordo com o artigo escrito por Batta Mahesh para o *Journal of Science and Research, "Machine Learning* é um ramo dentro da inteligência artificial e da ciência computacional que se foca na utilização de dados e algoritmos de modo a imitar a aprendizagem humana, tornando-se cada vez mais semelhante" (Batta, 2020).

Desde o início da existência humana que se utilizam várias ferramentas para auxiliar na realização das mais diversas tarefas. A criatividade do ser humano não tem limites e esta, alinhada aos materiais fornecidos pelo ambiente que os rodeia, originou a criação de diferentes máquinas. Estas máquinas surgiram efetivamente para facilitar a vida humana atendendo às mais diversificadas necessidades humanas, desde o auxílio na agricultura ao auxílio em viagens. O *Machine Learning* não se abstém. Este veio ensinar o humano a lidar com dados de forma mais eficiente, como por exemplo, depois de visualizar os dados, interpretá-los mais facilmente.

g) Data Visualization

Conforme Manuela Aparicio e Carlos Costa no artigo "*Data visualization*", *"Data Visualization* é a representação gráfica dos dados. Esta tem o propósito de fornecer uma maneira acessível para entender tendências, discrepâncias e padrões nos dados através de tabelas, mapas, gráficos e outros" (Aparicio & Costa, 2014).

Desde sempre que a mente humana é extremamente visual. Segundo o mesmo artigo nomeado anteriormente "ainda na pré-história, através da arte rupestre, isto é, das representações visualmente artísticas realizadas em cavernas, tetos, rochas e paredes, viam-se representados animais, plantas, pessoas e até alguns sinais abstratos. Na altura, e ainda agora, se considera difícil a interpretação destes, mas pensa-se que estas representam estratégias de caça ou a quantificação dos animais capturados. Também mais tarde, mais ou menos em 2500 a.C., começaram a ser usados mapas para representar a riqueza de cada reino. Igualmente, no século XVI, nos descobrimentos portugueses, a riqueza e os territórios dos reinos foram representados através do mapa Cantino. Mais tarde, já no século XVIII, William Playfair, um engenheiro e economista escocês, criou diversos tipos de diagramas para representar informações estatísticas. Mais atualmente, nos anos 70 surgiram os primeiros infográficos em jornais e revistas com o objetivo de sintetizar a informação" (Aparicio & Costa,

2014). E assim surgiram as representações gráficas e visuais dos dados, de maneira a simplificar a interpretação dos mesmos e a tirar melhores conclusões.

1.3. Objetivos e Resultados esperados

No âmbito deste estudo, pode-se identificar como principal questão de investigação à qual esta dissertação procura dar resposta a seguinte:

- Existe alguma ferramenta totalmente abrangente que se consiga adaptar a uma metodologia que permite ajudar no desenvolvimento de projetos de *Data Science* com componentes analíticas e preditivas?

De modo a responder a esta questão surge como objetivo principal a descoberta da metodologia e da ferramenta mais abrangente o levantamento e a análise comparativa das metodologias que melhor permita ajudar no desenvolvimento de projeto *Data Science* e, consequentemente, a descoberta da ferramenta mais completa e que faça um par perfeito com esta. Assim sendo propõe-se como objetivos específicos desta dissertação:

- Levantar todas as soluções já existentes acerca das metodologias, acerca das ferramentas de *Data Science* e acerca das duas em comum;
- Levantar todas as ferramentas de *Data Science, Big Data, Machine-Learning, Data Mining, Business Intelligence, Data Analytics, Data Warehousing e Data Visualization* já existentes, bem como a descrição de cada uma;
- Realizar uma matriz comparativa entre todas as ferramentas de *Data Science*, através de um *Benchmarking*;
- Identificar os pontos fulcrais de uma ferramenta de *Data Science*;
- Levantar todas as metodologias de *Data Science* que ajudem na implementação das soluções;
- Realizar uma matriz comparativa entre todas as metodologias de *Data Science*, através de um *Benchmarking*;
- Identificar as etapas obrigatórias de uma metodologia de *Data Science*.

Como resultado espera-se a identificação e o reconhecimento da metodologia e da ferramenta mais completa, caso esta não exista será proposta uma que reúna o melhor de todas as metodologias e ferramentas analisadas previamente.

1.4. Estrutura do documento

O presente documento está estruturado em 7 Capítulos e 1 Anexo, explicados de seguida:

- **Introdução** – Neste capítulo é apresentado ao leitor a contextualização e o enquadramento do tema estudado, bem como as motivações para o mesmo e os objetivos e resultados a atingir;
- **Revisão de Literatura** – Neste ponto é definida a estratégia de pesquisa e todos os temas abordados na dissertação. Também aqui será apresentado o estado da arte através de exemplos, projetos, artigos e casos concretos já existentes de acordo com o tema;
- **Abordagem Metodológica, Materiais e Métodos** – Neste capítulo são apresentadas e descritas todas as metodologias de investigação, bem como todas as ferramentas ou algoritmos que se estão ou se irão utilizar ao longo da dissertação;
- **Trabalho realizado** – Neste capítulo é feita uma apresentação do trabalho desenvolvido e dos resultados atingidos;
- **Plano de Atividades** – Será efetuada uma descrição detalhada do planeamento de toda a dissertação, incluindo o tempo estimado para cada tarefa, assim como uma tabela de riscos, para prevenir quaisquer problemas que possam influenciar negativamente o projeto;
- **Conclusão** – Tal como o nome indica, este capítulo destina-se a concluir tudo o que foi abordado anteriormente;
- **Referências** – Este capítulo contém todas as referências bibliográficas utilizadas para a realização do presente documento.
- **Anexo** – Este capítulo procura descrever aos leitores cada uma das ferramentas abordadas na Secção 4.

2. REVISÃO DE LITERATURA

Neste segundo capítulo é definida a estratégia de pesquisa e todos os temas abordados na dissertação. Também aqui é apresentado o estado da arte através de exemplos, projetos, artigos e casos concretos já existentes de acordo com o tema.

2.1. Estratégia de Pesquisa

No decorrer do desenvolvimento do presente documento foram usadas múltiplas plataformas como motor de pesquisa. Estas dão apoio a este documento através de artigos científicos publicados, livros ou capítulos de livros que, de alguma forma, foram relevantes para o tema. De entre estas plataformas as que mais se distinguem são:

- *Google;*
- *Google Scholar;*
- *Science Direct;*
- *Springer.*

Existe um leque de palavras-chave, termos e expressões que, regularmente, são usados/as como base de pesquisa. As seguintes destacam-se como as mais utilizadas:

- *Data Science;*
- *Data Science Methodology;*
- *Data Science Tools.*

Tal como estes termos, também os mesmos foram traduzidos à língua portuguesa e pesquisados, no entanto não retribuíram resultados tão relevantes para a realização da dissertação.

Antes de abrir qualquer livro ou artigo o primeiro fator de escolha fora o título e se este se enquadrava com o conteúdo procurado, posteriormente, outro requisito foi o ano em que fora publicado, sendo que na maioria dos casos se optaria por estes terem sete ou menos anos, a menos que fosse realmente importante saber os pontos de vista com mais de sete anos.

O método posterior de escolha dos artigos científicos e livros a ler foi a leitura do *Abstract* e da Introdução de cada um deles. A leitura do *Abstract* é fundamental e obrigatória para que haja uma noção do conteúdo abordado ao longo do artigo ou livro. A leitura da Introdução ocorre somente quando, para além do tema, é necessária uma metodologia ou um processo concreto e existe a necessidade entender se está em uso ou não naquele documento.

Com o decorrer do desenvolvimento da presente dissertação vários foram os artigos, projetos e livros lidos de modo a que estes auxiliassem a realização da mesma, mas também para perceber o que já existia e/ou ainda poderia ser feito. São vários os artigos existentes acerca de *Data Science*, desde temas como "*Data Science and Prediction*" até "*Data Science for Business*", no entanto o que é especificamente procurado neste subcapítulo é a existência de qualquer material que tenha uma junção entre *Data Science* e as suas metodologias, *Data Science* e as suas ferramentas, e ainda se já existe algum material que remeta para aquele que é o tema deste documento, ou seja, *Data Science*, as suas metodologias e ferramentas.

2.2. Data Analytics

Após a descrição de *Data Analytics* na Secção 1.2.2., eis que existem quatro tipos principais de análise de dados, nomeadamente, a análise descritiva, a análise diagnóstica, a análise preditiva e a análise prescritiva. Cada uma destes tipos tem um propósito diferente, sendo que a:

- **Análise descritiva:** De acordo com o artigo "Inteligência artificial e *big data* no entretenimento televisivo: aplicação e regulação na União Europeia", José Cordeiro, o autor, diz que análise descritiva caracteriza a fase preliminar de processamento dos dados e tem como objetivo fornecer probabilidades e tendências futuras dando ideias sobre o que pode acontecer no futuro" (Sotero, 2022), portanto é nesta análise que, geralmente, são feitas as primeiras manipulações que têm como propósito resumir, sumarizar e explorar o comportamento dos dados. Através do desenvolvimento de indicadores-chave de desempenho, isto é, KPIs, a análise descritiva procura identificar fraquezas ou forças;

- **Análise diagnóstica**: Segundo o mesmo artigo, análise diagnóstica "é utilizada para identificar possíveis causas de acontecimento de determinado evento, respondendo à questão "porque é que aconteceu?"" (Sotero, 2022), esta análise concentra-se no que já aconteceu tal como a anterior. Esta foca-se no uso de dados passados para criar uma relação onde é possível identificar processos baseados em probabilidades, sendo esta responsável por traçar um perfil do comportamento do consumidor;
- **Análise preditiva**: Em concordância com o mesmo artigo, José Cordeiro diz que a análise preditiva "em oposto à anterior, permite prever situações futuras, com recurso à data passada e atual para formular cenários" (Sotero, 2022). Esta procura responder a perguntas futuras. Através do uso de dados históricos, esta concentra-se na identificação de tendências e em perceber se estas se repetirão ou não. Esta análise normalmente inclui técnicas de estatística e de *Machine Learning*;
- **Análise prescritiva**: Ainda com base no artigo mencionado anteriormente, a análise prescritiva "tem como objetivo de encontrar a ação certa a tomar" (Sotero, 2022). Esta procura informar o que deve ser efetivamente feito através de *insights* da análise preditiva. Esta utiliza ferramentas estatísticas alinhadas à gestão de negócios de forma a recomendar ações que devem ser otimizadas, mudando e melhorando o futuro das organizações.

2.3. Business Intelligence

Na Secção 1.2.2. encontra-se descrito o que é *Business Intelligence*. Nesta secção este tema será abordado mais aprofundadamente.

2.3.1. Componentes de um *Business Intelligence*

Segundo Negash e Gray, autores do livro "*Business intelligence"* (Negash, 2004) declaram que os componentes essenciais de qualquer *Business Intelligence* são:

- "Armazenamento de dados em tempo real;
- *Data Mining*;
- Descoberta automática de anomalias ou exceções;

- Fluxo de trabalho continuo;
- Aprendizagem automática;
- Sistemas de informação geográfica;
- Visualização de dados".

2.3.2. Arquiteturas de um *Business Intelligence*

Ainda segundo o livro abordado na Secção 2.3.1., existem dois tipos de arquitetura de *Business Intelligence,* nomeadas de Arquitetura para Dados Estruturados e Arquitetura para Dados Semi-Estruturados.

a) Arquitetura para Dados Estruturados

Negash e Gray dizem que este tipo de arquitetura é normalmente usada em "*data centers* estruturados num data warehouse" (Negash, 2004), e tal como a Figura 1 indica, "os dados são extraídos de sistemas operacionais e distribuídos através de *browsers*. Os dados específicos necessários para *Business Intelligence* são retirados para um *data mart* usado por executivos. As saídas são obtidas a partir do envio frequente de dados do *data mart* e de respostas de utilizadores da *Web* e analistas *OLAP (Online Analytical Processing*). As saídas podem assumir várias formas, nomeadamente relatórios de exceção, relatórios de rotina e respostas a solicitações específicas. As saídas são enviadas sempre que os parâmetros estão fora dos limites pré-especificados" (Negash, 2004).

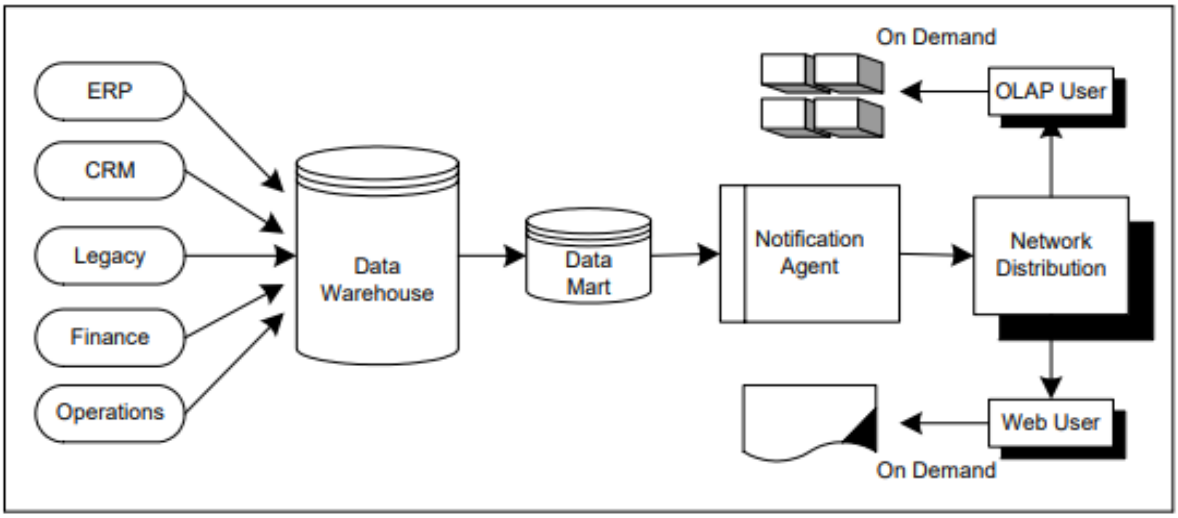

Figura 1-Arquitetura Business Intelligence para Dados Estruturados

b) Arquitetura para Dados Semi-Estruturados

De acordo com o livro *Business Intelligence,* a arquitetura para Dados Semi-Estruturados, como ilustrado na Figura 2, inclui um *Business Function Model*, um *Business Process Model*, um *Business Data Model*, uma *Application Inventory* e *Metadata Repository*.

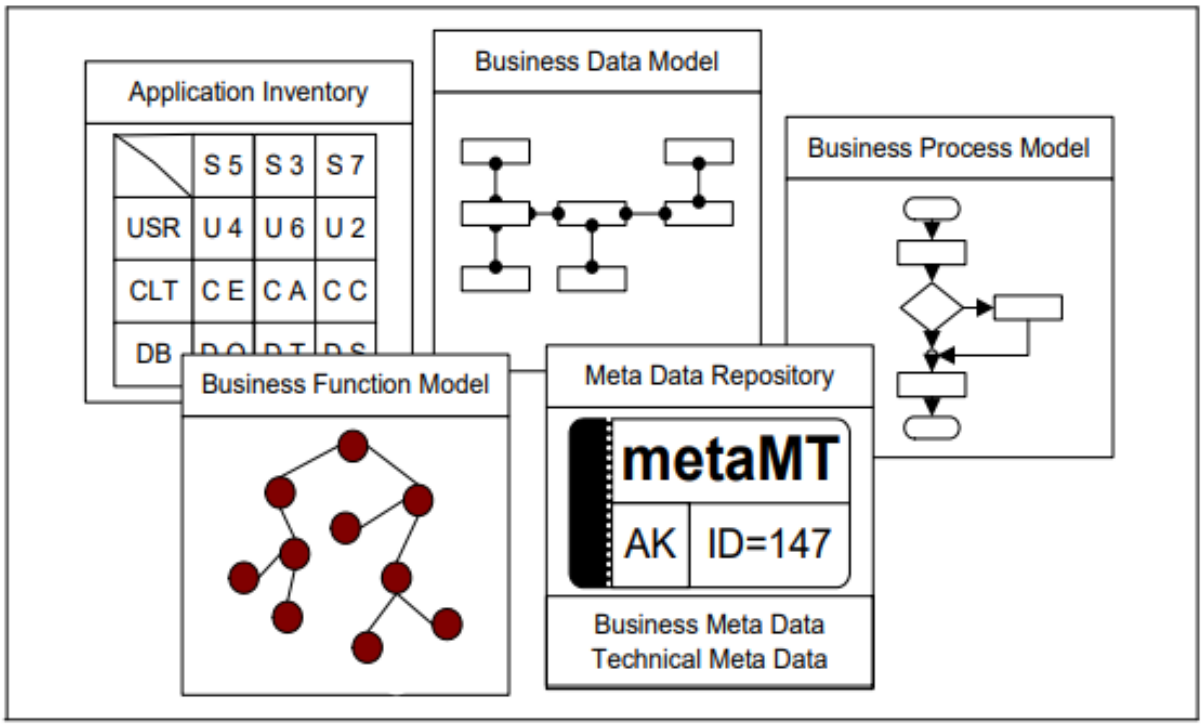

Figura 2-Arquitetura Business Intelligence para Dados Semi-Estruturados

- ***Business Function Model:*** *"*Este representa a hierarquia negocial da organização, mostrando o que a organização faz" (Negash, 2004);
- ***Business Process Model****: "*Este representa os processos implementados para funções empresariais e mostra como a organização desempenha as suas funções nos negócios" (Negash, 2004);
- ***Business Data Model****: "*Descreve os dados, os relacionamentos que conectam os dados com base nas atividades de negócio, mostrando ainda quais dados descrevem a organização" (Negash, 2004);
- ***Application Inventory****:* Trata da "contabilidade dos componentes de implementação física de funções de negócios, processos de negócios e dados de negócios", para além disto "mostra onde residem as peças arquitetónicas" (Negash, 2004);

- ***Metadata Repository**:* Faz o "detalhe descritivo dos modelos de negócios, através do uso de metadados" (Negash, 2004).

2.4. Data Warehousing

Após estar descrito um *Data Warehousing* na Secção 1.2.2., eis que as principais características segundo Márcio Silva e Marco Sartori (Silva & Sartori, 2015) que compõe este são:

- "**Orientado por tema**: Um *Data Warehouse* está orientado à volta dos assuntos específicos da organização.
- **Integrado**: A própria organização de um *Data Warehouse* exige a integração dos dados que nele estão contidos para que os objetivos sejam alcançados. Essa integração é obtida com o processo de ETL.
- **Não volátil**: Em teoria, os dados são inseridos no *Data Warehouse*, no entanto nunca podem ser excluídos ou alterados. O *Data Warehouse* já é preparado para otimizar os processos de introdução e seleção. Mas há casos específicos que exigem que o *Data Warehouse* seja atualizado.
- **Variante no tempo**: Um *Data Warehouse* mantém os dados históricos, onde todos assumem uma qualidade temporal, assim, o tempo é a dimensão fulcral que todos os *Data Warehouse* têm que suportar, uma vez que permite identificar tendências e desvios, assim como a previsão e comparação".

2.4.1. Componentes de um *Data Warehouse*

Segundo Kimball e Ross (Kimball & Ross, 2002), a primeira fase de construção do *Data Warehouse* consiste na seleção das fontes de dados, tanto internas como externas à organização. Sendo selecionadas as fontes, os dados são extraídos, limpos, transformados e posteriormente, preparados para serem carregados no modelo do *Data Warehouse*. Estes, de seguida, são reformulados num formato dimensional, de acordo com as necessidades da organização, e carregados para o *Data Warehouse*.

Na Figura 3 estão representados os principais componentes de um *Data Warehouse,* que segundo Jorge Sá na sua tese de doutoramento cujo tema é "Metodologias de Sistemas

Data Warehouse" (Vaz de Oliveira e Sá, 2009) e Kimball e Ross (Kimball & Ross, 2002) são os seguintes:

- **Fontes de Dados**: as fontes de dados podem ser internas ou externas à organização, sendo procedentes de múltiplos sistemas operacionais como sistemas legados ou outros.
- **ETL**: num primeiro momento, os dados são extraídos e copiados para o *Data Warehouse* para posteriormente serem manipulados. De seguida ocorrem as transformações, nomeadamente a limpeza de dados ou a combinação dos dados das múltiplas fontes. Por fim, é efetuada a estruturação física e o carregamento dos dados transformados para os modelos de dimensão do *Data Warehouse*.

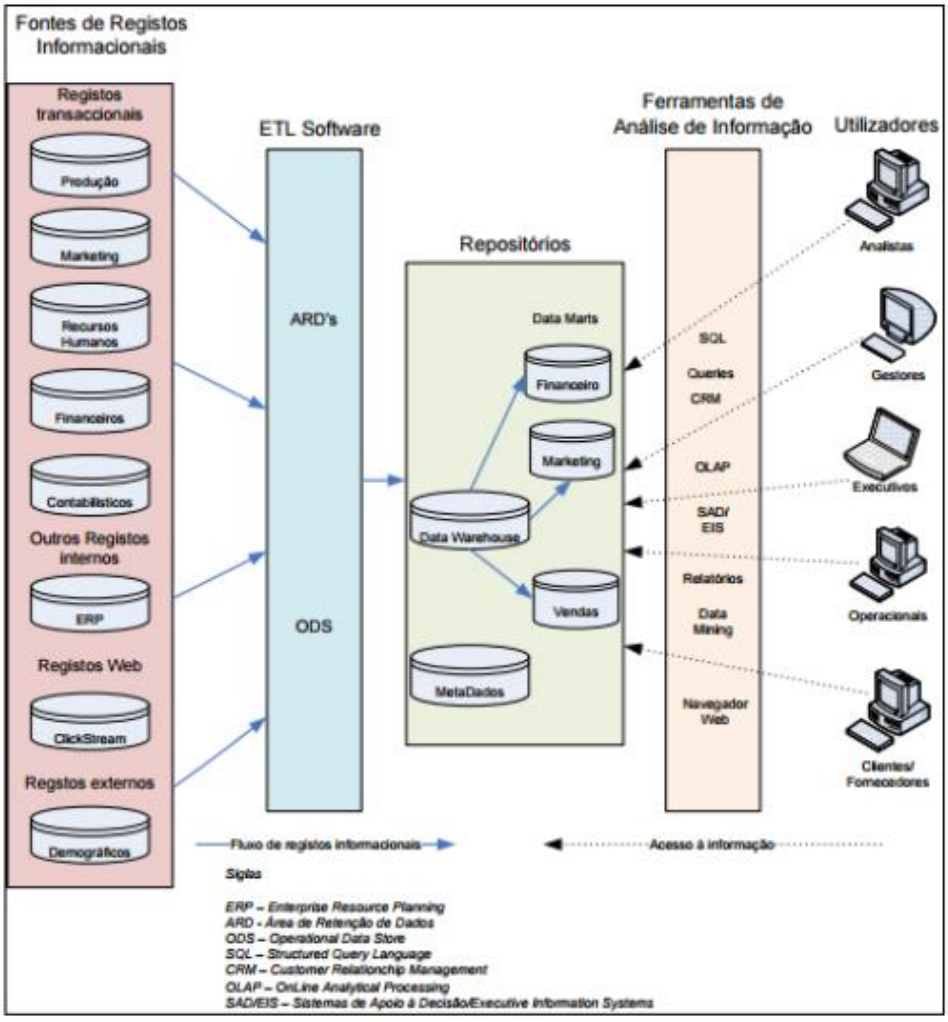

Figura 3--Exemplo de uma arquitetura de Data Warehouse

- **Repositórios:** os repositórios são constituídos pelo *Data Warehouse*, por *Data Marts* e metadados. O *Data Warehouse* fornece informações pertinentes e detalhadas com o intuito de apoiar a decisão. Os *Data Marts*, face aos *Data Warehouse*, são repositórios de menor dimensão, exibindo por norma uma área especifica da organização. Os metadados representam a informação dos registos informacionais existentes no Data Warehouse e no Data Mart.
- **Ferramentas de Análise de Informação**: a informação existente nos *Data Warehouses* pode ser acedida através de distintas ferramentas e aplicações, como *Data Mining*, relatórios ou sistemas de visualização.

2.4.2. Arquiteturas de Data Warehouse

Data Warehouse consta com várias arquiteturas básicas, sendo que as de duas e três camadas são as mais comuns. Apesar disto, existe também a arquitetura de uma camada. De seguida são apresentadas as diversas arquiteturas de Data Warehouse.

a) Arquitetura de uma camada

De acordo com Golfarelli e Rizzi (Golfarelli & Rizzi, 2009), a arquitetura de uma camada tem como propósito diminuir a quantidade de dados armazenamos através da remoção de dados redundantes. Como é percetível na Figura 4, a camada física é composta pela fonte dos dados e pelo *Data Warehouse*. Este tipo de arquitetura não permite o processamento de questões, uma vez que afeta os carregamentos transacionais.

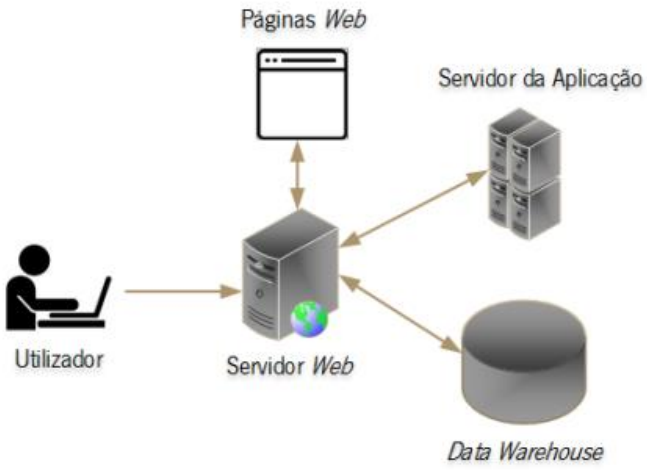

Figura 4-Exemplo de arquitetura de uma camada

b) Arquitetura de duas camadas

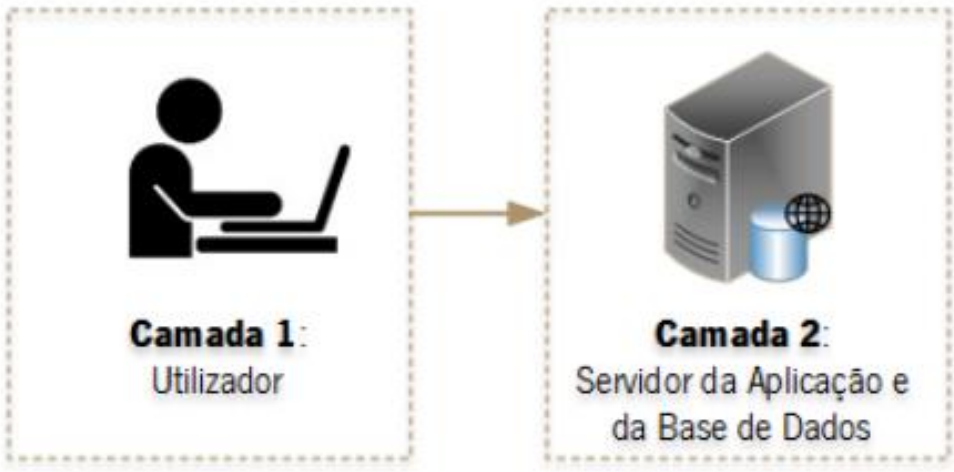

Figura 5-Exemplo de arquitetura de duas camadas

Na Figura 5 é o cliente que se encontra na primeira camada e a aplicação do sistema de apoio à decisão que se encontra na segunda, sendo que neste tipo de arquitetura este funciona na mesma plataforma de *hardware* que o *Data Warehouse.* Desta forma, esta arquitetura acaba por ser mais rentável economicamente do que a arquitetura de três camadas. Segundo Cristiano Castro (Castro, 2001), a arquitetura de duas camadas "simplesmente não suporta um determinado número de usuários simultaneamente, pois estimulam uma grande carga de processamento para processar nesta estação".

c) Arquiteturas de três camadas

Figura 6-Exemplo de arquitetura de três camadas

Tal como ilustrado na Figura 6, na primeira camada encontra-se o utilizador final, na segunda o *Data Warehouse* e na terceira camada a recolha de dados. Nesta arquitetura os dados são processados duas vezes e inseridos numa base de dados multidimensional complementar.

