CENTRO ESTADUAL DE EDUCAÇÃO TECNOLÓGICA PAULA SOUZA
Faculdade de Tecnologia de Jundiaí – "Deputado Ary Fossen"
Curso Superior de Tecnologia em Gestão da Tecnologia da Informação

Daniel Gervilla Rosa

O uso da ciência e da análise de dados nos processos de tomada de decisão

Jundiaí
2022

Daniel Gervilla Rosa

O uso da ciência e analise de dados nos processos de tomada de decisão

Trabalho de Graduação apresentado à Faculdade de Tecnologia de Jundiaí - “Deputado Ary Fossen” como requisito parcial para a obtenção do título de Tecnólogo em Gestão da Tecnologia da Informação, sob a orientação do Professor Me. Aimar Martins Lopes

Jundiaí
2022

(SUBSTITUIDA ESTA PÁGINA PELA FOLHA DE APROVAÇÃO DIGITALIZADA)

Dedico este trabalho
aos professores e alunos
da Fatec – Jundiaí que fizeram
parte dessa jornada de aprendizado.

AGRADECIMENTOS

Agradeço ao apoio dos meus familiares durante esse período.

Ao Prof. Me. Aimar Martins Lopes por ter sido meu orientador nesse trabalho.

A todos outros professores e professoras que fizeram parte da formação

A FATEC Jundiaí e sua equipe por oferecer a oportunidade de realizar esse curso.

ROSA, Daniel Gervilla. **O uso da ciência e da análise de dados nos processos de tomada de decisão**. 41 páginas Trabalho de Conclusão de Curso de Tecnólogo em Gestão da Tecnologia da Informação. Faculdade de Tecnologia de Jundiaí - “Deputado Ary Fossen”. Centro Estadual de Educação Tecnológica Paula Souza. Jundiaí. 2022

RESUMO

Esse trabalho tem o objetivo de discorrer sobre assuntos relacionados a big data, ciência de dados, como nos últimos anos a produção de dados aumentou exponencialmente devido aos avanços tecnológicos e suas aplicações no auxílio a tomada de decisão. Descrevendo conceitos fundamentais envolvendo a armazenagem, como tipos de bancos de dados tanto os relacionais quanto os não relacionais, suas propriedades e mineração de dados. Além de tratar acerca da análise de dados, conceitos envolvidos e suas aplicações, cujo objetivo é retirar informações e conhecimentos importantes a partir de dados brutos, utilizando ferramentas para a limpeza, transformação e modelagem, havendo diferentes formas de analisar de acordo com o objetivo desejado. Também comentando a importância de saber como expor tais informações através da visualização de dados, tipos de gráficos usados em variadas situações que trazem melhor compreensão sobre o assunto tratado. Todos esses conjuntos de ferramentas e técnicas são utilizadas nas empresas para a tomada de decisões baseadas em dados, oferecendo opções de escolha e embasamento para se definir algo, trazendo benefícios para as companhias e diminuindo riscos. Essas aplicações e ferramentas são englobadas no Business Intelligence que tornou importante dentro das empresas por auxiliar nas estratégias de negócio, aumentando a competividade e ganhos.

Palavras-chave: *big data*, ciência de dados, análise, decisão, *business intelligence*.

ROSA, Daniel Gervilla. **The use of data Science and analysis in decision-making process**. 41 p. End-of-course paper in Technologist Degree in Information Technology Management. Faculdade de Tecnologia de Jundiaí - “Deputado Ary Fossen”. Centro Estadual de Educação Tecnológica Paula Souza. Jundiaí. 2022.

ABSTRACT

This papper aims to discuss issues related to big data, data science, as in recent years data production has increased exponentially due to technological advances and their applications in aiding decision making. Describing fundamental concepts involving storage, such as types of databases both relational and non-relational, their properties and data mining. In addition to dealing with data analysis, concepts involved and their applications, whose objective is to extract important information and knowledge from raw data, using tools for cleaning, transformation and modeling, with different ways of analyzing according to the objective wanted. Also commenting on the importance of knowing how to expose such information through data visualization, types of graphics used in various situations that bring a better understanding of the subject addressed. All these sets of tools and techniques are used in companies to make decisions based on data, offering choices and basis for defining something, bringing benefits to companies and reducing risks. These applications and tools are encompassed in Business Intelligence that has become important within companies for helping in business strategies, increasing competitiveness and gains.

Keywords: big data, data Science, analysis, decision, Business Intelligence.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

SQL	Structured Query Language
NoSQL	Not Only Structured Query Language
ACID	Atomicidade, Consistência, Isolamento e Durabilidade
JSON	Javascript Object Notation
XML	Extensible Markup Language
KVS	Key Value Store
KDD	Knowledge Discovery in Databases
CRISP-DM	Cross Industry Standard Process for Data Mining
ETL	Extract, Transform, Load
DDD	Data-driven decision making
BI	Business Intelligence

SUMÁRIO

1 INTRODUÇÃO

A análise de dados é um fator relevante quando se trata sobre tomada de decisões em um processo de gestão. A ciência e a análise de dados vêm sendo cada vez mais utilizadas em diversos segmentos, possuindo ferramentas para coletar as informações, além de manipular, filtrá-las e criar representações visuais para facilitar o entendimento dos dados.

Esse tema vem ganhando importância na área da tecnologia da informação, devido ao grande volume de dados sendo produzidos diariamente, gerando também a necessidade de haver maneiras de expor esses dados de maneira clara e objetiva e assim os gestores das empresas responsáveis pelas tomadas de decisão possam ter um embasamento para definir os rumos que o negócio irá tomar a partir de dados e informações precisas e fidedignas.

Tendo como justificativa para se realizar esse trabalho foi a necessidade de buscar conhecimentos mais profundos e reuni-los para se ter uma melhor compreensão sobre o tema, visando demostrar como uma boa aplicação pode ser feita. Contribuindo para agregar conhecimentos para quem também possui interesse sobre a área de análise e ciência de dados.

O volume de dados criados tem aumentado nos últimos anos, grande parte deles já são digitais. Isso acontece graças ao crescimento do uso de dispositivos eletrônicos e da internet pela população, o que tem causado uma grande mudança na forma como as informações geradas são tratadas. (GALDINO, 2016)

Tais dados são muito diversos e complexos, sendo oriundos de fontes variadas, o armazenamento e a coleta de tantas informações é chamado de *Big Data*. Esse termo abrange vários processos relacionados ao tratamento e análise de dados, sendo utilizado em diversas áreas da ciência, tanto exatas quanto humanas. Porém também vem sendo bastante utilizado nos processos de gestão empresarial para aumentar vantagens competitivas. (FURLAN, LAURINDO,2017)

A manipulação dos dados coletados no *Big Data* é chamada de análise de dados, ou *data analytics*. Essas informações podem estar no formato de banco de dados relacionais e estruturados, mas também podem ser obtidos de fontes não estruturadas, como imagens, vídeos, informações de sensores entre outras fontes que os dados não estão em registros típicos e sendo de fácil pesquisa. O *Big Data* é um

fator importante que contribui para a tomada de melhores decisões dentro das corporações. A análise de dados, ou *big data analytics,* permite que os combinar um grande volume de dados levando em consideração fatores internos e externos da empresa, gerando informações que deem um melhor direcionamento nas tomadas de decisão. Para isso são usadas várias ferramentas que coletam e reúnem dados. Além de haver vários tipos de análises que podem ser efetuadas. (JUNIOR, PINTO, 2018)

O objetivo desse trabalho é discorrer acerca dos assuntos relacionados a ciência e a análise de dados buscando elucidar um pouco sobre a história, seus conceitos, tecnologias, processos e técnicas envolvidas.

Para atingir o objetivo, é apresentado a conceituação sobre ciência e análise de dados, como a retirada de informação a partir da manipulação e análise de dados, como o processo de extração ocorre e as diferentes maneiras de se armazenar dados. Além de comentar sobre conceitos relacionados aos processos de tomada de decisão, como o *Business Intelligence*, e sua importância para auxiliar os gestores a compreenderam a situação atual e guiar os rumos a serem seguidos para se atingir os objetivos.

1.2 Metodologia

Essa dissertação foi realizada através revisão bibliográfica, usando mecanismos de buscas, tais como Google Scholar e o site Scielo para a realização das pesquisas. Fez um levantamento de informações sobre o tema em diversas fontes, organizando o conhecimento em forma de capítulos, com o objetivo de levantar as características dos tópicos selecionados, apresentando os conceitos da temática escolhida e os expondo de maneira esclarecedora e objetiva. Sendo utilizado diferentes fontes, como artigos disponíveis na internet e livros relacionados com o assunto tratado no trabalho.

