INTERFACE TECNOLÓGICA

DATA LAKE: Suas Funcionalidades e Aplicações

DATA LAKE: Its Functionalities and Applications

Denis Henrique Pazini da Silva – hdenis32@gmail.com
Faculdade de Tecnologia de Catanduva (Fatec) – Catanduva – São Paulo – Brasil

Miriam Francieli Paes – franciellipaes@gmail.com
Faculdade de Tecnologia de Catanduva (Fatec) – Catanduva – São Paulo – Brasil

Eder Carlos Salazar Sotto – eder.sotto@fatec.sp.gov.br
Faculdade de Tecnologia de Catanduva (Fatec) – Catanduva – São Paulo – Brasil

Liriane Soares de Araújo – lirianearaujo@hotmail.com
Faculdade de Tecnologia de Catanduva (Fatec) – Catanduva – São Paulo – Brasil

DOI: 10.31510/infa.v21i1.1960
Data de submissão: 15/04/2024
Data do aceite: 10/03/2024
Data da publicação: 20/06/2024

RESUMO

Em uma era onde a voracidade por dados é insaciável, o conceito de *Data Lake* emerge como um reservatório robusto e inovador para a retenção e análise de informações. Inspirado por pesquisas pioneiras de autores como James Dixon em seu blog em 2010, fundador da Pentaho, e Thomas H. Davenport, renomado especialista em análise de dados, o *Data Lake* se destaca como uma abordagem disruptiva no cenário do gerenciamento de dados. Este artigo tem como objetivo explorar esse conceito, examinando a arquitetura flexível e expansível proposta por Dixon e as principais abordagens tradicionais ao preservar a integridade dos dados brutos independente de sua fonte ou formato, tudo em um único local, considerando a escassez de literatura ainda existente por ser um assunto novo. Ao abordar sobre o Lago de Dados (*Data Lake*), pretende-se abranger não apenas sua estrutura, mas também suas implicações que esse ambiente de armazenamento de dados brutos pode ter nas pesquisas científicas, mostrando ainda o que é o *Data Lake*, a fim de contribuir para a compreensão desse conceito. Além disso, apresenta-se dois *cases* de empresas em que o *Data Lake* foi utilizado a fim de demonstrar sua aplicabilidade. Espera-se contribuir para a literatura ao enfocar as características do *Data Lake* e seu impacto positivo nas empresas, que inclui eficiência, unificação de dados e aumento da lucratividade.

Palavras-chave: Data Lake. Banco de Dados. Dados Brutos.

ABSTRACT

In an age where the voracity for data is insatiable, the concept of the Data Lake emerges as a robust and innovative reservoir for the retention and analysis of information. Inspired by pioneering research by authors such as James Dixon on his blog in 2010, founder of Pentaho,

and Thomas H. Davenport, renowned data analytics expert, the Data Lake stands out as a disruptive approach in the data management landscape. This article aims to explore this concept, examining the flexible and scalable architecture proposed by Dixon and the main traditional approaches to preserving the integrity of raw data regardless of its source or format all in a single place, Considering the scarcity of literature that still exists because it is a new subject. When addressing the Data Lake, it is intended to cover not only its structure, but also its implications that this raw data storage environment can have on scientific research, also showing what the Data Lake is, in order to contribute to the understanding of this concept. In addition, two cases of companies in which the Data lake has been used are presented in order to demonstrate its applicabiblity. We hope to contribute to the literature by focusing on the characteristics of the Data Lake and its positive impact on companies, which includes efficiency, data unification and increased profitability.

Keywords: Data Lake. Database. Raw Data.

1 INTRODUÇÃO

As organizações procuram constantemente formas inovadoras de aproveitar os seus dados para a tomada de decisões inteligentes. Com o passar do tempo, a quantidade de dados armazenados em empresas para obter informações úteis para melhorar negócios e tomadas de decisões vem crescendo muito. O *Data Lake* é um repositório onde se guarda todos esses dados, sejam eles tratados ou não, geralmente o armazenamento desses dados ocorre de forma bruta, sem estrutura, processamento ou análise nenhuma. As empresas têm como ideia manter esses dados dentro de seus bancos de dados, sejam eles dados úteis ou não, e que possam ou não ser requisitados em algum momento.