2.5. Big Data

Na Secção 1.2.2. encontra-se a definição de *Big Data*, no entanto para entender melhor esta, é útil definir os componentes da arquitetura. Uma arquitetura de gestão de *big data* deve incluir uma variedade de serviços que permitem às empresas fazer uso de inúmeras fontes de dados de forma rápida e eficaz.

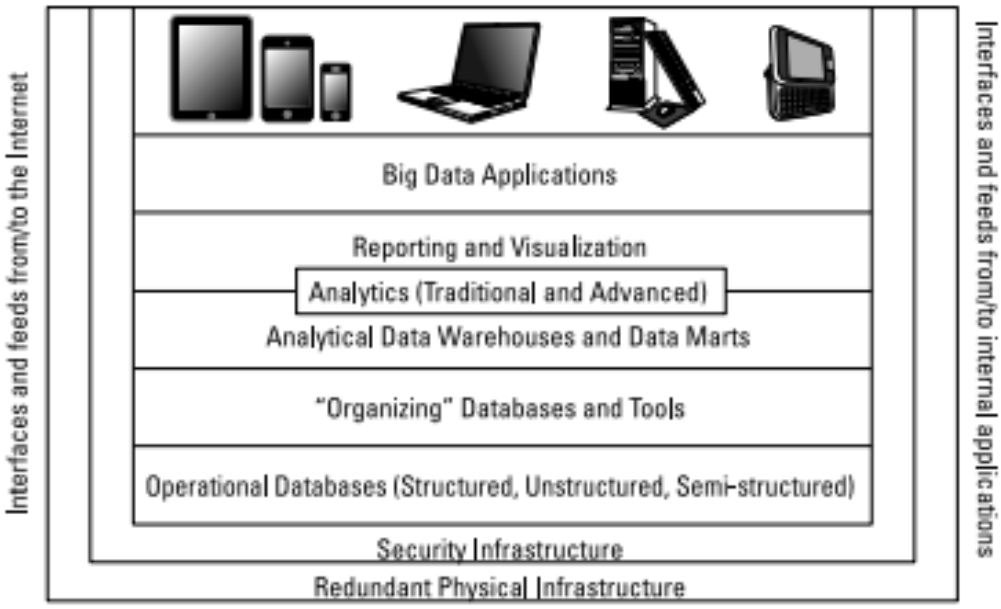

Figura 7-Arquitetura de Big Data

Como ilustrado na Figura 7, a arquitetura do *Big Data* é composta pelo *Redudant Physical Infrastructure*, pela *Security Infrastructure, Operational Databases, Organizing Databases and Tools, Analytical Data Warehouses and Data Marts, Reporting and visualization* e *Big Data Applications*.

Segundo Nugent, Halper, Hurwitz e Kaufman, no livro "*Big Data for Dummies*" cada uma destas composições se define da seguinte forma.

- ***Redudant Physical Infrastructure*** – "A infraestrutura física de suporte é fundamental para a operação e escalabilidade de uma arquitetura de *big data*. Para dar suporte a um volume de dados inesperado ou imprevisível, uma infraestrutura física para *big data* deve ser diferente daquela para dados tradicionais. A infraestrutura física é baseada num modelo de computação distribuída. Isso significa que os dados podem ser armazenados fisicamente em muitos locais diferentes e podem ser vinculados através de redes, do uso de um sistema de arquivos distribuído e de várias ferramentas de *big data*. A redundância é importante devido à quantidade de dados de fontes diferentes" (Hurwitz et al., 2013);
- ***Security Infrastructure*** – "Quanto mais importante for a análise de *big data* para as organizações, mais importante será proteger os dados. Por exemplo, se se tratar de uma empresa de assistência médica, provavelmente será usado o *big data* para identificar mudanças na demografia ou mudanças nas necessidades dos pacientes. Todos esses dados precisam de ser protegidos tanto para atender aos requisitos de conformidade como para proteger a privacidade dos pacientes" (Hurwitz et al., 2013);
- ***Operational Databases*** – "Tradicionalmente, fontes de dados operacionais consistem em dados altamente estruturados geridos pela linha de negócios numa base de dados relacional. No entanto é importante entender que os dados operacionais atualmente necessitam de abranger um conjunto mais amplo de fontes de dados, incluindo fontes não estruturadas, como dados de clientes em todas as suas formas" (Hurwitz et al., 2013);
- ***Organizing Databases and Tools*** – "Existe uma inúmera quantidade de dados que surge de fontes que não são tão organizadas ou diretas, incluindo dados provenientes de máquinas ou sensores ou de grandes fontes de dados públicas e privadas. No passado, a maioria das organizações não conseguia capturar ou armazenar essa grande quantidade de dados. Isto porque era simplesmente caro, a quantidade de dados era muito grande e não existiam ferramentas para fazer nada a respeito" (Hurwitz et al., 2013);

- ***Analytical Data Warehouses and Data Marts –*** "Depois de classificar as enormes quantidades de dados disponíveis, uma organização geralmente deve formar um subconjunto de dados com padrões e colocá-lo ao dispor da mesma" (Hurwitz et al., 2013);
- ***Reporting and visualization*** – "As organizações podem coletar, gerir e analisar dados suficientes, através de ferramentas que entendam verdadeiramente o impacto não apenas de uma coleção de elementos de dados, mas também a forma como esses elementos contextualizam os problemas de negócios que estão a ser abordados. Com o *big data*, relatórios e visualização de dados tornam-se ferramentas fundamentais para analisar o contexto dos dados, como estes se relacionam e o impacto que terão no futuro" (Hurwitz et al., 2013);
- ***Big Data Applications*** – "Tradicionalmente, uma organização espera que os dados sejam usados para responder a perguntas sobre o que e quando fazer. Com o aparecimento do *big data* isto mudou. Atualmente, são desenvolvidas ferramentas e aplicações para trabalhar e aproveitar exclusivamente as características do *big data*" (Hurwitz et al., 2013).

2.6. Machine Learning

Após a descrição do que é *Machine Learning* na Secção 1.2.2., esta Secção pretende dar outras informações ou leitores, nomeadamente um esquema muito interessante retirado do livro escrito por Swamynathan (Swamynathan, 2017) que se vê representado na Figura 8. Este esquema é composto por três níveis, um primeiro nomeado de *Machine Learning*, um segundo que se refere aos tipos de aprendizagem que é possível adotar no *Machine Learning*, e um terceiro nível que ilustra os tipos de modelos mais comuns para os respetivos tipos de aprendizagem do segundo nível.

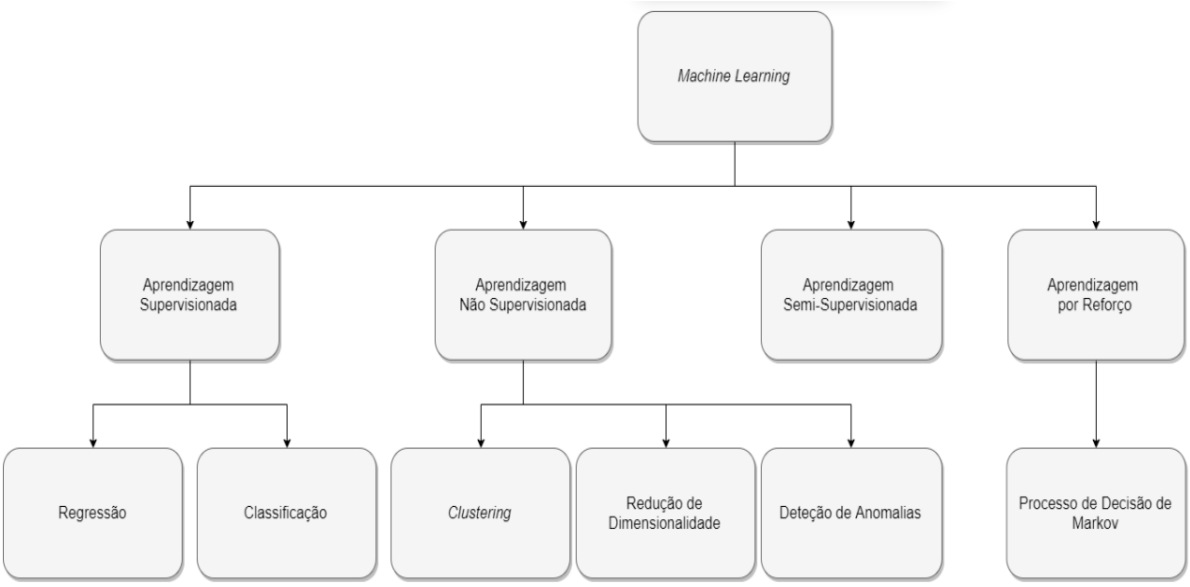

Figura 8-Tipos de aprendizagem e modelos Machine Learning segundo Swamynathan

2.6.1. Tipos de Aprendizagem

Como ilustrado na Figura 8, existem quatro tipos de abordagem de *Machine Learning*, sendo este a Aprendizagem Supervisionada, a Aprendizagem Não Supervisionada, a Aprendizagem Semi-Supervisionada e a Aprendizagem por Reforço. Seguidamente serão apresentados e descritos cada um destes tipos de aprendizagem.

a) Aprendizagem supervisionada

Segundo Mohammed, aprendizagem supervisionada é "inferir uma função ou mapeamento de dados de treino", este acrescenta ainda que "a máquina é ensinada com recurso a exemplos. Sendo fornecidos os exemplos das entradas e das saídas necessárias através de dados rotulados" (Mohammed et al., 2017).

b) Aprendizagem Não Supervisionada

De acordo com o artigo nomeado anteriormente, aprendizagem não supervisionada "contém dois grupos ou categorias de algoritmo, sendo estes a regressão e a classificação" (Mohammed et al., 2017), acrescenta ainda que "as técnicas mais comuns são regressão linear, regressão logística, redes neurais artificiais, máquina de suporte vetorial, árvores de decisão, Estatística Bayesiana, entre outras" (Mohammed et al., 2017). Em conformidade com

o que Kimberly Nevala escreveu no artigo "*The machine learning primer*", este tipo de aprendizagem pode ser aplicado de várias formas que de seguida serão nomeadas:

- "**Determinação de riscos** – isto acontece para determinar se um doente está em risco de ter algum problema de saúde, para determinar se um determinado equipamento está em risco de falhar tendo por base o histórico ou para determinar a probabilidade de um indivíduo deixar de pagar uma prestação de empréstimo;
- **Interação Personalizada** – Ocorre para determinar qual o conteúdo a mostrar a um individuo tendo por base o seu comportamento, intenções e gostos;
- **Deteção de Fraudes** – Ocorre de modo a identificar transações potencialmente fraudulentas;
- **Reconhecimento de imagem, texto e voz** – Acontece para identificação de objetos e pessoas em imagens; Sistemas que agem mediante indicações verbais;
- **Segmentação de clientes** – Ocorre de forma a segmentar clientes com base no seu histórico, preferências e intenções" (Practices & Nevala, 2017).

c) Aprendizagem Semi-Supervisionada

Segundo Nevala, a aprendizagem semi-supervisionada "é utilizada para resolver problemas idênticos aos da aprendizagem supervisionada, sendo necessários dados rotulados e não rotulados, isto é, alguns dados são dadas as respostas e outros não. Isto gera um modelo apropriado de classificação de dados" (Practices & Nevala, 2017). De acordo com o ponto de vista de Mohammed, "o propósito da aprendizagem semi-supervisionada é otimizar a forma de previsão de classes de dados futuros comparativamente aos modelos em que unicamente se usam dados rotulados" (Mohammed et al., 2017).

d) Aprendizagem por Reforço

De acordo com o artigo de Nevala a aprendizagem por reforço "é semelhante ao próprio ato de ensinar alguém a jogar um jogo. As regras e objetivos estão definidos, no entanto o resultado depende do jogador que ajusta a abordagem ao ambiente, à sua habilidade e às ações de determinado adversário" (Practices & Nevala, 2017). Dito de outra forma, é fornecido à máquina um conjunto de ações, regras e estados finais possíveis. Deste

modo, a máquina aprende a explorar as regras para criar o resultado desejado, isto é, determinar uma série de ações que levarão a um resultado ótimo ou desejável.

2.6.2. Tipos de Modelos

Como está ilustrado na Figura 8 da Secção 2.6.1., os tipos de modelos dependem de qual é o tipo de aprendizagem usado. Relativamente à aprendizagem supervisionada há dois modelos comummente usados, sendo este a regressão e a classificação, de acordo com a aprendizagem não supervisionada são normalmente usados o *clustering*, a redução de dimensionalidade e a deteção de anomalias. Relativamente à aprendizagem por reforço o tipo de modelo mais usado é o processo de decisão de Markov. Todos estes tipos de modelos serão descritos seguidamente.

a) Regressão

Segundo Brownlee, "neste tipo de modelos, espera-se uma variável contínua, isto é, um valor real" (Jason Brownlee, 2017). Swamynathan nomeia ainda como exemplos de casos de uso "a previsão das vendas de retalho, previsão do número de funcionários para cada turno, número de lugares de estacionamento necessários para uma loja, entre outras" (Swamynathan, 2017).

b) Classificação

De acordo com Brownlee "neste tipo de modelos espera-se uma variável discreta ou categórica, sendo necessário um número de classes igual ou superior a 2" (Jason Brownlee, 2017), acrescenta ainda que "o algoritmo deve aprender as características dos dados de entrada de cada classe com base em dados históricos e prever a classe nos novos dados de entrada" (Swamynathan, 2017). Como exemplos Swamunathan dá "se é ou não spam a identificação de um email" (Swamynathan, 2017).

c) *Clustering*

Conforme o artigo de Brownlee "*clustering* ocorre quando se pretende definir um agrupamento de dados considerando similaridades entre eles" (Jason Brownlee, 2017), Swamynathan dá como exemplos "artigos ou noticias semelhantes, agrupamentos de clientes tendo em conta o seu perfil, entre outros" (Swamynathan, 2017).

d) Redução de Dimensionalidade

De acordo com o artigo de Brownlee "neste tipo de técnicas, o objetivo passa por simplificar o conjunto de dados de entrada, de modo a mapear num espaço inferior dimensionalmente" (Jason Brownlee, 2017), Swamynathan dá como exemplos "a análise de conjuntos de dados de grande dimensão que requerem esforço em termos computacionais" (Swamynathan, 2017).

2.7. Soluções que envolvem *Data Science*, as suas metodologias e ferramentas

De acordo com o que foi dito na secção 2.1, existe um artigo escrito pelo *Journal of Information Systems Applied Research*, em 2016 (Hunsinger & Waguespaack, 2016), que procura comparar e explorar as diferentes ferramentas de código aberto de *Data Science*, de modo a selecionar uma ferramenta que atenda aos requisitos de projetos específicos de *Data Science*. Este artigo foca-se em doze ferramentas, sendo estas *ADAM, Alpha Miner, ESOM, Gnome Data Miner, KNIME, Mining Mart, MLC++, Orange, Rattle, Tanagra, Weka* e *Rapid Miner*. Destas doze ferramentas foram escolhidas seis para avançar para a comparação, sendo que *Tanagra, Orange, KNIME, Weka* e *Rapid Miner* foram escolhidas devido às suas várias avaliações positivas, e a ferramenta *R* foi também adicionada uma vez que esta é muito usada pela EMC, organização cujo negócio principal é armazenar e analisar dados. As variáveis de avaliação para a comparação das seis ferramentas foram:

- Agrupamento *K-means*;
- Mineração de regras de associação;
- Regressão linear;
- Regressão Logística;
- Classificadores *Bayesianos Naïve*;
- Árvore de decisão;
- Análise de Séries Temporais;
- Análise de texto;
- Processamento de *Big Data*;

- Fluxos de trabalho visuais.

Tal como a Tabela 1 indica, destas variáveis, a ferramenta que se mostrou com um maior suporte em código aberto foi a *Weka*, no entanto cada ferramenta possui pontos fortes e exclusivos. Por exemplo, para a ferramenta R são necessárias habilidades mais aprofundadas para realizar tarefas básicas, as outras quatro, ou seja, a *Tanagra, Orange, KNIME* e *Rapid Miner* fornecem abordagens visuais com um custo associado.

Tabela 1-Conclusões do artigo "A Comparison of Open Source Tools for Data Science"

	Orange	**Tanagra**	**Rapid Miner**	**KNIME**	**R**	**Weka**
Agrupamento *K-means*	Sim	Sim	Sim	Sim	Sim	Sim
Mineração de regras de associação	Sim	Sim	Sim	Sim	Sim	Sim
Regressão linear	Sim	Sim	Sim	Sim	Sim	Sim
Regressão Logística	Sim	Sim	Sim	Sim	Sim	Sim
Classificadores *Bayesianos Naïve*	Sim	Sim	Sim	Sim	Sim	Sim
Árvore de decisão	Sim	Sim	Sim	Sim	Sim	Sim
Análise de Séries Temporais	Não	Não	Alguma	Sim	Sim	Sim
Análise de texto	Sim	Não	Sim	Sim	Sim	Sim
Processamento de *Big Data*	Não	Não	Não	Não	Não	Sim
Fluxos de trabalho visuais	Sim	Sim	Sim	Sim	Não	Sim

Existe também um artigo escrito por João Ferreira cujo nome é "O processo ETL em sistemas *Data Warehouse*" (Ferreira et al., 2016) que, para além de descrever precisamente como funciona na totalidade um data warehouse e também o processo *ETL*, seleciona várias ferramentas de *ETL* de modo a encontrar a ferramenta adequada para a tomada de decisão. Segundo João Ferreira no mesmo artigo "As ferramentas de *ETL* disponíveis atualmente encontram-se bem preparadas para o processo de extração, transformação e carga. Tem-se

assistido a inúmeros avanços nestas ferramentas desde 1990, estando atualmente mais direcionadas para o utilizador. Uma boa ferramenta de *ETL* deve ser capaz de comunicar com as diversas Bases de Dados e ler diferentes formatos" (Ferreira et al., 2016). Este autor traz ao de cima uma lista de ferramentas *ETL* cujos nomes são: *Oracle Warehouse Builder, Data Integrator & Data Services, IBM Information Server, PowerCenter, Elixir, Data Migrator, SQL Server Integration Services, Talend Open Studio & Integration Suite, DataFlow Manager, Data Integrator, Open Text Integration Center, Transformation Manager, Data Manager/Decision Stream, Clover ETL, ETL4ALL, DB2 Warehouse, Pentaho Data Integration* e *Adeptia Integration Server*. De modo a escolher a ferramenta adequada, João Ferreira teve em conta dos seguintes parâmetros:

- S**uporte à plataforma** – Deve ser independente de plataforma, podendo assim correr em qualquer uma;
- **Tipo de fonte independente** – Deve ser capaz de ler diretamente da fonte de dados, independentemente do seu tipo, saber se é uma fonte de *RDBMS (Relational Database Management System*), ficheiro simples ou um ficheiro *XML*;
- **Apoio funcional** – Deve apoiar na extração de dados de múltiplas fontes, na limpeza de dados, e na transformação, agregação, reorganização e operações de carga;
- **Facilidade de uso** – Deve ser facilmente usada pelo utilizador;
- **Paralelismo** – Deve apoiar as operações de vários segmentos e execução de código paralelo, internamente, de modo que um determinado processo pode tirar proveito do paralelismo inerente da plataforma que está sendo executada. Também deve suportar a carga e equilíbrio entre os servidores e capacidade de lidar com grandes volumes de dados. Quando confrontados com cargas muito elevadas de trabalho, a ferramenta deve ser capaz de distribuir tarefas entre múltiplos servidores;
- **Apoio ao nível do *debugging*** – Deve apoiar o tempo de execução e a limpeza da lógica de transformação. O utilizador deve ser capaz de ver os dados antes e depois da transformação;
- **Programação** – Deve apoiar o agendamento de tarefas *ETL* aproveitando, assim, melhor o tempo não necessitando de intervenção humana para completar uma tarefa particular. Deve também ter suporte para programação em linha de comandos usando programação externa;

- **Implementação** – Deve suportar a capacidade de agrupar os objetos *ETL* e implementá-los em ambiente de teste ou de produção, sem a intervenção de um administrador de *ETL*;
- **Reutilização** – Deve apoiar a reutilização da lógica de transformação para que o utilizador não precise reescrever, várias vezes, a mesma lógica de transformação outra vez.

Na conclusão deste artigo nomeado de "O processo ETL em sistemas *Data Warehouse*", as ferramentas com mais destaque foram a *Oracle Warehouse Builder* e a SQL Server Integration Services. Segundo João Ferreira "As suas capacidades de tratamento e manipulação de informação, aliadas à facilidade e simplicidade de utilização, tornam-nas uma referência entre as ferramentas ETL abordadas" (Ferreira et al., 2016).

Os artigos nomeados de "*A Comparison of Open Source Tools for Data Science*" e "O processo ETL em sistemas *Data Warehouse*" referem-se mais propriamente às ferramentas de *Data Science*, e ainda, dentro destes, não foram encontrados outros que remetam mais diretamente ao tema desta dissertação. Relativamente a artigos, livros e projetos que se foquem na relação entre *Data Science* e as suas metodologias, o material encontrado foca-se essencialmente na distinção entre as metodologias CRISP-DM, KDD e SEMMA, o que será uma grande ajuda para a descoberta da melhor metodologia, no entanto não nos dão conclusões suficientemente concretas para que seja possível concluir algo nesse sentido. No que se refere ao que é efetivamente o tema do presente documento, isto é, "*Data Science* – Metodologias e Ferramentas", não foi encontrado qualquer material que vá de acordo com os objetivos concretos deste.

3. ABORDAGEM METODOLÓGICA, MATERIAIS E MÉTODOS

Uma vez que o objetivo desta dissertação é encontrar as ferramentas mais versáteis e a que melhor se adaptaria aos vários tipos de *Data Science*, ou seja, as mais funcionais de modo a que qualquer utilizador que as procure consiga realizar tudo o que pretende, decidiu-se usar o *benchmarking* para decidir quais seriam essas, comparando todas as ferramentas. Este *benchmarking* também seria feito com as metodologias, mas de forma mais simples, uma vez que são todas idênticas e tem um número muito mais reduzido.

Benchmarking é atualmente considerado umas das abordagens mais eficazes para melhorar o desempenho de qualquer organização. Segundo Madeira no artigo "*Benchmarking*: a arte de copiar" (Madeira, 1999), do *benchmarking* surgem quatro sob abordagens, o *benchmarking* interno que "representa uma comparação organizacional de desempenhos dentro do mesmo grupo", e o *benchmarking* competitivo que "envolve a comparação dos produtos, serviços e processo de trabalho da organização com os concorrentes diretos", o *benchmarking* funcional que "consistem em identificar as melhores praticas de qualquer tipo de organização que tenha uma reputação de excelência, e o *benchmarking* estratégico que "é uma tipo de *benchmarking* competitivo cujo principal objetivo é estabelecer uma determinada ação estratégica ou mudança organizacional". De acordo com o tema da presente dissertação, serão usadas ambas as vertentes do *benchmarking*. O *benchmarking* interno e o competitivo, serão usados uma vez que existem metodologias e ferramentas a serem comparadas e avaliadas que foram criadas pelas mesmas ou diferentes organizações, respetivamente. O *benchmarking* funcional e o estratégico visto que poderá não existir uma metodologia ou ferramenta que siga completamente os propósitos referidos anteriormente e, assim, será necessária a identificação das melhores práticas de organizações de excelência e uma nova ação estratégica.

Segundo o artigo "*Benchmarking*: a arte de copiar" (Madeira, 1999), e como ilustrado na Figura 9, o *benchmarking* é constituído por cinco passos e por cinco etapas. Este inicia-se com o:

- **Planeamento** – definindo que tipos de *data science* seriam interessantes e importantes conter nas ferramentas, que metodologia seria a mais indicada e quantas fases de análise existiriam;

- **Recolha de Dados** – reúne toda a informação importante destacada na etapa anterior, esta junção de informação será repartida em diferentes documentos, um inicial com vários resumos relativos às ferramentas e metodologias, um segundo em forma de tabela, estando tudo mais organizado e melhor preparado para ser usado na etapa posterior;
- **Análise de Dados** – nesta serão analisados esses documentos e remetidos à tabela nomeada anteriormente, para que a informação se mantenha sempre organizada;
- **Comparação de Dados** – é nesta onde se fazem os cálculos para se destacarem as melhores ferramentas, dando depois origem à última fase;
- **Escolha da Ferramenta** – etapa que reúne as ferramentas que mais se destacaram ao logo de toda a comparação.

Depois disto, inicia-se outra vez este processo desde o Planeamento até Escolha da Ferramenta, de modo a concluir a segunda fase do *benchmarking* e seja efetivamente possível chegar a uma ferramenta ideal.