A disposição dos capítulos se dará da seguinte forma.

- Primeiro capítulo será sobre a ciência de dados, o que seria, quais são seus componentes e aplicações, buscando explicar algumas de suas terminologias e como é realizada a coleta e armazenamento dos dados.

- Segundo capítulo tratará acerca da análise de dados, as metodologias usadas, os tipos de análises existentes e em quais situações pode ser utilizada.
- O terceiro capítulo será dedicado a discorrer a respeito dos processos de tomada de decisão nas corporações e como a ciência e análise de dados são importantes para auxiliar os gestores.

2 CIÊNCIA DE DADOS E BIG DATA

Desde o início do século XX, o volume de dados digitais produzidos vem crescendo de maneira exponencial. Esse fenômeno acontece devido ao aumento da disponibilidade de aparelhos eletrônicos e da utilização da internet. Esses dados são originados de diversas fontes diferentes, como o conteúdo da navegação na internet (redes sociais, páginas visitadas), transações financeiras (compras online, utilização de cartões de crédito entre outras formas de pagamento), dados de biometria (reconhecimento facial, impressão digital), dados gerados por pessoas que são privados além de informações geradas por máquinas para outras máquinas como sensores, GPS e medidores. O uso do *Big Data* se encontra na utilização desses grandes volumes de dados oriundos de diferentes fontes e necessitando de grande poder de processamento para obter um valor através das relações entre as informações. (GALDINO, 2016)

Nas décadas de 1980 e 1990, a maior parte do armazenamento de dados era feita por meios analógicos, fitas cassete e de VHS, discos de vinil e devido por serem frágeis seu compartilhamento era difícil. Porém com as novas tecnologias esse cenário mudou, os formatos digitais que representavam só 0,8% do armazenamento de dados em 1996 saltou para 94% em 2007. O aumento do uso de aparelhos móveis e da internet foram fatores que contribuíram para essa mudança. Com mais poder de processamento e armazenamento, somada com a facilidade de compartilhar dados e a redução dos custos para armazenar os dados, o volume de informações aumentou e diversos setores perceberam como a utilização dos dados poderiam trazer benefícios para seus processos. (MARQUESONE, 2018)

Alecrim (2015) define *Big Data* como conjuntos de dados tão amplos que possuem a necessidade da utilização de ferramentas capazes de lidar com grandes volumes de dados para assim as informações possam ser identificadas e analisadas. Ele também diz que a principal mudança em relação a geração dados foi a ampliação do uso da internet, vários sistemas diferentes conectados através da rede aumentaram a quantidade de informação circulando no mundo. O *Big Data* surge para auxiliar na tarefa de tratar a imensa quantidade de dados gerados.

Para Junior e Pinto (2021) *Big data* é atribuído a grandes conjuntos de dados ou que são modificados com grande rapidez que ferramentas e técnicas de análises de

bancos de dados relacionais tradicionais ou multidimensionais não seriam capazes de processar em tempo hábil. *Big data* também vai além de dados estruturais, incluindo também dados adquiridos de fontes que não geram dados estruturados, como vídeos, imagens, sensores entre outros.

O termo *Big Data* surge devido a utilização de recursos tecnológicos que gera um alto volume de dados produzidos heterogêneos, cujo métodos tradicionais de processamento não são capazes lidar. Inicialmente o *Big Data* se utilizava de três atributos básicos dos dados, com o passar do tempo apareceu outros atributos foram incluídos, formando assim o que ficou conhecido como os 6 Vs. (Rautenberg; Carmo,2019)

Os 6 Vs são:

- **Volume.** A quantidade de dados produzidos pelas mais diversas fontes.
- **Velocidade.** Com o aumento da capacidade computacional, a velocidade na geração de novos dados também cresce.
- **Variedade.** A diversidade dos tipos de dados dependendo de sua origem, seja e-mail, publicações em redes sociais, fotos, vídeos, dados gerados por dispositivos.
- **Veracidade.** Para ser um dado confiável, sua origem e armazenagem precisam ser íntegros e precisos. Quando isso não acontece, pode haver ruídos e ambiguidade gerando problemas na análise.
- **Variabilidade.** Trata-se de compreender fenômenos que podem interferir nos padrões gerais das informações, podendo causar interpretações erradas.
- **Valor.** Sendo uma das características mais importante, o valor de uma informação é notado a partir da análise de dados concisos.

Junto com o *Big Data*, surge também a Ciência de dados, como um campo interdisciplinar que tem como objetivo extrair de grandes e complexas bases de dados informações úteis a partir de dados brutos, através do uso de ferramentas metodológicas, tendo como objetivo sua utilização no processo de tomada de decisões. A Ciência de dados está inserida na inter-relação de três áreas de conhecimento. (Rautenberg e Carmo,2019)

- Programação de Computadores. Habilidades para o desenvolvimento de ferramentas para a extração, manipulação e armazenagem de dados. Além de criar algoritmos de aprendizado de máquina e de visualização de dados.
- Estatística e Matemática. Ter conhecimentos estatísticos e matemáticos é fundamental para a análise e interpretação dos resultados. Além de também ser necessário para a criação de representações gráficas.
- Domínio do Conhecimento. Na resolução de um problema é necessário ter domínio acerca do assunto tratado, para assim hipóteses possam ser criadas e seja possível saber onde se buscar as informações no meio dos dados.

O termo Ciência de dados surgiu na década de 1960, porém é uma ciência nova que busca coletar, organizar, padronizar informação e conhecimento através de metodologias sistemáticas além de estudar todo o ciclo de vida dos dados, desde sua produção até seu descarte. A Ciência de dados é erroneamente associada apenas com os processos de análise de dados, sendo considerada uma parte da estatística, contudo ela é mais abrangente, sendo a conciliação de outras ciências, modelagens, tecnologias e processos referentes aos dados. (AMARAL,2016)

Provost e Fawcett (2013) definem a Ciência de Dados é um conjunto de princípios básicos de boas práticas usados para auxiliar e orientar na tarefa de extrair informações e conhecimento a partir de dados. Incorporando esses princípios nas tecnologias de mineração de dados que seria a parte do processo na qual a extrair dos dados é efetuada empregando algoritmos. A Ciência de Dados pode ser usada em diversos campos de negócios, como no marketing, relação com o consumidor e na análise financeira. Porém cientistas de dados não se limitam a somente utilizar algoritmos de mineração de dados, mas também de enxergar e tratar os problemas de um negócio através dos dados. Compreender os princípios, métodos e metodologias das áreas de análise, estatística e visualização de dados são essenciais para a Ciência de Dados.

Há alguns conceitos fundamentais nessa área originados de vários campos do conhecimento que envolvem o estudo da análise de dados, tanto nas partes teóricas quanto nas partes empíricas do assunto. Alguns desses conceitos se formam a partir da relação entre a Ciência de Dados e os problemas de negócio encontrados, outros

se originam dos conhecimentos adquiridos e das soluções técnicas criadas. (Provost e Fawcett, 2013).

Alguns desses conceitos fundamentais são:

- Seguir um processo com estágios predefinidos para resolver problemas de negócios a partir da extração de conhecimento dos dados.
- Utilizar soluções para dividir um problema em partes menores para melhor analisar a relação entre eles e a análise, recombinar seus componentes pode ser útil para avaliar outras possibilidades e valores dos dados.
- Usar a tecnologia para encontrar correlações entre as informações e dados encontrados para assim diminuir as incertezas existentes na análise.
- Encontrar semelhanças entre aquilo que se conhece e aquilo que é desconhecido.
- Ao se formar conclusões é preciso tomar cuidado com a presença de fatores que possam criar equívocos ao se analisar os dados.

Como a quantidade de dados dos mais diversos tipos é gigantesca, as aplicações de Big Data necessitam trabalhar com elasticidade, saber atuar com volumes de informações crescendo constantemente e saber distribuir o processamento usado nos recursos. Os bancos de dados tradicionais, como o MySQL, PostgreSQL e Oracle, por serem do modelo relacional demostram não ser os mais adequados para atender esses requisitos por serem mais inflexíveis. Isso acontece devido a quatro propriedades conhecidas como ACID: (ALECRIM, 2015)

- Atomicidade: transições atômicas, a tarefa só é marcada como realizada quando é completamente executada.
- Consistência: é necessário seguir todas as regras adotadas pelo banco de dados.
- Isolamento: as transações não podem interferir em outras transações que estiverem sendo executadas.
- Durabilidade: após terminada a transação, seus dados não podem ser perdidos.

