INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO SERTÃO PERNAMBUCANO
COORDENAÇÃO DO CURSO DE SISTEMAS PARA INTERNET

LEIDIANE ANGELICA NUNES DA SILVA

CIÊNCIA DE DADOS COMO MÉTODO DE TRANSFORMAÇÃO DE DADOS EM INFORMAÇÃO

SALGUEIRO

2022
LEIDIANE ANGELICA NUNES DA SILVA

CIÊNCIA DE DADOS COMO MÉTODO DE TRANSFORMAÇÃO DE DADOS EM INFORMAÇÃO

Trabalho de Conclusão de Curso apresentado a Coordenação do curso de Sistemas Para Internet do Instituto Federal de Educação, Ciência e Tecnologia do Sertão Pernambucano, campus Salgueiro, como requisito parcial à obtenção do título de Tecnoloco em Sistemas Para Internet.

Orientador(a): Prof. Francenila Rodrigues.

SALGUEIRO

Dados Internacionais de Catalogação na Publicação (CIP)

S586 Silva, Leidiane Angelica Nunes da.

Ciência de dados como método de transformação de dados em informação / Leidiane Angelica Nunes da Silva. - Salgueiro, 2022.
23 f. : il.

Trabalho de Conclusão de Curso (Sistemas para Internet) -Instituto Federal de Educação, Ciência e Tecnologia do Sertão Pernambucano, Campus Salgueiro, 2022.
Orientação: Prof. Msc. Francenila Rodrigues.

1. Ciência da Computação. 2. Ciência de dados. 3. Dado-Informação. 4. Tecnologia da Informação. I. Título.

CDD 004

Gerado automaticamente pelo sistema Geficat, mediante dados fornecidos pelo(a) autor(a)

LEIDIANE ANGELICA NUNES DA SILVA

CIÊNCIA DE DADOS COMO MÉTODO DE TRANSFORMAÇÃO DE DADOS EM INFORMAÇÃO

Trabalho de Conclusão de Curso apresentado a Coordenação do curso de Sistemas Para Internet do Instituto Federal de Educação, Ciência e Tecnologia do Sertão Pernambucano, campus Salgueiro, como requisito parcial à obtenção do título de Tecnoloco em Sistemas Para Internet.

Aprovado em: 09/03/2022

BANCA EXAMINADORA

_____
Prof. Francenila Rodrigues Orientador(a)
IF Sertão PE – Campus Salgueiro

_____
Prof. Marcelo Anderson
IF Sertão PE – Campus Salgueiro

_____
Prof. Francisco Junior
IF Sertão PE – Campus Salgueiro

SALGUEIRO

2022

Dedicatória.

Aos meus pais, Lucia e Argemiro por ter agreditado em mim e terem feito de tudo para que eu conseguisse chegar até aqui.

AGRADECIMENTOS

Ao Prof. Francenila, por toda a paciência e comprometimento em me orientar.
Aos professores participantes da banca examinadora Marcelo e Francisco Junior pelo tempo, pelas valiosas colaborações e sugestões.
Aos colegas da turma, pelas reflexões, críticas e sugestões recebidas,a todos meus irmão que estiveram comigo nessa caminhada, e a todos que diretamente e indiretamente fizeram esse trabalho acontecer.

"A tecnologia move o mundo." Steve Jobs

Ciência de Dados como método de transformação de dados em informação

Leidiane Angelica Nunes da Silva[1], Francenila Rodrigues

[1] Instituto Federal de Educação Ciência e Tecnologia -- IF SERTÃO-PE

leidiane.angelica@aluno.ifsertao-pe.br,@aluno.ifsertao-pe.br

Abstract. *In a world increasingly based on the culture of data, the exploration, collection, production, circulation and understanding of data become crucial. In this article, in addition to presenting the concepts of Big Data and Data Science, we sought to understand what data is; and data science as a strategy for transforming data into knowledge, with emphasis on its use in everyday applications, presenting its benefits in specific areas. The methodological strategies involve exploratory and descriptive research in theoretical and conceptual exploration. The deliberate results demonstrate the effectiveness of data science in relation to the evolution of data and the conversion of information into knowledge, aiming at improvements for the processing of large volumes of data.*

Resumo. *Em um mundo cada vez mais baseado na cultura dos dados, tornam-se cruciais a exploração, coleta, produção, circulação e compreensão dos dados. Nesse artigo, além de apresentar os conceitos de Big Data e Data Science, buscou-se a compreensão do que são dados; e a ciência de dados como estratégia para a transformação de dados em conhecimento, com ênfase no seu uso em aplicações cotidianas, apresentando seus benefícios em áreas específicas. As estratégias metodológicas envolvem pesquisa exploratória e descritiva na exploração teórica conceitual. Os resultados deliberados demonstram a eficácia da ciência de dados em relação à evolução dos dados e conversão das informações em conhecimento, objetivando melhorias para o processamento sobre grande volume de dados.*

1. INTRODUÇÃO

É notório que a tecnologia da informação já faz parte do cotidiano da sociedade em geral, e seus impactos são visíveis e expressivos, os quais inevitavelmente repercutem na vida das pessoas; consequentemente são gerados um volume desenfreado de dados. Com o crescimento acelerado de tecnologias, a essencialidade do uso da internet em especial no período pandêmico, a necessidade dos usuários e cada vez mais dispositivos conectados entre si mediante a internet das coisas(IOT)), conceitos como Big Date e Data Scienc são cada vez mais visto; tendendo a serem denominados fenômenos tecnológicos, isso se dá pelo fato de muitos dados serem produzidos por cada usuário.

Segundo Raudemberg e Carmo (2019) “Big Data se trata de um conjunto de dados impeditivo de captura, armazenamento, gerenciamento e análise por parte de ferramentas computacionais tradicionais, requer formas inovadoras de processamento de grandes volumes de dados heterogêneos[...]”. Em contrapartida temos Data Science ou em português Ciência de dados, cujo, a Oracle (ORACLE, 2021) Definiu como: “é um subconjunto da inteligência artificial (IA) e se refere mais às áreas sobrepostas de

estatísticas, métodos científicos e análise de dados - todas as quais são usadas para extrair significado e percepções dos dados.” Logo de acordo com esses termos ter um grande volume de dados não significa que todos esses dados sejam conhecimento, podemos assim dizer que ciência de dados é a extração de informações através da base de dados do big data, resultando em dados complementares que se tornam informação. Partindo dessa discussão, levando em consideração a evolução desses conceitos,buscamos responder ao seguinte questionamento: “Como transformar dados em conhecimento, utilizando data scienc?”. Visando para tal utilizar os conceitos teóricos e metodológicos da ciência de dados para a composição e revisão da informação, bem como, a importância da ciência de dados no cotidiano.

