UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO SOCIOECONÔMICO
DEPARTAMENTO DE CIÊNCIAS DA ADMINISTRAÇÃO

Aruan Boritiyça Silva
Prof(a). Orientador(a): Ana Luiza Paraboni

IMPLEMENTAÇÃO DE UM DATA LAKEHOUSE: UM ESTUDO DE CASO NO SETOR DE TECNOLOGIA

Florianópolis

2023

Aruan Boritiyça Silva

IMPLEMENTAÇÃO DE UM DATA LAKEHOUSE: UM ESTUDO DE CASO NO SETOR DE TECNOLOGIA

Trabalho de Curso apresentado à disciplina CAD7305 Laboratório de Gestão: Trabalho de Curso como requisito parcial para a obtenção do grau de Bacharel em Administração pela Universidade Federal de Santa Catarina.
Enfoque: Monográfico – Artigo
Orientador(a): Prof[a]. Dr[a]. Ana Luiza Paraboni

Florianópolis

2023

Catalogação na fonte elaborada pela biblioteca da Universidade Federal de Santa Catarina

Silva, Aruan Boritiyça
IMPLEMENTAÇÃO DE UM DATA LAKEHOUSE: UM ESTUDO DE CASO NO SETOR DE TECNOLOGIA / Aruan Boritiyça Silva ; orientadora, Ana Luiza Paraboni, 2023.
36 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Socioeconômico, Graduação em Administração, Florianópolis, 2023.

Inclui referências.

1. Administração. 2. Big Data. 3. Gerenciamento de Dados. 4. Projetos de Tecnologia. I. Paraboni, Ana Luiza. II. Universidade Federal de Santa Catarina. Graduação em Administração. III. Título.

Aruan Boritiyça Silva

IMPLEMENTAÇÃO DE UM DATA LAKEHOUSE: UM ESTUDO DE CASO NO SETOR DE TECNOLOGIA

Este Trabalho de Curso foi julgado adequado e aprovado na sua forma final pela Coordenadoria Trabalho de Curso do Departamento de Ciências da Administração da Universidade Federal de Santa Catarina.

Florianópolis, 04 de dezembro de 2023.

_____

Profª. Drª. Ana Luiza Paraboni
Coordenadora de Trabalho de Curso

Avaliadores:

Documento assinado digitalmente
Ana Luiza Paraboni
Data: 13/12/2023 19:12:12-0300
CPF: ***.031.270-**
Verifique as assinaturas em https://v.ufsc.br

_____

Profª. Ana Luiza Paraboni, Drª.
Orientadora
Universidade Federal de Santa Catarina

_____

Prof. Leandro Dorneles dos Santos, Dr.
Avaliador
Universidade Federal de Santa Catarina

_____

Prof. Matheus Fernando Moro, Dr.
Avaliador
Universidade Federal de Santa Catarina

RESUMO

Este estudo aborda a implementação de um Data Lakehouse em uma empresa de tecnologia, sob a perspectiva de gerenciamento de dados, visando compreender um processo ideal de implementação e os benefícios alcançados com esta tecnologia. Foram realizadas entrevistas com 8 profissionais envolvidos com o projeto de implementação e observaram-se desafios em três dimensões: compreensão tecnológica, engajamento de stakeholders e gerenciamento de projetos. Os resultados destacam a importância da mensuração, planejamento refinado, comunicação contínua, migração gradual de ativos de dados e validação de dados para garantir o sucesso. O estudo contribui oferecendo percepções de uma aplicação bem sucedida para organizações que consideram adotar essa tecnologia, servindo como guia para antecipar desafios. Além disso, visa contribuir para a formação de gestores e analistas de negócios que desejam potencializar sua capacidade analítica.

Palavras-chave: Delta Lake. Data Lakehouse. Data Warehouse. Big Data. Implementação. Business Intelligence.

1 INTRODUÇÃO

O crescimento das “ponto.com” nos anos 2000 deram início também para a era dos dados, onde a informação se tornou o principal ativo de geração de valor estratégico para as organizações. Indivíduos e corporações detém e geram uma quantidade exuberante de dados, e quanto maior o volume, variedade e velocidade associado a essas informações, maiores serão os desafios de coleta, armazenamento, processamento e análise (PROVOST et al., 2013). Este fenômeno também é conhecido como “Big Data” (BD) e quando aplicadas tecnologias que tornam as informações interpretáveis e acionáveis obtemos o “Big Data Analytics” (BDA), o qual tem sido utilizado como um forte recurso para ganhos econômicos e sociais (AJAH, NWEKE, 2019).