Alecrim (2015) diz que essas propriedades podem restringir as soluções de Big Data, por isso surge a noção de NoSQL, que viria da expressão em inglês “*Not only*

SQL". Structured Query Language, SQL, é uma linguagem usada em bancos de dados relacionais. Enquanto o NoSQL trabalha com outros tipos de armazenamento que não se limitam ao modelo relacional, sendo mais flexível.

Marquesone (2018) define que bancos de dados relacionais garantem a integridade das informações por causa das suas propriedades, chamadas de ACID (Atomicidade, Consistência, Isolamento e Durabilidade) e se mostram eficientes em várias situações diferentes, contudo como são utilizados para armazenar dados estruturais que possuem estruturas rígidas dificulta sua utilização na *Big data* já que também são usados dados semi-estruturados como arquivos nos formatos *JSON* (*Javascript Object Notation*) e *XML* (*eXtensible Markup Language*). Além de também utilizar dados não estruturados, como vídeos, imagens e alguns tipos de textos, o que torna o processamento por ferramentas tradicionais uma tarefa complexa de ser realizada.

2.1 Banco de dados relacional

Nas últimas décadas o modelo de banco de dados mais usado é o relacional, segundo Silva e Ferreira (2017), sendo um paradigma que garante fidedignidade e organização dos dados, que se utilizado de maneira correta impede a duplicidade de informações nos registros de dados. Sua estrutura fundamental é a tabela formada pelos campos com atributos que possuem os tipos de dados armazenados e as linhas como instâncias do esquema que guardam os registros.

Foi criado nos anos 70 por Edgar Frank Codd, com o objetivo de substituir os modelos hierárquico e de redes utilizados até então. Se apresentando eficiente ao ser utilizado nos negócios por manter a integridade dos dados no decorrer das operações. Possuindo estrutura fixa e rígida em forma de tabelas por onde os dados são divididos, cada tabela possui um registro único, a chave primária, que é usado para interligar tabelas diferentes e evitar redundâncias. (AMARAL,2016)

Os Sistemas de Gerenciamento de Bancos de Dados Relacionais (SGBDR) utilizam a *Structured Query Language* (SQL), uma linguagem de programação voltada para criar e manipular dados. Garantem a integridade dos dados por seguir as propriedades ACID e permitem consultas com grande grau de complexidade. Porém com o advento da internet e aumento da produção de grandes volumes de dados

fizeram com que limitações surgissem nesse modelo devido aos fatores de escalabilidade, disponibilidade e flexibilidade. (MARQUESONE, 2018)

2.2 Banco de dados não relacional

Devido à grande quantidade de informações geradas e diferentes tipos de dados reunidos, novas necessidades surgiram e o modelo relacional não era capaz de suprir tais necessidades. Surge então novos modelos que usam dados não estruturados e normalizados, são tantas restrições de integridade, sendo mais flexíveis e com menor custo. É usado o termo NoSQL para definir banco de dados que não se apoiam no modelo relacional e ao invés de usar tabelas utilizam o conceito de chave-valor, também chamada de KVS, Key-Value-Store, que inserem somente um valor e uma chave ao arquivo e não uma série de atributos. (AMARAL,2016)

Os bancos de dados não relacionais contam com características e necessidades que os tornam diferentes dos bancos de dados relacionais com a capacidade de operar com volumes muito grandes de dados não estruturados ou semiestruturados. São elas (OLIVEIRA,2014):

- Escalabilidade: possuir a capacidade de operar uma crescente quantidade de dados dentro de um sistema de armazenamento de maneira uniforme.
- Alta disponibilidade: ser resistente a falhas, sustentando o funcionamento dos serviços, precisando estar pronto para responder as requisições rapidamente o que necessita um grande poder de processamento e de memória.
- Esquema flexível: ao contrário dos bancos relacionais que possuem estruturas de dados diagramadas divididas por tabelas, bancos de dados não relacionais não possuem um esquema forte, facilitando com que os dados sejam distribuídos em diversos servidores.
- Simples manipulação: esse tipo de banco costuma ser simples de manipular e configurar, facilitando seu desenvolvimento e não necessitando várias especialistas para gerenciar o sistema de banco de dados.

De acordo com Marquesone (2018) é possível dividir os bancos de dados NoSQL em quatro modelos principais levando em consideração a estrutura usada para armazenar os dados e apresentam alguns aspectos em comum além de não ser relacional, não possuem um esquema rígido, criados para serem utilizadas em estruturas de cluster e seguirem a tendência de usaram softwares livres. Esses modelos são:

- Banco de dados orientado a chave valor: tendo uma estrutura simples que possui chaves para identificar as informações e o campo valor que contém os dados armazenados de diferentes tipos não sendo necessário um esquema predefinido como o modelo relacional, possibilitando acesso rápido às informações. Porém só é possível acessar os dados utilizando a chave tornando difícil realizar consultas mais complexas, contudo é útil em diversas situações como armazenar imagens, documentos, dados de sessões de usuários.
- Banco relacional orientado a documentos: no quesito de gerenciamento de dados, esse modelo se mostra mais simples e flexível do que o orientado a chave valor, permite criar índices dentro das informações facilitando as consultas. Esses documentos são dados semiestruturados, como arquivo nos formatos XML e JSON, e eles não possuem esquemas pré-definidos para acrescentar registros, possui alta disponibilidade. Sendo um modelo ideal para armazenar dados de páginas da internet, catalogação de documentos gerenciamento de inventários entre outros.
- Banco de dados orientados a colunas: um dos modelos mais complexos, tem similaridades com o modelo relacional, mas com diferenças substanciais, como possuir mais flexibilidade e escalabilidade. Ao invés de definir colunas, a responsabilidade de modelar os dados é do chamado “famílias de colunas” que são uma maneira de organizar as informações em grupos de dados utilizados, possibilitando quantidades diferentes de colunas para cada registro, oferecendo flexibilidade e escalabilidade para o banco de dados.
- Banco de dados orientado a grafos: o mais especializado dos tipos de armazenamento NoSQL, esse tipo tem como foco os relacionamentos existentes entre os dados utilizando a estrutura da teoria dos grafos.

2.3 Data mining

Com o advento dos sistemas de computadores, as organizações tem tido com umas de suas principais prioridades o armazenamento de dados e com o avanço e barateamento das tecnologias, há uma tendência de aumentar cada vez mais a quantidade de dados coletados. Como isso as técnicas tradicionais que eram usadas para se explorar os dados se tornaram ineficientes, para solucionar esse problema no final da década de 80 surgiu Mineração de Dados ou *Data Mining*. Se tornando umas das tecnologias mais relevantes atualmente por auxiliar descoberta de informações valiosas a partir de um grande volume de dados. A Mineração de Dados tem sido aplicada em diversos campos de negócio, como bancos, operadores de cartões de crédito, auxiliar a tomada de decisões, marketing entre outros. (CAMILO e SILVA, 2009)

Na busca por solucionar o problema de sobrecarga de dados existente devido ao grande volume de dados produzidos que inviabilizam o processamento manual dos mesmos surgiu o *KDD* (*Knowledge Discovery in Databases* ou Descoberta de Conhecimento nas Bases de dados) que é atribuído ao processo que leva a descoberta de um conhecimento extraídos de dados através da definição de padrões que sejam validos possam ser utilizados e compreensíveis, sendo o *Data Mining* ou Mineração de dados um de seus processos, como ilustrado na figura 1. (Camilo e Silva, 2009)

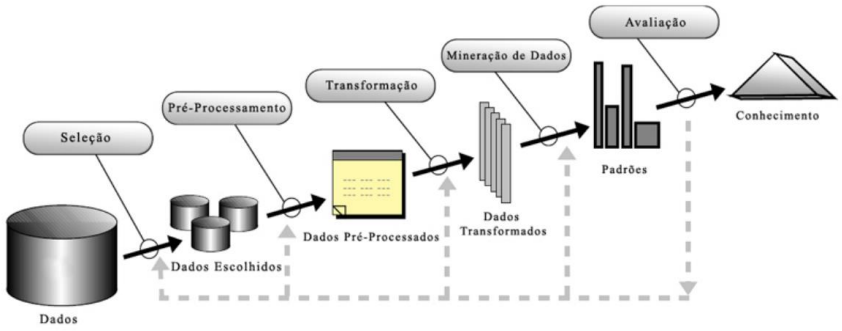


Figura 1– Representação do processo de KDD (Camilo e Silva, 2009)

Segundo Amaral (2016) a mineração de dados é composta de diversas fases, havendo dois padrões mais utilizados, o já citado KDD e o CRISP-DM (*Cross Industry*

Standard Process for Data Mining ou Processo Padrão Genérico para Mineração de Dados), sendo um padrão amplamente usado composto por seis fases:

- Entendimento do Negócio: aprender quais são os atributos do negócio a ser analisado, é uma etapa essencial para todo o processo.
- Entendimento dos dados: analisar os dados, suas estruturas, relacionamentos entre si, o volume, quantidade e como serão acessados.
- Preparação dos dados: tratar os dados pra organiza-los, limpar e selecionar o que será analisado por uma ferramenta de *machine learning.*
- Modelagem: um modelo é produzido para ser usado nos dados ainda não analisados.
- Avaliação: a performance do modelo é avaliada
- Implementação: o modelo criado começa a ser utilizado.