Para isso o presente estudo além dessa seção introdutória está estruturada da seguinte forma: seção 2 será apresentado o referencial teórico abordando conceitos de dados, informações e conhecimento, big data e ciência de dados; na seção 3 está descrito o método abordado; na seção 4 as principais aplicações no dia a dia; seção 5 resultados e discussões e ,por fim, na seção 6 as considerações finais.

2. REFERENCIAL TEÓRICO

Nessa seção são abordados alguns conceitos que inicialmente faz-se necessário para o entendimento, que são: Big Data e Data Science bem como conceitos que estejam relacionados a eles, a fim de aprofundar os estudos teóricos desses campos.

2.1 DADOS, INFORMAÇÕES E CONHECIMENTO

De acordo com Moreira, Beira e Oliveira (2020) a concepção de dado, informação e conhecimento são tidos como elementos essenciais para a comunicação. Os autores descrevem de forma resumida que os dados são comumente descritos como a matéria-prima para a informação, que é concebida como matéria-prima para o conhecimento, seguindo a hierarquia. Assim, o conhecimento pode ser compreendido como percepção da realidade de determinada informação.

Os dados podem ser definidos como informações brutas, que não possuem nem um significado antes de serem tratados. “Assim são observações específicas ou resultados de uma medição que não conseguem, por si só, transmitir uma mensagem; com isso vendo-o de maneira isolada não será compreendido o real significado.” (SANTOS-D’AMORIM et al., 2020)

Dando progressão, a ideia de informação integra-se ao conceito de dados, haja vista que a informação se dá quando os dados são estruturados, organizados, processados, contextualizados ou interpretados. Nesse contexto, quando um ou um conjunto de dados são dispostos, de modo a transmitir uma mensagem dentro de uma circunstância real, temos as informações. “As quais são providas de propósito, significado e relevância, podendo ser utilizadas pelo ser humano durante a tomada de decisão, por meio da sua compreensão e análise.” (SOMASUNDARAM; SHRIVASTA, 2011).

Fundamentando o conhecimento é gerado através da habilidade em analisar as informações encontradas. “[...] o mesmo ocorre quando é aplicado a informação, gerando uma ideia ou noção, resultando em aprendizagem; que pode-se dizer que é quando somos expostos a diversas informações novas, as quais, se tornam consistentes e somadas a nossa

experiência.” (DIAS; RODRIGUES 2017) . Em outros termos, o conhecimento ocorre no momento em que as informações são integradas e processadas, assim pode-se dizer que o ele constitui raciocínio e argumentação que a informação por si só, não conseguem conceber.

Desse modo, o método de prospecção em conhecimento tem a finalidade de dar maior segurança às direções perseguidas pela organização, dos quais, são produzidos internamente e externamente à organização, conseguindo assim consolidar valores para esse processo. Uma outra questão importante: “é a validade dos dados, informações e conhecimento, isto é, realmente eles respondem as perguntas críticas do negócio da organização ou população quanto a consistência e confiabilidade, utilidade e obsolescência e, finalmente, a confidencialidade exigida”. (PATRICIO; MAGNONI, 2018).

2.2 BIG DATA

Big Data é uma área de pesquisa apontada como de alto impacto para todas as áreas de conhecimento, tornando-se eficazes no armazenamento de dados. Com relação ao conceito e surgimento do termo Big Data o site da CETAX (2020) descreve “O termo Big Data nasceu no início da década de 1990, na Nasa, para descrever grandes conjuntos de dados complexos que desafiam os limites computacionais tradicionais de captura, processamento, análise e armazenamento informacional[...]”.

Com base nesses fatos demanda um conceito inovador para lidar com dados, para Dias e Rodrigues(2017), retratam Big Data como:

> *[...] o conceito de inovação disruptiva como pertinente para problematizar o horizonte das produções em visualização de dados na perspectiva de novo paradigma que representa o Big Data, tendo em vista tratar-se de um contexto ou de uma tecnologia disruptiva[...].*

Baseando-se nesses conceitos, apesar de um pouco distintos, podemos condensar o conceito como sendo um grande volume de dados estruturados ou não estruturados, de diversas fontes, que devem ser gerenciados e analisados de forma peculiar (JUNIOR et al., 2016).

Admitindo que encontram-se significativos conceitos com relação ao tema, tal qual, é o objetivo desta pesquisa, existem inúmeras definições para as características do Big Data; no Quadro 1 foram condensadas definições propostas por Raudemberg e Carmo(2019), cujo indicou seis características principais.

Quadro 1 - Definição dos seis Vs

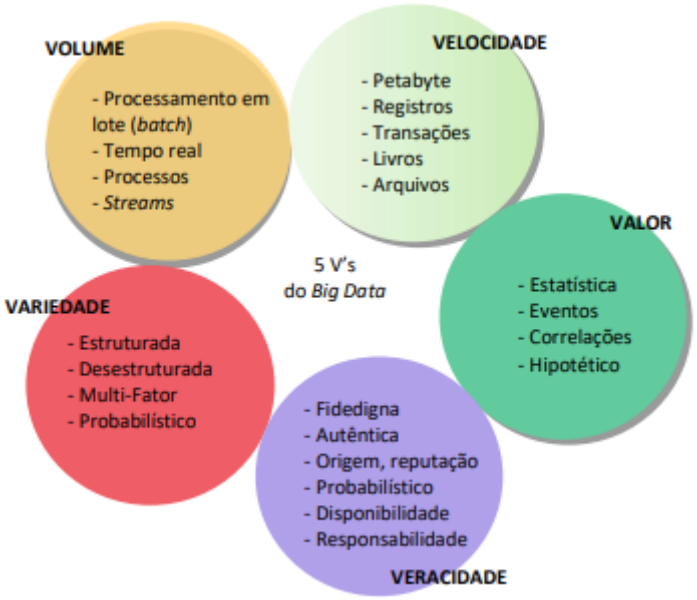
Caracteristica	Definição
Volume	É a característica mais evidenciada. Grandes volumes de dados são gerados mediante o uso de recursos computacionais abundantes. Com a evolução das mídias sociais e outros recursos e serviços da Internet, as pessoas produzem mais e mais conteúdo, vídeos, fotos, tweets, entre outros tipos de dados.

