A implicação do Data Lake nas empresas: uma Análise mais profunda

Jaspreet Singh1
Professor Associado
Ciência da Computação e Engenharia
Chandigarh University
Mohali, Índia
cec.jaspreet@gmail.com

Gurpreet Singh2
Professor Assistente
Ciência da Computação e Engenharia
Chandigarh University
Mohali, Índia
aiet.cse.gurpreet@gmail.com

Bhoopesh Singh Bhati3
Professor Associado
Ciência da Computação e Engenharia
Chandigarh University
Mohali, Índia
bhoopesh.e11458@cumail.in

Resumo—Todos os dias enormes quantidades de informações são produzidas a partir dos avanços da informática e lidar com esses dados complexos gigantescos requer um conhecimento decente sobre o método mais proficiente para lidar com esses dados. Com o objetivo de aproveitar ao máximo esses dados multiformes para determinados benefícios, o data lake surge como ideia para maior adaptabilidade e forte análise de dados. A terminologia Data Lake significa um espaço de armazenamento para armazenar dados heterogêneos, tanto organizados quanto não estruturados, trazendo uma associação adaptável que permite que os clientes do data lake incorporem dinamicamente os dados que eles solicitam. A inovação de Big Data oferece ajuda às empresas no processo de inteligência de negócios, mas existe uma falta de estudo empírico sobre a utilização da técnica de data lake nas empresas. Este artigo faz uma revisão exploratória sobre a implicação do data lake, retratando seu conceito, arquitetura funcional, estágios de desenvolvimento envolvidos e numerosos desafios e direções de pesquisa; o que melhorará a utilização efetiva da abordagem de data lake nas empresas.

Termos do índice — Data Lake, Data Lake vs Data Warehouse, Desafios de pesquisa em Data Lake, Data Lake na empresa, Estágios para a criação de Data Lake, Necessidade de Data Lake.

I. INTRODUÇÃO

A inteligência de negócios (BI) sempre encontra novos potenciais, destaca ameaças potenciais, revela novos insights de negócios e melhora os processos de tomada de decisão no processo de BI existente da empresa com mais frequência dependem de métodos de Data Warehouse e fluxo de dados entre vários componentes de negócios. .Por exemplo, os aplicativos de Internet das Coisas (IoT) capacitam a coleta constante de informações de forma eficaz a partir da linha de fabricação.Normalmente, as informações utilizadas para aplicativos de inteligência de negócios (BI) e análises são heterogêneas, complexas e excepcionalmente grandes.Conhecimento significativo para inteligência de negócios (BI) e sistemas analíticos vem em uma infinidade de formatos, mesmo de uma variedade de fontes, ou seja, tanto internas quanto externas. Hoje, as organizações se concentram em vários números de aprendizado de máquina de tecnologia, análise de dados, inteligência artificial etc a fim de criar inovação disruptiva e alterar seus negócios. As informações estão no centro de como essas organizações de ponta estão usando IA para mudar suas operações. O ambiente de negócios atual está em constante evolução e há necessidade de dados econômicos e tecnologicamente viáveis solução de design orientada para obter uma ampla gama de formatos de dados e armazenar tudo no mesmo repositório. arquitetura, banco de dados, ferramentas analíticas e aplicativos são todos importantes para este sistema. A visão de negócios, em conjunto com uma forte gestão financeira, ajuda a associação a progredir nos negócios [1,8]. Big data e análise de negócios são dois desvios de negócios que estão tendo um efeito decente no processo operacional das organizações. Ao revelar novas experiências de negócios , e desenvolvendo ainda mais modelos de lucro de ciclos dinâmicos, o BI pode ajudar as organizações a melhorar seu desempenho [2]. Os data lakes foram adotados pelas organizações, pois separam os produtores de dados (como estruturas funcionais) dos consumidores de dados. Os data lakes são uma camada de depósito útil para informações de teste em ciência de dados. Os dados podem ser produzidos e utilizados livremente, sem a necessidade de coordenação com diferentes estruturas ou especialistas[17].