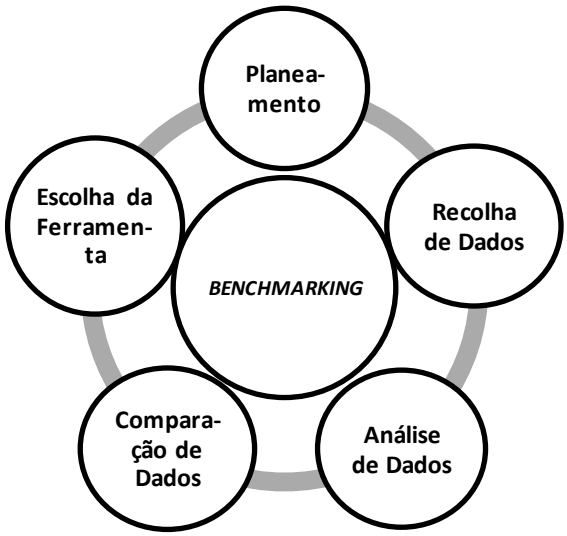

Figura 9-Modelagem do Benchmarking

As cinco etapas do *benchmarking* são:

- **Estudo preliminar de análise** – Consiste na "determinação das áreas chave de estudo e do âmbito e significância do mesmo" (Madeira, 1999);
- **Desenvolvimento de um compromisso de longo prazo com o projeto benchmarking** – "Definição de claro conjunto de objetivos" (Madeira, 1999);

- **Identificação de parcerias do benchmarking** – "identificar parcerias, bem como a sua dimensão e posição dentro da indústria" (Madeira, 1999);
- **Recolha de informação à escolha do método** – "Recolha de informações do produto, estratégicas, entre outras" (Madeira, 1999).

O *benchmarking* das metodologias servirá para eleger ou idealizar uma metodologia ágil e completa, que vá de encontro com aquilo que se pretende.

O *benchmarking* das ferramentas vem atuar nesta dissertação através de duas fases, isto é, dois momentos de comparação, um inicial que servirá apenas como comparação e filtro das ferramentas para a fase seguinte e tem em conta quais as ferramentas que se adequam aos diferentes tipos de *Data Science: Data Analytics, Data Mining, Business Intelligence, Data Warehousing, Big Data, Machine Learning* e *Data Visualization*, e um segundo, sendo mais importante, que remete a uma análise mais focada no que se pretende retirar da dissertação, isto é, a ferramenta mais útil e facilmente utilizada, que deverá ir de acordo com a melhor metodologia encontrada ou esquematizada.

4. TRABALHO REALIZADO

Este capítulo demonstra todo trabalho desenvolvido até ao momento, bem como as conclusões retiradas do mesmo.

Sendo o tema do presente documento "*Data Science* - Metodologias e Ferramentas", numa primeira etapa o que se procurou fazer foi um levantamento das metodologias e das ferramentas que serão descritas nas seguintes secções, sendo que no final ocorre uma comparação entre estas:

4.1. Levantamento das Metodologias de *Data Science*

No seguimento do método de pesquisa realizado e presente na secção 2.1., até ao momento foram analisadas cinco metodologias de *Data Science*, sendo estas:

- *CRISP-DM;*
- *Knowledge Discovery in Database;*

- *SEMMA;*
- *OSEAM;*
- IBM – Metodologia de Base para *Data Science*.

4.1.1. CRISP-DM

A metodologia CRISP-DM significa *CRoss-Industry Standard Process for Data Mining*. Esta consiste num ciclo que compreende seis etapas:

1. **Compreensão do negócio** – Esta fase foca-se na compreensão dos objetivos do projeto e requisitos na perspetiva do negócio, convertendo esse conhecimento na definição de um problema de *Data Mining* e num plano preliminar projetado para atingir os objetivos;
2. **Compreensão dos dados** – A fase de compreensão dos dados começa com a recolha dos dados e prossegue com as atividades de familiarização com os dados, nomeadamente com a identificação de problemas e de conjuntos de dados interessantes;
3. **Preparação dos dados** – A fase de preparação de dados costuma ocorrer várias vezes ao longo do processo. Esta constrói conjuntos de dados finais a partir dos dados iniciais;
4. **Modelação** – Nesta fase, várias técnicas de modelagem são selecionadas e aplicadas, e os parâmetros são calibrados para valores ótimos;
5. **Avaliação** – Nesta fase o modelo obtido é avaliado mais detalhadamente, de modo a garantir toda a qualidade. É nesta fase que se verifica se os objetivos do negócio foram atingidos ou não;
6. **Aplicação** – Nesta fase o modelo deverá ser organizado de modo a ser apresentado ao cliente.

Tal como a Figura 10 indica, as transições entre diferentes etapas podem ser revertidas no *CRISP-DM*, isto acontece por exemplo na Modelagem quando um especialista não encontra dados suficientes para ir de encontro com o objetivo do projeto, devendo retornar à etapa Preparação dos Dados.

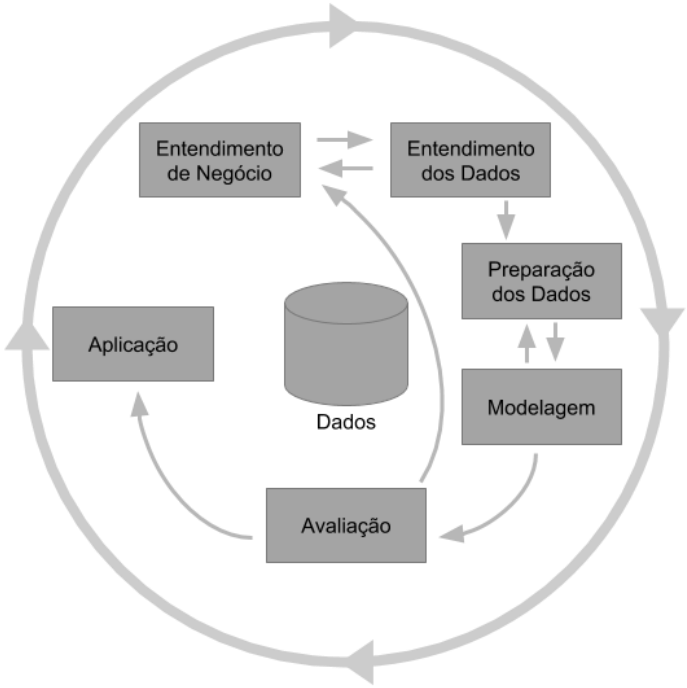

Figura 10-Ciclo CRISP-DM

4.1.2. Knowledge Discovery in Databases

Esta metodologia procura extrair padrões e informações relevantes de dados. Esta é composta por cinco etapas, sendo estas:

1. **Seleção** – Esta fase foca-se na criação de um conjunto de dados ou na identificação de dados que requerem uma exploração adicional;
2. **Pré-Processamento** – Nesta fase dá-se a limpeza e o um processamento prévio dos dados de modo a identificar dados consistentes;
3. **Transformação** – Tal como o nome indica, dá-se a transformação dos dados, através de métodos de redução do número de variáveis;
4. ***Data Mining*** – Nesta fase dá-se a identificação de padrões através de regras de classificação, regressão e agrupamento;
5. **Interpretação/Avaliação** – Esta fase serve para interpretar e avaliar os padrões encontrados na fase anterior.

Através da seguinte Figura 11 é possível percecionar mais facilmente as etapas descritas anteriormente.

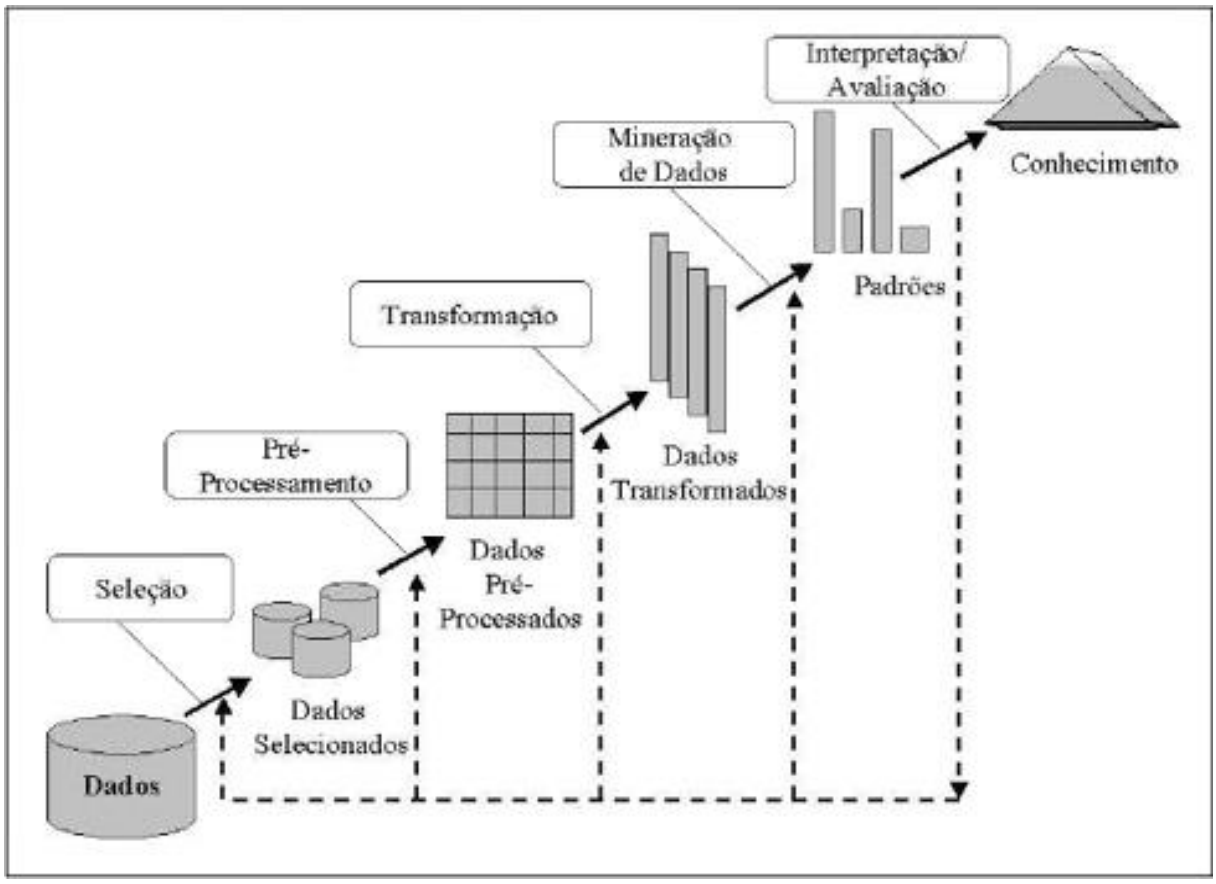

Figura 11-Ciclo do KDD

SEMMA

A metodologia *SEMMA* foi desenvolvida pelo *SAS Institute*. A sigla *SEMMA* significa *Sample, Explore, Modifiy, Model, Assess* e refere-se ao processo de condução de um projeto de *Data Mining*. Esta é constituída por 5 etapas para o processo:

1. ***Sample*/Amostra** – Esta etapa consiste na extração de uma amostra de um grande conjunto de dados que seja grande suficiente para conter informações significativas, mas pequeno o suficiente para ser rapidamente manipulada;
2. ***Explore*/Exploração** – Esta etapa consiste na exploração dos dados de modo a encontrar tendências e anomalias, obtendo compreensão e ideias;

3. ***Modify*/Modificação** - Esta etapa consiste em criar, selecionar e transformar variáveis na preparação para a modelagem dos dados;
4. ***Model*/Modelação** – Esta etapa consiste na modelação dos dados permitindo que o software pesquise automaticamente uma combinação de dados que preveja de forma confiável um resultado desejado;
5. ***Assess*/Avaliação** – Esta etapa consiste em avaliar a utilidade e a confiabilidade dos resultados da modelagem e dos modelos criados.

A metodologia *SEMMA* oferece um processo de fácil compreensão, permitindo um desenvolvimento organizado e adequado à manutenção de projetos de *Data Mining*. Confere assim uma estrutura para a sua conceção, criação e evolução, ajudando a apresentar soluções para problemas de negócios, bem como encontrar os objetivos de negócios do *Data Mining*.

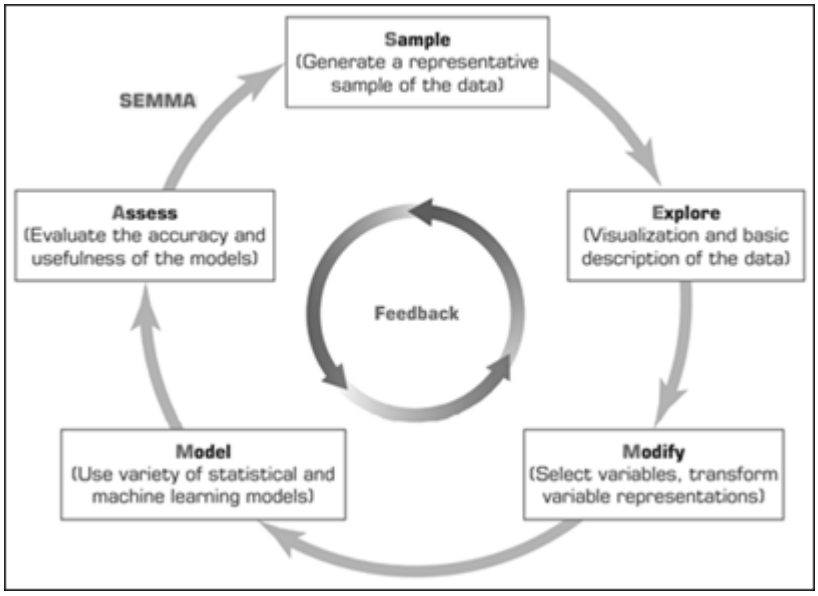

Figura 12-Ciclo do SEMMA

Na Figura 12 estão sintetizadas as etapas descritas anteriormente.

4.1.3. CRISP-DM vs. KDD vs. SEMMA

Sendo estas três metodologias idênticas viu-se a necessidade de comparar as mesmas antes ainda de prosseguir para novas pesquisas.

Estas três metodologias têm a mesma natureza cíclica, a principal diferença encontra-se no *CRISP-DM* uma vez que as transições entre as etapas podem ser revertidas. Posto isto, o *CRISP-DM* tem uma vantagem perante as outras metodologias, sendo esta a desnecessidade

de aguardar pelo término do ciclo quando já é evidente que os dados não trarão nenhum resultado.

A metodologia *KDD* e *SEMMA* são idênticas, uma vez que cada etapa da *KDD* corresponde diretamente a uma etapa da *SEMMA*, a Seleção corresponde à Amostra, o Pré-Processamento à Exploração, a Transformação à Modificação, o *Data Mining* à Modelação, e a Interpretação/Avaliação à Avaliação.

Comparando as etapas da metodologia *KDD* com as do *CRISP-DM* não existe uma correspondência tão encaixável quanto às etapas da *SEMMA*, no entanto podemos identificar semelhanças entre todo o processo Pré-*KDD*, isto é, a identificação dos objetivos e a compreensão do negócio, com a etapa Compreensão do Negócio do *CRISP-DM*. A etapa Aplicação do *CRISP-DM* pode também ser comparada com a fase posterior a todo o ciclo do *KDD*, ou seja, o Pós-*KDD*, onde todo o conhecimento é posto em prática. A etapa Compreensão dos Dados compreende a junção da etapa Seleção e Pré-Processamento, a etapa Preparação dos Dados é semelhante à Transformação, a Modelação ao *Data Mining* e a Avaliação à Interpretação/Avaliação.

Na tabela 2 é possível ver estas correspondências de forma muito mais simplificada.

Tabela 2-Comparação entre CRISP-DM, SEMMA e KDD

CRISP-DM	***KDD***	***SEMMA***
Compreensão do Negócio	Pré-*KDD*	------
Compreensão dos Dados	Seleção	Amostra
	Pré-Processamento	Exploração
Preparação dos Dados	Transformação	Modificação
Modelção	*Data Mining*	Modelação
Avaliação	Interpretação/Avaliação	Avaliação
Aplicação	Pós-*KDD*	------

4.1.4. OSEMN

A metodologia *OSEMN*, nomeada por *MSEMiN*, é popularmente conhecida pela sua simplicidade que vem facilitar a passagem entre etapas. Havendo qualquer problema durante etapas é sempre possível retroceder e adequar novamente os dados aos objetivos do projeto.

Esta metodologia é constituída por 5 etapas:

1. ***Obtain*/Obter** – Esta fase foca-se na compreensão dos objetivos do projeto e requisitos na perspetiva do negócio. Também é nesta onde são obtidos os dados dos *stakeholders;*
2. ***Scrub*/Esfregar** – O foco desta fase é a limpeza dos dados. O objetivo é preparar todos os dados para que possam ser analisados posteriormente;
3. ***Explore*/Explorar** – Nesta fase o suposto é aumentar o conhecimento relativamente aos dados, nomeadamente com a identificação de problemas e de conjuntos de dados interessantes. Esta etapa é essencialmente visual, uma vez que é a partir de histogramas, mapas de calor e gráficos de pontos que é mais fácil a interpretação e a compreensão de dados;
4. ***Model*/Modelar** – Nesta fase são feitos os modelos para o sucesso do projeto. São aplicados algoritmos de *Machine Learning* com maior foco nos resultados favoráveis. No caso de existir a necessidade de mais dados para esta etapa é possível regredir duas etapas para que a reestruturação dos dados possa produzir melhores resultados;
5. ***Interpret*/Interpretar** – Nesta fase são recolhidos os resultados e partilhados com os *stakeholders*. Através da comunicação com estes, o modelo pode sofrer alterações, isto é, melhorias. Se, por outro lado, os *stakeholders* estiverem satisfeitos com os resultados, então nesse momento pode-se prosseguir para a produção dos modelos e para a automação.

4.1.5. *IBM* – Metodologia de Base para *Data Science*

Esta metodologia segue uma abordagem "*top-down*", isto é, inicialmente é identificado o problema de negócio, e seguidamente os dados são analisados de modo a encontrar uma solução. Contrariamente às metodologias anteriores que tem cerca de seis etapas focadas em *Data Mining*, esta consta com 10. A razão deve-se à introdução de outras técnicas, como o uso de grandes volumes de dados, a incorporação de análise de texto na modelagem preditiva e a automatização de alguns processos.

Na Figura 13 estão exibidas todas as etapas da Metodologia de Base para *Data Science* segundo a *IBM*.

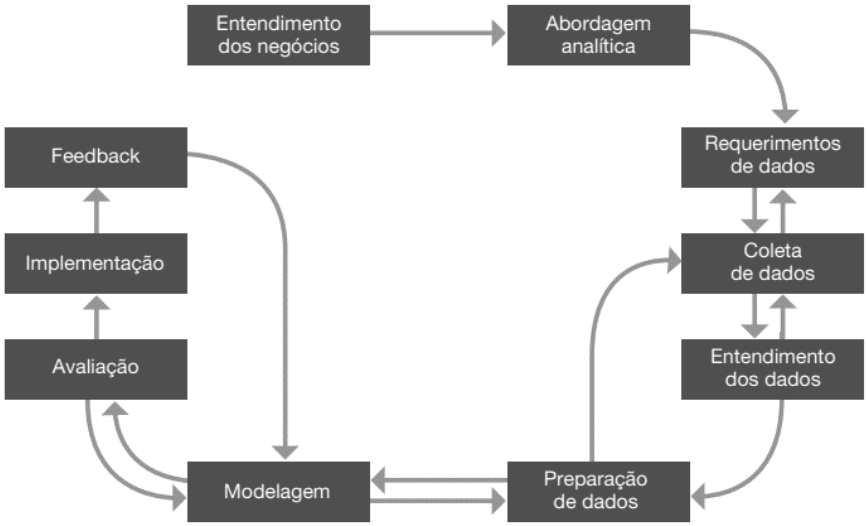

Figura 13-Ciclo da IBM – Metodologia de Base para Data Science

1. **Entendimento dos negócios** – Esta fase serve para identificar o problema e definir os objetivos do projeto e requisitos da solução. Esta fase é crucial para uma solução bem-sucedida;
2. **Abordagem analítica** – Nesta fase o objetivo é definir a abordagem analítica para resolver o problema através de técnicas de estatística e de *Machine Learning*;
3. **Requisitos de dados** – Tal como o nome indica esta fase serve para determinar os requisitos de dados, nomeadamente os formatos, as representações dos dados e os próprios conteúdos;
4. **Coleta de dados** – Nesta fase são recolhidos todos os dados relevantes para a resolução do problema. Se houver lacunas na recolha de dados, o cientista pode voltar a verificar os requisitos;
5. **Entendimento dos dados** – Nesta fase os cientistas de dados procuram identificar informações a partir dos dados recolhidos, normalmente são usadas técnicas de visualização para entender o conteúdo dos mesmos. Se necessário os cientistas poderão voltar à etapa anterior para coletar dados adicionais;
6. **Preparação dos dados** – Esta fase inclui a limpeza dos dados, a combinação de dados de várias fontes e a transformação dos mesmo em variáveis mais úteis. Esta etapa tende a ser a mais demorada de todo o processo;

7. **Modelagem** – Esta fase inicia-se quando os conjuntos de dados já estão feitos. A modelagem destina-se ao desenvolvimento de modelos preditivos ou descritivos de acordo com a abordagem definida anteriormente;
8. **Avaliação** – Esta etapa consiste na avaliação do modelo feito anteriormente. A avaliação é feita de diversas formas, nomeadamente através de tabelas e gráficos, permitindo que os cientistas consigam verificar a qualidade e eficácia do modelo;
9. **Implementação** – Após a aprovação e o desenvolvimento do modelo, este é implementado no ambiente de produção. É nesta fase que isso ocorre. A implementação do modelo é feita de forma limitada para que seja possível fazer uma avaliação completa;
10. ***Feedback*** – Após a recolha dos resultados do modelo implementado, a organização obtém feedback de toda a performance do modelo.

4.1.6. Conclusão das metodologias descritas

Relativamente às metodologias abordadas anteriormente, e como se pode confirmar na Tabela 3, têm todas um ciclo idêntico. A maior diferença entre estas está no facto de que evidentemente a metodologia do IBM tem um ciclo mais longo e completo do que as outras metodologias. Outra diferença é o facto de que algumas poderão retroceder imediatamente quando alguma etapa não tem o material necessário ou adequado para prosseguir, enquanto outras só podem fazê-lo no final do ciclo, ou no final de um certo número de etapas.

Tabela 3-Comparação de Metodologias

CRISP-DM	**KDD**	**SEMMA**	**OSEMN**	**IBM**
Compreensão do Negócio	------	------	------	Entendimento dos negócios
------	------	------	------	Abordagem analítica
------	------	------	------	Requisitos de dados
------	------	------	*Obtain* / Obter	Coleta de dados
Compreensão dos Dados	Seleção	Amostra	*Scrub* / Esfregar	Entendimento dos dados
------	Pré-Processamento	Exploração	*Explore* / Explorar	------
Preparação dos Dados	Transformação	Modificação	------	Preparação dos dados
Modelagem	*Data Mining*	Modelagem	*Model* / Modelar	Modelagem
Avaliação	Interpretação / Avaliação	Avaliação	*Interpret* / Interpretar	Avaliação
Aplicação	------	------	------	Implementação
------	------	------	------	*Feedback*

Existem ainda poucas conclusões a retirar das metodologias, uma vez que esse será um ponto importante no trabalho a fazer futuramente e, também, porque seria interessante aumentar o número de metodologias para que haja diversidade de ideias, de modo a ter a certeza que será encontrada ou idealizada a metodologia adequada e às necessidades.

4.2. Levantamentos das ferramentas de *Data Science*

Até ao momento foram analisadas diversas ferramentas de diferentes tipos de *Data Science*, nomeadamente, *Data Analytics, Data Mining, Business Intelligence, Data Warehousing, Big Data, Machine Learning* e *Data Visualization.*

4.2.1. Ferramentas *Data Analytics*

Relativamente às ferramentas de *Data Analytics* estas estão descritas no Anexo I e foram analisadas 46, nomeadamente: *Apache Spark*[1], *Adverity*[2], *Apache Storm*[3], *Chartio*[4], *Dataddo*[5], *Domo*[6], *Google Data Studio*[7], *Google Fusion Tables*[8], *Grafana*[9], HubSpot[10], *IBM Cognos*[11], *Juicebox*[12], *Jupyter Notebook*[13], *KNIME*[14], *Looker*[15], *Metabase*[16], *Microsoft Excel*[17], *Microsoft Power BI*[18], *Mode*[19], *Oracle Analytics Cloud*[20], *Oribi*[21], *Orange*[22], *OpenRefine*[23],

[1] https://spark.apache.org/
[2] https://www.adverity.com/
[3] https://storm.apache.org/
[4] https://chartio.com/
[5] https://www.dataddo.com/pt
[6] https://www.domo.com/
[7] https://datastudio.google.com/u/0/
[8] https://support.google.com/fusiontables/answer/2571232?hl=en
[9] https://grafana.com/
[10] https://www.hubspot.com
[11] https://www.ibm.com/products/cognos-analytics
[12] https://www.juiceanalytics.com
[13] https://jupyter.org/
[14] https://www.knime.com/
[15] https://looker.com/
[16] https://www.metabase.com/
[17] https://www.microsoft.com/en-us/microsoft-365/excel
[18] https://powerbi.microsoft.com/pt-pt/
[19] https://mode.com/
[20] https://www.oracle.com/pt/business-analytics/analytics-platform/
[21] https://oribi.io/
[22] https://orangedatamining.com/
[23] https://openrefine.org/

Periscope Data[24], *Python*[25], *Qlik*[26], *Query.me*[27], *R*[28], *RapidMiner*[29], *Redash*[30], *SAP BusinessObjects*[31], *SAS Business Intelligence*[32], *SAS*[33], *Sprinkle*[34], *Structured query Language*[35], *Sisense*[36], *Splunk*[37], *Tableau*[38], *Talend* [39], T*houghtspot*[40], *TIDAMI*[41], *TIBCO Spotfire*[42], *Whatagraph*[43], *Weka*[44], *Xplenty*[45] e *Zoho Analytics*[46].