O KDD pode ser dividido em cinco fases que são semelhantes ao CRISP-DM (Amaral, 2016):

- Entendimento do negócio;
- Pré-processamento;
- Transformação;
- Mineração de Dados;
- Interpretação e Avaliação.

Os dados podem ser classificados em várias categorias distintas que afeta nas características e funções que podem e devem ser levadas em consideração durante um projeto. Duas categorias básicas são os dados gerados por humanos e dados gerados por máquinas. O primeiro tipo é composto por dados criados a partir do pensamento de uma pessoa que foi transformado em um dado ou da interação humana com meios digitais, como o uso de mídias sociais, aplicativos de comunicação, criação de documentos, e-mails entre outras formas tanto no nível pessoal quanto empresarial. Esses dados por utilizados para fazer diversos tipos de analises em seus respectivos contextos. Além das informações geradas ativamente

pelas pessoas, há dados que são registrados de maneira implícita, como os sites visitados, tipo de equipamento usado, localização. (MARQUESONE, 2018)

Mas além dos dados criados por humanos, diretamente ou indiretamente, Marquesone (2018) explica haver outro tipo de dado gerado a partir de processos computacionais sem uma pessoa intervindo, como a geração de registros de log que armazenam a utilização de servidores, também dados gerados em interações entre máquinas, que vem crescendo com o paradigma da internet das coisas (*Internet of Things – IoT*) que conectada vários dispositivos aumentando a interação das pessoas e máquinas nos ambientes físicos e virtuais e a quantidade de dados gerados a partir disso.

Extração, Transformação e Carga (*Extract, Transform, Load, ETL*) são processos utilizados na Ciência de Dados que consiste em coletar as informações de uma ou diversas fontes de dados, passando por etapas para transforma-los e armazena-los em *data warehouses*. Os dados são copiados e guardados de maneira temporária para depois passar pelas transformações necessárias e ao final ser armazenados definitivamente em um local especifico. Os processos de ETL podem converter um modelo relacional em um multidimensional, promover a limpeza dos dados, eliminar duplicatas e aumentar a qualidade dos dados. ETL é um processo custoso e suscetível a falhas, sendo necessário um programa que consiga lidar com imprevistos. ETL é usado para integrar os dados sendo algo essencial para uma companhia com diversos setores. (AMARAL, 2016)

2.4 Data Warehouse

Data warehouses são repositórios de dados cujo o objetivo é armazenar os dados de forma a tornar os processos de análise mais fácies, guardando os dados históricos da empresa originados de diversos bancos de dados transacionais diferentes. Dentro de um *data warehouse* pode haver os *data marts* que são bancos de dados menores onde são armazenadas informações de departamentos específicos, podendo estar ou não integrados com outros *data marts* e *data warehouse*. (AMARAL, 2016)

Amaral (2016) também explica que um *data warehouse* usa o modelo multidimensional, onde um fato é o centro da estrutura a ser analisada, enquanto as dimensões são as características a serem analisadas e são organizadas em formato

de hierarquias. No *data warehouse* também é possível definir o nível de detalhamento dos dados, definido pela granularidade, quanto menor a granularidade, mais detalhada a informação.

2.5 Data Lake

Com o objetivo de tomar o acesso aos dados dentro da corporação, algumas empresas utilizam data lakes, que é um local onde diferentes tipos de dados, estruturados ou não-estruturados, são armazenados juntos, esses dados podem ser originados de fontes diversas e como estão em uma mesma plataforma proporciona a integração e a análise desses dados através de ferramentas de *big data*. (MARQUESONE,2018)

Segundo Tito et al (2020) um *data lake* possibilita a armazenagem de diversos dados em seus formatos originais sem perder nada, com o objetivo de serem explorados futuramente. Sendo mais vantajoso do que um *data warehouse* que têm a tendência de seguir um esquema de modelação para todas as consultas. O funcionamento de um *data lake* acontece através das seguintes etapas:

- Ingestão de dados: a importação de dados em tempo real oriundos de diversas fontes.
- Armazenamento de dados no formato original: são guardados de maneira que torne seu acesso fácil e sem passar por nenhum tipo de transformação.
- Processamento e análise: os dados passam por processos de transformação para se adequarem ao formato necessário para a análise.
- Exploração e visualização: converte os resultados da análise em formatos que tornem a visualização e a extração de informações mais fácil.

3 ANÁLISE DE DADOS

Considera-se a análise de dados como um processo para obter informações que sejam utilizáveis usando técnicas para inspecionar, limpar, transformar e modelar os dados. Sendo empregada em diversos setores, como a área de negócios e tomada de decisões, ciências extas e humanas. Uma vez coletados, os dados são analisados com objetivo de responder perguntas e testar ou refutar hipóteses. A análise de dados pode ser dividida nas seguintes fases: (WIKIPEDIA)

- Requisitos de dados
- Coleta de dados
- Processamento de dados
- Limpeza de dados
- Análise exploratória de dados
- Modelagem e algoritmos
- Produtos de dados
- Comunicação

Amaral (2016) define que a análise de dados como a aplicação de processos de transformação nos dados com o objetivo de obter informação. Distinguindo dois tipos de análises diferentes, as explicitas quando as informações desejadas estão claramente expostas nos dados, sendo necessários somente algumas operações básicas para evidenciar os dados e informações, como a utilização de filtros, colocar os registros em ordem, colunas com alguns cálculos e o uso de comados SQL. Enquanto nas análises implícitas, as informações desejas não estão expostas de maneira clara, sendo necessário utilizar métodos mais avançados para a extração do conhecimento, como o uso de estatística ou aprendizado de máquina (*machine learning*). Sendo essencial conhecer os dados antes de analisa-los.

A análise de *big data* envolve grandes volumes de dados coletados sendo esses estruturados ou não os trabalhando de maneira analítica e inteligente, podendo haver diversos caminhos diferentes a serem seguidos no processo de tomada de decisão. Entregando informações valorosas para o aperfeiçoamento do desempenho de uma empresa em qualquer segmento de mercado ao torna-la mais eficiente a partir

das análises de dados realizadas, trazendo benefícios financeiros para a companhia. (JUNIOR e PINTO, 2018)

Segundo Junior e Pinto (2018) ao se levar em consideração a quantidade de dados gerados é possível analisar as informações de diversas formas para a tomada de decisão, mas quatro tipos de análises se sobressaem devidos aos resultados que oferecem:

- Análise preditiva: utilizando as informações já existentes nos bancos de dados é feita uma análise sobre possíveis cenários futuros a partir do reconhecimento de padrões nas informações existentes. Não confiando unicamente na intuição, usando de *data mining* para informações importantes, dados estatísticos e históricos para se ter noção das predisposições futuras.
- Análise prescritiva: criando cenários com base nos dados, a análise busca oferecer qual melhor solução para a resolução de um problema.
- Análise descritiva: busca descrever os dados no presente, sem relacionar com informações do passado ou suposições futuras. É feita e acompanhada em tempo real.
- Análise diagnostica: esse tipo de análise tem como objetivo avaliar as razões que levaram certos fatos acontecerem, identificando possíveis falhas e quais pontos são ineficazes, mostrando também estratégias que estão funcionando.