A. Data Lake: necessidade e conceito

Telefones inteligentes, mídia on-line, objetos conectados e diferentes geradores de informações produzem um enorme volume de dados estruturados, semiestruturados e não estruturados de forma impressionantemente mais rápida do que antes na era do big data. Esses dados são incrivelmente significativos para os Sistemas de Apoio à Decisão das empresas (DSS), que dependem vigorosamente de dados como seu estabelecimento. Em qualquer caso, lidar com quantidades heterogêneas e grandes de dados é especialmente difícil para DSS. Data Warehouse (DW) é uma solução amplamente utilizada em DSS atualmente.Técnicas ETL foram usadas para extrair, transformar e carregar dados de acordo com o esquema atual, mas dados substanciais são danificados como resultado de operações ETL.A prevalência de DW pode ser creditada a sua resposta rápida, execução estável e exame prático cruzado, mas o custo de um DW pode aumentar drasticamente conforme as solicitações de melhoria de desempenho, maior volume de dados e aumento da complexidade do banco de dados [3]. Para abordar a falha do big data e a fraqueza do data warehouse, os autores de [4] propuseram o conceito de data lake (DL). dados estruturados, não estruturados e binários, independentemente de seu tipo, formato ou forma. os dados serão salvos em seu arranjo exclusivo no lago. Empilhamento

978-1-6654-0816-5/22/$31,00 ©2022 IEEE

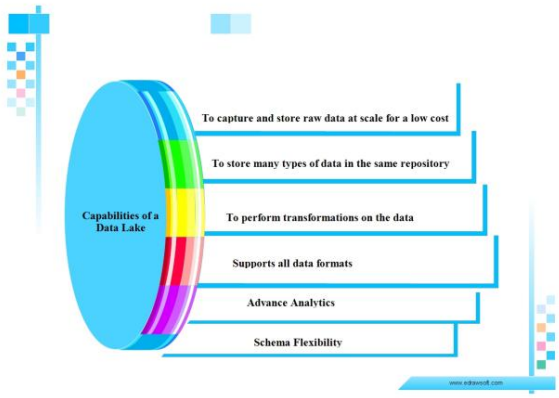

Fig. 1. Capacidades do Data Lake

os dados em centros de armazenamento de dados nunca exigirão pré-manipulação e alteração complexas. As despesas de ingestão de dados também podem ser reduzidas diretamente. Sempre que os dados são armazenados no lago, eles estão abertos a todos na empresa para investigação. O objetivo fundamental de um data lake é armazenar cada uma das informações, mantendo um inegável grau de flexibilidade. descrições. Assim, é necessária a configuração de metadados da estrutura da placa para DL [6]. Os recursos de um Data Lake podem ser visualizados na figura 1 mencionada acima:

B. Necessidade/popularidade do Data Lake

1) Os Data Lakes tentam resolver dois problemas - Os silos de dados (antigo problema) e os desafios impostos pelas iniciativas de big data (novo problema) [5]. Em vez de ter coletas de dados autônomas, todos os dados a serem armazenados são reunidos no Data Lake para lidar com o antigo problema dos silos. A nova questão é lidar com as dificuldades do período de big data, por exemplo, os data lakes tentam resolver as dificuldades impostas pelas qualidades de big data V - volume, velocidade, veracidade, variedade e valor.

2) O Data Lake reconhece qualquer volume de dados, bem como a estrutura de dados. Cada dado pode ser armazenado no data lake de maneira simplificada. Informações críticas suficientes podem ser fornecidas ao lake (por exemplo, mais nós serão adicionados em a solução Hadoop garantindo escalabilidade) [5].

3) DLs ingerem uma ampla gama de dados em seu formato nativo com avanços mínimos de despesas para dar maior capacidade de adaptação e escalabilidade [3].

4) Os Data Lakes também são adequados para realocação de processos ETL que ocupam padrões de gerenciamento de centros de distribuição de informações de grandes negócios que podem ser utilizados para aplicativos lógicos e funcionais. Os dados podem ser realocados das estruturas de origem para o data lake e o ETL pode acontecer lá.