Velocidade	Os dados são gerados em grande velocidade, à medida que os recursos computacionais têm sua capacidade de produção, captura e processamento de dados aumentada.
Variedade	Os dados advêm de variadas fontes (sistemas legados, e-mails, posts em mídias sociais, arquivos de vídeo/áudio, gráficos, dispositivos ou sensores), as quais implementam tecnologias distintas para representação e armazenamento de recursos digitais.
Veracidade	Refere-se à integridade e à precisão dos dados, contrapondo o fenômeno GIGO (garbage-in, garbage-out – lixo entra, lixo sai) na recuperação da informação. Neste sentido, deve-se evitar ruídos e incertezas no armazenamento dos dados de modo a não interferir, consequentemente, na análise da informação e no Processo de Tomada de Decisão
Variabilidade	Relaciona-se à compreensão e ao tratamento dos fenômenos subliminares e temporariamente presentes nos dados. Por exemplo, sazonalmente, alguns eventos específicos (virais nas mídias sociais, como a estreia de um filme muito aguardado ou o acontecimento de um fato midiático) podem refletir em padrões de comportamento que não se sustentam ao longo do tempo.
Valor	É a característica mais importante em termos dos dados, independente das demais dimensões (volume, velocidade, variedade, variabilidade e veracidade). O valor em Big Data é, principalmente, percebido mediante a análise com dados precisos na aquisição das informações úteis.

Fonte: Elaborado pelo autor com base em Raudemberg e Carmo(2019)

Dentre as definições já descritas de Big Data, entende-se como: um mecanismo estratégico de análise, que tem como foco a análise de grande volumes de dados. No trabalho de SANTOS-D'AMORIM et al. (2020) é descrito apenas cinco Vs os autores fazem um esquema representando tais características.

Figura 1: Os 5 V's do big data e suas associações



Fonte: SANTOS-D'AMORIM et al. (2020)

2.3. DATA SCIENCE

Ciência de Dados (*Data Science*, em inglês), é um termo utilizado para descrever a presença direta e constante da transformação de dados em informações. Atribui-se à Ciência de Dados a extração de informação útil a partir de imensas bases de dados complexas, dinâmicas, heterogêneas e distribuídas (Raudemberg e Carmo 2019).

Conforme dialogado no livro de Amaral (2016), a expressão "Data Science" vem sendo proposta desde a década de 1960, todavia é uma ciência nova embora controversa e mal interpretada, essa ciência obtém de forma sistemática conhecimento e informações, tal qual organizada, processa e modela esses dados. A fim de elucidar a aplicação e relação desses conceitos dentro da Ciência da Informação, a ciência de dados busca compreender o dado em toda sua existência, desde de sua geração até seu desgaste.

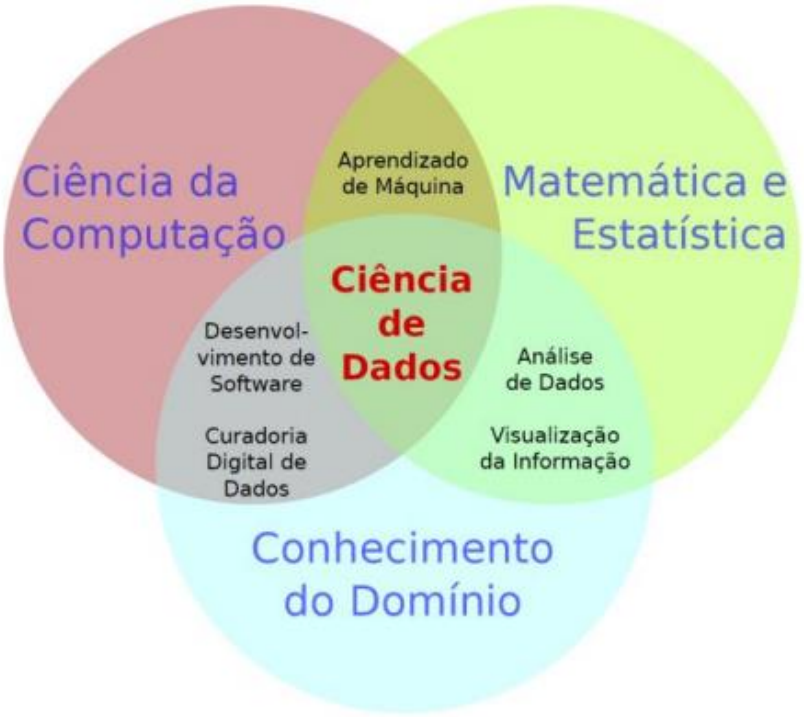
Amaral(2016) destaca ainda que:

> *"Normalmente a ciência de dados é comparada de forma equivocada apenas aos processos de análises dos dados, onde com o uso de estatística aprendizado de máquina ou a simples aplicação de filtro se produz informação e conhecimento. Nessa visão, "míope", a ciência de dados passa a ser vista apenas como um nome mais elegante para estatística[...]."*

O autor entende que a ciência de dados é bem mais complexa que deixa transparecer, pois a mesma usa de técnicas semelhantes a ciência de computação, que vai desde a obtenção dos dados, armazenamento, gerenciamento, visualização, compartilhamento, privacidade, modelagem, análise, e segurança dos dados, bem como a

integração de todos esses serviços. Na figura abaixo percebe-se que a ciência de dados engloba muitas outras áreas, sendo ela um núcleo interdisciplinar que contribui para a transformação dos dados.