O processo de analisar dados é complexo, busca-se através dele organizar os dados recolhidos com o objetivo de extrair informações apropriadas para se chegar ao objetivo delimitado no início do processo. É preciso aplicar corretamente as ferramentas disponíveis para realizar a análise, extraindo os dados necessários. Parte importante desse processo é elaborar as perguntas corretas que deverão ser respondidas e que tipo de análise será feita. (ASSIS e SILVA, 2019)

A maior parte do tempo é usado para o processamento e limpeza dos dados com o objetivo de prepara-los para a análise, pois em uma base de dados haverá dados faltando, incompletos, corrompidos, duplicados entre outros fatores que atrapalham a análise, sendo necessário que o profissional tenha habilidade para tratar os dados evitando resultados inconsistentes. Por isso é importante validar a qualidade dos

dados ao processa-los e refina-los, sem isso dados errados podem ser usados na construção de modelos analíticos cujo resultados se demostrem incorretos na realidade prejudicando a tomada de decisão. (MARQUESONE, 2018)

A análise exploratória de dados é um processo concebida pelo estatístico John Wilder Tukey e utilizado até os dias atuais, cujo o foco é buscar compreender os dados antes de aplicar outras técnicas de análise que visam extrair conclusões deles. Análises exploratórias podem usar técnicas quantitativas, como calcular medidas de dispersão e média, mediana e desvio padrão e também técnicas visuais nas seguintes formas: (AMARAL, 2016)

- Gráficos de dispersão nos quais é possível fazer relações entre valores numéricos distribuídos no gráfico;
- Em diagramas de caixa, *boxplots*, onde são destacados a média, mediana, os maiores e menores valores;
- Histogramas usado para exibir a frequência dos dados dentro intervalos dentro de um gráfico de barras;
- Nuvem de palavras utilizando principalmente na mineração de dados, coleta diversas palavras, as exibindo dentro de um gráfico de acordo com a frequência que aparecem.
- Caras de Chernoff no qual são exibidas figuras de rostos humanos para exibir informações através deles.

Há diferenças entre analises exploratórias e explicitas, enquanto a primeira busca mais entender os dados, a segunda tem foco em um objetivo mais especifico, apesar de possuem diversas técnicas em comum. Uma técnica usada na análise explicita é o uso de junções que visam unir os conteúdos de tabelas diferentes utilizando suas chaves primárias e chaves estrangeiras e as relações entre os dados presentes nas tabelas. Além das junções, há outras técnicas como o uso de condições lógicas para criar um subconjunto de dados, chamado de predicados, os resumos que agregam valores e geram informações como médias, desvios padrões, frequências por exemplo. Enquanto os resumos lidam com valores nominais, a estratificação usa valores numéricos. Durante o processo de análise explicita também são verificados a existências de dados duplicados ou semelhantes, os padrões e lacunas e distorções. (AMARAL, 2016)

Amaral (2016) continua ao falar sobre técnicas de analises implícitas, como o aprendizado de máquina computacional que busca descobrir os padrões presentes nos dados, porém ocultos, para isso são usadas ferramentas estatísticas e de inteligência artificial, esse processo também está relacionado com a mineração de dados, sendo empregado em áreas como negócios, medicinas, detecção de fraudes. Há vários produtos ofertados para o uso em mineração de dados com diversos recursos diferentes, como o R e Weka que são *open source* e gratuitos e os de grandes empresas como Microsoft, SAS e Oracle.

3.1 Visualização de dados

A visualização de informações é uma área cujo estudos vem se ampliando no cenário onde há cada mais informações disponíveis para ajudar os usuários a analisarem e compreender informações. O excesso de informações pode atrapalhar na interpretação dos dados e na tomada de decisão, então a tecnologia de visualização de dados é um diferencial que auxilia na transparência dos dados e no processo de tomadas de decisão. (DE PAULA, et al, 2011)

O objetivo da visualização de dados é transmitir informações de modo preciso e eficiente para que os usuários consigam compreender os dados informados utilizando tabelas e diversos gráficos diferentes que podem ser manipulados com uso de softwares, criando maneiras de expor os dados. Não é somente sobre obter e analisar dados, mas ter a capacidade de criar sentido e explicações a partir deles. O grande volume de informações gerados com o advento do *Big Data* não terão valor se não for possível tirar as informações necessárias e a visualização de dados se torna importante ao apresentar os elementos extraídos de maneira gráfica assim facilitando a compreensão e a tomada de decisão. (SILVA,2019)

A visualização de dados atua como um meio eficiente de transmitir as mensagens, utilizando de representações gráficas para a melhor compreensão das informações. O ser humano possui grande capacidade de entender padrões através de estímulos visuais, tornando a assimilação de informações eficiente quando apresentada de maneira gráfica, mas para isso é preciso uma visualização de dados clara acerca do que é dito. (MARQUESONE, 2018)

Silva (2019) comenta que com a *big data* e popularização da visualização de dados, os tomadores de decisão das empresas ganharam a possibilidade de escolher quais informações são importantes para o negócio, pois proporcionou a obtenção de insights com as novas maneiras de apresentar os dados e informações. Com o aumento da geração de dados precisar passar por processos de coleta, exploração, processados para depois serem armazenados e usados em análises para encontrar informações importantes. Tendo quatro pontos importantes para a visualização de dados:

- O conjunto de dados ser limpo, possuir um formato adequado para ser utilizado em um programa de visualização;
- A comunicação ser feito através de uma mensagem somente que ganhará destaque nos gráficos;
- Ter um gráfico que se encaixe na maneira como a informação precisa ser passada;
- Possuir um design e cores que deem destaque às informações passadas.

O processo de visualização de dados passa por processos onde os dados brutos são transformados para depois haver um mapeamento visual onde sã escolhidas as maneiras de apresentar os dados e as transformações visuais, onde o usuário pode interagir com a representação gráfica de informações podendo modifica-la para ter outros pontos de vista sobre os dados. (DE PAULA, et al, 2011)

Para Marquesone (2018) é necessário determinar o objetivo da visualização no início do processo. Para a autora, a visualização pode ser dividida em dois tipos, visualização exploratória e explanação de dados. Na visualização exploratória, os dados são analisados para se decidir como serão usados, verificando as tendências, relacionamentos e anomalias presentes, sendo usados algumas representações visuais para expor essas informações, sendo duas delas o histograma que mostra a distribuição dos dados e frequência e o diagrama de caixa, ou *boxplot* (Figura 2), usando para encontrar dados anômalos e fazer paralelos entre grupos de dados diferentes. Enquanto a visualização explanatória é criada não somente para a compreensão do analista, mas sim para um público diverso que desejam saber os resultados da análise, sendo necessário criar uma visualização clara que dê ênfase nas informações desejadas.

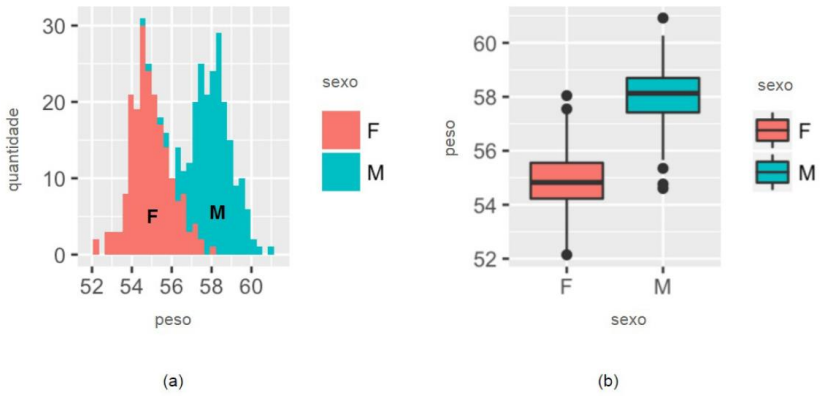


Figura 2– Exemplo de histograma(a) e boxplot(b). (MARQUESONE,2018)

O computador é usado para expor padrões, tendencias e relações entre as informações. Para permitir a interação do usuário com os dados são usadas várias técnicas, algumas delas são os gráficos de linha, de barra, gráficos de dispersão e de pizza, além das cores para retratar cada informação. Sendo essencial compreender qual melhor representação gráfica é a melhor para expor as informações desejadas com o objetivo de auxiliar a tomada de decisão e servir como ferramentas de avaliação. (SADIKU et al, 2016)

Sadiku (2016) continua ao dizer que o design de uma visualização serve como ferramenta para auxiliar na tomada de decisão aumentando o entendimento sobre os dados, sendo necessário saber como a visualização de dados será usada, pois ela não é somente sobre mostrar os números, mas também selecionar os dados exibidos gerando reflexão sobre eles. A visualização de dados é parte importante das ciências da computação, possuindo várias ferramentas desenvolvidas para a análise de dados e aplicadas em várias áreas, como saúde pública, detecção de fraudes e otimização do uso de energia elétrica.