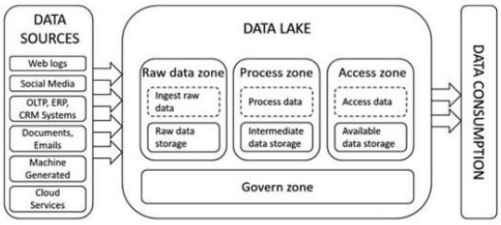

Fig. 2. Arquitetura funcional do Data Lake

II. DATA LAKE: ARQUITETURA FUNCIONAL

A figura 2 acima mencionada indica que parece haver quatro zonas básicas, cada uma com um retângulo pontilhado e uma área de armazenamento de estatísticas que retém o resultado dos processos (além da zona governamental) [3], a capacidade de cada elemento é listada debaixo de:

1) Zona de dados brutos: uma ampla variedade de dados é ingerida e armazenada em seu arranjo nativo sem ser processada. Ingestão em lote, em tempo real ou híbrida são todas as opções. Os clientes podem usar esta zona para encontrar a primeira forma de dados para seu exame, tornando os tratamentos posteriores mais simples. A disposição dos dados brutos salvos pode diferir da forma original dos dados existentes.
2) Zona de processo: Os clientes podem alterar os dados de acordo com seus requisitos e manter todos os dados intermediários nesta zona. Da mesma forma, dados em lote e/ou em tempo real são concebíveis. Esta zona permite que os clientes lide com informações para sua análise de dados (projeção , junção, seleção, agregação, etc).
3) Zona de acesso: A zona de acesso a dados armazena e oferece acesso a todos os dados para análise de dados. Essa região permite que os clientes façam uso de dados de autoatendimento para uma variedade de análises, como cálculos de IA, análise de inteligência de negócios, etc.
4) Zona de governança: todas as outras zonas diferentes dependem da governança de dados. É diretamente responsável pela segurança dos dados, qualidade dos dados, ciclo de vida das informações dos executivos, acesso aos dados e gerenciamento dos metadados.

III. DATA WAREHOUSE VS DATA LAKE

Os armazéns de dados são enormes espaços de armazenamento para dados coletados de uma variedade de fontes. Os armazéns de dados têm sido a base para o conhecimento corporativo e descoberta/armazenamento de informações por muito tempo. De acordo com um ponto de vista empresarial, os proponentes dos conceitos de data lake [14] resumem as distinções entre Data Lake e Data Warehouse. Um data lake armazena um enorme volume de informações naturais até que seja necessário. Um data lake utiliza uma arquitetura plana para armazenar informações, enquanto um data warehouse hierárquico armazena dados em arquivos ou pastas. Uma identificação exclusiva é fornecida a cada componente de dados em um lago e é marcada com vários rótulos de metadados estendidos. O data lake pode ser acessado para obter dados relevantes quando surge uma questão comercial. Abaixo

Comparsion	Data Warehouse	Data Lake
Schema Style	Schema on-Write	Schema on-Read
Agility	Fixed configuration so less agile	Configurable as desired so highly agile
Data	Structured Processed Data	Structured Data,Raw Data
Users	Business Professional	Data Scientists
Storage	Expensive	Low-Cost Storage
Applications	Business Intelligence,Enterprice Reporting	Data Science,Machine Learning,etc.
Scale	Scale to moderate volumes at high cost	Scale to moderate volumes at low cost
Best Fit In	Historical Data Analysis	Advance Data Analysis

Fig. 3. DATA WAREHOUSE VS DATA LAKE

A figura 3 mencionada acima representa as principais distinções entre um data warehouse e um data lake.

4. ARMAZENAMENTO DE DATA LAKE

O problema de armazenamento de data lake envolve a determinação de quais tecnologias de armazenamento de dados devem ser empregadas para manter os conjuntos de dados geridos seguros. Alguns métodos dependem de bancos de dados relacionais ou NoSQL típicos, enquanto outros (Polystore) criaram sistemas ou combinações de armazenamento exclusivos. Dividimos as soluções em três categorias com base em como os dados ingeridos são mantidos no data lake: como arquivos, em um único formato de banco de dados ou em polystores.