4.2.2. Ferramentas Data Mining

De acordo com o *Data Mining* recolheram-se 43 ferramentas, sendo que estas estão descritas no Anexo I e analisadas: *Advanced Miner*[47], *Analytic Solver*[48], *Alteryx*[49], *Apache Mahout*[50], *Apache Spark*[51], *Board*[52], *Civis*[53], *Datawatch*[54], *DataMelt*[55], *Dundas*[56], *ELKI*[57], *Enterprise Miner*[58], *H2O*[59], *H3O*[60], *Hadoop*[61], *IBM Cognos*[62], *IBM SPSS Modeller*[63], *Inetsoft*

[24] https://www.sisense.com/blog/periscope-data-is-now-sisense-for-cloud-data-teams/
[25] https://www.python.org/
[26] https://www.qlik.com/us/
[27] https://query.me/
[28] https://www.r-project.org/about.html
[29] https://rapidminer.com/
[30] https://redash.io/
[31] https://www.sap.com/products/bi-platform.html
[32] https://www.sas.com/en_us/solutions/business-intelligence.html
[33] https://www.sas.com/en_us/home.html
[34] https://www.sprinkledata.com/
[35] https://www.mysql.com/
[36] https://www.sisense.com/
[37] https://www.splunk.com/
[38] https://www.tableau.com/
[39] https://www.talend.com/
[40] https://www.thoughtspot.com/
[41] https://www.tidami.com/
[42] https://www.tibco.com/products/tibco-spotfire
[43] https://whatagraph.com/
[44] https://www.cs.waikato.ac.nz/ml/weka/
[45] https://try.integrate.io/software-testing-help/
[46] https://www.zoho.com/analytics/
[47] https://algolytics.com/products/advancedminer/
[48] https://www.solver.com/products-overview
[49] https://www.alteryx.com/
[50] https://mahout.apache.org//
[51] https://spark.apache.org/
[52] https://www.board.com/en
[53] https://www.civisanalytics.com/platform/
[54] https://www.altair.com/panopticon/
[55] https://datamelt.org/
[56] https://www.dundas.com/support/dundas-bi-free-trial
[57] https://elki-project.github.io/
[58] https://www.sas.com/en_us/software/enterprise-miner.html
[59] https://h2o.ai/
[60] https://h2o.ai/
[61] https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html
[62] https://www.ibm.com/products/cognos-analytics
[63] https://www.ibm.com/products/spss-modeler

Intelligence[64], *Kaggle*[65], *KNIME*[66], *MonkeyLearn*[67], *Oracle*[68], *Oracle BI*[69], *Orange*[70], *PolyAnalyst*[71], *Python*[72], *Qlik*[73], *R*[74], *RapidMiner*[75], *Rattle*[76], *SAS Enterprise Miner*[77], *SAS Data Mining*[78], *Scikit-learn*[79], *Sisense*[80], *Solver*[81], *SPMF*[82], *SQL Server Data Tools*[83], *Teradata*[84], *Tanagra*[85], V*iscovery*[86], *Weka*[87], *Xplenty*[88]e *Zoho Analytics*[89]

4.2.3. Ferramentas *Business Intelligence*

Relativamente às ferramentas de *Business Intelligence,* foram recolhidas e descritas no Anexo I 30 ferramentas, sendo estas: *Adaptive Discovery*[90], *Alteryx*[91], *Board*[92], *Clear Analytics*[93], *Datapine*[94], *Domo*[95], *Dundas*[96], *GoodData*[97], HubSpot[98], *IBM Cognos*[99], *Infor Birst*[100], *Jaspersoft*[101], *Looker*[102], *Microsoft Power BI*[103], *MicroStrategy*[104], *Oracle BI*[105],

[64] https://www.inetsoft.com/products/StyleIntelligence/
[65] https://www.kaggle.com/
[66] https://www.knime.com/
[67] https://monkeylearn.com/
[68] https://www.oracle.com/database/technologies/datawarehouse-bigdata/dataminer.html
[69] https://www.oracle.com/pt/business-analytics/business-intelligence/
[70] https://orangedatamining.com/
[71] https://www.megaputer.com/polyanalyst/
[72] https://www.python.org/
[73] https://www.qlik.com/us/
[74] https://www.r-project.org/about.html
[75] https://rapidminer.com/
[76] https://rattle.togaware.com/
[77] https://www.sas.com/en_us/software/enterprise-miner.html
[78] https://www.sas.com/en_us/insights/analytics/data-mining.html
[79] https://scikit-learn.org/stable/
[80] https://www.sisense.com/
[81] https://www.solver.com/xlminer-data-mining
[82] http://www.philippe-fournier-viger.com/spmf/
[83] https://docs.microsoft.com/en-us/previous-versions/sql/
[84] https://www.teradata.com/
[85] https://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html
[86] https://www.viscovery.net/somine/
[87] https://www.cs.waikato.ac.nz/ml/weka/
[88] https://try.integrate.io/software-testing-help/
[89] https://www.zoho.com/analytics/
[90] https://www.adaptiveinsights.com/products/adaptive-discovery
[91] https://www.alteryx.com/
[92] https://www.board.com/en
[93] https://www.clearanalyticsbi.com/
[94] https://www.datapine.com/
[95] https://www.domo.com/
[96] https://www.dundas.com/support/dundas-bi-free-trial
[97] https://www.gooddata.com/
[98] https://www.hubspot.com
[99] https://www.ibm.com/products/cognos-analytics
[100] https://www.infor.com/solutions/advanced-analytics/business-intelligence/birst
[101] https://www.jaspersoft.com/
[102] https://looker.com/
[103] https://powerbi.microsoft.com/pt-pt/
[104] https://www.microstrategy.com/en
[105] https://orangedatamining.com/

Pentaho[106], *Qlik*[107], *Qlik View*[108], *Query.me*[109], *SAP BusinessObjects*[110], *SAS Business Intelligence*[111], *Sisense*[112], *Style Intelligence*[113], *Tableau*[114], *Targit BI Suite*[115], *TIBCO Spotfire*[116], *Yellowfin BI*[117], *Xplenty*[118] e *Zoho Analytics*[119].

4.2.4. Ferramentas Data Warehousing

No que se refere às ferramentas de *Business Intelligence,* foram recolhidas e descritas no Anexo I 35 ferramentas, sendo estas: *Ab Initio*[120], *Alteryx*[121], *DynamoDB*[122], *Redshift*[123], *Astera DW Builder*[124], *BigQuery*[125], *CData Sync*[126], *Cloudera*[127], *Domo*[128], *Dundas*[129], *Exadata*[130], *Greenplum*[131], *IBM Db2 Warehouse*[132], *Informatica*[133], *Infor Birst*[134], *MariaDB*[135], *MarkLogic*[136], *Azure SQL*[137], *Vertica*[138], *SQL Server Integration Services*[139], *MicroStrategy*[140],

[106] https://www.hitachivantara.com/en-us/home.html
[107] https://www.qlik.com/us/products/qlik-sense
[108] https://www.qlik.com/us/
[109] https://query.me/
[110] https://www.sap.com/products/bi-platform.html
[111] https://www.sas.com/en_us/solutions/business-intelligence.html
[112] https://www.sisense.com/
[113] https://www.inetsoft.com/products/StyleIntelligence/
[114] https://www.tableau.com/
[115] https://www.targit.com/targit-decision-suite
[116] https://www.tibco.com/products/tibco-spotfire
[117] https://www.yellowfinbi.com/
[118] https://try.integrate.io/software-testing-help/
[119] https://www.zoho.com/analytics/
[120] https://www.abinitio.com/en/
[121] https://www.alteryx.com/
[122] https://aws.amazon.com/pt/dynamodb/
[123] https://aws.amazon.com/pt/redshift/
[124] https://www.astera.com/products/data-warehouse-builder/
[125] https://cloud.google.com/bigquery/
[126] https://www.cdata.com/sync/?utm_source=guru99&utm_medium=web
[127] https://www.cloudera.com/
[128] https://www.domo.com/
[129] https://www.dundas.com/support/dundas-bi-free-trial
[130] https://www.oracle.com/pt/engineered-systems/exadata/
[131] https://greenplum.org/
[132] https://www.ibm.com/products/db2-warehouse
[133] https://www.informatica.com/solutions/initiatives.html#fbid=de0ejW9AbjI
[134] https://www.infor.com/solutions/advanced-analytics/business-intelligence/birst
[135] https://mariadb.org/
[136] https://www.marklogic.com/
[137] https://azure.microsoft.com
[138] https://www.vertica.com/training/
[139] https://www.microsoft.com/en-us/download/
[140] https://www.microstrategy.com/en

Netezza[141], *Numetric*[142], *Panoply*[143], *Pentaho*[144], *PostgreSQL*[145], *QuerySurge*[146], *SAP HANA*[147], *SAS*[148], *Sisense*[149], *Snowflake*[150], *Tableau*[151], *Talend*[152], *Teradata*[153] e *Xplenty*[154]

4.2.5. Ferramentas *Big Data*

Relativamente às ferramentas de *Big Data,* foram recolhidas e descritas no Anexo I 35 ferramentas, sendo estas: *Adverity*[155], *Cassandra*[156], *Apache Flink*[157], *Apache Hadoop*[158], *SAMOA*[159], *Apache Spark*[160], *Apache Storm*[161], *Cloudera*[162], *CouchDB*[163], *DataCleaner*[164], *Dataddo*[165], *Datawrapper*[166], *Drill*[167], *Hadoop*[168], *HCATALOG*[169], *HPCC*[170], *Iceberg*[171], *Kafka*[172], *Kaggle*[173], *KNIME*[174], *Lumify*[175], *MongoDB*[176], *OpenRefine*[177], *Oozie*[178], *Pentaho*[179], *Presto*[180],

[141] https://www.ibm.com/products?types[0]=software
[142] https://numetric.com/
[143] https://panoply.io/
[144] https://www.hitachivantara.com/en-us/products
[145] https://www.postgresql.org/
[146] https://www.querysurge.com
[147] https://www.sap.com/products/hana.html
[148] https://www.sas.com/en_us/home.html
[149] https://www.sisense.com/
[150] https://www.snowflake.com/
[151] https://www.tableau.com/
[152] https://www.talend.com/
[153] https://www.teradata.com/
[154] https://try.integrate.io/software-testing-help/
[155] https://www.adverity.com/
[156] https://cassandra.apache.org/_/index.html
[157] https://flink.apache.org/
[158] https://hadoop.apache.org/
[159] https://incubator.apache.org/projects/samoa.html
[160] https://spark.apache.org/
[161] https://storm.apache.org/
[162] https://www.cloudera.com/
[163] https://couchdb.apache.org/
[164] https://github.com/datacleaner
[165] https://www.dataddo.com
[166] https://www.datawrapper.de/
[167] https://drill.apache.org/
[168] https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html
[169] https://hive.apache.org/hcatalog_downloads.html
[170] https://hpccsystems.com/
[171] https://iceberg.apache.org/
[172] https://www.redhat.com/
[173] https://www.kaggle.com/
[174] https://www.knime.com/
[175] http://www.altamiracorp.com/lumify/
[176] https://www.mongodb.com/
[177] https://openrefine.org/
[178] https://oozie.apache.org/
[179] https://www.hitachivantara.com
[180] https://prestodb.io/

Qubole[181], *R*[182], *RapidMiner*[183], *Samza*[184], *Stats iQ*[185], *Talend*[186], *Tableau*[187], *Xplenty*[188] e *Zoho Analytics*[189].

4.2.6. Ferramentas Machine Learning

No que se refere às ferramentas de *Machine Learning,* foram recolhidas e descritas no Anexo I 36 ferramentas, sendo estas: *Accord.net*[190], *Amazon Machine Learning*[191], *Apache Mahout*[192], *Apache Spark*[193], *Catalyst*[194], *CatBoost*[195], *Google Colab*[196], *CUV*[197], *Dask*[198], *DLIB*[199], *Fast.ai*[200], *Gensim*[201], *Glueviz*[202], *Cloud AutoML*[203], *H2O*[204],*Watson Machine Learning*[205], *Keras*[206], *KNIME*[207], *Jupyter Notebook*[208], *LightGBM*[209], *Mallet*[210], *Azure Machine*

[181] https://www.qubole.com/
[182] https://www.r-project.org/about.html
[183] https://rapidminer.com/
[184] https://samza.apache.org/
[185] https://www.qualtrics.com
[186] https://www.talend.com/
[187] https://www.tableau.com/
[188] https://try.integrate.io/software-testing-help/
[189] https://www.zoho.com/analytics
[190] http://accord-framework.net/
[191] https://aws.amazon.com/pt/sagemaker/
[192] https://mahout.apache.org//
[193] https://spark.apache.org/
[194] https://catalyst-team.github.io/catalyst/
[195] https://catboost.ai/
[196] https://colab.research.google.com/notebooks/welcome.ipynb
[197] https://www.ais.uni-bonn.de/deep_learning/doc/html/index.html
[198] https://dask.org/
[199] http://dlib.net/
[200] https://www.fast.ai/
[201] https://radimrehurek.com/gensim/
[202] https://glueviz.org/
[203] https://cloud.google.com/automl
[204] https://h2o.ai/
[205] https://www.ibm.com/cloud/watson-studio
[206] https://keras.io/
[207] https://www.knime.com/
[208] https://jupyter.org/
[209] https://lightgbm.readthedocs.io/en/latest/
[210] https://mallet.cs.umass.edu/index.php/Main_Page

Learning[211], *Numpy*[212], *Open NN*[213], *Pandas*[214], *PyTorch*[215], *R*[216], *Ramp*[217], *RapidMiner*[218], *Scikit-learn*[219], *Scipy*[220], *Shogun*[221], *TensorFlow*[222], *XGBoost*[223], *Weka*[224] e *Yellowfin BI*[225]

4.2.7. Ferramentas *Data Visualization*

De acordo com as ferramentas de *Data Visualization,* foram analisadas e descritas no Anexo I 32 ferramentas, sendo estas: *ChartBocks*[226], *DataBox*[227], *Datapine*[228], *Datawrapper*[229], *Domo*[230], *Dundas*[231], *Fusioncharts*[232], *GoodData*[233], *Google Charts*[234], *Google Data Studio*[235], *Highchats*[236], *IBM Cognos*[237], *Watson Machine Learning*[238], *Infogram*[239], *Jupyter Notebook*[240], *Klipfolio*[241], *Leaflet*[242], *Looker*[243], *Microdoft Excel*[244], *MicroStrategy*[245], *Microsoft Power*[246], *MicroStrategy*[247], *Qlik*[248], *Microsoft Power BI*[249], *OpenRefine*[250], *Ploty*[251], *RawGraphs*[252], *SAP*

[211] https://azure.microsoft.com/en-us/
[212] https://numpy.org/
[213] https://www.opennn.net/
[214] https://pandas.pydata.org/
[215] https://pytorch.org/
[216] https://www.r-project.org/about.html
[217] https://ramp.studio/
[218] https://rapidminer.com/
[219] https://scikit-learn.org/stable/
[220] https://scipy.org/
[221] https://www.shogun-toolbox.org/
[222] https://www.tensorflow.org/
[223] https://xgboost.ai/
[224] https://www.cs.waikato.ac.nz/ml/weka/
[225] https://www.yellowfinbi.com/
[226] https://www.chartblocks.com/en/
[227] https://www.databox.pt/
[228] https://www.datapine.com/
[229] https://www.datawrapper.de/
[230] https://www.domo.com/
[231] https://www.dundas.com/support/dundas-bi-free-trial
[232] https://www.fusioncharts.com/
[233] https://www.gooddata.com/
[234] https://developers.google.com/chart
[235] https://datastudio.google.com/u/0/
[236] https://www.highcharts.com/
[237] https://www.ibm.com/products/cognos-analytics
[238] https://www.ibm.com/cloud/watson-studio
[239] https://infogram.com/
[240] https://jupyter.org/
[241] https://www.klipfolio.com/
[242] https://leafletjs.com/
[243] https://looker.com/
[244] https://www.microsoft.com/en-us/microsoft-365/excel
[245] https://www.microstrategy.com/en
[246] https://powerbi.microsoft.com/pt-pt/
[247] https://www.microstrategy.com/en
[248] https://www.qlik.com/us/products/qlik-sense
[249] https://powerbi.microsoft.com/pt-pt/
[250] https://openrefine.org/
[251] https://plotly.com/
[252] https://rawgraphs.io/

Analytics Cloud[253], *Sisense*[254], *Tableau*[255], *Visme*[256], *Visual.ly*[257], *Whatagraph*[258] e *Zoho Analytics*[259]

4.2.8. Conclusão das ferramentas descritas

Como é visível na tabela 4 foram descritas 168 ferramentas, sendo que estas estão divididas por três cores:

- Vermelho – Ferramentas que se adequam a 1 ou 2 tipos de *data science*;
- Cor-de-Laranja – Ferramentas que se adequam a 3 ou 4 tipos de *data science*;
- Verde – Ferramentas que se adequam a 5 tipos *data science*.

Na integral e como se faz notar na tabela 4, não existem ferramentas adequadas a 6 ou mais tipos de *data science*, o que indica que não há nenhuma ferramenta 100% adequada.

Pode-se também constatar que 69% das ferramentas, isto é, 117, se adequam apenas a um tipo de *data science*, e que apenas 2% se adequam a cinco tipos. Tendo em conta isto, as quatro ferramentas mais abrangentes são *Sisence, Tableau, Xplenty* e *Zoho Analytics*.

Tabela 4-Ferramentas Data Science

Ordem	Ferramentas	*Data Analytics*	*Data Mining*	*Business Intelligence*	*Data Warehousin*	*Big Data*	*Machine Learning*	*Data Visualization*	Percentagens (%)
1.	*Ab Initio Software*				X				14.29
2.	*Accord.net*						X		14.29
3.	*Adaptive Discovery*			X					14.29
4.	*Advanced miner*		X						14.29
5.	*Adverity*	X				X			28.57

[253] https://www.sap.com/products/cloud-analytics.html
[254] https://www.sisense.com/
[255] https://www.tableau.com/
[256] https://www.visme.co/
[257] https://visual.ly/
[258] https://whatagraph.com/
[259] https://www.zoho.com/analytics

Ordem	Ferramentas	*Data Analytics*	*Data Mining*	*Business Intelligence*	*Data Warehousin*	*Big Data*	*Machine Learning*	*Data Visualization*	Percentagens (%)
6.	*Alteryx*		X	X	X				42.86
7.	*Amazon DynamoDB*				X				14.29
8.	*Amazon Machine Learning*						X		14.29
9.	*Amazon Redshift*				X				14.29
10.	*Analytic Solver*		X						14.29
11.	*Apache Cassandra*					X			14.29
12.	*Apache Flink*					X			14.29
13.	*Apache Hadoop*					X			14.29
14.	*Apache Mahout*		X			X	X		42.86
15.	*Apache SAMOA*					X			14.29
16.	*Apache Spark*	X	X			X	X		57.14
17.	*Apache Storm*	X				X			28.57
18.	*Astera DW Builder*				X				14.29
19.	*BigQuery*				X				14.29
20.	*Board*		X	X					28.57
21.	*Catalyst*						X		14.29
22.	*CData Sync*				X		X		28.57
23.	*ChartBlocks*							X	14.29
24.	*Chartio*	X							14.29
25.	*Civis*		X						14.29
26.	*Clear Analytics*			X					14.29
27.	*Cloudera*				X	X			28.57

Ordem	Ferramentas	*Data Analytics*	*Data Mining*	*Business Intelligence*	*Data Warehousin*	*Big Data*	*Machine Learning*	*Data Visualization*	Percentagens (%)
28.	*CouchDB*					X			14.29
29.	*CUV*						X		14.29
30.	*Dask*						X		14.29
31.	*DataBox*							X	14.29
32.	*Datadoo*	X				X			28.57
33.	*Datapine*			X				X	28.57
34.	*Datawatch*		X						14.29
35.	*Datawrapper*					X		x	28.57
36.	*Data Melt*		X						14.29
37.	*DLIB*						X		14.29
38.	*Domo*	X		X	X			X	57.14
39.	*Drill*					X			14.29
40.	*Dundas*		X	X	X			x	57.14
41.	*ELKI*		X						14.29
42.	*EnterpriseMiner*		X						14.29
43.	*Exadata*				X				14.29
44.	*Fast.ai*						X		14.29
45.	*Fusioncharts*							X	14.29
46.	*Gensim*						X		14.29
47.	*Glueviz*						X		14.29
48.	*GoodData*			X				X	28.57
49.	*Google Charts*							X	14.29
50.	*Google Colab*						X		14.29
51.	*Google Cloud AutoML*						X		14.29
52.	*Google Data Studio*	X						X	28.57

Ordem	Ferramentas	*Data Analytics*	*Data Mining*	*Business Intelligence*	*Data Warehousin*	*Big Data*	*Machine Learning*	*Data Visualization*	Percentagens (%)
53.	*Google Fusion Tables*	X							14.29
54.	*Grafana*	X							14.29
55.	*Greenplum*				X				14.29
56.	*Hadoop MapReduce*		X			X			28.57
57.	*HCATALOG*					X			14.29
58.	*Highchats*							X	14.29
59.	*HPCC*					X			14.29
60.	*HubSpot*	X		X					28.57
61.	*H2O*		X				X		28.57
62.	*H3O*		X						14.29
63.	*IBM Cognos*	X	X	X				X	42.86
64.	*IBM Db2 Warehouse*				X				14.29
65.	*IBM SPSS Modeler*		X						14.29
66.	*IBM Watson*						X	X	28.57
67.	*Iceberg*					X			14.29
68.	*Inetsoft*		X						14.29
69.	*Infogram*							X	14.29
70.	*Infor Birst*			X	X				28.57
71.	*Informatica*				X				14.29
72.	*Jaspersoft*			X					14.29
73.	*JuiceBox*	X							14.29
74.	*Jupyter Notebook*	X					X	X	42.86
75.	*Kafka*					X			14.29
76.	*Kaggle*		X			X			28.57

Ordem	Ferramentas	*Data Analytics*	*Data Mining*	*Business Intelligence*	*Data Warehousin*	*Big Data*	*Machine Learning*	*Data Visualization*	Percentagens (%)
77.	*Keras.io*						X		14.29
78.	*Klipfolio*							X	14.29
79.	*KNIME*	X	X			X	X		57.14
80.	*Leaflet*							X	14.29
81.	*LightGBM*						X		14.29
82.	*Looker*	X		X				X	42.86
83.	*Lumify*					X			14.29
84.	*Mallet*						X		14.29
85.	*MariaDB*				X				14.29
86.	*MarkLogic*				X				14.29
87.	*Metabase*	X							14.29
88.	*Micro Focus Vertica*				X				14.29
89.	*Microsoft Azure*				X				14.29
90.	*Microsoft Azure Machine Learning*						X		14.29
91.	*Microsoft Excel*	X						X	28.57
92.	*Microsoft Power BI*	X		X				X	42.86
93.	*Microsoft SQL Server Integration Services*				X				14.29
94.	*MicroStrategy*			X				X	28.57
95.	*Mode*	X							14.29
96.	*MongoDB*					X			14.29

Ordem	Ferramentas	*Data Analytics*	*Data Mining*	*Business Intelligence*	*Data Warehousin*	*Big Data*	*Machine Learning*	*Data Visualization*	Percentagens (%)
97.	*MonkeyLearn*		X						14.29
98.	*Netezza*				X				14.29
99.	*Numetric*				X				14.29
100.	*Numpy*						X		14.29
101.	*Oozie*					X			14.29
102.	*OpenNN*						X		14.29
103.	*OpenRefile*	X				X		X	42.86
104.	*Oracle Analytics Cloud*	X							14.29
105.	*Oracle BI*		X	X					28.57
106.	*Oracle Data Mining*		X						14.29
107.	*Orange*	X	X						28.57
108.	*Orobi*	X							14.29
109.	*Pandas*						X		14.29
110.	*Panoply*				X				14.29
111.	*Pentaho*			X	X	X			42.86
112.	*Periscope Data*	X							14.29
113.	*Ploty*							X	14.29
114.	*PolyAnalyst*		X						14.29
115.	*PostgreSQL*				X				14.29
116.	*Presto*					X			14.29
117.	*Python*	X	X						28.57
118.	*PyTorch*						X		14.29
119.	*Qlik*			X					14.29
120.	*QlikView*	X	X	X				X	57.14

Ordem	Ferramentas	*Data Analytics*	*Data Mining*	*Business Intelligence*	*Data Warehousin*	*Big Data*	*Machine Learning*	*Data Visualization*	Percentagens (%)
121.	*Qubole*					X			14.29
122.	*QuerySurge*				X				14.29
123.	*Query.me*	X		X					28.57
124.	*R*	X	X			X	X		57.14
125.	*Ramp*						X		14.29
126.	*RapidMiner*	X	X			X	X		57.14
127.	*Redash*	X							14.29
128.	*Rattle*		X						14.29
129.	*RawGraphs*							X	14.29
130.	*Samza*					X			14.29
131.	*SAP Analytics Cloud*							X	14.29
132.	*SAP BusinessObjects*	X		X					28.57
133.	*SAP HANA*				X				14.29
134.	*SAS Business Intelligence*	X		X					28.57
135.	*SAS Data Mining*		X						14.29
136.	*SAS Enterprise Miner*		X						14.29
137.	*SAS*	X			X				28.57
138.	*Scikit-learn*		X				X		28.57
139.	*SciPy*						X		14.29
140.	*Shogun*						X		14.29
141.	*Sisense*	X	X	X	X			X	71.43
142.	*Snowflake*				X				14.29
143.	*Sprinkle Data*	X							14.29

Ordem	Ferramentas	*Data Analytics*	*Data Mining*	*Business Intelligence*	*Data Warehousin*	*Big Data*	*Machine Learning*	*Data Visualization*	Percentagens (%)
144.	*SQL*	X							14.29
145.	*Solver*		X						14.29
146.	*Splunk*	X							14.29
147.	*SPMF*		X						14.29
148.	*SSDT*		X						14.29
149.	*Stats iQ*					X			14.29
150.	*Style Intelligence*			X					14.29
151.	*Tableau*	X		X	X	X		X	71.43
152.	*Talend*	X			X	X			42.86
153.	*Tanagra*		X						14.29
154.	*Targit BI*			X					14.29
155.	*TensorFlow*						X		14.29
156.	*Teradata*		X		X				28.57
157.	*Thoughtspot*	X							14.29
158.	*TIDAMI*	X							14.29
159.	*TIBCO Spotfire*	X		X					28.57
160.	*Viscovery*		X						14.29
161.	*Visme*							X	14.29
162.	*Visual.ly*							X	14.29
163.	*Yellowfin BI*			X			X		28.57
164.	*Whatagraph*	X						X	28.57
165.	*Weka*	X	X				X		42.86
166.	*XGBoost*						X		14.29
167.	*Xplenty*	X	X	X	X	X			71.43
168.	*Zoho Analytics*	X	X	X		X		X	71.43
	Número	45	43	30	35	35	36	32	

Ordem	Ferramentas	*Data Analytics*	*Data Mining*	*Business Intelligence*	*Data Warehousin*	*Big Data*	*Machine Learning*	*Data Visualization*	Percentagens (%)
	Percentagem	45/168 26.79%	43/168 25.60%	30/168 17.88%	35/168 20.83%	35/168 20.83%	36/168 21.43%	32/168 19.05%	

5. PLANO DE ATIVIDADES

Neste capítulo é apresentado o plano de atividades do projeto de dissertação. Na secção 5.1 estão representadas as principais atividades intrínsecas ao plano de trabalho, com as respetivas datas de início e fim e os seus precedentes. Na secção seguinte, ou seja, na secção 5.2 será ilustrada a lista com todos os riscos associados à dissertação, bem como possíveis impactos e ações de mitigação.

5.1. Planeamento das atividades

Na presente secção e mais propriamente na Figura 14 estão representadas todas as atividades intrínsecas à dissertação que se iniciou a 1 de novembro de 2021 e prevê a finalização no dia 17 de janeiro de 2023.