Para criar uma interface visual é necessário trabalhar aspectos de forma, cor, tamanho, saturação, área entre outros que orientam o leitor em quais pontos observar apresentando um conjunto com variáveis diferentes. Sendo preciso considerar alguns atributos de forma, cor e posição espacial que fazem o olhar da pessoa focalizar naquilo que é importante e na ordem correta, sendo essencial haver um profissional na equipe que entenda sobre design gráfico. A pessoa responsável por criar a visualização precisa ter em mente quais questões deseja responder com a apresentação para assim saber quais tipos de gráficos são mais adequados no projeto. (MARQUESONE,2018)

Há uma grande variedade de gráficos que podem ser utilizados, além dos tradicionais gráficos de barras, pizza ou linha, para representar informações de maneira eficiente, como os gráficos de mapas que possuem informações geográficas ou *word clouds,* nuvem de palavras, um tipo de gráfico que mostra informações em forma de texto, quanto maior a frequência de uma palavra, maior ela aparecerá na imagem (Figura 3). (MARQUESONE, 2018)

Figura 3– gráfico word cloud. (MARQUESONE,2018)

4 PROCESSOS DE TOMADA DE DECISÃO

A informação é o resultado da construção de sentido por um indivíduo dentro de um contexto especifico a partir dos conhecimentos que possui. Estando a tomada de decisão ligada a necessidade do ser humano em adquirir conhecimento e informação com o objetivo de responder perguntas, compreender uma situação e negociar algo. Essas características estão relacionadas com o processo de tomada de decisões. (FICHT et al, 2019)

Para um empreendimento ter sucesso é primordial haver planejamento onde as decisões são tomadas de modo estratégico, buscando atingir o melhor resultado independente da opção assumida, sendo necessário desenvolver processos, utilizar técnicas e ter atitudes coerentes e concisas com as decisões a serem tomadas, tirando proveito de algoritmos presentes nos métodos de apoio multicritério a decisão. Existindo várias técnicas que podem ser usadas, contudo é preciso escolher a que melhor se enquadre na situação. (FERREIRA et al, 2018)

Ficht et al (2019) cita em seu texto que o fator crucial para o processo de decisão é a informação. Sendo importante nesse processo o gerenciamento da informação e o resultado é a aquisição de um novo conhecimento para algo ser decidido de maneira racional. A tomada de decisão envolve a manipulação de informação e havendo a possibilidade de sofrer influência de diferentes fatores dependendo do contexto, como no empresarial no qual os dados podem ser classificados como quantitativos ou qualificativos.

Para haver uma tomada de decisão é preciso colocar em pratica o conhecimento adquirido nos processos de ciência e análise de dados, onde os dados são transformados em informação que se tornam conhecimento. Nesse ponto a ciência de dados fornece suporte aos processos de tomada de decisão através da curadoria dos dados e informações e permitindo a analise com base em dados acurados. (RAUTENBER e CARMO, 2019)

Provost e Fawcett (2013) dizem em seu trabalho que o principal objetivo da ciência de dados é aprimorar o processo de tomada de decisão de uma empresa. Ao invés de confiar na intuição e na experiência, a pessoa responsável pelas decisões pode usar os métodos de tomada de decisão baseada em dados (*data-driven decision making- DDD)* embasar uma escolha com análises das informações disponíveis.

Pesquisas realizadas mostraram como a DDD influencia a performance de uma empresa, a tornando mais competitiva.

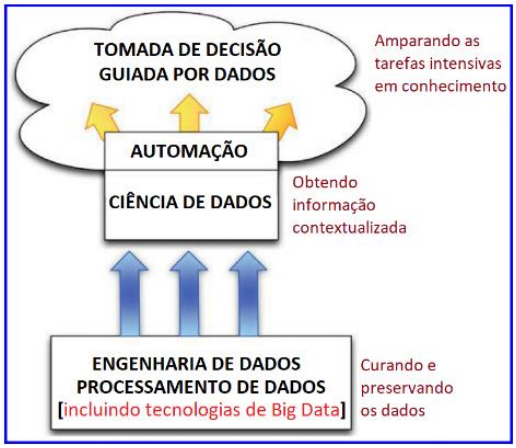


Figura 4– Big Data aplicada no processo de tomada de decisão por Provost e Fawcett, traduzido por Rautenberg e Carmo(2019)

Segundo Rautenberg e Carmo (2019) a ciência de dados oferece ajuda aos gestores durante a execução das tarefas relacionadas aos processos de tomada de decisão, aumentando a qualidade das escolhas e da produtividade como mostrado na figura 4. Algumas dessas tarefas são:

- Associação: tarefa na qual dois grupos de informações para se extrair comparações, relações de causa e efeito, combinações;
- Avaliação: as informações são examinadas para verificar se estão dentro dos parâmetros estabelecidos para a tomada de uma decisão;
- Diagnostico: é verificado o estado do objeto de estudo e se o mesmo está dentro dos padrões e comportamentos estabelecidos pela regra de negócio;
- Monitoramento: é o acompanhamento constante de um objeto ao longo do tempo, visando determinar se o comportamento do mesmo está dentro da normalidade e identificando qualquer mudança que ocorra;
- Predição: a partir da avaliação dos históricos de dados existentes é feita uma previsão de quais eventos e de que tipo podem acontecer futuramente.

O *Big Data* oferece um volume grande de informações e dados que ao passar por processos de análise oferecem amparo a tomada de decisão ao expor padrões e correlações úteis para se decidir algo, porém não estão explícitos nos dados. São usados softwares para realizar as correlações em uma ampla gama de dados e assim oferecer embasamento das opções a serem escolhidas no processo de tomada de

decisão, podendo afetar de maneira positiva áreas de negócio como desenvolvimento de produtos e de mercado, melhorar a experiência do cliente e seu relacionamento com a empresa e previsão de mercado. Contudo também há riscos ao usar a análise desse tipo de dados, aparecem questões como onde os dados serão armazenados, como protege-los de acordo com as exigências legais, lidando com as questões de segurança das informações, se a infraestrutura suporta lidar com o *Big Data*, além da qualidade dos dados obtidos. (JUNIOR e PINTO,2018)

4.1 Business Intelligence

O uso de *Business Intelligence* (*BI*) vem crescendo nos últimos anos com a grande demanda das empresas em analisar grandes volumes de informações e dados. *BI* pode ser definido como conjuntos de aplicações e ferramentas com o objetivo de auxiliar os processos de tomada de decisão das organizações através da análise de dados. Surgindo da necessidade dos administradores das empresas dos mais diversos tamanhos em trabalhar com amplas quantidades de dados para proporcionar uma visão holística da companhia baseada em conhecimentos internos e externos originados nos três níveis empresariais, o operacional, tático e estratégico. (SCHINAIDER et al, 2022)

Em 1958, um cientista da computação, Hans Peter Luhn, escreveu um artigo discorrendo sobre o potencial do *Business Intelligence* e como o desenvolvimento de sistemas automatizados seriam capazes de ampliar o alcance da informação em diversas áreas. Além de definir o que seria *BI*, Luhn, hoje considerado o pai do *Business Intelligence*, definiu métodos usados na construção de sistemas da IBM. Com a criação dos computadores, surgiu um novo meio para as organizações armazenarem seus dados, sendo a invenção do HD pela IBM em 1956 algo revolucionário para o armazenamento de informações e com o desenvolvimento tecnológico, o espaço para guardar dados aumentou consideravelmente. Com isso surgiu o primeiro sistema de gerenciamento de dados que ficou conhecido como *Decision Support System* (DSS) sendo o *BI* considerado uma evolução do *DSS*, sendo ferramentas para armazenar e analisar dados. A partir da década de 80 surgiram novas e mais simples ferramentas para serem usadas no *BI*, usadas produzir e

visualizar relatórios. Com o início dos anos 2000, dois problemas existentes, tempo e complexidade, foram resolvidos com as novas tecnologias, como a possibilidade de acompanhar informações em tempo real, aumentando a agilidade das tomadas de decisão. (ALASIRI e SALAMEH, 2020)

Alasiri e Salameh (2022) contam que o aumento de complexidade dos processos de negócios devido às mudanças nos ambientes de negócios faz com que se tornasse essencial os responsáveis da empresas estarem bem informados e aumentando a importância do *BI* como sendo um sistema capaz de armazenar, processar e analisar grandes volumes de informação com um conjunto de ferramentas e funções com o objetivo tornando as decisões estratégicas mais eficientes para ganhar competividade, além de melhorar as receitas da empresa e aperfeiçoar as operações.