Conforme Singh (2019), pode-se definir o *Data Lake* como um repositório ou local de armazenamento desses dados. A necessidade de obter melhor desempenho e velocidade na captação de dados fez com que surgisse o *Data Lake* para suprir essa demanda das empresas.

O conceito de *Data Lake* surgiu como componente principal na arquitetura de dados, oferecendo flexibilidade para armazenamento, processamento e análise de dados. O armazenamento estruturado e a organização de dados são essenciais para evitar acúmulo de dados não gerenciáveis.

Khine (2017), mostra que o *Data Lake* representa uma mudança na gestão e utilização de dados. Eles são projetados para armazenar grandes quantidades de dados brutos, não estruturados ou semiestruturados, permitindo que consolidem dados de diversas fontes, incluindo bancos de dados estruturados, mídias sociais, dispositivos IoT, entre outros.

Um *Data Lake* atua como um repositório centralizado que democratiza o acesso e a análise de dados, tornando-os um ativo inestimável para as organizações (CUTTING & CAFARELLA, 2015).

Como Cutting e Cafarella (2015) afirmam, um *Data Lake* atua como um repositório centralizado que democratiza o acesso e a análise de dados, tornando-os um ativo inestimável para as organizações (CUTTING & CAFARELLA, 2015).

Em seu artigo "*Unlocking Business Valuewith Data Lakes*", James Dixon discute como os *Data Lakes* facilitam análises avançadas, aprendizado de máquina e inteligência artificial, fornecendo uma visão abrangente e unificada dos dados (DIXON, 2010). Essa visão unificada é essencial para organizações que buscam descobrir padrões, correlações e tendências ocultas em seus dados, levando, em última análise, a decisões de negócios mais inteligentes.

Os *Data Lakes* também capacitam as organizações a realizarem análises de dados em tempo real, permitindo-lhes responder prontamente às mudanças na dinâmica do mercado. No seu livro "*Data Lakes in Action",* Alex Gorelik e Ted Malaska enfatizam a importância da análise em tempo real, mostrando como esta pode melhorar os processos de tomada de decisão e permitir respostas proativas às mudanças do mercado (GORELIK & MALASKA, 2017).

Considerando toda essa definição, este artigo possui como objetivo explorar a literatura sobre esse conceito por meio de uma análise bibliográfica e documental descritiva e exploratória a fim de auxiliar na literatura, sendo também realizada uma análise comparativa dos resultados da aplicação do *Data Lake* em empresas.

2 FUNDAMENTAÇÃO TEÓRICA

Conforme visto no artigo publicado pelo Ajit Singh (2019), um *Data Lake* é um repositório de dados centralizado que pode armazenar uma grande quantidade de dados estruturados ou semiestruturados, oferecendo um armazenamento escalável para lidar com uma quantidade crescente de dados, proporcionando agilidade para fornecer insights mais rápidos.

Fang (2015), esclarece que essa metodologia vai possibilitar que um repositório de dados massivo com baixo custo em tecnologias, tenha uma melhora na captação, refinamento, arquivamento e a exploração de dados brutos. Para Miloslavskaya e Tolstoy (2016), o *Data Lake* é um armazenamento compacto de uma grande quantidade de dados em seu formato real, que é inserido sem comprometer a estrutura de dados.