1) Sistemas de armazenamento baseados em arquivos: talvez as soluções de armazenamento de dados mais comumente indicadas para data lakes sejam o Hadoop Distributed File System (HDFS). O HDFS pode lidar com uma ampla variedade de tipos de registro. Ele suporta uma variedade de designs de pressão de informações, além de texto (por exemplo, CSV, XML, JSON) e registros binários (por exemplo, imagens). Ele também suporta tipos de capacidade colunar, facilitando a administração do esquema. O Hadoop por si só raramente atinge os objetivos de um data lake.
   O armazenamento do lago de dados do Azure é um sistema de armazenamento hierárquico baseado em arquivos de várias camadas da Microsoft [9,10].

2) Armazenamento de dados único: Algumas estruturas de DL se concentram em um tipo específico de dados e utilizam um único sistema de banco de dados para armazenamento de capacidade. O data lake pessoal, por exemplo, utiliza um modelo de informações baseado em gráficos (por exemplo, diagramas de propriedade) e armazena informações no Neo4j. As entradas do data lake pessoal, que são fragmentos de dados pessoais heterogêneos gerados a partir da interação usuário-web (estruturada, semiestruturada e não estruturada), são serializadas para objetos JSON especificamente definidos, que são simplificados para estruturas gráficas Neo4j com gerenciamento extensível de metadados no data lake , categorizando por tipos de dados: dados brutos, metadados, semântica adicional e identificadores de fragmentos de dados [9].

3) Data Lakes baseados em nuvem: Além de alguns aplicativos da vida real, a maior parte das estruturas de data lake mencionadas anteriormente está no local.
   A plataforma de computação em nuvem Google Infrastructure as a Service (IaaS), por exemplo, é utilizada para controlar o data lake. Devido ao excelente grau de dados em data lakes industriais, é mais popular criá-los em ambiente de nuvem. Alguns provedores significativos de base de informações em nuvem, incluindo Amazon Web Services (AWS), Data Cloud from Snowflake e outros, estão promovendo análise de dados sem servidor e plataformas nativas de nuvem para gerar data lakes[9,11].

4) Sistemas Polystore: A persistência poliglota é implementada por meio de sistemas polystore (ou multistore), que fornecem acesso integrado a uma configuração de vários armazenamentos de dados para dados heterogêneos. Constance, por exemplo, ingere dados brutos em bancos de dados relacionais (por exemplo, MySQL), baseados em documentos (por exemplo, MongoDB) e gráficos, dependendo de seu formato original (por exemplo, Neo4j). Um arquivo JSON, por exemplo, será mantido em MongoDB. Se um conjunto de dados de entrada não puder ser salvo diretamente em um banco de dados relacional ou NoSQL, ou se a escalabilidade para computação distribuída for uma preocupação, os dados podem ser armazenados em HDFS[12,13].

V. ETAPAS PARA CONSTRUÇÃO DO DATA LAKE

A maioria dos data lakes se desenvolveu ao longo do tempo devido à expansão e experimentação incrementais.
Projetar um data lake é uma ideia que poucos indivíduos investigaram em algum momento. A maneira mais eficaz de construir um data lake é explicada em várias etapas abaixo mencionadas na figura 4 contém [7]:

Estágio 1: Governando grandes quantidades de dados: O estágio principal é configurar a estrutura e descobrir como adquirir e controlar informações em um escopo enorme. O exame pode ser simples agora, mas muito se aprende sobre como fazer o Hadoop funcionar da maneira que queremos.

Fase 2: Desenvolvimento de capacidades analíticas e transformacionais: A capacidade de mudar e dissecar informações é aprimorada na segunda etapa. As organizações e as ferramentas mais adequadas ao seu conjunto de habilidades começam a obter dados adicionais e desenvolver aplicativos neste estágio. O data warehouse corporativo e os recursos do data lake são combinados.