Para a implementação de um sistema de *BI* em uma empresa é necessário a escolha de um sistema que atenda demandas da companhia e também haver treinamentos para capacitação dos funcionários a fim de que esses não tenham dificuldade na utilização do novo sistema, criando uma cultura voltada ao *BI*. A utilização de um sistema de *Business Intelligence* promove redução dos custos por oferecer tomadas de decisões mais acertadas e expor problemas antes escondidos, mas para isso o uso da ferramenta precisa estar alinhado com os objetivos e visão estratégica da empresa, além de haver bases de dados que forneçam informações com qualidade para as análises. (SCHINAIDER et al, 2022)

5 CONSIDERAÇÕES FINAIS

Os termos ciência de dados e *big data* ganharam destaque nos anos mais recentes com o crescimento do volume de dados gerados e a utilização das informações extraídas deles através de diversos tipos de análise nos processos de tomada de decisão. Esse aumento se deve à ampliação do uso de ferramentas tecnológicas que geram e captam dados nas mais diversas áreas.

Há variados tipos de processos envolvidos para lidar com esse grande volume de dados gerado, chamado também de *big data*, que exige poder de processamento, além de uma equipe multidisciplinar com profissões com conhecimentos nas áreas, como estatística, computação entre outras. Envolve também diferentes tipos de tecnologias, desde a parte de armazenamento com vários tipos de bancos de dados, relacionais que usam linguagem SQL e os não relacionais dos tipos chave-valor, colunas, grafos, documentos, armazenamento em nuvem, *data warehouse*, até questões ligadas a análise e visualização para tomada de decisão.

Após serem extraídos e transformados, os dados são usados para a realização de análises com o objetivo de descobrir informações acerca determinados temas para assim se obter conhecimento a ser usado em uma tomada de decisão. Pode ser feitas tipos de análises diferentes de acordo com a situação, uma análise descritiva para entender uma situação ou uma análise preditiva para ter uma visão de possíveis cenários futuros. Uma etapa importante da análise de dados é a maneira como as informações serão visualizadas, para isso é necessário escolher a maneira ideal para expor os dados de acordo com o que é pedido, qual o gráfico mais adequado, as cores utilizadas entre outros detalhes.

E todos esses processos, desde a extração até a visualização dos dados, tem como objetivo gerar indicadores para a aquisição de conhecimento auxiliando assim os processos de tomada de decisão baseada em dados fornecidos pelas mais diversas fontes. As ferramentas de *Business Intelligence* são de grande importância nos tempos atuais por facilitarem os processos que levam às tomadas de decisão.

Esses conjuntos de ferramentas e técnicas veem se tornando cada vez mais essenciais às corporações pois tornam a visão sobre o negócio mais clara, identificando os pontos fortes e fracos além de ajudar a determinar quais rumos possuem mais vantagens se seguidos. A tendência de se utilizar dados é aumentar

cada vez mais, assumindo uma posição central quando for tomada decisões, o que requer cada vez mais profissionais capacitados e aperfeiçoamento das técnicas e ferramentas usadas, desde a parte envolvendo a geração e armazenamento de dados até seu processamento e análise pois se tornará essencial para o crescimento e sobrevivência das organizações o domínio sobre os dados produzidas e das informações extraídas deles.

REFERÊNCIAS

ALASIRI, Mohanad M.; SALAMEH, Anas A., The Impact of Business Intelligence (BI) and Decision Support Systems (DSS): Exploratory Study (June 20, 2020). International Journal of Management, 11 (5), 2020, pp. 1001-1016, Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3631747. Acesso em: 21 jul. 2022.

ALECRIM, Emerson. O que é Big Data?, Jan, 2015 Disponível em: https://www.infowester.com/big-data.php. Acesso em 11 de nov. 2021.

AMARAL, Fernando. Introdução à Ciência de Dados: Mineração de dados e big data. Rio de Janeiro: Alta Books Editora,2016. Edição Kindle.3915 posições.

ASSIS, Geyzon Ferreira da Silva; SILVA, Rogério Oliveira da. As questões da análise de dados no contexto da ciência de dados. Revista Tecnologias em Projeção, v10, n°1, ano 2019. p.81. Disponível em http://revista.faculdadeprojecao.edu.br/index.php/Projecao4/article/view/1361. Acesso em 6 de mar. 2022

CAMILO, Cássio Oliveira; SILVA, João Carlos da. Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas. Technical Report - RT-INF_001-09 -Relatório Técnico Agosto, 2009. Disponível em: https://rozero.webcindario.com/disciplinas/fbmg/dm/RT-INF_001-09.pdf. Acesso em 20 de jun. 2022

DE PAULA, Melise M. V.; RIBEIRO, Fernanda C.; CHAVES, Miriam; RODRIGUES, Sergio A.; DE SOUZA, Jano M.. A Visualização de Informação e a Transparência de Dados Públicos. *In*: SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO (SBSI), 7. , 2011, Salvador. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2011. p. 384-395. DOI: https://doi.org/10.5753/sbsi.2011.14592. Acesso em 05 de mar. 2022

FERREIRA, Maria Madalena Guerra; JÚNIOR, Samuel Natividade Ferreira; SOUSA, Jordana Ramalho; EGUCHI, Thiago Yugi; SILVA, Ruy Gomes; SERRA, Cláudio Mauro Vieira. Escolha dos métodos de multicritério a tomada de decisão com o auxílio de um sistema especialista formulado a partir de um fluxograma. Braz. Ap.Sci. Rev., Curitiba, v. 2, n. 5,p. 1593-1608,out./dez.2018. Disponível em: https://brazilianjournals.com/ojs/index.php/BASR/article/view/545. Acesso em 14 de set. 2022.

FICHT, Nadia; ROGO, Gysele; LUNARDELLI, Rosane Suely Alvares; MOLINA, Letícia Gorri; PALETTA, Francisco Carlos. Busca e uso da informação para tomada de decisão. 4º Colóquio em organização, acesso e apropriação da informação e conhecimento, Agosto, 2019. Disponível em: https://www.eca.usp.br/acervo/producao-academica/002958696.pdf. Acesso em 04 de jul. 2022.

FURLAN, Patricia Kuzmenko e LAURINDO, Fernando José Barbin. Agrupamentos epistemológicos de artigos publicados sobre big data analytics. Transinformação [online]. 2017, v. 29, n. 1, pp. 91-100. Disponível em: https://www.scielo.br/j/tinf/a/8d56jfrqnQ9xNRXyxvGd3vR/?lang=pt&format=pdf.Acesso em 07 de nov. 2021.

GALDINO, Natanael. Big Data: Ferramentas e Aplicabilidade. Disponível em: https://www.aedb.br/seget/arquivos/artigos16/472427.pdf. Acesso em 08 de nov. 2021.

JUNIOR, Jander César Fernandes; PINTO, Giuliano Scombatti. BIG DATA ANALYTICS: apresentação do estudo de caso da webmotors. V SIMTEC –Simpósio de Tecnologia -Faculdade de Tecnologia de Taquaritinga –2018. Disponível em: https://simtec.fatectq.edu.br/index.php/simtec/article/view/400/243. Acesso em 07 de nov. 2021

MARQUESONE, Rosangela. Big Data, Técnicas e tecnologias para extração de valor dos dados. São Paulo. Casa do Código, 2018.

OLIVEIRA, Samuel Silva de. Bancos de dados não-relacionais: Um novo paradigma para armazenamento de dados em sistemas de ensino colaborativo. Revista Eletrônica da Escola de Administração Pública do Amapá. Macapá, v.2 n. 1, p. 184–194, ago.- dez. 2014. Disponível em https://www2.unifap.br/oliveira/files/2016/02/35-124-1-PB.pdf. Acesso em 18 de fev. 2022.

PROVOST, Foster; FAWCETT,Tom. Data Science and it Relationship to Big Data and Data-Driven Decision Making. **Big Data**, Mar, 2013, p. 51-59. Disponível em < https://www.liebertpub.com/doi/full/10.1089/big.2013.1508. Acesso em 10 nov. 2021.