Task Name	Start	Finish	Predece
Plano de Atividades	**01/11/21 08:00**	**17/01/23 17:00**	
Plano de Trabalho	**01/11/21 08:00**	**22/11/21 17:00**	
Elaboração do Enquadramento e Motivação	01/11/21 08:00	03/11/21 17:00	
Descrição dos Objetivos e Resultados Esperados	04/11/21 08:00	08/11/21 17:00	3
Desenvolvimento da Calendarização	09/11/21 08:00	12/11/21 17:00	4
Elaboração do Documento do Plano de Trabalho	15/11/21 08:00	18/11/21 17:00	5
Revisão do Documento do Plano de Trabalho	19/11/21 08:00	19/11/21 17:00	6
Submissão do Documento do Plano de Trabalho	22/11/21 08:00	22/11/21 17:00	7
Pré-Dissertação	**07/03/22 08:00**	**15/08/22 17:00**	**8**
Introdução	**07/03/22 08:00**	**24/03/22 17:00**	
Desevolvimento da Contextualização	07/03/22 08:00	09/03/22 17:00	8
Elaboração do Enquadramento e Motivação	10/03/22 08:00	14/03/22 17:00	11
Definição dos Objetivos e Resultados Esperados	15/03/22 08:00	18/03/22 17:00	12
Definição da Estrutura do Documento	21/03/22 08:00	24/03/22 17:00	13
Revisão de Leitura	**25/03/22 08:00**	**04/05/22 17:00**	**10**
Definição da Estratégia de Pesquisa	25/03/22 08:00	30/03/22 17:00	14
Descrição de Soluções envolvendo Data Science, as suas metodologias e ferramentas	31/03/22 08:00	05/04/22 17:00	16
Abordagem Metodológica, Materiais e Métodos	28/04/22 08:00	04/05/22 17:00	17
Trabalho realizado	**05/05/22 08:00**	**20/06/22 17:00**	**18**
Elaboração do Levantamento das Metodologias de Data Science	05/05/22 08:00	13/05/22 17:00	17
Descrição das Metodologias	16/05/22 08:00	20/05/22 17:00	20
Desenvolvimento da Conclusão das metodologias descritas	23/05/22 08:00	27/05/22 17:00	21
Elaboração do Levantamento das Ferramentas de Data Science	30/05/22 08:00	07/06/22 17:00	22
Descrição das Ferramentas	08/06/22 08:00	14/06/22 17:00	23
Desenvolvimento da Conclusão das Ferramentas descritas	15/06/22 08:00	20/06/22 17:00	24
Plano de Atividades	**21/06/22 08:00**	**15/08/22 17:00**	**19**
Desenvolvimento do Plano de Atividades	21/06/22 08:00	28/06/22 17:00	25
Desenvolvimento da Lista de Riscos	29/06/22 08:00	07/07/22 17:00	27
Elaboração do Relatório de Pré-Dissertação	08/07/22 08:00	15/07/22 17:00	28
Realização de alterações no Relatório da Pré Dissertação	18/07/22 08:00	10/08/22 17:00	29
Revisão do Relatório de Pré-Dissertação	11/08/22 08:00	12/08/22 17:00	30
Submissão do Relatório de Pré Dissertação	15/08/22 08:00	15/08/22 17:00	31
Dissertação	**22/11/21 08:00**	**17/01/23 17:00**	**7**
Levantamento das Ferramentas	22/11/21 08:00	14/01/22 17:00	7
Descrição das Ferramentas	17/01/22 08:00	01/04/22 17:00	34
Levantamento das Metodologias Data Science	04/04/22 08:00	15/04/22 17:00	35
Descrição das Metodologias Data Science	18/04/22 08:00	06/05/22 17:00	36
Análise dos dados	16/08/22 08:00	09/09/22 17:00	9
Benchmarking das Metodologias Data Science	12/09/22 08:00	30/09/22 17:00	38
Definição da melhor metodologia	03/10/22 08:00	06/10/22 17:00	39
Identificação das variáveis de comparação das Ferramentas Data Science	07/10/22 08:00	10/10/22 17:00	40
Benchmarking das Ferramentas Data Science	11/10/22 08:00	28/10/22 17:00	41
Identificação da Ferramenta que melhor se adequa à metodologia escolhida	31/10/22 08:00	11/11/22 17:00	42
Avaliação do Resultado	14/11/22 08:00	25/11/22 17:00	43
Produzir o Relatório Final	28/11/22 08:00	02/12/22 17:00	44
Revisão do Relatório de Dissertação	05/12/22 08:00	05/12/22 17:00	45
Submissão da Dissertação	16/01/23 08:00	16/01/23 17:00	49;46
Apresentação da Dissertação	17/01/23 08:00	17/01/23 17:00	47
Escrita de Artigos Cientificos	13/01/23 08:00	13/01/23 17:00	45

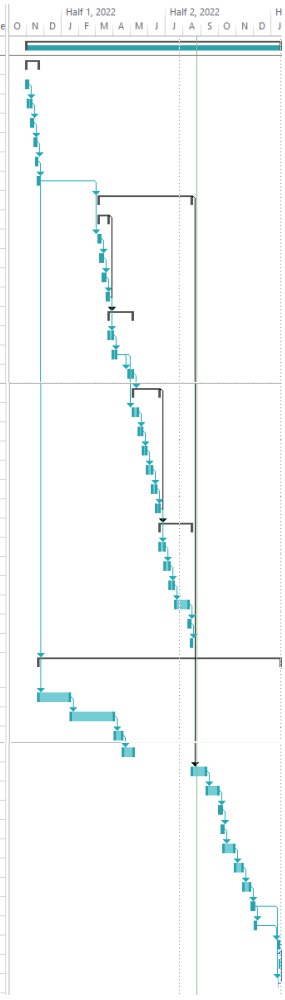

Figura 14-Plano de Atividades

5.2. Lista de riscos

Na Tabela 5 estão presentes os riscos que poderão ocorrer ao longo do desenvolvimento desta dissertação, com a respetiva ação de mitigação. Apresentam-se, também, estimativas para a sua probabilidade de ocorrência e impacto. A cada um dos riscos é atribuída um valor numa escala de 1 a 5, sendo que o 1 corresponde a uma pontuação baixa e o 5 corresponde uma alta. A seriedade de cada risco obtém-se multiplicando a probabilidade pelo impacto, permitindo identificar os riscos cujo impacto terá mais implicações na dissertação. Esta lista é preciosa para identificar riscos que podem ser prevenidos, de modo a provocarem o menor dano possível ao longo de todo o processo de desenvolvimento.

Tabela 5-Lista de Riscos

Risco	**Ação de mitigação**	**Impacto**	**Probabilidade**	**Seriedade**
Incumprimento dos prazos de entrega	Esforço redobrado no cumprimento do planeamento e reajuste de tarefas	5	2	10
Inexperiência com as metodologias a usar	Maior gasto de tempo para aprender o processo das metodologias necessárias para a execução do projeto	5	4	20
Planeamento da dissertação mal-executada	Revisão do planeamento, e alocação de mais tempo na execução das tarefas	4	3	12
Perda de todos os documentos	Guardar todos os dias todos os documentos importantes e atualizados na drive e no e-mail	5	1	5
Utilização de metodologias inadequadas ao projeto	Estudar aprofundadamente as metodologias que melhor se adequam finalidade da dissertação	4	2	8

Risco	Ação de mitigação	Impacto	Probabilidade	Seriedade
Incompreensão dos objetivos do projeto e dos resultados esperado	Manter constantemente o contacto e a comunicação ativa com o orientador	4	2	8
Alteração dos objetivos e Resultados esperados	Ajustar do plano de trabalho	3	2	6

6. CONCLUSÃO

Com o propósito de responder à questão de investigação "Existe alguma ferramenta totalmente abrangente que se consiga adaptar a uma metodologia que permite ajudar no desenvolvimento de projetos de *Data Science* com componentes analíticas e preditivas?", procedeu-se à realização do presente documento. Este documento engloba a contextualização, o enquadramento e as motivações do tema, definindo *Data Science* e as principais tecnologias que se aplicam neste, sendo estas, *Data Analytics, Data Mining, Business Intelligence, Data Warehousing, Big Data, Machine Learning* e *Data Visualization*.

Apesar de existirem inúmeras publicações, artigos e livros sobre *Data Science*, verificou-se a inexistência de algo semelhante na área, ou seja, algo que permita aos leitores identificar quais as ferramentas e metodologias mais adequadas ao *Data Science*, o que torna o tema em algo certamente mais motivacional.

Como ajuda no alinhamento da presente dissertação será seguida a metodologia *Benchmarking* que é atualmente considerada umas das abordagens mais eficazes na melhoria de desempenhos.

Também neste documento é apresentado o trabalho realizado, que, para além de bastante avançado, ainda se encontra numa fase preparatória daquele que será o resultado final.

Este documento é ainda uma parte inicial do projeto de dissertação, no entanto já se consegue ter uma ideia de como será o desenvolvimento esperado e quais os maiores desafios que irão surgir durante o desenvolvimento do mesmo.

7. REFERÊNCIAS

Aparicio, M., & Costa, C. J. (2014). Data visualization. In *Judicature* (Vol. 101, Issue 2, p. 11). https://doi.org/10.7551/mitpress/9780262062749.003.0011

Assaid, A., Cíntia, S., Andrade, R., Nacif, L. A., Ludmila, M., Gomes, E., Fernando De Oliveira, L., De, A., Pereira, C., Martins Da Mata, A. F., Andréa, S., Barra, O., Navarro, A., Bárbara, G., Toledo, D., Cíntia, L., Cristina, C., Lima, P., Felipe, C., … Pedroso, S. (n.d.). *Caderno de Resumos do I SER* (p. 76).

Barnes, T. J. (2013). Big data, little history. *Dialogues in Human Geography*, *3*(3), 297–302. https://doi.org/10.1177/2043820613514323

Batta, M. (2020). Machine Learning Algorithms - A Review. *International Journal of Science and Research (I)*, *9*(1), 1–7. https://doi.org/10.21275/ART20203995

Caroline, A. N. A., Rosa, D. A., & Itaquaquecetuba, F. (2020). *O Erp Como Ferramenta De Gestão De Estoque Nas Franquias De Fast-Food Na Cidade De Itaquaquecetuba*.

Castro, C. S. (2001). *Data Warehouse Estudo Comparativo das Ferramentas*.

Ferreira, J., Miranda, M., Abelha, A., & Machado, J. (2016). O Processo ETL em Sistemas Data Warehouse. *INForum 2010 - II Simpósio de Informática*, 10.

Golfarelli, M., & Rizzi, S. (2009). Data Warehouse Design: Modern Principles and Methodologies. In *Data Warehouse* (pp. 1–42, 420–423).

Hunsinger, S., & Waguespaack, L. (2016). A Comparison of Open Source Tools for Data Science. In *Journal of Information Systems Applied Research* (Vol. 9, Issue 2, p. 33).

Hurwitz, J., Nugent, A., Halper, F., & Kaufman, M. (2013). Big Data for Dummies. In *Wiley* (Issue 1).

Jason Brownlee. (2017). Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch. In *Master Machine Learning Algorithms: Vol. v1.12*.

Kimball, R., & Ross, M. (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. http://dl.acm.org/citation.cfm?id=560521

Madeira, P. J. (1999). Benchmarking: A arte de copiar. In *Jornal do Técnico de Contas e da Empresa* (Vol. 411, pp. 364–367).

Mohammed, M., Klan, M. B., & Bashier, E. B. M. (2017). *Machine Learning - Algorithms and Applications*.

Negash, S. (2004). *Business Intelligence* (Vol. 13). https://doi.org/10.17705/1CAIS.01315

Oussous, A., Benjelloun, F. Z., Ait Lahcen, A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences*, *30*(4), 431–448. https://doi.org/10.1016/j.jksuci.2017.06.001

Paiva, A. F. D. E., & Branco, P. (2018). *Aplicação De Data Science Como Ferramenta De Apoio A Tomada de Decisão Orientada por Dados*.

Practices, S. A. S. B., & Nevala, K. (2017). The MACHINE LEARNING Primer. In *SAS Institute Inc.*

Rollins, J. B. (2015). Metodologia de Base para Ciência de Dados. *IBM Analytics Route 100 Somers, NY 10589*. https://www.ibm.com/downloads/cas/B1WQ0GM2

Santos, N. H. S. dos. (2022). *Aplicação do Big Data e Data Science no Âmbito Jurídico*. *1*, 51–54.

Silva, M., & Sartori, M. (2015). *Data warehouse: A vantagem da modelagem dimensional de dados* (p. 10).

Sotero, J. (2022). *Inteligência Artificial e Big Data no Entretenimento Televisivo: aplicação e regulação na União Europeia* (p. 232).

Swamynathan, M. (2017). Mastering Machine Learning with Python in Six Steps. In *Scandinavian Journal of Information Systems* (Vol. 19, Issue 2). http://aisel.aisnet.org/sjis%0Ahttp://aisel.aisnet.org/sjis/vol19/iss2/4

Tripathi, A., Bagga, T., & Aggarwal, R. K. (2020). Strategic impact of business intelligence: A review of literature. In *Prabandhan: Indian Journal of Management* (Vol. 13, Issue 3, pp. 35–48). https://doi.org/10.17010/pijom/2020/v13i3/151175

Vaz de Oliveira e Sá, J. (2009). *Metodologia de Sistemas de Data Warehouse*. 24. http://repositorium.sdum.uminho.pt/bitstream/1822/10663/4/Tese de doutoramento_Jorge Vaz de Oliveira e Sá_2009.pdf

ANEXO I

Descrição das ferramentas *Data Analytics*

- ***Apache Spark*:** O *Apache Spark* é uma estrutura de software que permite que analistas e cientistas de dados processem rapidamente grandes conjuntos de dados. Este foi projetado para analisar *big data* não estruturado, o *Spark* distribui tarefas de análise computacionalmente pesadas em muitos computadores. Existem outras estruturas semelhantes do *Apache*, no entanto o *Spark* distingue-se pela sua rapidez, devido ao uso de RAM em vez de memória local. Este consta com uma biblioteca de algoritmos de *Machine learning, MLlib*, incluindo algoritmos de classificação, regressão e *clustering*. Um ponto menos benéfico é o facto de este consumir tanta memória, tornando-se computacionalmente caro. Este não dispõe de um sistema de gestão de arquivos, sendo que, é comum precisar da integração de outro *software*.
- ***Adverity:*** *Adverity* é uma plataforma de análise de marketing que permite que profissionais de marketing com o foco em dados tomem melhores decisões e melhorem o desempenho nas suas campanhas e canais. Esta acelera e simplifica relatórios de marketing através de painéis flexíveis, permitindo que o utilizador crie rapidamente relatórios avançados para os casos de uso de marketing mais comuns. Para além disto, a *Adverity* identifica as tendências e anomalias de modo a criar campanhas que tenham um impacto significativo no crescimento dos negócios com através de inteligência artificial.
- ***Apache Storm***: *Apache Storm* é uma ferramenta de *Big Data* no que se refere ao movimento de dados. Esta ferramenta funciona com dados estáticos e é faz as suas análises em tempo real ou processamento de fluxo.
- ***Chartio***: O *Chartio* é um sistema de inteligência de negócios *self-service* que se integra em vários *data warehouses* e permite a fácil importação de arquivos. O *Chartio* tem uma representação visual exclusiva do SQL que permite a construção de consultas do tipo apontar e clicar, o que permite que analistas de negócios que não estejam familiarizados com a sintaxe SQL modifiquem e experimentem consultas sem precisar se aprofundar na linguagem.

- ***Datadoo***: *Dataddo* não é propriamente uma ferramenta para a análise de dados, mas sim para a arquitetura de dados. Deste modo, fornece aos analistas dados limpos e integrados para a realização das suas tarefas de forma eficaz. O Dataddo é uma plataforma ETL baseada em nuvem sem codificação, esta tem uma vasta diversidade de conectores e métricas totalmente personalizáveis. Esta plataforma facilita ainda a criação de pipelines.Esta tem um custo de 20$ por fonte de dados por mês no mínimo.
- ***Domo***: O *Domo* fornece mais de mil integrações integradas que permitem que os usuários transfiram dados de e para sistemas locais e externos na nuvem. O *Domo* também oferece suporte à criação de aplicativos personalizados que se integram à plataforma, o que permite aos desenvolvedores estender o sistema com acesso imediato aos conectores e ferramentas de visualização. O *Domo* vem como uma plataforma única que inclui um *data warehouse* e *software ETL*, portanto, as empresas que já têm seu próprio data warehouse e pipeline de dados configurados podem querer procurar noutro lugar.
- ***Google Data Studio***: O *Google Data Studio* é uma ferramenta gratuita de painel e visualização de dados que se integra automaticamente à maioria das outras aplicações do *Google*, como *Google Analytics, Google Ads* e Google BigQuery. Graças à sua integração com outros serviços da *Google*, o *Data Studio* é ótimo para quem precisa analisar os seus dados do *Google*. Por exemplo, os profissionais de marketing podem criar painéis para os seus dados do *Google Ads* e *Analytics* para entenderem melhor a conversão e a retenção de clientes. O *Data Studio* também pode trabalhar com dados de várias outras fontes, desde que os dados sejam replicados primeiro para o *BigQuery* através de um pipeline de dados como o *Stitch*.
- ***Google Fusion Tables:*** *Google Fusion Tables* é uma plataforma gratuita que ajuda a coletar, visualizar e compartilhar as informações em tabelas de dados. Esta tem capacidade para trabalhar com grandes conjuntos de dados e estes podem ser visualizados através de gráficos, mapas e gráficos de rede.
- ***Grafana***: O *Grafana* é um software gratuito de análise de dados e de código aberto com o propósito de monitorar e observar métricas em diversas bases de dados e aplicações.
- ***HubSpot:*** HubSpot é um *software* de análise de marketing que tem como propósito o sucesso de campanhas de *marketing*. Uma das funcionalidades mais populares do *HubSpot* é o rastreamento de ações num dado *website*, sendo o objetivo conhecer o

usuário e poder proporcionar a este um conteúdo adequado. Este *software* fornece ainda um relatório detalhado com base no *website*, em *e-mails*, publicações, entre outros. O *hubSpot* fornece todos os dados que são necessários para a tomada de decisões inteligentes. Este é constituído por três planos com preços distintos, o primeiro, nomeado como *Starter* tem um preço de 40 $ por mês no mínimo, o segundo, o *Professional,* tem um custo de 800$ por mês no mínimo e o terceiro, o *Enterprise*, o seu valor será a partir dos 3200$ por mês. Todos os planos contem ferramentas de *marketing* gratuitas.

- ***IBM Cognos***: O *IBM Cognos* é uma plataforma de inteligência de negócios que apresenta ferramentas de Inteligência Artifical integradas para revelar *insights* ocultos em dados e explicá-los em linguagem simples. O *Cognos* também possui ferramentas automatizadas de preparação de dados para limpar e agregar fontes de dados automaticamente, o que permite integrar e experimentar rapidamente as fontes de dados para análise.
- ***JuiceBox:*** *Juicebox* cria visualizações e apresentações de dados interativas e visualmente atraentes. O modelo de preços é gratuito para indivíduos e acessível para equipes. Esta plataforma é intuitiva e tem um layout responsivo para visualização móvel. Esta plataforma consta com uma narrativa de dados inspirados no jornalismo de dados moderno e nas técnicas de visualização. *Juicebox* tem dois planos, o gratuito para até três usuários com uso ilimitado e o plano de equipa que tem um custo de 49$ por mês para 5 editores e 15 visualizadores.
- ***Jupyter Notebook***: O *Jupyter Notebook* é uma ferramenta *Web* gratuita e de código aberto. Esta permite que os desenvolvedores criem relatórios com dados e visualizações de código ao vivo. O sistema suporta mais de quarenta linguagens de programação. O *Jupyter Notebook* é usado principalmente para partilhar código, criar tutoriais e apresentar trabalhos.
- ***KNIME:*** *KNIME* é uma plataforma de análise de dados gratuita e de código aberto que oferece suporte à integração, processamento, visualização e geração de relatórios de dados. O *KNIME* é ótimo para cientistas de dados que precisam de integrar e processar dados para *machine learning* e outros modelos estatísticos, mas não necessariamente têm fortes habilidades de programação. A interface gráfica permite a análise e a modelagem de apontar e clicar.

- ***Looker:*** *Looker* é uma plataforma de *business intelligence* e análise de dados baseada em nuvem. Esta apresenta geração automática de modelo de dados que verifica esquemas de dados e infere relacionamentos entre tabelas e fontes de dados. Os engenheiros de dados podem modificar os modelos gerados por meio de um editor de código integrado.
- ***Metabase:*** *Metabase* é uma ferramenta de análise e inteligência de negócios gratuita e de código aberto. O *Metabase* permite que os usuários "façam perguntas" sobre os dados, que é uma maneira de usuários não técnicos usarem uma interface de apontar e clicar para a construção de consultas. Isso funciona bem para filtragem e agregações simples; usuários mais técnicos podem ir direto ao SQL bruto para análises mais complexas. O *Metabase* também tem a capacidade de enviar resultados de análise para sistemas externos como o *Slack*.
- ***Microsoft Excel***: *Microsoft Excel* é dos *software*s mais conhecidos do mundo. Além de que, contem cálculos e funções gráficas ideais para a análise de dados. Os seus recursos integram tabelas dinâmicas e ferramentas de criação de formulários. Apesar destas vantagens, este possui também algumas limitações, nomeadamente a lentidão com grandes conjuntos de dados e a tendência a fazer aproximações com grandes valores, levando a imprecisões.
- ***Microsoft Power BI***: O *Microsoft Power BI* é uma plataforma de *business intelligence* de topo com suporte para dezenas de fontes de dados. Esta plataforma permite que os usuários criem e compartilhem relatórios, visualizações e painéis. O *Microsoft Power BI* consegue ser usado em diferentes
- ***Mode***: O *Mode* é uma solução moderna de análise e *Business Intelligence* que combina *SQL, Python, R* e análise visual para responder a perguntas mais rapidamente. Ele fornece um editor SQL interativo e um ambiente de *Notebook* para análise, juntamente com ferramentas de visualização e colaboração para usuários menos técnicos. O modo tem um mecanismo de dados exclusivo chamado *Helix* que transmite dados de bases de dados externos e armazena-os na memória para permitir uma análise rápida e interativa. Ele suporta análise na memória de até 10 *GigaBytes* de dados.
- ***Oracle Analytics Cloud***: O *Oracle Analytics Cloud* uma plataforma de *business intelligence* e análise de dados baseadas em nuvem. Este tem como propósito ajudar grandes organizações a transformar os seus sistemas antiquados em plataformas de nuvem

digitais. Os utilizadores recorrem à sua vasta gama de recursos analíticos, como visualizações básicas ou algoritmos de *Machine Learning*, para obter *insights* de dados.

- ***Orobi:*** *Oribi* é uma plataforma de análise de *marketing* que pretende valorizar todo o potencial que um website tem para dar da contagem de cada clique de botão, visitas à página, envio de formulários, entre outras. Todas estas informações serão úteis aquando do alcance de objetivos.
- ***Orange***: *Orange* uma ferramenta de visualização de dados e *Machine Learning*. Esta possui um sistema de código aberto próprio para especialistas, mas também para iniciantes, e fornece vários algoritmos de classificação e regressão. Sendo esta também uma plataforma de visualização interativa, é possível selecionar pontos de dados de um gráfico de dispersão e nós em arvores. Esta ferramenta é gratuita.
- ***OpenRefine***: *OpenRefine* é um *software* de análise de dados de código aberto. Este *software* gratuito permitirá organizar, limpar e transformar dados, bem como a importação e exportação de diferentes formatos de arquivos. Este suporta ainda catorze idiomas.
- ***Periscope Data***: *Periscope Data* é agora propriedade da *Sisense* e é uma plataforma de inteligência de negócios que suporta integrações para uma variedade de data *warehouses* e bases de dados populares. Os analistas técnicos podem transformar dados através do uso de *SQL, Python* ou *R*, e usuários menos técnicos podem criar e compartilhar painéis com facilidade. O *Periscope Data* também possui várias certificações de segurança, como *HIPAA-HITECH*.
- **Python**: *Python* é uma linguagem de programação de alto nível com uma enorme gama de usos. Esta pretende valorizar o esforço humano sobre o computacional, sendo uma linguagem de programação acessível e popular. *Python* contem uma ampla variedade de bibliotecas de recursos adequadas para uma diversidade de tarefas de análise de dados. A principal desvantagem do *Python* é sua velocidade, sendo que consome muita memória e é mais lento do que muitas outras linguagens.
- ***QlikView:*** *Qlik* é uma plataforma de análise de dados para negócios inteligentes que oferece suporte à implantação na nuvem e no local. A ferramenta possui forte suporte para exploração e descoberta de dados por usuários técnicos e não técnicos. A *Qlik* oferece

suporte a muitos tipos de gráficos que os usuários podem personalizar com *SQL* incorporado e módulos de arrastar e soltar.

- ***Query.me***: O *Query.me* permite analisar e visualizar dados através de ferramentas simples que não exigem conhecimentos de programação para além do *SQL*. Estando o *SQL* cada vez mais popular, o *Query.me* procura criar uma solução que substitui ferramentas *SQL* que exigem muita mão de obra e tempo. Para além disto, procura também criar relatórios que ajudam as empresas a transformar dados em histórias.
- ***R***: *R* é uma linguagem de programação de código aberto. Esta é normalmente usada para criar software de análise estatística ou de dados. *R* foi criado especificamente para lidar com tarefas pesadas de computação estatística e é bastante popular no que diz respeito à visualização de dados. Um aspeto negativo desta linguagem é a sua péssima gestão de memória.
- ***RapidMiner***: O *RapidMiner* fornece toda a tecnologia que os usuários precisam para integrar, limpar e transformar dados antes de executar análises preditivas e modelos estatísticos. Os usuários podem realizar quase tudo isso por meio de uma interface gráfica simples. O *RapidMiner* também pode ser estendido usando *scripts R* e *Python*, e vários plugins de terceiros que estão disponíveis no mercado da empresa. No entanto, o produto é altamente otimizado para sua interface gráfica para que os analistas possam preparar dados e executar modelos por conta própria.
- ***Redash***: *Redash* é uma ferramenta para consultar fontes de dados e criar visualizações. O código é de código aberto e uma versão hospedada acessível está disponível para organizações que desejam começar rapidamente. O núcleo do *Redash* é o editor de consultas, que fornece uma interface simples para escrever consultas, explorar esquemas e gerenciar integrações. Os resultados da consulta são armazenados em cache no *Redash* e os usuários podem agendar atualizações para serem executadas automaticamente.
- ***SAP BusinessObjects***: *SAP BusinessObjects* fornece um conjunto de ferramentas de análise de dados que ajudam na descoberta, análise e geração de relatórios de dados. As ferramentas são intuitivas e por isso perfeitas para novos utilizadores, mas também úteis para a realização de análises complexas. Este permite análises preditivas de autoatendimento. O *SAP BusinessObjects* incorpora produtos *Microsoft Office*, permitindo

aos analistas de negócios reverter e alternar facilmente entre aplicações, como *Excel* e relatórios do *BusinessObjects*.