Singh (2019) ainda mostra que o *Data Lake* tem uma arquitetura plana para

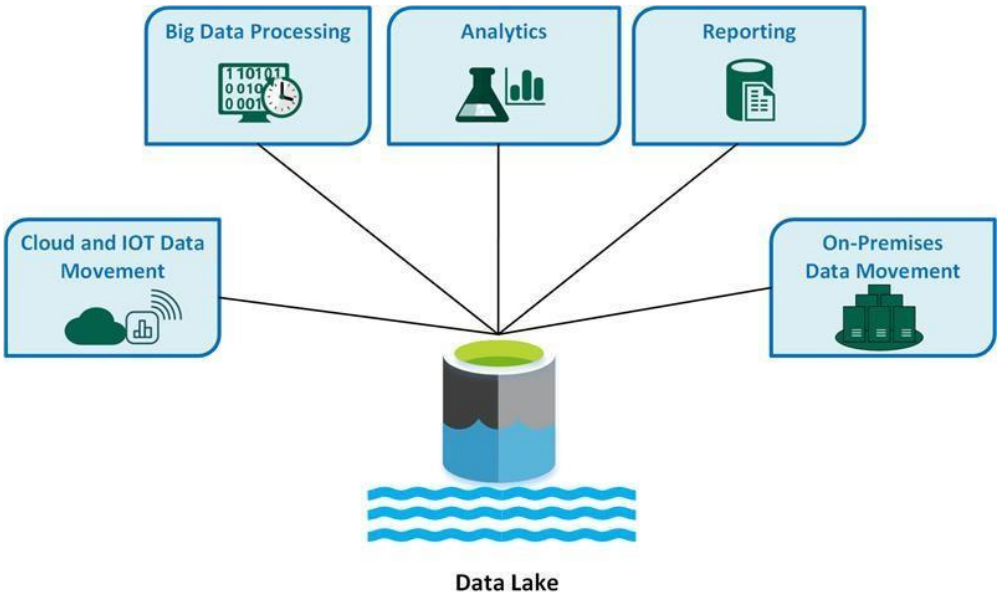
armazenamento de dados. Como afirmam no livro "*Data Lakes: The Definitive Guide*", um *Data Lake* atua como um repositório centralizado que democratiza o acesso e a análise de dados, tornando-os um ativo inestimável para as organizações (CUTTING & CAFARELLA, 2015).

Em seu artigo "*Unlocking Business Valuewith Data Lakes*", James Dixon discute como os *Data Lakes* facilitam análises avançadas, aprendizado de máquina e inteligência artificial, fornecendo uma visão abrangente e unificada dos dados. Essa visão unificada é essencial para organizações que buscam descobrir padrões, correlações e tendências ocultas em seus dados, levando, em última análise, a decisões de negócios mais inteligentes (DIXON, 2010).

Os *Data Lakes* também capacitam as organizações a realizarem análises de dados em tempo real, permitindo-lhes responder prontamente às mudanças na dinâmica do mercado. No seu livro "*Data Lakes in Action*", Alex Gorelik e Ted Malaska enfatizam a importância da análise em tempo real, mostrando como esta pode melhorar os processos de tomada de decisão e permitir respostas proativas às mudanças do mercado (GORELIK E MALASKA, 2017).

No livro "*Data Lake Architecture*", Inmon e Linstedt (2017) argumentam que "um *Data Lake* bem projetado é um componente fundamental do gerenciamento de dados moderno" (p. 24). Ressaltando assim, a importância do design adequado do *Data Lake* em ambientes de dados. Na Figura 1 pode-se entender que *Data Lake* também pode ser considerado como um lago de dados, onde ficam todos os dados brutos coletados e, posteriormente, algum *software* vai fazer a busca e processamento desses dados, estruturando-os da maneira que melhor atenda à empresa.

Figura 1 – Exemplo de Estrutura de um *Data Lake*.



Fonte: *Prazad et al.* (2023).

Na Figura 1 é possível verificar que os dados vêm de todos os lugares acessíveis e são armazenados em seu estado bruto, ou seja, no seu formato original, podendo utilizar o Big Data, realizar movimentações no sistema local ou em nuvem, gerando relatórios e análises.

Singh (2001), afirma que um *Data Warehouse* integra os dados de uma empresa em um único repositório, facilitando consultas, análises e a geração de relatórios, ele é um ambiente de suporte à decisão utilizando dados de diversas fontes. Ele ainda afirma que com o seu uso é possível identificar vários benefícios, economia de tempo e produtividade elevada.

A SAP mostra que um *Data Warehouse* é um sistema de armazenamento no qual armazena grandes volumes de dados de várias fontes diferentes, com a finalidade de alimentar relatórios, funções analíticas e *business intelligence* (BI), auxiliando na tomada de decisões, baseadas em dados, ele armazena dados atuais e históricos em um só local.

Ainda baseado nos dados apresentados pela SAP, Data *Warehouse* gerencia dados estruturados e não estruturados, como vídeos, arquivos de imagem e demais dados, sem esse armazenamento de dados, torna-se muito difícil combinar dados e garantir o formato certo para análises.