Figura 2: Interdisciplinaridade da Ciência de Dados



Fonte: RAUTENBERG E CARMO, 2019

Como visto na imagem anterior, a ciência de dados contém traços de várias disciplinas. Como destaca Cardoso (2019) "A ciência de dados é um novo campo transdisciplinar que constrói e sintetiza várias disciplinas e corpos de conhecimento relevantes, incluindo estatística, informática, computação, comunicação, gerenciamento e sociologia".

3. METODOLOGIA

Como recurso metodológico para essa pesquisa recorreu-se à exploratório-descritiva utilizando-se e dissertando uma abordagem qualitativa acerca da temática, com o intuito de fazer levantamento de todas as áreas que atualmente fazem uso da ciência de dados, bem como do processo de transformação dos dados em informação-conhecimento. Utilizando de uma análise de referencial teórico que são desenvolvidas a partir de materiais já elaborados, cujo, também o permite verificar a aplicação da ciência de dados em determinadas áreas, sob o aspecto teórico de todas cases de estudo em suas abordagens.

De acordo com Prodanov e Freitas(2013), "as pesquisas descritivas são similares às exploratórias, que geralmente trata-se do envolvimento de verdades e interesses

universais, que tem como objetivo principal o desenvolvimento, sem o envolvimento de uma aplicação direta de prática". Para a realização desta pesquisa adotou-se um levantamento bibliográfico e documental que possibilitou a conceitualização e utilização da ciência de dados a partir de como é feito o processo de transformação de dados-conhecimento.

No que se refere à análise elaborada, foram selecionados artigos e periódicos publicados em anais e eventos, preferenciais dos congressos e eventos da Sociedade Brasileira de Computação-SBC apartir do ano de 2014, e livros específicos sobre ciência de dados. O material foi selecionado buscando os seguintes termos: "big data", "gestão de dados", "ciência de dados", "ciência de dados aplicada ao "e-commerce/educação/saúde"; que foram reunidos observando título, resumo, e palavras-chaves como fonte inicial da coleta.

Por fim, o levantamento bibliográfico e documental possibilitou melhor elaboração do procedimento metodológico semelhantes; vindo a destacar os que se assemelham a essa pesquisa, que são os casos dos trabalhos de Raudemberg e Carmo(2019) que destacam método de caráter descritivo, o trabalho de D'AMORIM utiliza de processo metodológico exploratório utilizando-se de um levantamento como método de coleta de dados, por ultimo, o trabalho de Rodrigues e Dias(2017) que realizaram uma pesquisa exploratória-descritiva, adotando a abordagem qualitativa sobre a temática.

4. TRANFORMAÇÃO DE DADOS

As principais pesquisas sobre ciencia de dados e mineração de dados estão concentradas na implementação e otimização dos algoritmos, os chamados algoritmos de aprendizado. Toda via, devido a quantidade e qualidade dos bancos de dados, os resultados prodizidos não se mostram confiaveis. Nessa seção são abordados conceitos que ajudaram a compreender o processo de transformação de dados em informação e posteriomente em conhecimento; apesar de existirem mais metodos, nesse trabalho foi descrito dois dos mais conhecidos.

4.1 KDD

A mineração de dados passou a ser definida como um processo ainda mais amplo chamado knowledge Discovery in Databases (KDD) ou traduzindo ao portugues Descorbeta do Conhecimento em Grandes Bases de Dados. O processo de KDD é focado nos objetos do negocio; podendo ser dividido em até 06 fases interdependentes, que são elas: Descoberta da Tarefa, Descoberta dos Dados, Limpeza dos Dados, Desenvolvimento do Modelo, Analise de Dados e Geração de Saida.(PIMENTA et al, 2018).

Em um nivel abstrato, o KDD se preocupa com o desenvolvimento de metodos e tecnicas para da sentido aos dados. Por mais que frequentemente KDD e Data Miningsejam entendido como sinonimos, o KDD compreende todas as etapas do processo

desde a existencia de dados enquanto a mineração de dados é apenas uma das etapas do processo(.

Segundo Bartlmae e Riemenschneider(2016), atualmente projetos de dados com KDD por serem complexos e fortemente dependete do usuario, falta de documentação, erros cometidos acabam sendo acometidos novamente; com isso a maioria dos projetos de KDD fracasam, execede o custo ou tem qualidade comprometida.

4.2 ETL

ETL – Extract, Transform, Load é o processo de carga de dados, utilizando em integração de sistemas, normalmente baseado em softwares e programação. Também encontrado como ETT – Extrair, Transformar, Carregar (em português). Esses softwares, cuja função é a extração de dados de diversos sistemas, transformação desses dados conforme regras de negócios e por fim a carga dos dados em um Data Mart ou um Data Warehouse.

Existem muitas ferramentas de ETL disponíveis no mercado como IBM Information Server (Data Stage), o Oracle Data Integrator (ODI), o Informatica Power Center, o Microsoft Integration Services (SSIS). Existe também um conjunto de Ferramentas de ETL Open Source como o PDI – Pentaho Data Integrator e Talend ETL.

O BI – Business Intelligence juntamente com a Data Science veio para agregar agilidade à nova realidade das empresas, estreitando a relação entre gestão e técnica. No mundo globalizado em que se vive hoje, seja nos negocios, na saúde, na educação e em qualquer outro setor que exija grandes processamentos de dados, é necessário que o banco de dados seja estruturado, com as informações consistente e mapeadas, para em seguida ser aplicada as transformações de limpeza e consolidação dos dados e finalmente o carregamento desses dados.(BANSAL,2014).

Para Ferreira et al(2010), “O processo de extração, transformação e carregamento (ETL) abrange alguns passos importantes, como exemplo, podemos considerar um Banco de dados de Clientes Especiais com todas as informações essenciais”. No mapeamento, a extração de origem deve conter a especificação da identidade e seus atributos detalhados, tudo armazenado numa zona temporária. Quando forem efetuadas as análises e filtragens dos dados, a nova versão poderá ser comparada com a cópia da versão prévia.

No trabalho de Bansal e Kagemann (2015), eles decrevem: “[...] A transformação inclui limpeza, racionalização e complementação dos registros, o processo de limpeza removerá erros e padronizará as informações e a complementação implicará no acréscimo de dados[...]”. Antes de esforçar-se em transformação de dados é fundamental diagnosticar e compreender os problemas, mais comuns são dados incompletos, formatação errada. A tabela a segui temos os conceitos das etapas do métdo ETL.