Fase 3: Amplo impacto operacional: A terceira etapa envolve colocar o máximo de dados e análises nas mãos de quantos indivíduos for razoavelmente esperado. Agora, o data lake e o data warehouse corporativo começam a cooperar, cada um atendendo a uma necessidade específica. Quase toda empresa de big data que começou com um data lake logo adicionou um data warehouse corporativo para operacionalizar seus dados, como exemplo da necessidade dessa combinação. As organizações com data warehouses corporativos não os estão deixando para o Hadoop, todas as coisas

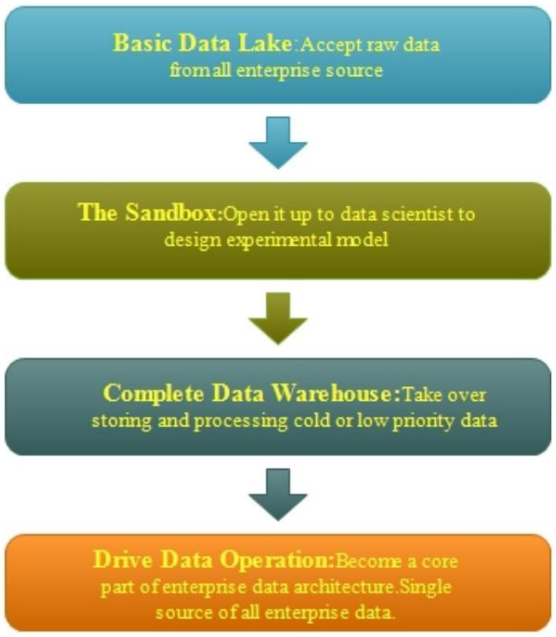

Fig. 4. Etapas para construção do Data Lake

considerado.

Estágio 4: Recursos corporativos: os recursos corporativos são adicionados ao data lake nesta fase. Poucas organizações chegaram a esse nível de desenvolvimento, mas à medida que cresce a utilização de enormes volumes de dados, mais organizações o farão, exigindo administração, consistência, segurança e revisão.

VI. DATA LAKES: DESAFIOS E ORIENTAÇÕES DE PESQUISA

1) **Ferramentas de aprendizado de máquina sobre data lakes:** as técnicas de aprendizado de máquina são suportadas nos níveis de fronteira de grandes data lakes, o que é uma característica distintiva. Isso visa, eventualmente, permitir a extração de valor das principais fontes de big data, que podem ser utilizadas para revelação de informações e tomada de decisões sensatas. Inúmeras ferramentas de IA foram propostas na literatura ao longo dos estudos anteriores. Infelizmente, eles usam um modelo rudimentar de dados brutos (entrada) que não consegue acompanhar os novos recursos do repositório de big data. Como resultado, construir uma ferramenta poderosa de aprendizado de máquina para ser configurada no front-end de grandes data lakes tornou-se uma das dificuldades mais prementes a serem abordadas, apesar do fato de seus resultados terem sido demonstrados em uma variedade de grandes projetos do mundo real. aplicações de dados[15].

2) **Metodologias de Governança de Big Data via Data Lakes:** Governança de dados refere-se a um conjunto de modelos, metodologias e padrões para apoiar a alegada sociedade avançada impulsionada pela informação, ou seja, uma visão de ponta onde os ciclos superculturais (por exemplo, demografia, serviços residentes , administração de políticas, etc) são conduzidos pelo exame de informações registradas e atuais, que é utilizado como um ponto de apoio dinâmico. De fato, um cenário claro que exemplifica esse conceito é claramente visível hoje: a epidemia da pandemia de COVID-19 e as estratégias que os países estão implementando para contê-la, sendo essas políticas inteiramente conduzidas pela análise diária dos dados do surto[15].

3) **Ingestão de dados:** A ingestão freqüentemente requer comunicação com fontes de dados externas com capacidade de transmissão restrita, bem como alto paralelismo e baixa latência. A ingestão não realiza nenhuma análise extensiva dos dados baixados dessa maneira.A reprodução e a formação múltipla de conjuntos de dados dinâmicos também devem ser possíveis com esboços básicos de dados, como somas de verificação.Apoiar a entrada contínua de dados de alta velocidade com indexação mais refinada para tornar essas informações mais prontamente acessível para fins analíticos é uma das dificuldades contínuas na ingestão de dados[17].