A *AWS Amazon*, mostra que um *Data Warehouse*, é um repositório central de informações que podem ser analisadas para tomar decisões mais adequadas, esses dados são captados de diversas fontes.

3 PROCEDIMENTOS METODOLÓGICOS

A metodologia descritiva desenvolvida neste trabalho apresenta os conceitos básicos para compreensão do tema abordado, junto com métodos e funcionalidades, e benefícios.

Gil (1991), mostra que pesquisa descritiva têm como objetivo características determinadas, tendo com uma das características principais a utilização de técnicas padronizadas de coleta de dados, tais como questionário e observação sistemática. Ele também nos explica que a pesquisa exploratória tem como objetivo principal o aprimoramento de ideias, com um planejamento flexível, possibilitando variados aspectos, como um levantamento bibliográfico, entrevistas e análises de exemplos e situações.

Falar sobre *Data Lake* nos mostra a evolução da tecnologia e suas implicações comparadas às abordagens mais antigas, e compreende-se que é importante para ter resultados cada vez mais positivos para os negócios, seja em pequenas, medias ou grandes empresas.

Discutir sobre o assunto ajuda na análise de dados, a tomada de decisão, inovações e destacar e entender nossas vantagens e desvantagens.

Além da pesquisa bibliográfica é realizada a pesquisa de levantamento de dois casos na Internet, em que se apresenta a implantação do *Data Lake*, mostrando algumas funcionalidades e desafios.

4 APLICAÇÕES E FUNCIONAMENTO DO *DATA LAKE*

Em um mundo onde a constante de metadados cresce a cada segundo e cada dado desse pode ser útil em algum momento para empresa, os Bancos de Dados Relacionais (*Data WareHouse*) são muito uteis para dados relacionais, porém é necessário ir além, para que seja possível armazenar todos os dados que uma organização produz ao logo do tempo, e para isso foi criado o *Data Lake*, onde é possível armazenar todos os dados em seu estado bruto. O *Data Lake* veio para auxiliar as organizações no gerenciamento de grande volume de dados e suas diversas fontes.

Para a realização da implantação de um *Data Lake* existem várias etapas. É preciso levar em consideração todas elas, desde a concepção até a implantação, além da necessidade de manutenção constante.

É importante discutir a utilização do *Data Lake* com o *Data Warehouse*. Ambas as soluções podem funcionar em conjunto, permitindo a obtenção de informações precisas. Em um *Data Warehouse*, os dados são armazenados já estruturados e estáticos, de modo que sua criação já é toda estruturada desde o início. Já no *Data Lake* os dados são armazenados de forma bruta, onde todos os dados que obtidos, não importa seu tipo, sejam arquivos doc, pdf, vídeo, imagens e outros em seus estados mais crus e guarda-os no *Data Lake* para posteriormente tratá-los e serem utilizados quando necessário. Com isso, não é preciso estruturar e analisar cada dado desde o início, e sim somente quando for necessário, sem o risco de perder qualquer detalhe destes dados, o que poderia ocorrer em caso de tratamento. Entretanto, os dados extraídos e tratados obtidos originalmente no *Data Lake* podem ser salvos em um *Data Warehouse* permanentemente após seu uso. Este é um exemplo de uso conjunto das duas tecnologias.