- ***SAS Business Intelligence***: O *SAS Business Intelligence* fornece um conjunto de aplicações para análises *self-service*. Ele tem muitos recursos de colaboração integrados, como a capacidade de enviar relatórios para aplicações móveis. Embora o *SAS Business Intelligence* seja uma plataforma abrangente e flexível, pode ser mais caro do que alguns dos seus concorrentes. Empresas maiores podem achar que vale a pena o preço devido à sua versatilidade.
- ***SAS***: *SAS* significa *Statistical Analysis System* e é uma ferramenta de *Business Intelligence* e *Data Analytics*. Esta é principalmente usada para perfis de clientes, relatórios, *Data Mining* e modelagem preditiva. O *SAS* é um produto comercial, tendo, por isso, um preço alto. Este custo traz benefícios, nomeadamente os módulos que são adicionados regularmente, conforme a necessidade dos clientes.
- ***Sprinkle Data***: *Sprinkle* é uma plataforma que permite integrar, combinar e modelar dados sem escrever nenhum código. A capacidade da *Sprinkle* criar relatórios, segmentos e painéis com facilidade, mesmo para usuários sem experiência, diferencia-a das restantes ferramentas de *Business Intelligence* do mercado.
- ***SQL***: *Structured query Language* é uma linguagem padrão para bases de dados relacionais, sendo foi baseada em álgebra relacional. Esta é popular pela sua simplicidade e fácil uso. *SQL* é usada para qualquer tipo de manipulação no registo de bases de dados, e é possível criar, inserir, atualizar, excluir e consultar informações que estejam inseridas na base de dados.
- ***Sisense***: *Sisense* é uma plataforma de análise de dados destinada a ajudar os desenvolvedores técnicos e analistas de negócios a processar e visualizar todos os seus dados de negócios. Um aspeto exclusivo da plataforma *Sisense* é sua tecnologia *In-Chip* customizada, que otimiza a computação para utilizar o cache da *CPU* em vez de *RAM* mais lenta. Para alguns fluxos de trabalho, isso pode levar a um cálculo de dez a cem vezes mais rápido.
- ***Splunk***: O *Splunk* é uma ferramenta de análise de dados. Esta começou com o propósito de processar dados de arquivos de log de máquina, mas recentemente tornou-se também uma plataforma com opções de visualização e uma interface *web* de fácil uso.

- ***Tableau***: O *Tableau* é uma plataforma de análise e visualização de dados que permite aos usuários criar relatórios e compartilhá-los em plataformas móveis e de *desktop*, num navegador ou incorporados em aplicações. Ele pode ser executado na nuvem ou no local. Grande parte da plataforma do *Tableau* é executada na sua linguagem de consulta principal, o *VizQL*. Isso traduz o painel de arrastar e soltar e os componentes de visualização em consultas de back-end eficientes e minimiza a necessidade de otimizações de desempenho do usuário final. No entanto, o *Tableau* não oferece suporte para consultas SQL avançadas.
- ***Talend***: *Talend* é uma plataforma baseada em nuvem para integração de dados. Esta funciona em várias plataformas de computação em nuvem, nomeadamente, a *AWS, Google Cloud, Azure* e *Snowflake*. Ele suporta vários ambientes de nuvem, públicos, privados e híbridos. O *Talend* possui produtos gratuitos e comerciais, sendo que os produtos gratuitos podem ser usados em *Windows* e *Mac*. A análise desta plataforma é feita em tempo real e *Internet of Things*.
- ***Thoughtspot***: O *Thoughtspot* é uma plataforma de análise de dados que permite aos utilizadores explorar dados de diferentes fontes através de relatórios e pesquisas em linguagem natural. O *SpotIQ* é o sistema com inteligência artificial desta plataforma, este procura automaticamente insights de modo a ajudar os utilizadores a descobrir tendências que eles não sabiam pesquisar.
- ***TIDAMI***: O *TIDAMI* é um software com o propósito de ajudar engenheiros e cientistas a analisar dados numéricos. Este é útil para os utilizadores que pretendem modificar ou criar parâmetros, integrar modelos, visualizar grandes quantidades de dados, selecionar fases de interesse em particular, e comparar medidas com modelos. *Tidami* conta com uma versão gratuita com funcionalidades limitadas e com duas versões pagas, uma com o custo de 250€ por ano para pequenas empresas ou estudantes, e outra com o valor de 500€ por ano para empresas.
- ***TIBCO Spotfire***: *TIBCO Spotfire* é uma plataforma que permite que todos explorem e visualizem novas descobertas em dados através de painéis imersivos e análises avançadas. A análise *Spotfire* oferece recursos em escala, incluindo análise preditiva, análise de geolocalização e análise de *streaming*.

- ***Whatagraph:*** O *Whatagraph* é uma plataforma que oferece a análise visual de dados com entrada de fontes de dados automáticas e funcionalidades de arrastar e soltar para criar relatórios. A maior vantagem do *Whatagraph* é a sua fácil configuração e interface visual intuitiva. Esta plataforma consta com opções de visualização de dados através de formas gráficas, tabelas*, widgets* de rastreamento de *KPI,* entre outras.
- ***Weka:*** A *Weka* é uma plataforma que fornece algoritmos de *Machine Learning* para *Data Mining*. Esta plataforma gratuita fornece ferramentas para a preparação de dados, classificação, regressão, *clustering*, mineração de regras de associação e visualização. E pode ser utilizada em *Microsoft Windows, Mac* e *Linux*.
    - ***Xplenty***: A *Xplenty* é uma solução de extração, transformação e carregamento (*ETL*) e integração de dados. Esta possui ferramentas de transformação na plataforma que ajudam a limpar, a normalizar e a transformar os seus através de ótimas práticas. Através da sua interface gráfica intuitiva é possível implementar *ETL (Extract, Transform, Load), ELT (Extract, Load, Transform), ETLT* (o melhor de *ETL* e *ELT*) ou replicação. Este faz a transferência de dados entre bases de dados, *Data Warehouse* e *Data Lakes*. O *Xplenty* contem um conector de *API Rest*, de modo a conseguir extrair dados que os usuários necessitem de qualquer *API Rest*.
- ***Zoho Analytics***: O *Zoho Analytics* é uma plataforma de fácil acesso que permite analisar dados de uma ampla variedade de fontes. Esta destaca-se pela oferta de uma vasta diversidade de ferramentas de visualização de dados, na forma de tabelas dinâmicas, gráficos, painéis temáticos, entre outros. Esta plataforma fornece ainda um assistente com inteligência artificial que permite que os usuários obtenham respostas inteligentes no formato de relatórios relevantes. O *Zoho Analytics* consta com cinco planos, um primeiro gratuito, os restantes têm um preço mensal, sendo que o *Basic* que tem um custo de 22$, o *Standard* 45$, o Premium 112$, e o *Enterprise* 445$.

Descrição das ferramentas *Data Mining*

- ***Advanced miner***: *Advanced Miner* é uma ferramenta para processamento, análise e modelagem de dados. Esta tem uma interface amigável que permite a exploração de vários tipos de dados. A *Advanced Miner* constrói modelos estatísticos, faz uma análise de importância variável, e de agrupamento.

- ***Analytic Solver***: O *Analytic Solver* é uma ferramenta gratuita que permite fazer análises de risco e análises prescritivas no *browser*. Esta oferece trabalhos de *Data Mining* de potência total. A ferramenta *Analytic Solver* disponibiliza simulações *Monte Carlo*, análises de risco, árvores de decisão e otimização de simulações.
- ***Alteryx:*** *Alteryx* é uma ferramenta de *Business Intelligence* e *Analytics* para empresas. Esta ferramenta é fundamentalmente projetada para analistas de dados e líderes de negócios. *Alteryx* disponibiliza um processamento analítico online rápido, relatórios automáticos e um painel altamente personalizável.
- ***Apache Mahout***: O *Apache Mahout* é uma plataforma gratuita de código aberto com *Machine Learning*. O seu principal propósito é ajudar os cientistas ou pesquisadores a implementar os seus próprios algoritmos. Esta plataforma foca-se em três áreas principais, nomeadamente, em mecanismos de recomendação, *clustering* e classificação. O *Apache Mahout* é ótimo para projetos de *Data Mining* complexos e que envolvam grandes quantidades de dados.
- ***Apache Spark*:** O *Apache Spark* é uma estrutura de software que permite que analistas e cientistas de dados processem rapidamente grandes conjuntos de dados. Este foi projetado para analisar *big data* não estruturado, o *Spark* distribui tarefas de análise computacionalmente pesadas em muitos computadores. Existem outras estruturas semelhantes do *Apache*, no entanto o *Spark* distingue-se pela sua rapidez, devido ao uso de RAM em vez de memória local. Este consta com uma biblioteca de algoritmos de *Machine learning, MLlib*, incluindo algoritmos de classificação, regressão e *clustering*. Um ponto menos benéfico é o facto de este consumir tanta memória, tornando-se computacionalmente caro. Este não dispõe de um sistema de gestão de arquivos, sendo que, é comum precisar da integração de outro *software*.
- ***Board***: *Board* é um *software* de tomada de decisão com foco em *Business Intelligence*, Gestão de Performance e Análise Preditiva. Através desta plataforma é possível analisar, simular, planear e prever preferências futuras.
- ***Civis***: *Civis* é uma ferramenta que oferece soluções rápidas e disponibiliza arquitetura, produtos e processos que ajudam os clientes a proteger dados.
- ***Datawatch***: O *Datawatch Desktop* é uma plataforma de *Data Mining* e *Business Intellingence*, que permite a visualização de dados em tempo real, oferecendo

ferramentas que permitem a construção e a implantação de sistemas de análise sem a necessidade de escrever uma única linha de código.

- ***Data Melt:*** *DataMelt* é uma ferramenta gratuita de computação numérica, matemática, análise de dados e visualização de dados. Esta dispõe linguagens de script, nomeadamente*, Python, Ruby* e *Groovy* com o poder de inúmeros pacotes *Java*. É possível usar o *DaraMelt* em diferentes Sistemas Operativos.
- ***Dundas***: *Dundas* é uma ferramenta de *Data Mining* elaborada para empresas que pretendem construir e visualizar painéis interativos e personalizáveis, relatórios, entre outros. Esta ferramenta faz uma análise preditiva e avançada.
- ***ELKI***: *ELKI* é uma ferramenta de *Data Mining* de código aberto escrita em *Java*. Esta permite pesquisar algoritmos, com base em métodos que não tem supervisão em análise de cluster e deteção de *outliers*. A *ELKI* oferece uma avaliação e um *benchmarking* fácil e justo através de algoritmos.
- ***Enterprise Miner***: O *Enterprise Miner* é um software produzido pela *SAS* para *Data mining* que ajuda a melhorar a precisão de previsões. Este fornece uma modelagem preditiva e descritiva, e disponibiliza os resultados através de gráficos de fácil entendimento.
- ***H2O***: *H2O* é uma plataforma de *Machine Learning* de código aberto que disponibiliza tecnologias de inteligência artificial. Esta oferece suporte a algoritmos de *Machine Learning* de forma mais rápida e fácil, até para os que não se consideram especialistas. O H2O pode ser integrado por meio de uma *API* e emprega computação distribuída na memória, o que o torna ideal para analisar grandes conjuntos de dados.
- ***H3O***: O *H3O* é um *software* de *Data Mining* de código aberto que é usado especialmente por organizações para analisar dados armazenados na infraestrutura de nuvem. Esta ferramenta é escrita em linguagem *R*, mas também é compatível com *Python* para construção de modelos. Uma das maiores vantagens é que o *H3O* permite uma implementação rápida e fácil em produção devido ao suporte à linguagem Java.
- ***Hadoop MapReduce***: *Hadoop* é uma ferramenta de código aberto que lida com *big data*. *Hadoop* é escrito em *Java*, no entanto qualquer linguagem de programação pode ser utilizada com o *Hadoop Streaming. MapReduce* que é uma implementação do Hadoop e um modelo de programação. Esta tem sido uma solução adotada para *Data Mining* em *Big Data*.

- ***IBM Cognos***: O *IBM Cognos* é uma plataforma de inteligência de negócios que apresenta ferramentas de Inteligência Artifical integradas para revelar *insights* ocultos em dados e explicá-los em linguagem simples. O *Cognos* também possui ferramentas automatizadas de preparação de dados para limpar e agregar fontes de dados automaticamente, o que permite integrar e experimentar rapidamente as fontes de dados para análise.
- ***IBM SPSS Modeler***: *IBM SPSS Modeller* é uma solução visual de *Data Science* e *Machine Learning*, que ajuda a reduzir tempos acelerando tarefas operacionais para cientistas de dados. Este *software* é principalmente usado para preparar dados, fazer análises preditivas e gerir modelos.
- ***Inetsoft:*** *Inetsoft Intelligence* é uma plataforma de *Data Mining* que fornece um software de *Business Intelligence* fácil, ágil e robusto, permitindo a transformação rápida e flexível de dados de várias fontes. Esta plataforma oferece painéis interativos e visualmente atraentes sendo apelativos ao utilizador final.
- ***Kaggle***: A *Kaggle* é uma plataforma que se iniciou em competições de *Machine Learning*, no entanto, agora, está também presente como uma plataforma de *Data Science* baseada em nuvem pública. Esta fornece mais de cinquenta mil conjuntos de dados públicos úteis para *Data Mining*.
- ***KNIME:*** *KNIME* é uma plataforma de análise de dados gratuita e de código aberto que oferece suporte à integração, processamento, visualização e geração de relatórios de dados. O *KNIME* é ótimo para cientistas de dados que precisam de integrar e processar dados para *machine learning* e outros modelos estatísticos, mas não necessariamente têm fortes habilidades de programação. A interface gráfica permite a análise e a modelagem de apontar e clicar.
- ***MonkeyLearn***: *MonkeyLearn* é uma plataforma de *Machine Learning* especializada em *Data Mining*, que disponibiliza uma interface amigável e intuitiva. As ferramentas de *Data Mining* da *MonkeyLearn* e o seu painel de visualização de dados personalizável são constantemente usados para uma melhoria constante na tomada de decisões de várias organizações, através deteção de tendências e de padrões.
- ***Oracle Data Mining***: A *Oracle* é um *software* bem popular de bases de dados de Tecnologias da Informação. Este contém vários algoritmos de *Data Mining* para classificação, regressão, previsão, entre outras. A Oracle *Data Mining* conta com a ajuda

da equipa técnica da *Oracle* para construir uma infraestrutura de *Data Mining* mais robusta.

- ***Oracle BI***: O *Oracle BI* é uma plataforma *Machine Learning* de código aberto e visualização de dados para especialistas e não só. Esta oferece uma exploração de dados interativa para a análise qualitativa rápida e, também, uma vasta gama de complementos para *Data Mining* de fontes de dados externas.
- ***Orange***: *Orange* uma ferramenta de visualização de dados e *Machine Learning*. Esta possui um sistema de código aberto próprio para especialistas, mas também para iniciantes, e fornece vários algoritmos de classificação e regressão. Sendo esta também uma plataforma de visualização interativa, é possível selecionar pontos de dados de um gráfico de dispersão e nós em arvores. Esta ferramenta é gratuita.
- ***PolyAnalyst***: *PolyAnalyst* é uma ferramenta Data Mining que permite aceder a dados de várias fontes, incorporando dados de diferentes fontes. Com esta é possível selecionar uma vasta seleção de algoritmos estatísticos e de *Machine Learning*.
- **Python**: *Python* é uma linguagem de programação de alto nível com uma enorme gama de usos. Esta pretende valorizar o esforço humano sobre o computacional, sendo uma linguagem de programação acessível e popular. *Python* contem uma ampla variedade de bibliotecas de recursos adequadas para uma diversidade de tarefas de análise de dados. A principal desvantagem do *Python* é sua velocidade, sendo que consome muita memória e é mais lento do que muitas outras linguagens.
- ***QlikView:*** *Qlik* é uma plataforma de análise de dados para negócios inteligentes que oferece suporte à implantação na nuvem e no local. A ferramenta possui forte suporte para exploração e descoberta de dados por usuários técnicos e não técnicos. A *Qlik* oferece suporte a muitos tipos de gráficos que os usuários podem personalizar com *SQL* incorporado e módulos de arrastar e soltar.
- ***R***: *R* é uma linguagem de programação de código aberto. Esta é normalmente usada para criar software de análise estatística ou de dados. *R* foi criado especificamente para lidar com tarefas pesadas de computação estatística e é bastante popular no que diz respeito à visualização de dados. Um aspeto negativo desta linguagem é a sua péssima gestão de memória.

- ***RapidMiner***: O *RapidMiner* fornece toda a tecnologia que os usuários precisam para integrar, limpar e transformar dados antes de executar análises preditivas e modelos estatísticos. Os usuários podem realizar quase tudo isso por meio de uma interface gráfica simples. O *RapidMiner* também pode ser estendido usando *scripts R* e *Python*, e vários plugins de terceiros que estão disponíveis no mercado da empresa. No entanto, o produto é altamente otimizado para sua interface gráfica para que os analistas possam preparar dados e executar modelos por conta própria.
- ***Rattle***: *Rattle* é uma ferramenta gratuita e de código aberto baseada na linguagem *R* para *Data Mining*. Esta ferramenta pode ser usada para obter resumos estatísticos e visuais de dados, para criar modelos de *Machine Learning*, e para comparar o desempenho de modelos graficamente.
- ***SAS Enterprise Miner***: *SAS Enterprise Miner* é uma ferramenta de *Data Mining* que cria modelos para um melhor desempenho usando algoritmos inovadores e métodos específicos do setor. Esta verifica os resultados com avaliação visual e métricas de validação.
- ***SAS Data Mining***: O *Statistical Analysis System Data Mining* é um produto desenvolvido pela *SAS*, de modo a analisar e gerir dados. Esta plataforma é bastante popular em *Data Mining* e oferece uma interface gráfica intuitiva, sendo assim útil para utilizadores não especializados.
- ***Scikit-learn***: O *Scikit-learn* é uma ferramenta de software gratuita para *Machine Learning* em *Python*. Esta disponibiliza ótimos recursos de *Data Mining* e *Data Analytics*. Esta ferramenta oferece um vasto número de recursos como classificação, regressão, agrupamento, pré-processamento, seleção de modelo e redução de dimensão.
- ***Sisense***: *Sisense* é uma plataforma de análise de dados destinada a ajudar os desenvolvedores técnicos e analistas de negócios a processar e visualizar todos os seus dados de negócios. Um aspeto exclusivo da plataforma *Sisense* é sua tecnologia *In-Chip* customizada, que otimiza a computação para utilizar o cache da *CPU* em vez de *RAM* mais lenta. Para alguns fluxos de trabalho, isso pode levar a um cálculo de dez a cem vezes mais rápido.

- ***Solver***: O *Solver* é uma ferramenta de *Data Mining* de nível profissional para visualização de dados, previsão e *Data Mining* em *Excel*. Esta é de usabilidade fácil e oferece um conjunto amplo de recursos de baseados em métodos estatísticos e *Machine Learning*.
- ***SPMF***: *SPMF* é uma biblioteca de *Data Mining* de código aberto escrita em *Java*. Esta permite que o utilizador integre o código-fonte com outro *software Java*. *SPMF* dá apoio ao processo complexo de *clustering* e classificação.
- ***SSDT (SQL Server Data Tools)***: O *SQL Server Data Tools (SSDT)* é uma ferramenta que permite criar bases de dados relacionais do *SQL Server*, bases de dados no *Azure SQL*, modelos de dados do *Analysis Services*, pacotes do *Integration Services* e relatórios do *Reporting Services*. Com o *SSDT* é possível projetar e implantar qualquer tipo de conteúdo do *SQL Server* com a mesma facilidade com que o desenvolveria no *Visual Studio*.
- ***Teradata***: *Teradata* é uma plataforma de análise de dados em nuvem. Através desta plataforma não são necessários conhecimentos de programação para codificar algoritmos complexos de *Machine Learning*.
- ***Tanagra:*** *Tanagra* é uma ferramenta de *Data Mining* gratuita para fins de estudo e pesquisa. Esta oferece inúmeros métodos de *Data Mining*, análise de dados e *Machine Learning*. E ainda disponibiliza um software de *Data Mining* fácil e intuitivo para cientistas de dados, pesquisadores e estudantes.
- ***Viscovery***: *Viscovery* é um software de usabilidade fácil que se baseia em mapas auto-organizados e estatísticas multivariadas para *Data Mining* e modelagem preditiva. Esta ferramenta foi das primeiras de *Data Mining* na Europa, deste modo as suas soluções destacam-se pela orientação intuitiva e implementação madura.
- ***Weka:*** A *Weka* é uma plataforma que fornece algoritmos de *Machine Learning* para *Data Mining*. Esta plataforma gratuita fornece ferramentas para a preparação de dados, classificação, regressão, *clustering*, mineração de regras de associação e visualização. E pode ser utilizada em *Microsoft Windows, Mac* e *Linux*.
- ***Xplenty***: A *Xplenty* é uma solução de extração, transformação e carregamento (*ETL*) e integração de dados. Esta possui ferramentas de transformação na plataforma que ajudam a limpar, a normalizar e a transformar os seus através de ótimas práticas. Através da sua interface gráfica intuitiva é possível implementar *ETL (Extract, Transform, Load), ELT (Extract, Load, Transform), ETLT* (o melhor de *ETL* e *ELT*) ou replicação. Este faz a

transferência de dados entre bases de dados, *Data Warehouse* e *Data Lakes*. O *Xplenty* contem um conector de *API Rest*, de modo a conseguir extrair dados que os usuários necessitem de qualquer *API Rest*.

- ***Zoho Analytics***: O *Zoho Analytics* é uma plataforma de fácil acesso que permite analisar dados de uma ampla variedade de fontes. Esta destaca-se pela oferta de uma vasta diversidade de ferramentas de visualização de dados, na forma de tabelas dinâmicas, gráficos, painéis temáticos, entre outros. Esta plataforma fornece ainda um assistente com inteligência artificial que permite que os usuários obtenham respostas inteligentes no formato de relatórios relevantes. O *Zoho Analytics* consta com cinco planos, um primeiro gratuito, os restantes têm um preço mensal, sendo que o *Basic* que tem um custo de 22$, o *Standard* 45$, o Premium 112$, e o *Enterprise* 445$.

Descrição das ferramentas *Business Intelligence*

- ***Adaptative Discovery***: *Adaptive Discovery* é uma ferramenta baseada em nuvem que integra fontes de dados e facilita a análise de dados visualizados no painel. Contem análise, consultas e relatórios ad hoc, visualização de dados, planeamento estratégico, entre outras.
- ***Alteryx***: *Alteryx* é uma ferramenta de *Business Intelligence* e *Analytics* para empresas. Esta ferramenta é fundamentalmente projetada para analistas de dados e líderes de negócios. *Alteryx* disponibiliza um processamento analítico online rápido, relatórios automáticos e um painel altamente personalizável.
- ***Board***: *Board* é uma plataforma de *software* de tomada de decisão com foco em *Business Intelligence*, Gestão de Performance e Análise Preditiva. Através desta plataforma é possível analisar, simular, planear e prever preferências futuras.
- ***Clear Analytics***: *Clear Analytics* é uma ferramenta de *Business Intelligence* fácil e intuitiva baseada em Excel. Esta fornece um sistema de *Business Intelligence* de autoatendimento que oferece diversos recursos como criar, automatizar, analisar e visualizar os dados da organização em questão. Este *software* também funciona com a ferramenta *Microsoft Power BI*, através do uso de *Power Query* e *Power Pivot* para limpar e modelar diferentes conjuntos de dados.

- ***Datapine:*** *Datapine* é uma ferramenta de *Business Intelligence* que permite que analistas de dados integrem facilmente diversas fontes de dados, realizando análises avançadas de dados e criando painéis de negócios interativos que giram insights de negócios acionáveis.
- ***Domo***: O *Domo* fornece mais de mil integrações integradas que permitem que os usuários transfiram dados de e para sistemas locais e externos na nuvem. O *Domo* também oferece suporte à criação de aplicativos personalizados que se integram à plataforma, o que permite aos desenvolvedores estender o sistema com acesso imediato aos conectores e ferramentas de visualização. O *Domo* vem como uma plataforma única que inclui um *data warehouse* e *software ETL*, portanto, as empresas que já têm seu próprio data warehouse e pipeline de dados configurados podem querer procurar noutro lugar.
- ***Dundas***: *Dundas* é uma ferramenta de *Data Mining* elaborada para empresas que pretendem construir e visualizar painéis interativos e personalizáveis, relatórios, entre outros. Esta ferramenta faz uma análise preditiva e avançada.
- ***GoodData***: *GoodData* é uma plataforma de *Business Intelligence* que disponibiliza ferramentas de armazenamento, consultas analíticas, visualizações e integração de aplicações, onde se podem criar painéis e relatórios.
- ***HubSpot:*** HubSpot é um *software* de análise de marketing que tem como propósito o sucesso de campanhas de *marketing*. Uma das funcionalidades mais populares do *HubSpot* é o rastreamento de ações num dado *website*, sendo o objetivo conhecer o usuário e poder proporcionar a este um conteúdo adequado. Este *software* fornece ainda um relatório detalhado com base no *website*, em *e-mails*, publicações, entre outros. O *hubSpot* fornece todos os dados que são necessários para a tomada de decisões inteligentes. Este é constituído por três planos com preços distintos, o primeiro, nomeado como *Starter* tem um preço de 40 $ por mês no mínimo, o segundo, o *Professional,* tem um custo de 800$ por mês no mínimo e o terceiro, o *Enterprise*, o seu valor será a partir dos 3200$ por mês. Todos os planos contem ferramentas de *marketing* gratuitas.
- ***IBM Cognos***: O *IBM Cognos* é uma plataforma de inteligência de negócios que apresenta ferramentas de Inteligência Artifical integradas para revelar *insights* ocultos em dados e explicá-los em linguagem simples. O *Cognos* também possui ferramentas automatizadas de preparação de dados para limpar e agregar fontes de dados automaticamente, o que permite integrar e experimentar rapidamente as fontes de dados para análise.