Tabela 1: Etapas do método ETL

Extract	A primeira fase do processo é destinada à extração de dados SQL. Nesse estágio, é possível fazer uma análise preliminar dos dados, organizando-os em uma área de transição. No processo de extração, os dados são organizados e convertidos em um formato único, o que torna possível manipulá-los nas

	próximas etapas. Como os dados são muito diferentes entre si, é necessário adotar essa medida inicial, fazendo a padronização massiva deles.
Transform	Na fase de transformação, ocorre a adaptação das informações que foram analisadas e padronizadas no estágio da extração. Aqui, os dados são transformados, fazendo o que se chama de higienização. O objetivo é levar para a análise apenas aquilo que será efetivamente aproveitado. Também são criados nessa etapa os filtros para agrupar informações de critérios como idade, localização, tempo, cargo, nível hierárquico ou qualquer outro que seja útil para a realização de futuras análises.
Load	No terceiro e último passo do processo, é preciso fazer o carregamento dos dados já organizados em um novo repositório.Isso ocorre em um ambiente corporativo (data warehouse) ou em um ambiente departamental (data mart). Para essa fase, novamente duplicamos a tabela com a informação tratada e realizamos os ajustes necessários para corrigir novos desvios de fluxo informacional. Mantendo um modelo dos dados organizados, é possível criar um mapeamento de todos os padrões, tornando-os sempre acessíveis para a utilização futura.

Fonte: Elaborado pelo autor com base em cetax(2020)

Importante destacar que Não necessariamente, executado em um único ambiente de tratamento informacional, podemos utilizar diversas aplicações para o processo todo, seja em nuvem ou não (AGRAWAL et al., 2008). Em um estágio mais avançado e com o trabalho concluído, é possível também fazer a mineração de dados, de forma que seja viável estabelecer e identificar novos padrões de comportamento de usuários, compradores ou, até mesmo, fornecedores.

5. ÁREAS DE APLICAÇÃO

Nesta seção, são apresentados as aplicações da ciência de dados nas áreas de educação, comércio e saúde, bem como suas limitações e desafios.

5.1 EDUCAÇÃO

O uso da ciência de dados na educação pode auxiliar a atuação na formulação de melhora nas políticas educacionais que favorecem as melhorias nos ganhos de

aprendizagem, contribuindo aos educadores e gestores facilidades e possibilitando o máximo de proveito do que pode oferecer as tecnologias computacionais.

Os autores Scaico, Queiroz e Scaico (2014), descrevem o uso do big data com a ciência de dados na educação como:

> *"Sua utilização na educação através do incentivo de governos, universidades e empresas pode estabelecer novas tecnologias, ferramentas e recursos que são capazes de apoiar uma cultura orientada ao conhecimento, à eficiência, à aprendizagem adaptativa e personalizada e promovedora de novas experiências de aprendizagem capazes de melhorar a maneira como os professores ensinam, os estudantes aprendem e a escola funciona."*

Conseguindo assim agrupar e explorar um imenso volume de dados, cujos, são compostos pelo elo de ensino-aprendizagem. Essas tecnologias emergem muitas informações que abrangem vários graus de especializações, com os quais, métodos manuais e convencionais não são capazes de acumular.

Ainda segundo Scaico, Queiroz e Scaico (2014): "A capacidade de processar massas de dados em escala, através da análise e da comparação de comportamento de milhares de estudantes,[...]oferecendo estímulo compatíveis com o seu nível de proficiência e dificuldade, gerando um ciclo de feedbacks contínuos." Além disso, é muito importante na aquisição de conhecimentos estabelecidos em meio ao processo de aprendizado, que especificam dificuldades a assentos deliberados ou parâmetros pedagógicos.

A EDS ou ciência de dados educacional teve origem no início do século 21 nas conferências sobre mineração de dados educacionais *(Education Data Mining-EDM).* A ciência de dados quando aplicada ao campo da educação engloba ciência da computação, educação, estatísticas e outras ciências para examinar, compreender e desmistificar a área da educação. No livro de Filatro(2021) ela descreve: "[...]pode ser definido como um campo orientado a dados, sistêmico, transdisciplinar e dinâmico, que combina habilidades técnicas e sociais à compreensão profunda de práticas educacionais em diferentes ambientes de aprendizagem". Fazendo assim termos uma visão vasta sobre educação inteligente, aprendizagem inteligente, ambientes inteligentes e singulares.

Na educação temos alguns exemplos que utilizam os recursos disponibilizados pela ciência de dados, temos universidades que reúnem informações sobre seus alunos por meio de sistemas personalizar a assistência a alunos por meio de sistemas que permitem personalizar a assistência ao aluno, auxiliando assim o combate à evasão escolar e agregando aos estudantes, a exemplo da universidade comunitária de Saddleback na Califórnia.

Nesse contexto também temos empresas de livros digitais capazes de rastrear a interação dos alunos pelos textos digitados, como a Smart Course, "[...]startup EdTech e HealthTech que usa produtos e serviços digitais para educar crianças com necessidades especiais e seus cuidadores, usam big data e instrutores líderes mundiais para oferecer

cursos online envolventes".(CAPSOURCE,2022). Outras plataformas comerciais de aprendizagem adaptativa utilizam conceitos de Big Data e Ciência de Dados a fim de auxiliar o aprendizado e melhora na distribuição dos conteúdos, que são os casos da DreamBox e da Knewton.

A khan Academy desenvolveu um método de modelo de previsão de comportamento, por meio de uma avaliação diária de mais de oito milhões de pontos de dados, possibilitando avaliar os ganhos de aprendizagem e conteúdo individual para cada usuário, modelo que utiliza da tratativa de dados da Data Science e vem sendo estudado e aprimorado desde 2014.

5.2 ECOMERCIE

A ciência de dados é um campo de estudos que busca obter conhecimento e informação por meio da tratativa de dados estruturados ou não, a mesma faz uso de modelos estatísticos e científicos com o propósito de gerar informações e solucionar problemas.