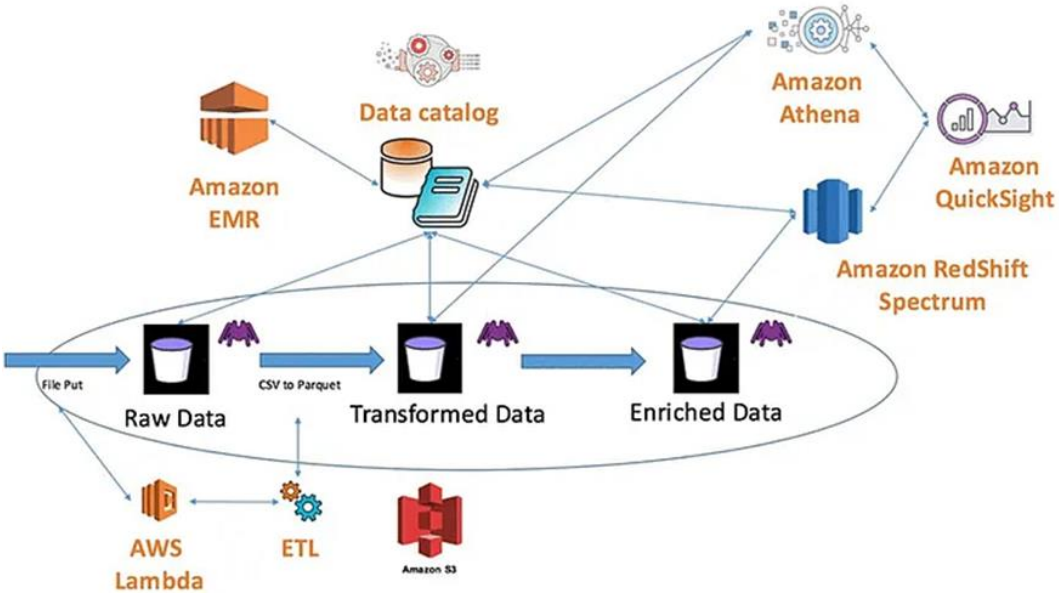
4.1 Implementação De Um *Data Lake* na *Amazon AWS*

A AWS utilizou os principais serviços utilizados para implantação do *Data Lake*, que geralmente são *Amazon S3, AWS Glue, AWS Lake Formation e Amazon EMR.*

A primeira etapa foi a criação dos *buckets*. O primeiro é responsável por armazenar todos os dados em seu estado original. O segundo *bucket* é para onde vão os dados depois que forem processados, tratados e convertidos para um melhor formato (como o *Apache Parquet*). Por último, o *bucket* final é para onde os dados vão depois de prontos, para que possam ser utilizados para a produção de informações.

A Figura 2 apresenta uma pequena demonstração das etapas dos dados para a criação de *buckets*, que são os recipientes básicos de armazenamento de dados.

Figura 2 – Criação dos Buckets.



Fonte: Amazon (2021).

Foram criados dois *buckets*, um para os dados crus e outro para os processados, o terceiro tem como função ser um banco de dados responsável por disponibilizar os dados prontos e enriquecidos.

Vale ressaltar que é preciso criar uma política bem definida do ciclo de vida para esses dados. Isso é crucial quando existe uma numerosa massa de dados, fazendo com que os dados menos utilizados sejam salvos de uma maneira menos custosa, como Lourenço (2021) mostra em seu artigo "Como criamos nosso *Data Lake* utilizando a AWS".

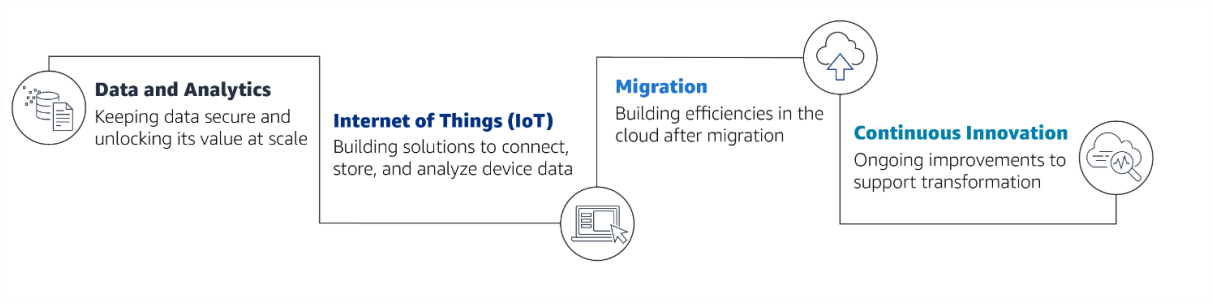
Como havia um banco de dados onde seria necessário transferir todo o seu conteúdo para o *Data Lake*, utilizaram todos esses dados históricos já existentes na empresa, mas para

isso precisaram utilizar ferramentas como o *Blueprint do Data Lake Information*, que é um *workflow* pré-escrito para facilitar a obtenção dos dados de outras fontes. No caso do banco de dados já existente ou dados de *CloudTrail* para os *buckets* específicos, a utilização do *CloudTrail* foi necessária para gerenciamento de um *Data Lake*, o qual permite capturar, armazenar, acessar e analisar a atividade de usuários e APIs na AWS, permitindo maior segurança e auditoria. Os dados foram inseridos no *bucket* destinado a dados processados, porém antes foi criada uma cópia de seu estado original no *bucket* destinado a dados crus, permitindo o *Data Lake* ser reconstruído, quando necessário.