RAUTENBERG, S.; CARMO, P. R. V. do. Big data e ciência de dados: complementariedade conceitual no processo de tomada de decisão. Brazilian Journal of Information Science: research trends, [S. l.], v. 13, n. 1, p. 56–67, 2019. DOI: 10.36311/1981-1640.2019.v13n1.06.p56. Disponível em: https://revistas.marilia.unesp.br/index.php/bjis/article/view/8315. Acesso em: 08 de nov. 2021.

SADIKU, Matthew N. O.; SHADARE, Adebowale E.; MUSA, Sarhan M.; AKUJUBI, Cajetan M.. Data Visualization, International Journal of Engineering Research And Advanced Technology(IJERAT),Volume. 02 Issue.12, December– 2016. Disponível em: https://www.researchgate.net/profile/Adebowale-Shadare/publication/311597028_DATA_VISUALIZATION/links/5851945608aef7d0309f20a7/DATA-VISUALIZATION.pdf. Acesso em 12 de jul. 2022.

SCHINAIDER, M. A. A. .; LEE, V. N. T. .; SERVARE JUNIOR, M. W. J. BUSINESS INTELLIGENCE COMO SUPORTE À TOMADA DE DECISÃO: O ESTADO DA ARTE POR MEIO DO PROKNOW-C. Brazilian Journal of Production Engineering, [S. l.], v. 8, n. 2, p. 79–98, 2022. DOI: 10.47456/bjpe.v8i2.37106. Disponível em: https://periodicos.ufes.br/bjpe/article/view/37106. Acesso em 21 de jul. 2022.

SILVA, Fabiano Couto Corrêa da. Visualização de dados: passado, presente e futuro. LIINC em revista. Rio de Janeiro, RJ. Vol. 15, n. 2 (nov. 2019), p. 205-223. Disponível em https://www.lume.ufrgs.br/handle/10183/204001. Acesso em 06 de mar. 2022

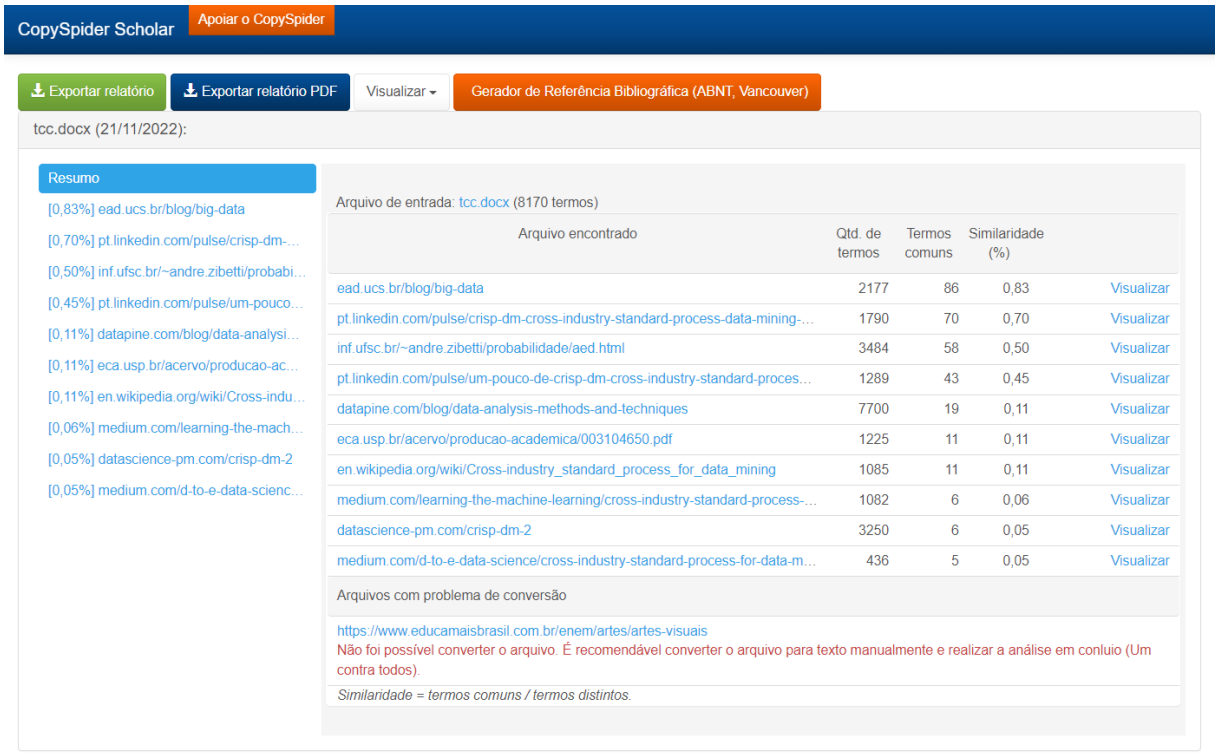
SILVA, Gilmar José da; FERREIRA, Júlio Cesar Oliveira. Análise comparativa de desempenho de consultas entre um bando de dados relacional e um banco de dados não relacional. Jun,2017. Disponível em https://repositorio.uniube.br/handle/123456789/178. Acesso em 17 de fev. 2022.

TITO, Lucas; MOTINHA, Cristina; SANTIAGO, Filipe; OCAÑA, Kary; BEDO, Marcos; DE OLIVEIRA, Daniel. Xi-DL: um Sistema de Gerência de Data Lake para Monitoramento de Dados da Saúde. *In*: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS (SBBD), 35., 2020, Evento Online. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2020. p. 151-156. ISSN 2763-8979. DOI: https://doi.org/10.5753/sbbd.2020.13633. Acesso em 19 de nov. 2022

WIKIPÉDIA, Análise de Dados. WIKIPÉDIA, a enciclopédia livre. Flórida: Wikimedia Foundation, 2021. Disponível em: https://pt.wikipedia.org/w/index.php?title=An%C3%A1lise_de_dados&oldid=62543054. Acesso em: 18 de fev. 2022.

APÊNDICE

Relatório de avaliação do CopySpider



CopySpider Scholar | Apoiar o CopySpider

Exportar relatório | Exportar relatório PDF | Visualizar | Gerador de Referência Bibliográfica (ABNT, Vancouver)

tcc.docx (21/11/2022):

Resumo

[0,83%] ead.ucs.br/blog/big-data
[0,70%] pt.linkedin.com/pulse/crisp-dm-...
[0,50%] inf.ufsc.br/~andre.zibetti/probabi...
[0,45%] pt.linkedin.com/pulse/um-pouco...
[0,11%] datapine.com/blog/data-analysi...
[0,11%] eca.usp.br/acervo/producao-ac...
[0,11%] en.wikipedia.org/wiki/Cross-indu...
[0,06%] medium.com/learning-the-mach...
[0,05%] datascience-pm.com/crisp-dm-2
[0,05%] medium.com/d-to-e-data-scienc...

Arquivo de entrada: tcc.docx (8170 termos)

Arquivo encontrado	Qtd. de termos	Termos comuns	Similaridade (%)	
ead.ucs.br/blog/big-data	2177	86	0,83	Visualizar
pt.linkedin.com/pulse/crisp-dm-cross-industry-standard-process-data-mining-...	1790	70	0,70	Visualizar
inf.ufsc.br/~andre.zibetti/probabilidade/aed.html	3484	58	0,50	Visualizar
pt.linkedin.com/pulse/um-pouco-de-crisp-dm-cross-industry-standard-proces...	1289	43	0,45	Visualizar
datapine.com/blog/data-analysis-methods-and-techniques	7700	19	0,11	Visualizar
eca.usp.br/acervo/producao-academica/003104650.pdf	1225	11	0,11	Visualizar
en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining	1085	11	0,11	Visualizar
medium.com/learning-the-machine-learning/cross-industry-standard-process-...	1082	6	0,06	Visualizar
datascience-pm.com/crisp-dm-2	3250	6	0,05	Visualizar
medium.com/d-to-e-data-science/cross-industry-standard-process-for-data-m...	436	5	0,05	Visualizar

Arquivos com problema de conversão

https://www.educamaisbrasil.com.br/enem/artes/artes-visuais
Não foi possível converter o arquivo. É recomendável converter o arquivo para texto manualmente e realizar a análise em conluio (Um contra todos).

Similaridade = termos comuns / termos distintos.