As vendas online na internet ou varejo online se caracterizam pela comercialização de serviços ou produtos por meio da internet, como o foco no consumidor final. As compras e vendas pela internet tornaram-se ainda mais corriqueiras nos últimos anos devido ao período pandêmico, segundo a Associação Brasileira de Comércio Eletrônico (ABCOMM), que afirma que o comércio eletrônico ou E-commerce são englobados pelo varejo online.

No E-commerce há uma grande produção de dados, consequentemente tem sido visto como tendo grande potencial para aplicação das técnicas de coleta de dados disponibilizados pela ciência de dados, sendo assim possível estruturar práticas dos usuários a partir do interesse até a compra em determinado site; todos esses dados são utilizados para auxiliar na potencialização das vendas, planejar publicidade e etc.

De acordo com Mata(2021), "[...]as empresas captam informações para auxílio em tomadas de decisão eficazes, sendo feitas análises descritivas, diagnósticas e prescritivas". Esse grande volume de dados são etapas da obtenção, incluindo a coleta, a exploração e a verificação da qualidade dos dados obtidos. O E-commerce disponibiliza imensa variedade de produtos 24 h por dia em qualquer lugar, com isso cada usuário gera dados em uma simples pesquisa.

Levando em consideração que os dados obtidos através da Data Science no e-commerce ajuda os vendedores a construir uma imagem de seus consumidores. A exemplo temos a Airbnb, que a ciência de dados ajudou-a a renovar completamente sua função de pesquisa, melhorando assim a experiência do usuário.(ILUMEO,2018). Um outro exemplo clássico da utilização dos dados aplicada ao e-commerce são as redes

sociais do Facebook que tornaram-se polos de vendas online, e os vários dados gerados por cada usuário permite ao algoritmo melhor alcance dos comerciais.

5.3 SAÚDE

Estudos recentes mostram que, “um paciente crônico gera em torno de 80 Mbytes por ano em dados, entre imagens e outras informações médicas e testes clínicos”. (NETTO, 2021). Notoriamente todos esses dados têm um valor clínico e operacional, todavia o setor de saúde brasileiro utiliza pouco esses dados, precisando ainda que haja um registro dessas informações investindo na análise de dados e determinando esses dados como essenciais.

Netto(2021) descreve a utilização do Big Data na saúde como:

> *“Big data na área de saúde se refere a conjuntos de dados eletrônicos de saúde tão grandes e complexos que são difíceis de gerenciar com software e/ou hardware tradicionais; nem podem ser facilmente manipulados com ferramentas e métodos tradicionais ou comuns de gerenciamento de dados”.*

As fontes de dados aplicadas a Big Data Analytics, podemos citar: administrativo, biomarcadores, biométricos, registros clínicos, registros eletrônicos de saúde, internet, relatório de pacientes e imagens médicas. Assim a aplicação da técnicas de ciência de dados na saúde possibilita ampla vertente para soluções envolvendo informática, onde a sugestão é lidar com dados estruturados e não estruturados de fontes diversas.

Na saúde a ferramenta permite o cruzamento de dados, facilitando avaliações e análises precisas, como o Big Data engloba dados estruturados e não estruturados, a ciência de dados contribui para a adesão do público a telemedicina, colaborando para estudos, monitoramento e intervenção. Segundo estudo da IBM(2020), projeta-se cerca de 25 mil petabytes de informações para o setor de saúde, essas são utilizadas no rastreamento em saúde, assistência ao paciente, prevenção a falhas, reduções de custos e auxílio nas pesquisas, simplificando a busca por informações.

5.3.1 Pandemia

No ano de 2019 o mundo foi assolado por uma pandemia que mudou a realidade da humanidade, com isso a ciência de dados é empregada visando projetar elementos como dados provenientes do vírus intitulado COVID-19 apelido para Sars-CoV-2; dados como números de casos, número de mortos, período de pico de infectados, quantidade de infectados e dentre outros.

Em concordância com Silva et al. (2020), todos os volumosos dados gerados produzidos pela pandemia são características para a ciência de dados, por meio dela modelos de predições de dados acerca do Sars-CoV-2 foram desenvolvidos, para mostrar

o impacto pandemia, que vão desde redução da população ao impacto na economia mundial.

Considerando que a estatística descritiva aplica técnicas para descrever e organizar os dados com média e variância. “[...] suas técnicas empregadas na ciência de dados permitiu a criação de repositórios de pesquisas e *datasets* que são utilizados diariamente para a criação, desenvolvimento e implantação de gráficos que permitem a elaboração estratégica de combate ao vírus.” (SILVA et al., 2020).

Com isso o ministério da saúde junto a fundações e instituições de pesquisa acadêmicas e científicas disponibilizam conjunto de dados que formam a plataforma openDataSUS. “A plataforma disponibiliza informações que podem servir para subsidiar análises objetivas da situação sanitária, tomadas de decisão baseada em evidências e elaboração de programas de ações de saúde”.(BRASIL,2022). Todavia nele também são encontrados informações sobre assistência em saúde da população, os cadastros, redes hospitalares e ambulatórios, bem como informações demográficas e informações socioeconômicas. O openDataSUS tomou destaque durante a pandemia disponibilizando *datasets* de registros de cunho sanitário, como distribuição de equipamentos, ocupações hospitalares, dados sobre a Síndrome respiratória aguda grave, vacinômetro dentre outros bancos de dados que auxiliam na elaboração de programas de saúde públicas, acesso a serviços e tomada de decisões referente ao combate a Sars-CoV-2.

Como outro exemplo de transparência de dados no período pandémico temos o integraSUS Analytics que é uma ferramenta com a qual é mantido pela secretaria de saúde do estado do Ceará, onde pesquisadores, profissionais e estudantes têm acesso ao cenário da pandemia no estado, bem como tudo sobre a gestão de saúde do estado; a plataforma oferece datasets com códigos e modelos, utilizando ferramentas de Inteligência Artificial para processar e armazenar dados.

6. RESULTADOS E DISCUSSÕES

Com base nos artigos estudados, as tecnologias versadas neste artigo encontram sentido principalmente pelo espantoso crescimento dos volumes de dados, em contrapartida, temos o surgimento da necessidade de tratar esses dados, tornando-se imprescindível a aplicação de meios tecnológicos capazes de manipulá-los; desencadeado pela massiva quantidade de dados, que muitas vezes são desperdiçados se não forem tratados e utilizados. Essa demanda pode ser suprida pelo Big Data com a ciência de dados, para auxiliar na transformação de dados.