4.1.1 Estudo De Caso – Coca-Cola *Company* Na *AWS*

Com base no caso de uso apresentado pela *Amazon Web Services* (2021), que trouxe a Coca-Cola *Company*, uma empresa global de bebidas que atua em mais de 200 países e territórios. Após duas décadas operando em datacenters próprios, a Coca-Cola aderiu à AWS (*Amazon Web Services*) devido a um comercial no Super Bowl, onde houve um aumento e acúmulo de dados que sobrecarregou seus servidores, sendo este um dos fatores que fez com que a empresa passasse a admitir essa migração. Na figura 3 é possível verificar a jornada para a nuvem que a Coca-Cola utilizou na AWS.

Figura 3 – Jornada para a nuvem.



Fonte: AWS (2021).

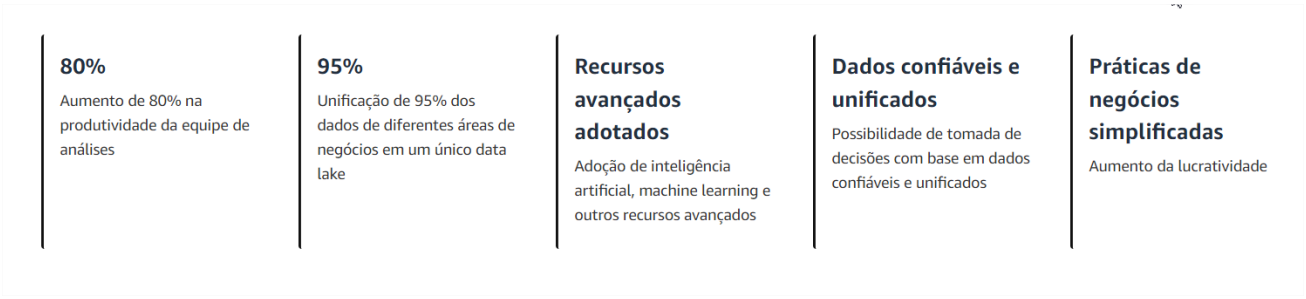
Através da análise da situação, constatou-se uma eficiente gestão da migração, garantindo a segurança dos dados e implementando soluções integradas de conectividade, análise e armazenamento. Este processo promoveu uma melhoria contínua e eficaz pós-migração.

De acordo com informações obtidas junto à AWS, a Coca-Cola Andina optou por adotar o *Data Lake*, visando impulsionar o crescimento de seus empreendimentos e assegurar uma experiência superior para seus clientes.

O estudo de caso realizado pela *Amazon* em 2021, identificou que após a implementação do *Data Lake*, observou-se uma significativa transformação. Valendo-se dos recursos de armazenamento, bancos de dados, computação e análise proporcionados pela tecnologia AWS, a Coca-Cola Andina conseguiu elevar em 80% a eficiência de sua equipe de análise. Isso permitiu que tanto a empresa quanto seus clientes embasassem suas decisões em dados confiáveis, fomentando o crescimento conjunto de todo o ecossistema, preservando a vantagem competitiva e incrementando a receita da organização.

Valderrama (2021) afirmou que a AWS se mostrou a solução em nuvem capaz de atender a todas as expectativas estabelecidas para o *Data Lake* da empresa, enfatizando a necessidade de uma arquitetura que incorporasse uma plataforma como serviço (PaaS), viabilizando o desenvolvimento e a desmontagem ágeis e econômicos das soluções. Ele expressou sua satisfação com a decisão, destacando a cultura de aprendizado prático da empresa. Na Figura 4 é possível identificar algumas melhorias realizadas com o *Data Lake.*

Figura 4 – Lista de Melhorias.



Fonte: *AWS – Coca-Cola* (2021).

Dessa forma, a Coca-Cola Andina possibilitou que tanto a empresa quanto seus clientes tenham uma tomada de decisão com base em dados confiáveis, obtendo assim um crescimento conjunto e com uma vasta vantagem competitiva e aumento significativo de lucros.