- ***Infor Birst:*** O *Infor Birst* é uma plataforma de análise *Business Intelligence* baseada em nuvem que ajuda os utilizadores a compreender e otimizar processos complexos. Esta é uma solução avançada de ponta a ponta com *Data Warehouse*, que fornece uma plataforma de visualização e relatórios.
- ***Jaspersoft***: *Jaspersoft* é uma ferramenta de *Business Intelligence* de código aberto que oferece uma análise interativa, relatórios, *OLAP*, visualização de dados e integração de dados aos seus utilizadores. Esta fornece suporte para o processo de tomada de decisão por meio de indicadores-chave de desempenho e indicadores de tendência.
- ***Looker:*** *Looker* é uma plataforma de *business intelligence* e análise de dados baseada em nuvem. Esta apresenta geração automática de modelo de dados que verifica esquemas de dados e infere relacionamentos entre tabelas e fontes de dados. Os engenheiros de dados podem modificar os modelos gerados por meio de um editor de código integrado.
- ***Microsoft Power BI***: O *Microsoft Power BI* é uma plataforma de *business intelligence* de topo com suporte para dezenas de fontes de dados. Esta plataforma permite que os usuários criem e compartilhem relatórios, visualizações e painéis. O *Microsoft Power BI* consegue ser usado em diferentes.
- ***MicroStrategy***: A *MicroStrategy* é uma ferramenta de *business intelligence* para empresas que disponibiliza painéis poderosos, análise de dados e soluções em nuvem. Esta permite a identificação de tendências, o reconhecimento de oportunidades, a melhoraria da produtividade, entre outras. Para iniciantes é algo complexa.
- ***Oracle BI***: O *Oracle BI* é uma plataforma *Machine Learning* de código aberto e visualização de dados para especialistas e não só. Esta oferece uma exploração de dados interativa para a análise qualitativa rápida e, também, uma vasta gama de complementos para *Data Mining* de fontes de dados externas.
- ***Pentaho:*** *Pentaho* é uma ferramenta de código aberto que se baseia na tomada de decisões de negócios precisas e orientadas por dados. Esta fornece análises interativas, navegação e visualização de dados e possibilita a integração de *Big Data*, *Data Mining* e análise preditiva de dados.
- ***Qlik***: *Qlik* é uma ferramenta de *Data Mining* e visualização de dados, que oferece painéis e suporte a diferentes fontes de dados e tipos de arquivo. Esta disponibiliza também interfaces de arrastar e soltar para criar visualizações de dados flexíveis e interativas.

- ***QlikView:*** *Qlik* é uma plataforma de análise de dados para negócios inteligentes que oferece suporte à implantação na nuvem e no local. A ferramenta possui forte suporte para exploração e descoberta de dados por usuários técnicos e não técnicos. A *Qlik* oferece suporte a muitos tipos de gráficos que os usuários podem personalizar com *SQL* incorporado e módulos de arrastar e soltar.
- ***Query.me***: O *Query.me* permite analisar e visualizar dados através de ferramentas simples que não exigem conhecimentos de programação para além do *SQL*. Estando o *SQL* cada vez mais popular, o *Query.me* procura criar uma solução que substitui ferramentas *SQL* que exigem muita mão de obra e tempo. Para além disto, procura também criar relatórios que ajudam as empresas a transformar dados em histórias.
- ***SAP BusinessObjects***: *SAP BusinessObjects* fornece um conjunto de ferramentas de análise de dados que ajudam na descoberta, análise e geração de relatórios de dados. As ferramentas são intuitivas e por isso perfeitas para novos utilizadores, mas também úteis para a realização de análises complexas. Este permite análises preditivas de autoatendimento. O *SAP BusinessObjects* incorpora produtos *Microsoft Office*, permitindo aos analistas de negócios reverter e alternar facilmente entre aplicações, como *Excel* e relatórios do *BusinessObjects*.
- ***SAS Business Intelligence***: O *SAS Business Intelligence* fornece um conjunto de aplicações para análises *self-service*. Ele tem muitos recursos de colaboração integrados, como a capacidade de enviar relatórios para aplicações móveis. Embora o *SAS Business Intelligence* seja uma plataforma abrangente e flexível, pode ser mais caro do que alguns dos seus concorrentes. Empresas maiores podem achar que vale a pena o preço devido à sua versatilidade.
- ***Sisense***: *Sisense* é uma plataforma de análise de dados destinada a ajudar os desenvolvedores técnicos e analistas de negócios a processar e visualizar todos os seus dados de negócios. Um aspeto exclusivo da plataforma *Sisense* é sua tecnologia *In-Chip* customizada, que otimiza a computação para utilizar o cache da *CPU* em vez de *RAM* mais lenta. Para alguns fluxos de trabalho, isso pode levar a um cálculo de dez a cem vezes mais rápido.
- ***Style Intelligence***: *Style Intelligence* é uma ferramenta gratuita de *Business Intelligence* de código aberto que ajuda na exploração de dados, conectando bases de dados relacionais

e multidimensionais. Esta ferramenta de desenvolvimento ágil, robusta e de autoatendimento fornece suporte à nuvem e segurança granular.

- ***Tableau***: O *Tableau* é uma plataforma de análise e visualização de dados que permite aos usuários criar relatórios e compartilhá-los em plataformas móveis e de *desktop*, num navegador ou incorporados em aplicações. Ele pode ser executado na nuvem ou no local. Grande parte da plataforma do *Tableau* é executada na sua linguagem de consulta principal, o *VizQL.* Isso traduz o painel de arrastar e soltar e os componentes de visualização em consultas de back-end eficientes e minimiza a necessidade de otimizações de desempenho do usuário final. No entanto, o *Tableau* não oferece suporte para consultas *SQL* avançadas.
- ***Targit BI***: O *Targit BI Suite* é uma ferramenta de tomada de decisão poderosa e de fácil usabilidade. Fornece um pacote de autoatendimento de painéis e disponibiliza geração fácil de relatórios de *insights*.
- ***TIBCO Spotfire***: *TIBCO Spotfire* é uma plataforma que permite que todos explorem e visualizem novas descobertas em dados através de painéis imersivos e análises avançadas. A análise *Spotfire* oferece recursos em escala, incluindo análise preditiva, análise de geolocalização e análise de *streaming*.
- ***Yellowfin BI***: O *Yellowfin BI* é uma ferramenta de *Business Intelligence* que concilia visualização, *Machine Learning* e colaboração. É ainda possível filtrar facilmente uma vasta quantidade de dados através da filtragem intuitiva, bem como abrir painéis em praticamente qualquer lugar. Uma vantagem desta ferramenta é a facilidade de desenvolvimento sem necessidade de código ou usando muito pouco.
    - ***Xplenty***: A *Xplenty* é uma solução de extração, transformação e carregamento (*ETL*) e integração de dados. Esta possui ferramentas de transformação na plataforma que ajudam a limpar, a normalizar e a transformar os seus através de ótimas práticas. Através da sua interface gráfica intuitiva é possível implementar *ETL (Extract, Transform, Load), ELT (Extract, Load, Transform), ETLT* (o melhor de *ETL* e *ELT*) ou replicação. Este faz a transferência de dados entre bases de dados, *Data Warehouse* e *Data Lakes*. O *Xplenty* contem um conector de *API Rest*, de modo a conseguir extrair dados que os usuários necessitem de qualquer *API Rest*.

- ***Zoho Analytics***: O *Zoho Analytics* é uma plataforma de fácil acesso que permite analisar dados de uma ampla variedade de fontes. Esta destaca-se pela oferta de uma vasta diversidade de ferramentas de visualização de dados, na forma de tabelas dinâmicas, gráficos, painéis temáticos, entre outros. Esta plataforma fornece ainda um assistente com inteligência artificial que permite que os usuários obtenham respostas inteligentes no formato de relatórios relevantes. O *Zoho Analytics* consta com cinco planos, um primeiro gratuito, os restantes têm um preço mensal, sendo que o *Basic* que tem um custo de 22$, o *Standard* 45$, o Premium 112$, e o *Enterprise* 445$.

Descrição das ferramentas *Data Warehousing*

- ***Ab Initio Software***: *Ab Initio* é uma ferramenta especializada em processamento e integração de dados de alto volume que fornece produtos de armazenamento de dados fáceis de usar. Esta manipula e processa dados tanto quantitativos como qualitativos.
- ***Alteryx***: *Alteryx* é uma ferramenta de *Business Intelligence* e *Analytics* para empresas. Esta ferramenta é fundamentalmente projetada para analistas de dados e líderes de negócios. *Alteryx* disponibiliza um processamento analítico online rápido, relatórios automáticos e um painel altamente personalizável.
- ***Amazon DynamoDB***: O *DynamoDB* é uma ferramenta de base de dados *NoSQL* baseada em nuvem para empresas que consegue processar *petabytes* de dados em segundos. Assim, as tabelas podem ser dimensionadas automaticamente adicionando novas colunas com base nos requisitos crescentes.
- ***Amazon Redshift***: *Redshift* é uma ferramenta de armazenamento de dados baseada em nuvem para empresas que consegue processar *petabytes* de dados em segundos, daí ser a ferramenta adequada para análises de dados de alta velocidade. Esta ferramenta pretende otimizar o desempenho do Data Warehouse e reduzir os custos operacionais.
- ***Astera DW Builder***: *Astera DW Builder* é uma ferramenta de armazenamento de dados de ponta a ponta que permite aos utilizadores projetar, desenvolver e implantar data *warehouses* de alto volume usando uma abordagem orientada a metadados.
- ***BigQuery:*** O *BigQuery* é uma ferramenta de armazenamento de dados económica com recursos integrados de *Machine Learning*. Esta executa consultas com *petabytes* de dados

em segundos para análises em tempo real. Este data warehouse nativo da nuvem permite analisar dados baseados em localização.

- ***CData Sync***: *CData Sync* é uma ferramenta que replica dados *Cloud/SaaS* para bases de dados e *data warehouse*. Através desta é possível conectar os dados com *Business Intelligence, Analytics* e *Machine Learning*.
- ***Cloudera***: *Cloudera* é uma ferramenta que fornece uma base de dados operacional localizada na nuvem com baixa latência e alta simultaneidade. Esta plataforma é ótima para a analise de *big data* e para a extração de *Business Intelligence* em tempo real.
- ***Domo***: O *Domo* fornece mais de mil integrações integradas que permitem que os usuários transfiram dados de e para sistemas locais e externos na nuvem. O *Domo* também oferece suporte à criação de aplicativos personalizados que se integram à plataforma, o que permite aos desenvolvedores estender o sistema com acesso imediato aos conectores e ferramentas de visualização. O *Domo* vem como uma plataforma única que inclui um *data warehouse* e *software ETL*, portanto, as empresas que já têm seu próprio data warehouse e pipeline de dados configurados podem querer procurar noutro lugar.
- ***Dundas***: *Dundas* é uma ferramenta de *Data Mining* elaborada para empresas que pretendem construir e visualizar painéis interativos e personalizáveis, relatórios, entre outros. Esta ferramenta faz uma análise preditiva e avançada.
- ***Exadata***: *Exadata* é uma ferramenta autónoma de *Data Warehousing* que através do *Machine Learning* automatiza tarefas administrativas. A *Exadata* fornece métodos fáceis para o armazenamento dos dados.
- ***Greenplum***: *Greenplum* é uma ferramenta pouco dispendiosa de *Big Data*. Esta usa técnicas de processamento paralelo maciço, consistindo em nós mestres, nós de espera e nós de segmento.
- ***IBM Db2 Warehouse***: *IBM Db2 Warehouse* é uma ferramenta escalável para armazenar dados em nuvem totalmente. Esta para além de fornecer ferramentas de *Machine Learning* é, também, realmente útil para análise de dados e inteligência artificial.
- ***Informatica:*** *Informatica* é uma ferramenta de *data Warehousing* bem popular. Esta possui um portfólio em integração de dados, *ETL,* integração de dados, entre outras.
- ***Infor Birst:*** O *Infor Birst* é uma plataforma de análise *Business Intelligence* baseada em nuvem que ajuda os utilizadores a compreender e otimizar processos complexos. Esta é

uma solução avançada de ponta a ponta com *Data Warehouse*, que fornece uma plataforma de visualização e relatórios.

- ***MariaDB***: *MariaDB* é uma ferramenta de base de dados empresarial em tempo real que aplica um processamento paralelo massivo.
- ***MarkLogic:*** *MarkLogic* é uma ferramenta que fornece um sistema de base de dados *NoSQL* com consultas e serviços poderosos. Esta plataforma permite a inserção de diversos tipos de dados de vido ao seu armazenamento nativo para esquemas predefinidos. Os formatos suportados incluem dados geoespaciais, *JSON* e binários massivos.
- ***Microsoft Azure:*** O *Azure SQL* é uma base de dados relacional baseada em nuvem da *Microsoft*. Esta permite otimizações para carregamento e processamento de dados em escala de *petabytes* e relatórios em tempo real.
- ***Micro Focus Vertica***: *Vertica* é uma ferramenta data warehouse SQL disponível na nuvem em plataformas como *AWS* e *Azure*. Esta oferece recursos integrados para análise, nomeadamente *Machine Learning*, correspondência de padrões e séries temporais. A *Vertica* usa compactação para otimizar o armazenamento.
- ***Microsoft SQL Server Integration Services***: O *SQL Server Integration Services* é uma ferramenta de armazenamento de dados usada para realizar operações de *ETL*; isto é, extrair, transformar e carregar dados. O *SQL Server Integration* também inclui um vasto conjunto de tarefas internas.
- ***MicroStrategy***: A *MicroStrategy* é uma ferramenta de *business intelligence* para empresas que disponibiliza painéis poderosos, análise de dados e soluções em nuvem. Esta permite a identificação de tendências, o reconhecimento de oportunidades, a melhoraria da produtividade, entre outras. Para iniciantes é algo complexa.
- ***Netezza***: *Netezza* é uma ferramenta da *IBM* que fornece um sistema integrado especializado e possui os principais recursos de design de velocidade, simplicidade, escalabilidade e poder analítico.
- ***Numetric***: *Numetric* é uma ferramenta de *business intelligence* que se conecta automaticamente, limpa e filtra dados, fornecendo dados importantes para os utilizadores. Esta filtra instantaneamente grandes quantidades de linhas de dados e fornece um *data warehouse* pessoal.

- ***Panoply***: *Panoply* é uma ferramenta de fácil usabilidade que combina *ETL*, ou seja, a extração, a transformação e o carregamento, juntamente com o *Data Warehousing*.
- ***Pentaho:*** *Pentaho* é uma ferramenta de código aberto que se baseia na tomada de decisões de negócios precisas e orientadas por dados. Esta fornece análises interativas, navegação e visualização de dados e possibilita a integração de *Big Data*, *Data Mining* e análise preditiva de dados.
- ***PostgreSQL***: *PostgreSQL* é uma ferramenta de gestão de bases de dados de código aberto disponível na nuvem. Esta pode ser usada para trabalhar com dados geoespaciais com a integração de *PostgreSQL* e com a extensão *PostGIS*. Esta integração permitirá oferecer soluções de negócios baseadas em localização.
- ***QuerySurge***: *QuerySurge* é uma ferramenta produzida especificamente para automatizar o teste de *Data Warehouses e Big Data*. Esta garante que os dados extraídos das fontes de dados permaneçam intactos nos sistemas de destino também.
- ***SAP HANA***: *SAP HANA* é uma ferramenta baseada em nuvem com recursos de cache na memória. Esta contem um processamento rápido de transações em tempo real e a análise de dados. *SAP HANA* fornece uma interface simples e centralizada para acesso, integração e virtualização de dados.
- ***SAS***: *SAS* significa *Statistical Analysis System* e é uma ferramenta de *Business Intelligence* e *Data Analytics*. Esta é principalmente usada para perfis de clientes, relatórios, *Data Mining* e modelagem preditiva. O *SAS* é um produto comercial, tendo, por isso, um preço alto. Este custo traz benefícios, nomeadamente os módulos que são adicionados regularmente, conforme a necessidade dos clientes.
- ***Sisense***: *Sisense* é uma plataforma de análise de dados destinada a ajudar os desenvolvedores técnicos e analistas de negócios a processar e visualizar todos os seus dados de negócios. Um aspeto exclusivo da plataforma *Sisense* é sua tecnologia *In-Chip* customizada, que otimiza a computação para utilizar o cache da *CPU* em vez de *RAM* mais lenta. Para alguns fluxos de trabalho, isso pode levar a um cálculo de dez a cem vezes mais rápido.
- ***Snowflake:*** *Snowflake* é uma ferramenta usada para configurar um *data warehouse* em nuvem a nível empresarial. Com esta ferramenta é possível analisar dados de várias fontes

não estruturadas e estruturadas. A escalabilidade desta ferramenta acelera o desempenho de consultas para fornecer insights acionáveis mais rapidamente.

- ***Tableau***: O *Tableau* é uma plataforma de análise e visualização de dados que permite aos usuários criar relatórios e compartilhá-los em plataformas móveis e de *desktop*, num navegador ou incorporados em aplicações. Ele pode ser executado na nuvem ou no local. Grande parte da plataforma do *Tableau* é executada na sua linguagem de consulta principal, o *VizQL.* Isso traduz o painel de arrastar e soltar e os componentes de visualização em consultas de back-end eficientes e minimiza a necessidade de otimizações de desempenho do usuário final. No entanto, o *Tableau* não oferece suporte para consultas *SQL* avançadas.
- ***Talend***: *Talend* é uma plataforma baseada em nuvem para integração de dados. Esta funciona em várias plataformas de computação em nuvem, nomeadamente, a *AWS, Google Cloud, Azure e Snowflake*. Ele suporta vários ambientes de nuvem, públicos, privados e híbridos. O *Talend* possui produtos gratuitos e comerciais, sendo que os produtos gratuitos podem ser usados em *Windows e Mac*. A análise desta plataforma é feita em tempo real *e IoT.*
- ***Teradata:*** *Teradata* é uma plataforma de análise de dados em nuvem. Através desta plataforma não são necessários conhecimentos de programação para codificar algoritmos complexos de *Machine Learning*.
    - ***Xplenty***: A *Xplenty* é uma solução de extração, transformação e carregamento (*ETL*) e integração de dados. Esta possui ferramentas de transformação na plataforma que ajudam a limpar, a normalizar e a transformar os seus através de ótimas práticas. Através da sua interface gráfica intuitiva é possível implementar *ETL (Extract, Transform, Load), ELT (Extract, Load, Transform), ETLT* (o melhor de *ETL* e *ELT*) ou replicação. Este faz a transferência de dados entre bases de dados, *Data Warehouse* e *Data Lakes*. O *Xplenty* contem um conector de *API Rest*, de modo a conseguir extrair dados que os usuários necessitem de qualquer *API Rest*.

Descrição das ferramentas *Big Data*

- ***Adverity:*** *Adverity* é uma plataforma de análise de marketing que permite que profissionais de *marketing* com o foco em dados tomem melhores decisões e melhorem

o desempenho nas suas campanhas e canais. Esta acelera e simplifica relatórios de marketing através de painéis flexíveis, permitindo que o utilizador crie rapidamente relatórios avançados para os casos de uso de *marketing* mais comuns. Para além disto, a *Adverity* identifica as tendências e anomalias de modo a criar campanhas que tenham um impacto significativo no crescimento dos negócios com através de inteligência artificial.

- ***Apache Cassandra***: *Cassandra* é uma ferramenta de gestão de bases de dados distribuída que é usado para lidar com grandes volumes de dados em vários servidores. Esta é uma das tecnologias de *Big Data* mais populares, e é frequentemente usada para o processamento de conjuntos de dados estruturados.
- ***Apache Flink***: *Apache Flink* é uma ferramenta de análise de dados de código aberto para processamento de *big data*. Esta disponibiliza dados distribuídos, de alto desempenho, sempre disponíveis e precisos.
- ***Apache Hadoop***: *Apache Hadoop* é uma ferramenta de *big data* que permite o processamento distribuído de grandes conjuntos de dados em clusters de computadores. Esta é uma ferramenta popular no que diz respeito a ferramentas de *big data* projetadas para escalar de servidores únicos para milhares de máquinas.
- ***Apache SAMOA***: *SAMOA* é uma ferramenta de código aberto para *Data Mining, Big Data* e *Machine Learning* que permite a criação de algoritmos de *Machine Learning*. Esta é uma ferramenta de fácil usabilidade, rápida e escalável e com *streaming* em tempo real.
- ***Apache Spark***: O *Apache Spark* é uma estrutura de *software* que permite que analistas e cientistas de dados processem rapidamente grandes conjuntos de dados. Este foi projetado para analisar *big data* não estruturado, o *Spark* distribui tarefas de análise computacionalmente pesadas em muitos computadores. Existem outras estruturas semelhantes do *Apache*, no entanto o *Spark* distingue-se pela sua rapidez, devido ao uso de *RAM* em vez de memória local. Este consta com uma biblioteca de algoritmos de *Machine learning, MLlib*, incluindo algoritmos de classificação, regressão e *clustering*. Um ponto menos benéfico é o facto de este consumir tanta memória, tornando-se computacionalmente caro. Este não dispõe de um sistema de gestão de arquivos, sendo que, é comum precisar da integração de outro software.

- ***Apache Storm***: *Apache Storm* é uma ferramenta de *Big Data* no que se refere ao movimento de dados. Esta ferramenta funciona com dados estáticos e é faz as suas análises em tempo real ou processamento de fluxo.
- ***Cloudera***: *Cloudera* é uma ferramenta que fornece uma base de dados operacional localizada na nuvem com baixa latência e alta simultaneidade. Esta plataforma é ótima para a analise de *big data* e para a extração de *Business Intelligence* em tempo real.
- ***CouchDB***: *CouchDB* é uma ferramenta de armazenamento dados em documentos *JSON* que podem ser consultados na *web* ou usando *JavaScript*.
- ***DataCleaner:*** *DataCleaner* é uma ferramenta de análise de qualidade de dados e uma plataforma de solução. Disponibiliza um motor de perfil de dados forte, interativo e exploratório. Esta ferramenta transforma, padroniza e valida dados e relatórios.
- ***Datadoo:*** *Dataddo* não é propriamente uma ferramenta para a análise de dados, mas sim para a arquitetura de dados. Deste modo, fornece aos analistas dados limpos e integrados para a realização das suas tarefas de forma eficaz. O *Dataddo* é uma plataforma *ETL* baseada em nuvem sem codificação, esta tem uma vasta diversidade de conectores e métricas totalmente personalizáveis. Esta plataforma facilita ainda a criação de *pipelines.* Esta tem um custo de 20$ por fonte de dados por mês no mínimo.
- ***Datawrapper***: *Datawrapper* é uma ferramenta de código aberto para visualização de dados que permite aos seus utilizadores a criação de gráficos simples, precisos e incorporáveis muito rapidamente. Esta é uma ferramenta amigável, responsiva, rápida e interativa e que não requer a utilização de código.
- ***Drill:*** *Drill* é uma ferramenta de *Big Data Analytics* de código aberto que permite que especialistas trabalhem em grande escala. Esta foi desenvolvida pela *Apache*, sendo projetada para escalar mais de dez mil servidores e processar em segundos quantidades absurdas de dados e milhões de registos.
- ***Hadoop MapReduce***: *Hadoop* é uma ferramenta de código aberto que lida com *big data. Hadoop* é escrito em Java, no entanto qualquer linguagem de programação pode ser utilizada com o *Hadoop Streaming. MapReduce* que é uma implementação do *Hadoop* e um modelo de programação. Esta tem sido uma solução adotada para *Data Mining* em *Big Data.*

- ***HCATALOG:*** *HCATALOG* é uma ferramenta de *Big Data Analytics* de código aberto que permite que especialistas trabalhem em grande escala. Esta foi desenvolvida pela *Apache*, sendo projetada para escalar mais de dez mil servidores e processar em segundos quantidades absurdas de dados e milhões de registos.
- ***HPCC***: *HPCC* é uma ferramenta de *big data* que realiza uma única plataforma, uma arquitetura singular e uma única linguagem de programação para processamento de dados.
- ***Iceberg:*** *Iceberg* é uma ferramenta para gerir dados em *data lakes*. Este foi criado pela *Netflix* para guardar *petabytes* de dados numa só tabela, mas agora é um projeto *Apache.*
- ***Kafka:*** *Kafka* é uma ferramenta para armazenar, ler e analisar dados de *streaming*. Esta ferramenta é tolerante a falhas e consta com um conjunto de cinco *APIs* principais para *Java* e linguagem de programação *Scala*.
- ***Kaggle:*** A *Kaggle* é uma plataforma que se iniciou em competições de *Machine Learning*, no entanto, agora, está também presente como uma plataforma de *Data Science* baseada em nuvem pública. Esta fornece mais de cinquenta mil conjuntos de dados públicos úteis para *Data Mining*.
- ***KNIME:*** *KNIME* é uma plataforma de análise de dados gratuita e de código aberto que oferece suporte à integração, processamento, visualização e geração de relatórios de dados. O *KNIME* é ótimo para cientistas de dados que precisam de integrar e processar dados para *machine learning* e outros modelos estatísticos, mas não necessariamente têm fortes habilidades de programação. A interface gráfica permite a análise e a modelagem de apontar e clicar.
- ***Lumify***: O *Lumify* é uma ferramenta gratuita e de código aberto de *big data*, análise e visualização de dados. Esta fornece pesquisa de texto completo, visualizações de gráficos 2D e 3D, *layouts* automáticos, análise de *links* entre entidades gráficas, integração com sistemas de mapeamento, análise geoespacial, análise multimídia e colaboração em tempo real.
- ***MongoDB:*** *MongoDB* é uma base de dados *NoSQL* de código aberto orientada a documentos que é usada para armazenar grandes volumes de dados. O *MongoDB* é uma boa ferramenta para empresas que precisam tomar decisões rápidas e querem trabalhar com dados em tempo real.

- ***OpenRefine:*** *OpenRefine* é um *software* de análise de dados de código aberto. Este *software* gratuito permitirá organizar, limpar e transformar dados, bem como a importação e exportação de diferentes formatos de arquivos. Este suporta ainda catorze idiomas.
- ***Oozie***: *Oozie* é uma ferramenta de *Big Data Analytics* de fluxos de trabalho que permite definir uma gama diversificada de trabalhos escritos ou programados em diversos idiomas.
- ***Pentaho:*** *Pentaho* é uma ferramenta de código aberto que se baseia na tomada de decisões de negócios precisas e orientadas por dados. Esta fornece análises interativas, navegação e visualização de dados e possibilita a integração de *Big Data, Data Mining* e análise preditiva de dados.
- ***Presto***: *Presto* é uma ferramenta de código aberto que lida simultaneamente com consultas rápidas e grandes volumes de dados em conjuntos de dados distribuídos. Esta foi concretizada para consultas de baixa latência e para oferecer suporte a aplicações de análise com um vasto número de *petabytes* de dados em *Data Warehouses*.
- ***Qubole***: *Qubole* é uma ferramenta de *Big Data* de fácil usabilidade que gere e otimiza dados por conta própria a partir do seu uso.
- ***R***: *R* é uma linguagem de programação de código aberto. Esta é normalmente usada para criar *software* de análise estatística ou de dados. *R* foi criado especificamente para lidar com tarefas pesadas de computação estatística e é bastante popular no que diz respeito à visualização de dados. Um aspeto negativo desta linguagem é a sua péssima gestão de memória.
- ***RapidMiner***: O *RapidMiner* fornece toda a tecnologia que os usuários precisam para integrar, limpar e transformar dados antes de executar análises preditivas e modelos estatísticos. Os usuários podem realizar quase tudo isso por meio de uma interface gráfica simples. O *RapidMiner* também pode ser estendido usando scripts *R* e *Python*, e vários *plugins* de terceiros que estão disponíveis no mercado da empresa. No entanto, o produto é altamente otimizado para sua interface gráfica para que os analistas possam preparar dados e executar modelos por conta própria.
- ***Samza***: *Samza* é uma ferramenta de processamento de fluxo distribuído e código aberto gerido pelo Apache. Esta pode lidar com enormes quantidades de dados, usando baixa latência, alto rendimento e rapidez na análise dos dados.