Perante a análise, os resultados mostram que ao mesmo tempo em que temos um grande volume de dados gerados em várias áreas do nosso dia a dia, em suma, boa parte desses dados não recebem o devido tratamento, embora as tecnologias emergidas pela ciência de dados tenham potencial suficiente para lidar com megadados.

A partir dessa discussão empreendida, pode-se assim dizer que essas tecnologias exigem cada vez mais capacidade de processamento para a execução dessas tarefas. Os autores Pacheco e Disconzi(2019) afirmam que a “[...]ciência de dados fazendo uso da tecnologia e arquitetura do Big Data, gera conhecimento a fim de potencializar negócios, visando valores competitivos de mercado, políticos e auxiliando a tomada de decisões”; com base nas informações extraída dos megadados .Tais operações por sua vez tornam-

se impossíveis de serem realizadas por máquinas comuns, com isso , “[...] a infraestrutura de Big Data deve suportar o gerenciamento, providência e a curadoria dos dados e seus megadados.” (RAUDENBERG E CARMO, 2019).

Os autores Raudenberg e Carmo afirmam ainda que: “A Ciência de dados por sua vez é considerada a camada dos dados métodos, voltados à tomada de decisões”. Ela em si é a camada de transformação de dados, dados esses advindos do Big Data, com base nisso a ciência de dados com o auxílio do profissional cientista de dados é responsável pelo ciclo evolutivo dos dados que vai até o conhecimento.

DADOS -> INFORMAÇÕES ->CONHECIMENTO

Deste modo essa pesquisa permitiu pensar em como a ciência de dados e suas competências incluindo Hardware e Software, bem como todas as suas plataformas que produzem soluções e dados de ponta, ao mesmo tempo em que ajudam a comunidade, tornando-se vital para o sistema e levando em consideração estarmos em uma era tudo cadas vez mais digitalizado.

7. CONSIDERAÇÕES FINAIS

Os avanços da tecnologia da informação apresentam grande utilidade no dia a dia da humanidade, auxiliando de maneira significativa o armazenamento e processamento de dados. A ciência de dados tornou-se parte vital de uma série de atividades que exigem grande poder de processamento, por ser uma ciência que utiliza a mineração e análise de dados objetivando estruturar, capturar, transformar, gerar e analisar esses dados.

Tendo em vista que a pesquisa foi desenvolvida mostrando as aplicações da ciência de dados como método de transformação de dados em conhecimento. Os resultados apresentados na análise contribuem tanto para pesquisas acadêmicas descritivas quanto qualitativas.

Na concepção da ciência de dados, os elementos estruturais abordados em volta do Big Data mostram uma nova percepção para a heterogeneidade dos dados, levando em consideração que os dados em grande escala tornaram-se desafio para as tecnologias convencionais, sendo assim necessário utilizar técnicas da ciência de dados. Diante da pesquisa desenvolvida pode-se concluir que são recursos importantes,cujo, a partir dos dados já existentes e considerando a interdisciplinaridade supracitada a transformar evolutiva de Dados →Informação →Conhecimento, dissertando a diferença entre eles e a complementaridade dos conceitos Big Data e Ciência de Dados, sendo possível previsões por meio desses dados que auxiliam na tomada de decisões.

Apesar de ficar evidente o caráter exploratório e descritivo predominantemente teórico do artigo, por ser um tema em evolução, destaca-se a importância de estudos teóricos sobre essa linha de pesquisa; onde contribui para um posicionamento sobre o tema proposto, contudo, para trabalhos futuros pretende-se explorar *datasets* aplicando

algoritmos que possam realizar a tratativas de dados gerando a informação propriamente dita, usufruindo dos mecanismos disponibilizados pela ciência de dados.

6. REFERENCES

AGRAWAL, Himanshu; CHAFLE, Girish; GOYAL, Sunil; MITTAL, Sumit; MUKHERJEA, Sougata. An Enhanced Extract-Transform-Load System for Migrating Data in Telecom Billing. **2008 IEEE 24th International Conference on Data Engineering**, [*s. l.*], 2008.

AMARAL, Fernando. **Introdução à ciência de dados**: mineração de dados e big data. 1. ed. rev. [*S. l.*]: Starlin Altas book, 2016.

BANSAL, Srividya K. Towards a Semantic Extract-Transform-Load (ETL) Framework for Big Data Integration. **2014 IEEE International Congress on Big Data**, [*s. l.*], 2014.

BRASIL, Governo do. **OpenDataSus**. [*S. l.*], 2022. Disponível em: https://opendatasus.saude.gov.br/. Acesso em: 12 jan. 2022.

CARDOSO, Pedro Henrique. **Ciência de dados aplicada a dados governamentais abertos sob a ótica da Ciência da Informação**. Dissertação de Mestrado, [s. l.], 2019.

CEARÁ, Governo do Estado do. **IntegraSus**. [*S. l.*], 2022. Disponível em: https://integrasus.saude.ce.gov.br/#/home. Acesso em: 04 mar. 2022.

CETAX, BLOG. DATA ANALYTICS, BIG DATA, DATA SCIENCE: Artigos, materiais e tutoriais de Business Intelligence, Big Data, Data Warehouse e ETL. *In*: **O que é ETL – Extract Transform Load?**. [*S. l.*], 2020. Disponível em: https://www.cetax.com.br/blog/etl-extract-transform-load/. Acesso em: 20 mar. 2022

FERREIRA, J. (2013) **Big data in education: The five types that matter**. Disponível em http://www.knewton.com/blog/knewton/from-jose/2013/07/18/big-data-in-education

FERREIRA, João; MIRANDA, Miguel; ABELHA , António; MACHADO, José. O Processo ETL em Sistemas Data Warehouse. **INForum 2010 - II Simposio de Informatica**, [*s. l.*], 2010.

GODOI, Douglas. Big Data: Tudo o que você precisa saber. *In*: **Big Data: Tudo o que você precisa saber**. [*S. l.*], 7 nov. 2020. Disponível em: https://www.cetax.com.br/big-data-tudo-o-que-voce-precisa-saber/. Acesso em: 14 nov. 2021.