4) **Versão do conjunto de dados:** os data lakes estão em constante evolução. Na fase de ingestão, novos registros e declaração atualizada de documentos atuais são injetados no lago. Os dispositivos também podem mudar a longo prazo, trazendo declarações atualizadas de dados disponíveis. À medida que a quantidade de variantes se desenvolve, permitir armazenamento e recuperação de versões eficientes e econômicos se tornará cada vez mais crucial em um sistema de data lake bem-sucedido. A evolução do esquema é um desafio para os sistemas de versionamento de data lake[17].

5) **Limpeza de dados:** A limpeza de dados foi amplamente investigada para dados corporativos, mas pouco foi feito em relação ao data lake. A limpeza de dados lógicos e relacionais requer dados de composição exatos, bem como restrições de integridade. Usar o insight do lago e fazer a limpeza coletiva de dados é uma possibilidade cativante na limpeza de dados do lago. Além disso, à luz do fato de que a metodologia do data lake, como a extração, pode infundir erros metódicos no lago, é fundamental observar as condições e atividades fundamentais que resultam nesses problemas [18].

6) **Gerenciamento de metadados:** os data lakes geralmente não são acompanhados por catálogos de dados abrangentes. Um data lake sem muito espaço pode assumir a forma de um pântano de dados se os conjuntos de dados não tiverem informações específicas vinculadas a eles. Os sistemas de gerenciamento de metadados devem permitir metadados eficientes armazenamento e consulta reagindo sobre metadados, bem como reunindo metadados de fontes de dados e avançando dados com metadados úteis (como extensas descrições de dados e requisitos de integridade). Apesar da maneira como a divulgação de metadados oferece a deliberação de dados vitais para a interpretação e descoberta de dados, ainda há oportunidades para extrair dados das informações do lago

e consolidá-lo em bases de informação existentes (área explícita) [19]. O pântano de dados resulta da ausência de metadados descritivos e de uma maneira de armazenar metadados. Cada vez que os dados são divididos, isso deve ser feito sem qualquer preparação. É difícil garantir resultados[16].

7) **Pântano de dados:** De fato, até mesmo os defensores dos data lakes conhecem as desvantagens do data lake. Uma das maiores é se transformar em um pântano de dados. Ninguém imagina o que será descartado no lake. Além disso, não há metodologia definida para evitar que eles aconteçam, como inserir dados errados, repetir dados ou inserir dados incorretos. A veracidade das informações enviadas para o data lake não pode ser assegurada pelo fato de terem sido extraídas. Supondo que ninguém saiba que tipo de dados é colocado no lago, é concebível que ninguém veja que uma parte dos dados está arruinada até passar do ponto sem retorno. Como as empresas começaram a usar isso sem procedimentos de proteção de última geração, essas falhas se tornaram mais aparentes. Além disso, as falhas de segurança ainda precisam ser resolvidas [20].

VII. CONCLUSÃO

Atualmente, as empresas estão se concentrando mais nos dados para fazer julgamentos fundamentados. Em termos de receita, desenvolvimento e crescimento, as empresas que podem usar dados com sucesso são pioneiras mundiais. De fato, mesmo para sobreviver, trabalhar e competir nesta era, as organizações devem ter a opção de utilizar seus dados de maneira viável. Uma quantidade imensa de investimento é feita no tratamento de muitos dados para fazer as melhores escolhas. Este artigo explorou as habilidades dos data lakes e esclareceu como os data lakes são vistos em termos de suas vantagens e utilizações. Além disso, vários desafios relacionados aos data lakes também são identificados.

Ao mesmo tempo em que atende aos requisitos de BI e Big Data, o Data Lake fornece fontes de dados abundantes e abundantes para pesquisadores, especialistas, analistas de dados e consumidores de dados de autoatendimento.

REFERÊNCIAS

[1] Ajah, IA e Nweke, HF, 2019. Big data e análise de negócios: tendências, plataformas, fatores de sucesso e aplicações. Big Data e Computação Cognitiva, 3(2), p.32.

[2] Kowalczyk, Martin e Peter Buxmann. (2014) ”Big data e processamento de informações em processos de decisão organizacional.” Negócios e Engenharia de Sistemas de Informação 6 (5): 267-278.