4.2 Estudo De Caso – Neighborly

De acordo com o estudo de caso realizado pela *Ipsense* na Neighborly, empresa líder em serviços domésticos especializados, a análise de dados da empresa apresentava desafios

significativos, principalmente a dependência do uso de um sistema baseado em *SQL Server*, no qual possui ferramentas antiquadas e não escaláveis. Diante do grande volume de dados, ficou claro que a empresa precisava de uma solução alternativa para gestão e análise eficiente de suas informações.

A implantação da solução de *Data Lake* em nuvem pela *IPsense* foi fundamental para que a Neighborly superasse os desafios enfrentados e aprimorasse consideravelmente sua capacidade de análise de dados. Essa solução possibilitou à empresa migrar seus dados para a nuvem, onde poderiam ser armazenados e gerenciados com maior eficiência, resultando em uma análise de dados mais ágil e precisa. Além disso, o *Data Lake* em nuvem permitiu à Neighborly dimensionar sua infraestrutura de dados de acordo com as demandas de negócios em constante evolução, o que possibilitou à empresa se adaptar rapidamente às exigências dos clientes e do mercado. Como plataforma, a Neighborly escolheu a AWS, modernizando seu ambiente analítico e focando em seus valores, tendo como parceria especializada a *IPsense*, na qual atua com expertise e soluções personalizadas para cada cliente.

A solução desenvolvida pela *IPsense* validou resultados e benefícios significativos, como o sucesso na migração de dados do *SQL Server on-premises* para o AWS S3, com a utilização do *AWS Data Migration Service* (DMS). Foi identificada uma significativa melhoria em eficiência e desempenho do ambiente analítico utilizando os serviços da AWS, permitindo a Neighborly uma transformação de dados mais ágil e a realização de consultas aperfeiçoadas.

5 CONCLUSÃO

Algo essencial no mundo de hoje é ficar atualizado sobre assuntos novos que surgem na área de tecnologia. Neste contexto, identifica-se que *Data Lake* além de ser relativamente novo, é um assunto com pouco estudo e muito a ser explorado. Com o intuito de obter e levar conhecimento a todos, foram apresentadas pesquisas sobre o *Data Lake*, que representam uma nova ferramenta para gerenciamentos dos dados, análises e que permite a extração de *insights*. No *Data Lake*, os dados ficam armazenados em dados brutos e em grandes volumes sem a necessidade de tratamento, desempenhando um papel fundamental na era da transformação digital e de análises de dados em tempo real.

O potencial do *Data Lake* para transformar dados e a eficiência nos negócios é muito significativa e certamente continuará a crescer conforme surgirem novas tecnologias para a estratégias de dados.

Os estudos focados em processamento em tempo real e em lote permitem a análise retrospectiva em grandes conjuntos de dados. Já as ferramentas de consultas e visualizações integradas permitem que os usuários explorarem os dados de forma intuitiva e tenham *insights* valiosos para impulsionar inovações e a eficiência dos negócios.

Como estudo apresentado das empresas Coca Cola e Neighborly, foi verificado que o *Data Lake* é muito eficiente e um diferencial para as organizações era digital, onde o volume de dados cresce a cada segundo e entende-se que todo dado seja importante em algum momento para uma tomada de decisão. Com a implantação do *Data Lake,* as empresas apresentadas no estudo obtiveram resultados substanciais na produtividade da equipe de análise (aumento de 80%) e 95% dos dados de diferentes áreas de negócios foram unificados em um único *Data Lake*. Com a implantação do *Data Lake* unificando todos os dados, pôde ser feita a adoção de Inteligência Artificial e outros recursos que possibilitam a tomada de decisão com base nos dados unificados do *Data Lake,* possibilitando o aumento considerável da competitividade destas empresas.

O objetivo neste trabalho foi mostrar que, mesmo sendo algo novo e desconhecido por boa parte dos profissionais dentro das organizações, o *Data Lake* já está sendo utilizado em grandes e conhecidas empresas, mostrando resultados satisfatórios e sinalizando que é um assunto com muito a ser explorado para o futuro tecnológico, possibilitando ganhos e funcionalidade para a utilização de todos os dados que obtemos no dia a dia.