- ***Stats iQ:*** O *Stats iQ da Qualtrics* é uma ferramenta estatística de fácil usabilidade idealizada para analistas de *big data*. Esta permite criar histogramas, gráficos de dispersão, mapas de calor e gráficos de barras que exportam para *excel*.
- ***Talend:*** *Talend* é uma plataforma baseada em nuvem para integração de dados. Esta funciona em várias plataformas de computação em nuvem, nomeadamente, a *AWS, Google Cloud, Azure e Snowflake*. Ele suporta vários ambientes de nuvem, públicos, privados e híbridos. O *Talend* possui produtos gratuitos e comerciais, sendo que os produtos gratuitos podem ser usados em *Windows e Mac*. A análise desta plataforma é feita em tempo real e *IoT*.
- ***Tableau:*** O *Tableau* é uma plataforma de análise e visualização de dados que permite aos usuários criar relatórios e compartilhá-los em plataformas móveis e de *desktop*, num navegador ou incorporados em aplicações. Ele pode ser executado na nuvem ou no local. Grande parte da plataforma do *Tableau* é executada na sua linguagem de consulta principal, o *VizQL*. Isso traduz o painel de arrastar e soltar e os componentes de visualização em consultas de back-end eficientes e minimiza a necessidade de otimizações de desempenho do usuário final. No entanto, o *Tableau* não oferece suporte para consultas *SQL* avançadas.
- ***Xplenty:*** A *Xplenty* é uma solução de extração, transformação e carregamento (*ETL*) e integração de dados. Esta possui ferramentas de transformação na plataforma que ajudam a limpar, a normalizar e a transformar os seus através de ótimas práticas. Através da sua interface gráfica intuitiva é possível implementar *ETL (Extract, Transform, Load), ELT (Extract, Load, Transform), ETLT* (o melhor de *ETL e ELT*) ou replicação. Este faz a transferência de dados entre bases de dados, *Data Warehouse e Data Lakes*. O *Xplenty* contem um conector de *API Rest*, de modo a conseguir extrair dados que os usuários necessitem de qualquer *API Rest*.
- ***Zoho Analytics:*** *O Zoho Analytics* é uma plataforma de fácil acesso que permite analisar dados de uma ampla variedade de fontes. Esta destaca-se pela oferta de uma vasta diversidade de ferramentas de visualização de dados, na forma de tabelas dinâmicas, gráficos, painéis temáticos, entre outros. Esta plataforma fornece ainda um assistente com inteligência artificial que permite que os usuários obtenham respostas inteligentes no formato de relatórios relevantes. O *Zoho Analytics* consta com cinco planos, um primeiro

gratuito, os restantes têm um preço mensal, sendo que o *Basic* que tem um custo de 22$, o *Standard* 45$, o *Premium* 112$, e o *Enterprise* 445$.

Descrição de ferramentas *Machine Learning*

- ***Accord.NET***: *Accord.net* é uma ferramenta que disponibiliza bibliotecas de *Machine Learning* para processamento de imagem e áudio. Esta fornece algoritmos para álgebra linear numérica, otimização numérica, estatística e redes neurais artificiais.
- ***Amazon Machine Learning***: *Amazon Machine Learning* é uma ferramenta com o propósito de produzir modelos de *Machine Learning* e gerar previsões.
- ***Apache Mahout***: O *Apache Mahout* é uma plataforma gratuita de código aberto com *Machine Learning*. O seu principal propósito é ajudar os cientistas ou pesquisadores a implementar os seus próprios algoritmos. Esta plataforma foca-se em três áreas principais, nomeadamente, em mecanismos de recomendação, *clustering* e classificação. O *Apache Mahout* é ótimo para projetos de *Data Mining* complexos e que envolvam grandes quantidades de dados.
- ***Apache Spark***: O *Apache Spark* é uma estrutura de software que permite que analistas e cientistas de dados processem rapidamente grandes conjuntos de dados. Este foi projetado para analisar *big data* não estruturado, o *Spark* distribui tarefas de análise computacionalmente pesadas em muitos computadores. Existem outras estruturas semelhantes do *Apache*, no entanto o *Spark* distingue-se pela sua rapidez, devido ao uso de *RAM* em vez de memória local. Este consta com uma biblioteca de algoritmos de *Machine learning, MLlib*, incluindo algoritmos de classificação, regressão e *clustering*. Um ponto menos benéfico é o facto de este consumir tanta memória, tornando-se computacionalmente caro. Este não dispõe de um sistema de gestão de arquivos, sendo que, é comum precisar da integração de outro *software*.
- ***Catalyst:*** *Catalyst* é uma ferramenta amigável de *Deep Learning* que desempenha tarefas de engenharia, como reutilização de código. *Deep Learning* é algo complexo e o *Catalyst* tenta contrariar isso através da execução de *Deep Learning* com linhas de código.
- ***CatBoost***: *CatBoost* é uma ferramenta de código aberto de fácil usabilidade. Esta reduz os esforços de pré-processamento, uma vez que consegue manipular dados categóricos de

maneira direta e ideal, através de codificações numéricas e experimentando várias combinações em segundo plano.

- ***Colab***: O *Google Colab* é uma ferramenta gratuita de nuvem compatível com *Python*. Esta ajuda a construir aplicações de *Machine Learning* através do uso das bibliotecas de *PyTorch, Keras, TensorFlow e OpenCV*.
- ***CUV:*** *CUV* é uma biblioteca baseada na linguagem *C++*, que consiste em operações matriciais básicas e convoluções.
- ***Dask***: *Dask* é uma biblioteca em *Python* usada para computação paralela. Esta possui agendamento dinâmico de tarefas e coleções de *big data. O Dask* permite atividades de análise de dados multidimensionais. É rápido, flexível e confiável quando ocorrem processos simultâneos.
- ***DLIB***: *DLIB* é uma ferramenta *Machine Learning* baseada na linguagem *C++.* Esta ajuda a codificar em robótica, sistemas embarcados, telefones, entre outras, e é usada em vários algoritmos de *Machine Learning*, processamento de imagens e design de rede.
- ***Fast.ai***: *Fast.ai* é uma ferramenta que visa tornar o *Deep Learning* acessível em vários idiomas e sistemas operacionais. Esta oferece uma interface de alto nível e fácil de usar para modelos *Deep Learning*.
- ***Gensim***: *Gensim* é uma ferramenta de *Machine Learning* com uma biblioteca baseada em *Python*. Esta trabalha com a recuperação e extração de informações. E pode executar tarefas matemáticas e científicas também. *Gensim* pode processar grandes quantidades de dados é usualmente usada na modelagem de espaços vetoriais.
- ***Glueviz***: *Glueviz* é uma ferramenta que fornece uma interface com recursos gráficos. Além disto, o utilizador pode criar imagens 2D e 3D dos dados escolhidos. Esta ferramenta também é usada para visualização de dados através de gráficos e histogramas. *Glueviz* é uma biblioteca baseada em *Python*.
- ***Google Cloud AutoML***: O *Cloud AutoML* é uma ferramenta da Google que fornece modelos pré-treinados para criar vários serviços, como reconhecimento de texto, fala, entre outros. Esta ferramenta permite o treino de modelos *Machine Learning* personalizados e de alta qualidade com o mínimo de esforço e conhecimento.
- ***H2O:*** *H2O* é uma plataforma de *Machine Learning* de código aberto que disponibiliza tecnologias de inteligência artificial. Esta oferece suporte a algoritmos de *Machine*

Learning de forma mais rápida e fácil, até para os que não se consideram especialistas. O H2O pode ser integrado por meio de uma *API* e emprega computação distribuída na memória, o que o torna ideal para analisar grandes conjuntos de dados.

- ***IBM Watson***: *Watson Machine Learning* é uma ferramenta de nuvem da *IBM* usada para *Machine Learning e Deep Learning*. Esta ferramenta permite os utilizadores realizarem treinos e pontuações, sendo estas duas operações fundamentais de *Machine Learning*.
- ***Keras.io***: *Keras* é uma *API* gratuita para redes neurais, que ajuda a fazer pesquisas rápidas e é escrita em *Python*. Esta suporta a combinação de duas redes, mas consta com a desvantagem de ser dependente do uso simultâneo do *TensorFlow, Theano ou CNTK*.
- ***KNIME***: *KNIME* é uma plataforma de análise de dados gratuita e de código aberto que oferece suporte à integração, processamento, visualização e geração de relatórios de dados. O *KNIME* é ótimo para cientistas de dados que precisam de integrar e processar dados para *machine learning* e outros modelos estatísticos, mas não necessariamente têm fortes habilidades de programação. A interface gráfica permite a análise e a modelagem de apontar e clicar.
- ***Jupyter Notebook***: O *Jupyter Notebook* é uma ferramenta *Web* gratuita e de código aberto. Esta permite que os desenvolvedores criem relatórios com dados e visualizações de código ao vivo. O sistema suporta mais de quarenta linguagens de programação. O *Jupyter Notebook* é usado principalmente para partilhar código, criar tutoriais e apresentar trabalhos.
- ***LightGBM***: *LightGBM* é um algoritmo que usa uma técnica de aprendizagem em árvore, isto é, as árvores vão crescendo umas após as outras, sendo que as árvores seguintes memorizam a aprendizagem das anteriores. Também o *XGBoost* tem estas competências, no entanto o *LightGBM* é mais adequado para grandes conjuntos de dados.
- ***Mallet:*** *Mallet* é uma ferramenta *Machine Learning* baseada em *Java* usada para extrair informações. Esta ferramenta fornece meios para classificar documentos e extrair entidades de textos, sendo possível extrair conjuntos de dados específicos consoante a escolha dos utilizadores.
- ***Microsoft Azure Machine Learning***: *O Azure Machine Learning* é uma ferramenta de nuvem que permite que os desenvolvedores criem e implementem modelos de inteligência artificial.

- ***Numpy***: *Numpy* é uma ferramenta de *Machine Learning* usada em cálculos científicos. Esta é ótima para cálculos baseados em matemática, pois possui uma matriz com muitas dimensões e uma vasta quantidade de ferramentas para cálculo de dados. *Numpy* é uma ferramenta baseada em *Python*.
- ***OpenNN***: *Open NN* é uma ferramenta que implementa redes neurais. Usualmente esta é escrita em *C++* e foca-se em pesquisa de *Deep Learning. Open NN* tem ainda funções de *Data Mining* e faz análises preditivas.
- ***Pandas:*** *Pandas* é uma ferramenta básica e fácil de *Machine Learning,* sendo que usualmente é usada para *Python*. Esta lida com manipulação de dados e fornece estruturas de dados rápidas e eficientes. Essas estruturas de dados facilitam muito o trabalho com dados estruturados e de séries temporais.
- ***PyTorch:*** *PyTorch* é uma biblioteca gratuita de *Machine Learning Python* baseada em *Torch*. Esta fornece uma diversidade de algoritmos de otimização para a construção de redes neurais, treinamento distribuído e várias ferramentas e bibliotecas. Consta também com *um front-end* híbrido, facilitando o seu uso.
- ***R:*** *R* é uma linguagem de programação de código aberto. Esta é normalmente usada para criar software de análise estatística ou de dados. *R* foi criado especificamente para lidar com tarefas pesadas de computação estatística e é bastante popular no que diz respeito à visualização de dados. Um aspeto negativo desta linguagem é a sua péssima gestão de memória.
- ***Ramp***: *Ramp* é uma ferramenta de *Machine Learning* usada para prototipagem rápida. Esta é um módulo *Python. Ramp* é uma ferramenta facilmente extensível, isto é, através desta é possível o uso de pacotes de outras ferramentas. Além disto, o *Ramp* pode armazenar e recuperar dados do disco.
- ***RapidMiner***: O *RapidMiner* fornece toda a tecnologia que os usuários precisam para integrar, limpar e transformar dados antes de executar análises preditivas e modelos estatísticos. Os usuários podem realizar quase tudo isso por meio de uma interface gráfica simples. O *RapidMiner* também pode ser estendido usando scripts *R e Python*, e vários plugins de terceiros que estão disponíveis no mercado da empresa. No entanto, o produto é altamente otimizado para sua interface gráfica para que os analistas possam preparar dados e executar modelos por conta própria.

- ***Scikit-learn***: O *Scikit-learn* é uma ferramenta de software gratuita para *Machine Learning em Python*. Esta disponibiliza ótimos recursos de *Data Mining e Data Analytics*. Esta ferramenta oferece um vasto número de recursos como classificação, regressão, agrupamento, pré-processamento, seleção de modelo e redução de dimensão.
- ***SciPy***: *Scipy* é uma biblioteca baseada em *Python* que faz operações matemáticas científicas e de ordem superior, como cálculos de álgebra linear, integração, entre outras. A *SciPy* executa estas operações em alta velocidade e com muita facilidade.
- ***Shogun:*** *Shogun* é uma ferramenta que disponibiliza uma vasta gama de algoritmos e estruturas de dados para *Machine Learning* com foco em pesquisa e educação. Esta ferramenta de fácil usabilidade gratuita fornece ainda suporte para regressão e classificação.
- ***TensorFlow:*** *TensorFlow* é uma biblioteca *JavaScript* que ajuda no *Machine Learning*. Através de *APIs* fornecidas pelo *TensorFlow* é possível a construção de modelos, estes podem ser executados pelo *TensorFlow.js*, sendo que este é um conversor de modelos. Esta biblioteca é de difícil aprendizagem.
- ***XGBoost:*** O *XGBoost* é um algoritmo que usa uma técnica de aprendizagem em árvore, isto é, as árvores vão crescendo umas após as outras, sendo que as árvores seguintes memorizam a aprendizagem das anteriores. O *XGBoost* é adequado para grandes conjuntos de dados e combinações de recursos numéricos e categóricos.
- ***Weka:*** A *Weka* é uma plataforma que fornece algoritmos de *Machine Learning para Data Mining*. Esta plataforma gratuita fornece ferramentas para a preparação de dados, classificação, regressão, *clustering*, mineração de regras de associação e visualização. E pode ser utilizada em *Microsoft Windows, Mac e Linux*.
- ***Yellowfin BI***: O *Yellowfin BI* é uma ferramenta de *Business Intelligence* que concilia visualização, *Machine Learning* e colaboração. É ainda possível filtrar facilmente uma vasta quantidade de dados através da filtragem intuitiva, bem como abrir painéis em praticamente qualquer lugar. Uma vantagem desta ferramenta é a facilidade de desenvolvimento sem necessidade de código ou usando muito pouco.

Descrição das ferramentas *Data Visualization*

- ***ChartBlocks:*** *ChartBocks* é uma ferramenta de visualização de dados de fácil usabilidade que permite a criação de tabelas e gráficos totalmente responsivos para poderem ser partilhados em qualquer plataforma e visualizados em qualquer dispositivo.
- ***DataBox***: O *DataBox* é uma ferramenta de visualização de dados que oferece um número bem generoso de modelos. Esta ferramenta possui tutoriais para integrar muitas fontes de dados e distingue-se pelas cores ricas e profundas que disponibiliza.
- ***Datapine:*** *Datapine* é uma ferramenta de *Business Intelligence* que permite que analistas de dados integrem facilmente diversas fontes de dados, realizando análises avançadas de dados e criando painéis de negócios interativos que giram insights de negócios acionáveis.
- ***Datawrapper:*** *Datawrapper* é uma ferramenta de código aberto para visualização de dados que permite aos seus utilizadores a criação de gráficos simples, precisos e incorporáveis muito rapidamente. Esta é uma ferramenta amigável, responsiva, rápida e interativa e que não requer a utilização de código.
- ***Domo***: O *Domo* fornece mais de mil integrações integradas que permitem que os usuários transfiram dados de e para sistemas locais e externos na nuvem. O *Domo* também oferece suporte à criação de aplicativos personalizados que se integram à plataforma, o que permite aos desenvolvedores estender o sistema com acesso imediato aos conectores e ferramentas de visualização. O *Domo* vem como uma plataforma única que inclui um *data warehouse* e *software ETL*, portanto, as empresas que já têm seu próprio data warehouse e pipeline de dados configurados podem querer procurar noutro lugar.
- ***Dundas***: *Dundas* é uma ferramenta de *Data Mining* elaborada para empresas que pretendem construir e visualizar painéis interativos e personalizáveis, relatórios, entre outros. Esta ferramenta faz uma análise preditiva e avançada.
- ***Fusioncharts*****:** O *Fusioncharts* é uma ferramenta de visualização de dados baseada em *Javascript* que oferece noventa pacotes diferentes de construção de gráficos que se integram com os principais *frameworks* e plataformas, oferecendo aos utilizadores uma flexibilidade significativa.
- ***GoodData***: *GoodData* é uma plataforma de *Business Intelligence* que disponibiliza ferramentas de armazenamento, consultas analíticas, visualizações e integração de aplicações, onde se podem criar painéis e relatórios.

- ***Google Charts:*** *Google Charts* é uma das mais populares ferramentas de visualização de dados, esta é codificada com *SVG e HTML5* e é conhecida pela sua capacidade de produzir visualizações de dados gráficos. O *Google Charts* oferece funcionalidade de *zoom* e disponibiliza aos seus utilizadores compatibilidade entre plataformas com *iOS, Android* e até mesmo com versões anteriores do navegador *Internet Explorer.*
- ***Google Data Studio***: O *Google Data Studio* é uma ferramenta gratuita de painel e visualização de dados que se integra automaticamente à maioria das outras aplicações do *Google*, como *Google Analytics, Google Ads e Google BigQuery*. Graças à sua integração com outros serviços da *Google, o Data Studio* é ótimo para quem precisa analisar os seus dados do *Google*. Por exemplo, os profissionais de marketing podem criar painéis para os seus dados do Google *Ads e Analytics* para entenderem melhor a conversão e a retenção de clientes. O *Data* Studio também pode trabalhar com dados de várias outras fontes, desde que os dados sejam replicados primeiro para o *BigQuery* através de um pipeline de dados como *o Stitch*.
- ***Highcharts:*** *Highchats* é uma ferramenta de visualização de análises de *big data de streaming*. Esta é executada na *API Javascript* e oferece integração com *jQuery,. Highcharts* oferece suporte para funcionalidades entre navegadores que facilitam o acesso a visualizações interativas.
- ***IBM Cognos***: O *IBM Cognos* é uma plataforma de inteligência de negócios que apresenta ferramentas de Inteligência Artifical integradas para revelar *insights* ocultos em dados e explicá-los em linguagem simples. O *Cognos* também possui ferramentas automatizadas de preparação de dados para limpar e agregar fontes de dados automaticamente, o que permite integrar e experimentar rapidamente as fontes de dados para análise.
- ***IBM Watson***: *Watson Machine Learning* é uma ferramenta de nuvem da *IBM* usada para *Machine Learning e Deep Learning*. Esta ferramenta permite os utilizadores realizarem treinos e pontuações, sendo estas duas operações fundamentais de *Machine Learning*.
- ***Infogram***: O *Infogram* é uma ferramenta de visualização de dados online que oferece uma interface de fácil usabilidade com mãos de trintas tipos de gráficos e mapas.
- ***Jupyter Notebook***: O *Jupyter Notebook* é uma ferramenta *Web* gratuita e de código aberto. Esta permite que os desenvolvedores criem relatórios com dados e visualizações de código ao vivo. O sistema suporta mais de quarenta linguagens de programação. O

Jupyter Notebook é usado principalmente para partilhar código, criar tutoriais e apresentar trabalhos.

- ***Klipfolio***: A *Klipfolio* é uma empresa *Business Inteligence* que fornece uma ferramenta de visualização de dados. Nesta é possível aceder aos dados de centenas de fontes de dados diferentes, como bases de dados, arquivos, entre outros. O *Klipfolio* permite criar visualizações de dados personalizadas, nas quais é possível escolher entre diferentes opções, como tabelas e gráficos.
- ***Leaflet***: *Leaflet* é uma ferramenta complexa que disponibiliza uma biblioteca *JavaScript* para incorporar na estrutura de visualização de dados. Esta é reconhecida por ser uma ferramenta leve e é excelente para criar não apenas mapas, mas também visuais de mapeamento interativos feitos especificamente para dispositivos móveis.
- ***Looker:*** *Looker* é uma plataforma de *business intelligence* e análise de dados baseada em nuvem. Esta apresenta geração automática de modelo de dados que verifica esquemas de dados e infere relacionamentos entre tabelas e fontes de dados. Os engenheiros de dados podem modificar os modelos gerados por meio de um editor de código integrado.
- ***Microsoft Excel***: *Microdoft Excel* é dos softwares mais conhecidos do mundo. Além de que, contem cálculos e funções gráficas ideais para a análise de dados. Os seus recursos integram tabelas dinâmicas e ferramentas de criação de formulários. Apesar destas vantagens, este possui também algumas limitações, nomeadamente a lentidão com grandes conjuntos de dados e a tendência a fazer aproximações com grandes valores, levando a imprecisões.
- ***MicroStrategy***: A *MicroStrategy* é uma ferramenta de *business intelligence* para empresas que disponibiliza painéis poderosos, análise de dados e soluções em nuvem. Esta permite a identificação de tendências, o reconhecimento de oportunidades, a melhoraria da produtividade, entre outras. Para iniciantes é algo complexa.
- ***Microsoft Power BI***: O *Microsoft Power BI* é uma plataforma de *business intelligence* de topo com suporte para dezenas de fontes de dados. Esta plataforma permite que os usuários criem e compartilhem relatórios, visualizações e painéis. O *Microsoft Power BI* consegue ser usado em diferentes.
- ***MicroStrategy***: A *MicroStrategy* é uma ferramenta de *business intelligence* para empresas que disponibiliza painéis poderosos, análise de dados e soluções em nuvem. Esta permite

a identificação de tendências, o reconhecimento de oportunidades, a melhoraria da produtividade, entre outras. Para iniciantes é algo complexa.

- ***Qlik***: *Qlik* é uma ferramenta de *Data Mining* e visualização de dados, que oferece painéis e suporte a diferentes fontes de dados e tipos de arquivo. Esta disponibiliza também interfaces de arrastar e soltar para criar visualizações de dados flexíveis e interativas.
- ***Microsoft Power BI***: O *Microsoft Power BI* é uma plataforma de *business intelligence* de topo com suporte para dezenas de fontes de dados. Esta plataforma permite que os usuários criem e compartilhem relatórios, visualizações e painéis. O *Microsoft Power BI* consegue ser usado em diferentes
- ***OpenRefine***: *OpenRefine* é um *software* de análise de dados de código aberto. Este software gratuito permitirá organizar, limpar e transformar dados, bem como a importação e exportação de diferentes formatos de arquivos. Este suporta ainda catorze idiomas.
- ***Ploty:*** *Ploty* é uma ferramenta de visualização de dados de código aberto, este oferece integração total com linguagens de programação centradas em análises, como *Matlab, Python e R*, que permitem visualizações complexas. Esta é usualmente utilizada para trabalho colaborativo, disseminando, modificando, criando e compartilhando dados gráficos interativos.
- ***RAW Graphs***: *RawGraphs* é uma ferramenta que trabalha com dados delimitados, como arquivos *TSV* ou arquivos *CSV*. *RAWGraphs* apresenta uma variedade de *layouts* não convencionais e convencionais, oferecendo uma segurança de dados robusta, mesmo esta sendo uma aplicação *web*.
- ***SAP Analytics cloud***: O *SAP Analytics Cloud* usa recursos de *business intelligence* e análise de dados que permitem avaliar os dados de modo a criar visualizações para prever resultados de negócios. Este fornece ferramentas de modelagem que alertam sobre possíveis erros nos dados e categorizam diferentes medidas e dimensões de dados.
- ***Sisense***: *Sisense* é uma plataforma de análise de dados destinada a ajudar os desenvolvedores técnicos e analistas de negócios a processar e visualizar todos os seus dados de negócios. Um aspeto exclusivo da plataforma *Sisense* é sua tecnologia *In-Chip* customizada, que otimiza a computação para utilizar o cache da *CPU* em vez de *RAM* mais

lenta. Para alguns fluxos de trabalho, isso pode levar a um cálculo de dez a cem vezes mais rápido.

- ***Tableau***: O *Tableau* é uma plataforma de análise e visualização de dados que permite aos usuários criar relatórios e compartilhá-los em plataformas móveis e de *desktop*, num navegador ou incorporados em aplicações. Ele pode ser executado na nuvem ou no local. Grande parte da plataforma do *Tableau* é executada na sua linguagem de consulta principal, o *VizQL.* Isso traduz o painel de arrastar e soltar e os componentes de visualização em consultas de back-end eficientes e minimiza a necessidade de otimizações de desempenho do usuário final. No entanto, o *Tableau* não oferece suporte para consultas *SQL* avançadas.
- ***Visme***: *Visme* é uma ferramenta de visualização de dados que permite a visualização de dados de diferentes tipos. Esta permite a criação de gráficos, e fornece uma biblioteca de w*idg*ets de dados. Esta ferramenta integra *Business Intelligence* com design interativo, o que ajuda o utilizador a criar visualizações de dados para conteúdos que não são tão fáceis de ler e entender.
- ***Visual.ly***: *Visual.ly* é uma das ferramentas de visualização de dados, esta é reconhecida pela sua rede de distribuição que ilustra os resultados do projeto.
- ***Whatagraph:*** O *Whatagraph* é uma plataforma que oferece a análise visual de dados com entrada de fontes de dados automáticas e funcionalidades de arrastar e soltar para criar relatórios. A maior vantagem do *Whatagraph* é a sua fácil configuração e interface visual intuitiva. Esta plataforma consta com opções de visualização de dados através de formas gráficas, tabelas, *widgets* de rastreamento de *KPI*, entre outras.
- ***Zoho Analytics***: O *Zoho Analytics* é uma plataforma de fácil acesso que permite analisar dados de uma ampla variedade de fontes. Esta destaca-se pela oferta de uma vasta diversidade de ferramentas de visualização de dados, na forma de tabelas dinâmicas, gráficos, painéis temáticos, entre outros. Esta plataforma fornece ainda um assistente com inteligência artificial que permite que os usuários obtenham respostas inteligentes no formato de relatórios relevantes.O *Zoho Analytics* consta com cinco planos, um primeiro gratuito, os restantes têm um preço mensal, sendo que o *Basic* que tem um custo de 22$, o *Standard* 45$, o Premium 112$, e o *Enterprise* 445$.