[3] Ravat, F. e Zhao, Y., 2019, agosto. Data lakes: tendências e perspectivas. Na Conferência Internacional sobre Banco de Dados e Aplicações de Sistemas Especialistas (pp. 304-313). Springer, Cham.

[4] Dixon, J.: Pentaho, Hadoop e data lakes, outubro de 2010.

[5] Gartner.Inc, Gartner diz cuidado com a falácia do Data Lake, STAMFORD, Connecticut, 28 de julho de 2014, acesso em 29 de agosto de 2017, http://www.gartner.com/newsroom/id/2809117 [6] Miloslavskaya, N., Tolstoy, A.: Big data, fast data e conceitos de data lake. Procedia Comput. ciência 88, 300–305 (2016).

[7] CITO Research: o Data Lake to Work - Um Guia de Práticas [Online]. Disponível: https://hortonworks.com/wpcontent/uploads/2014/05

[8] Llave, MR, 2018. Data lakes in business intelligence: reporting from the trincheiras. Ciência da computação Procedia, 138, pp.516-524.

[9] Hai, R., Quix, C. e Jarke, M., 2021. Conceito e sistemas de data lake: uma pesquisa. pré-impressão arXiv arXiv:2106.09592.

[10] B. Stein e A. Morrison. O data lake corporativo: melhor integração e análises mais profundas. Previsão da PwC Technology: Repensando a integração, 1(1-9):18, 2014.

[11] AA Munshi e YA-RI Mohamed. Arquitetura Data Lake Lambda para Smart Grids Big Data Analytics. IEEE Access, 6:40463–40471, 2018 [12] R. Hai, C. Quix e C. Zhou. Reescrita de consulta para dados heterogêneos lagos. Em ADBIS, páginas 35–49, 2018.

[13] R. Hai, S. Geisler e C. Quix. Constance: Um Sistema Inteligente de Data Lake. Em SIGMOD, páginas 2097–2100. ACM, 2016.

[14] Tamara Dull, Data Lake Vs Data Warehouse: Key Differences, Recuperado em 26 de setembro de 2017 http://www.kdnuggets.com/2015/09/data-lake-vs data-warehouse-key-differences.html.

[15] Cuzzocrea, A., 2021, janeiro. Big Data Lakes: modelos, estruturas e técnicas. Em 2021 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 1-4). IEEE.

[16] Gartner.Inc, Gartner diz cuidado com a falácia do Data Lake, STAMFORD, Connecticut, 28 de julho de 2014, acessado em 29 de agosto de 2017. http://www.gartner.com/newsroom/id/2809117.

[17] Nargesian, F., Zhu, E., Miller, RJ, Pu, KQ e Arocena, PC, 2019. Gerenciamento de data lake: desafios e oportunidades. Proceedings of the VLDB Endowment, 12(12), pp.1986-1989.

[18] X. Wang, M. Feng, Y. Wang, XL Dong e A. Meliou. Diagnóstico de erros e perfil de dados com raio-x de dados. PVLDB, 8(12):1984–1987, 2015.

[19] A. Halevy, F. Korn, NF Noy, C. Olston, N. Polyzotis, S. Roy e SE Whang. Bens: Organizando os conjuntos de dados do Google. Em SIGMOD, páginas 795–806, 2016.

[20] Timothy King “The Emergence of Data Lake: Pros and Cons”, 3 de março de 2016, acessado em 15 de setembro de 2017: https://solutionsreview.com/data integration/the-emergenceof-data-lake-pros-and -cons/ [21]

Singh, J. e Gupta, D., 2017. Rumo à economia de energia com algoritmo de agendamento de tarefas multifila mais inteligente na computação em nuvem. J. Eng. Appl. Sci, 12(10), pp.8944-8948.

[22] Singh, J., Duhan, B., Gupta, D. e Sharma, N., 2020, junho. Otimização do Gerenciamento de Recursos em Nuvem: Taxonomia e Desafios de Pesquisa. Em 2020, 8ª Conferência Internacional sobre Confiabilidade, Tecnologias Infocom e Otimização (Tendências e Direções Futuras)(ICRITO) (pp. 1133-1138). IEEE.