IBM BRAZIL. *In*: **IBM lança novas tecnologias com IA para ajudar comunidade de saúde e pesquisa a acelerar descoberta de insights e tratamentos médicos para COVID-19**.[*S. l.*], 7 abril. 202. Disponível em: https://www.ibm.com/blogs/ibm-comunica/ibm-lanca-novas-tecnologias-com-ia-para-ajudar-comunidade-de-saude-e-pesquisa-a-acelerar-descoberta-de-insights-e-tratamentos-medicos-para-covid-19/ Acesso em: 04 mar. 2022.

ILUMEO. *In*:**Como funciona o departamento de Data Science do Airbnb?** [*S. l.*],fev. 2018Disponível em: https://ilumeo.com.br/todos-posts/2018/11/5/como-funciona-o-departamento-de-data-science-do-airbnb. Acesso em: 24 fev. 2022.

JUNIOR, Jose Carlos Da Silva Freitas; MAÇADA, Antonio Carlos Gastaud; OLIVEIRA, Mirian; BRINKHUES, Rafael Alfonso. BIG DATA E GESTÃO DO CONHECIMENTO: DEFINIÇÕES E DIRECIONAMENTOS DE PESQUISA. **REVISTA ALCANCE**, [*S. l.*], p. 04-22, 1 nov. 2016.

MATA, KESLEY BRENNER DA COSTA. E-COMMERCE: ANÁLISE DE DADOS SOBRE O COMÉRCIO ELETRÔNICO NO BRASIL. **Escola de Ciências Exatas e da Computação, da Pontifícia Universidade Católica de Goiás**, [*s. l.*], 2021.

MOREIRA, Cristiano; BEIRA, Joana Carlos; OLIVEIRA, Marlene. UM OLHAR DOS ESTUDANTES DO CURSO DE BIBLIOTECONOMIA ACERCA DO QUE SÃO DADOS, INFORMAÇÕES E CONHECIMENTOS. Http://www.uel.br/revistas/informacao/, [s. l.], p. 484 – 508, 3 jan. 2022.

NETTO, Antonio Valerio. **CIÊNCIA DE DADOS EM SAÚDE: CONTRIBUIÇÕES E TENDÊNCIAS PARA APLICAÇÕES**. Revista Saúde.Com, [s. l.], 2021.

ORACLE BRAZIL. *In*: **O que é Ciência de Dados?**. [*S. l.*], 2 fev. 2021. Disponível em: https://www.oracle.com/br/data-science/what-is-data-science/. Acesso em: 29 out. 2021.

PACHECO, Bornieque Brister Marcovit; DISCONZI, Marcelo Salton. Ciência de Dados: Enfoque no Desafio do Processamento. **Res., Soc. Dev. 2019; 8(11):e128111444 ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v8i11.1444**, [*s. l.*], 23 ago. 2019.

PATRICIO, Thiago Seti; MAGNONI, Maria da Graça Mello. **Mineração de dados e big data na educação.** Revista GEMInIS, São Carlos, UFSCar, v. 9, n. 1, pp57-75, jan. / abr. 2018. Enviado em: 01 de abril de 2018 / Aceito em: 05 de junho de 2018

Pimenta, C., Ribeiro, R., Sá, V., Belfo, F.P.: Fatores que Influenciam o Sucesso Escolar das Licenciaturas numa Instituição de Ensino Superior Portuguesa. **In: 18ª Conferência da Associação Portuguesa de Sistemas de Informação (CAPSI 2018)** Associação Portuguesa de Sistemas de Informação: Santarém, Portugal (2018)

PRODANOV, Cleber Cristiano; FREITAS, Enanir Cesar de. **METODOLOGIA DO TRABALHO CIENTÍFICO**: Métodos e técnicas da pesquisa e do trabalho acadêmico. [*S. l.*: *s. n.*], 2013.

RAUTENBERG , Sandro; CARMO , Paulo Ricardo V. BIG DATA E CIÊNCIA DE DADOS: COMPLEMENTARIDADE CONCEITUAL NO PROCESSO DE TOMADA DE DECISÃO. **Brazilian Journal of Information Studies: Research Trends**, [*s. l.*], 2019.

RODRIGUES, Adriana Alves; DIAS, Guilherme Ataíde. ESTUDOS SOBRE VISUALIZAÇÃO DE DADOS CIENTÍFICOS NO CONTEXTO DA DATA SCIENCE E DO BIG DATA. **Pesq. Bras. em Ci. da Inf. e Bib.**, João Pessoa, p. 219-228, 12 set. 2017.

RODRIGUES, Adriana Alves; NÓBREGA, Emeide; DIAS, Guilherme Ataíde. **DESAFIOS DA GESTÃO DE DADOS NA ERA DO BIG DATA: PERSPECTIVAS PROFISSIONAIS. XVIII ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2017**, [s. l.], 27 out. 2017.

SANTOS-D'AMORIM, Karen; CRUZ, Rúbia Wanessa dos Reis; SILVA, Marcela Lino da; CORREIA, Anna Elizabeth Galvão Coutinho. DOS DADOS AO CONHECIMENTO: TENDÊNCIAS DA PRODUÇÃO CIENTÍFICA SOBRE BIG

DATA NA CIÊNCIA DA INFORMAÇÃO NO BRASIL: Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, [*s. l.*], 2020.

SCAICO, Pasqueline Dantas; QUEIROZ, , Ruy José G. B. de; SCAICO, Alexandre. O conceito big data na educação. **3º Congresso Brasileiro de Informática na Educação (CBIE 2014)** , [*s. l.*], 26 jul. 2014.

SILVA, Gabriel Di iorio; STROELE, Victor; DANTAS, Mario; MENDONÇA, Fabricio. Ciência de Dados Aplicada á COVID-19: Os Dados Implícitos em Meio a Pandemia. **Departamento de Ciência da Computação - Universidade Federal de Juiz de Fora (UFJF)**, [*s. l.*], 2020.

SOMASUNDARAM, G.; SHRIVASTAVA, A. Armazenamento e gerenciamento de informações: como armazenar, gerenciar e proteger informações digitais. [s. l.]: Bookman Editora, 2011.