REFERÊNCIAS

AMAZON WEB SERVICES. **Estudo de Caso:** Coca-Cola. Disponível em: https://aws.amazon.com/pt/solutions/case-studies/innovators/coca-cola/. Acesso em: 27 fev. 2024.

AMAZON WEB SERVICES. **Estudo de Caso:** Coca-Cola Andina. Disponível em: https://aws.amazon.com/pt/solutions/case-studies/coca-cola-andina-case-study/. Acesso em: 27 fev. 2024.

AMAZON WEB SERVICES. **Data Lakes and Analytics:** Data Lakes. Disponível em: https://aws.amazon.com/pt/big-data/datalakes-and-analytics/datalakes/. Acesso em: 12 mar. 2024.

AMAZON WEB SERVICES. **AWS CloudTrail:** Guia do usuário. (s.d.). Recuperado de https://docs.aws.amazon.com/pt_br/aescloudtrail/latest/userguide/cloudtrail-user-guide.html. Acessado em: 12 mar. 2024.

AMAZON WEB SERVICES. **O que é um data warehouse?**. (s.d.). Recuperado de https://aws.amazon.com/pt/what-is/data-warehouse/ Acesso em: 13 jun. 2024.

CUTTING, D., & CAFARELLA, M. **Data Lakes**: The Definitive Guide. Data Lake Management: Challenges and Opportunities. 2015. Disponível em: http://www.vldb.org/pvldb/vol12/p1986-nargesian.pdf. Acesso em: 25 fev. 2024.

DIXON, J. **Pentaho, Hadoop, and Data Lakes.** 2010. Disponivel em: https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes. Acesso em: 25 fev. 2024.

FANG, H. Managing Data Lakes in Big Data Era**:** What's a data lake and why has it become popular in data management ecosystem. **In**: The 5th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, June 8-12, 2015, Shenyang, China. Disponivel em: https://www.researchgate.net/publication/308871019_Managing_data_lakes_in_big_data_era_What's_a_data_lake_and_why_has_it_became_popular_in_data_management_ecosystem. Acesso em: 20 fev. 2024.

GIL, A, C. **Como elaborar projetos de pesquisa**. 1991. Atlas.

INMON, B., LINSTEDT, D. **Data Lake Architecture:** Designing the Data Lake and Avoiding the Garbage Dump. 2017. Technics Publications.

IPSENSE. **Estudo de Caso:** AWS Neighborly Data Lake. Disponível em: https://www.ipsense.com.br/estudo-de-caso-aws-neighborly-data-lake/. Acesso em: 27 fev. 2024.

KHINE, P.P. Data **lake:** a new ideolçogy in big data era. Disponivel em: https://doi.org/10.1051/itmconf/20181703025. Acesso em: 12 mar. 2024.

MILOSLAVSKAYA, N., & TOLSTOY, A. Application of Big Data, Fast Data and Data Lake Concepts to Information Security Issues. **In:** 2016 4th International Conference on Future Internet of Things and Cloud Workshops. Disponível em: https://www.researchgate.net/publication/309183107_Big_Data_Fast_Data_and_Data_Lake_Concepts. Acesso em 20 fev. 2024.

SAP. **O que é um data warehouse?**. (s.d.). Recuperado de https://www.sap.com/brazil/products/technology-platform/datasphere/what-is-a-data-warehouse.html. Acessado em: 13 jun. 2024.

SINGH, A. Architecture of Data Lake. **Revista Internacional de Pesquisa Científica em Ciência da Computação.** Engenharia e Tecnologia da Informação (IJSRCSEIT), 2019, vol.5, n.2, p.411-414. Disponível em: https://doi.org/10.32628/CSEIT1952121. Acesso em: 27 fev. 2024.

SINGH, A. & AHMAD, S. Architecture of Data Lake. International Journal of Scientific Research in Computer Science, **Engineering and Information Technology**, 2019, vol. 5. Diponivel em: https://doi.org/10.32628/CSEIT1952121. Acesso em: 12 mar. 2024.

SINGH, H, S. **Data Warehouse:** Conceitos, Tecnologias, Implementação e Gerenciamento.1ª ed. São Paulo: Makron Books, 2001.