O volume de dados organizacionais já passa das centenas de “petabytes”, o que cria simultaneamente uma grande oportunidade de diferencial competitivo - a razão de ser das aplicações de “Big Data” - e um grande desafio de gerenciamento de dados. Os processos e métodos de coleta, armazenamento, proteção e uso de dados em uma organização, caracterizam o que entendemos como gerenciamento de dados, uma área de estudo que evolui

tão rapidamente quanto o desenvolvimento das tecnologias de “Big Data Analytics” (AJAH, NWEKE, 2019).

Uma vez que se trata de um valioso ativo, que requer investimento em tecnologia e pessoas capazes de gerar ganhos a partir dele, encontrar uma maneira eficaz e adequada de mensurar os retornos sobre o investimento é fundamental. Compreende-se que o retorno do investimento em BDA dá-se pela geração de “*insights*” e a aplicação destes. Com o avanço das práticas e estudos, otimizar o gerenciamento de dados também tornou-se parte da avaliação do valor de BDA. E neste caso, não apenas as ferramentas, mas a arquitetura de dados assumiu um papel importante, sendo responsável pelo armazenamento apropriado dos dados e por garantir o atendimento das necessidades comerciais de informação.

A arquitetura de dados é, ou deveria ser, a primeira etapa no processo de gerenciamento de dados, guiando a modelagem de dados e a implementação de regras de negócio. Os anos 80 marcaram o princípio das soluções de arquitetura de dados com a ampla adoção do “Data Warehouse” (DW) que foi a primeira solução que armazenava dados estruturados a fim de criar um ambiente de análise e geração de valor, mas assim que o BD ganhava escala, os DWs começaram a apresentar dificuldades em lidar com armazenamento de dados não estruturados, bem como com o alto volume de informações e com a eficiência de custos. O que levou a próxima geração, fruto do armazenamento em nuvem de baixo custo, o “Data Lake” (DL), que propunha um modelo mais flexível e eficiente. Este modelo, porém, também fracassou em ser uma solução absoluta e se tornou um complemento para ambientes com DW. Surge então um campo de oportunidade para atender as demandas de um ambiente analítico flexível, eficiente e confiável (JANSSEN, 2022).

As organizações modernas dependem de dados para o sucesso, sendo assim, é importante considerar uma arquitetura de dados que entregue eficiência de custos e escalabilidade (PROVOST et al., 2013). Em 2020, o “Data Lakehouse” ou “Lakehouse” foi introduzido no mercado como uma verdadeira solução que combina os benefícios de gerenciamento dos DW com a flexibilidade, eficiência em custos e escalabilidade dos DL. O “Lakehouse” é compreendido como:

> Um sistema de gerenciamento de dados baseado em armazenamento de baixo custo e acessível diretamente, que também fornece recursos tradicionais de gerenciamento e desempenho de um DBMS analítico, como transações ACID, versionamento de dados, auditoria, indexação, armazenamento em cache e otimização de consultas. Os lakehouses combinam os principais benefícios dos data lakes e dos data warehouses: armazenamento de baixo custo em um formato aberto acessível por uma variedade

> de sistemas do primeiro e recursos de gerenciamento e otimização poderosos do segundo (ARMBRUST et al., 2021).

Implementar uma tecnologia emergente reúne desafios tanto na esfera de implementação quanto na de avaliação do investimento. Em relação à implementação, a dificuldade se dá pela natural impopularidade da tecnologia, e pelo tempo necessário para o aprendizado da mesma. Já sobre o retorno do investimento, é particularmente difícil mensurar um ativo digital, que não é de fato consumível, é integrável, e que revela seu valor quando a combinação do seu uso e manejo produz diferencial estratégico e competitivo para organização (GROVER et al., 2018).

Diante disso, este trabalho visa analisar a implementação de um data lakehouse, a fim de compreender os desafios enfrentados durante o processo, quais elementos que levam ao sucesso e a falha, e como a implementação desta arquitetura de dados afeta materialmente a organização. Por fim, a concatenação dessas visões permitirão a elaboração de um guia destinado a gestores de tecnologia com o desejo de implementar essa arquitetura de dados.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Analisar os impactos e o processo de implementação de um data lakehouse em uma empresa de tecnologia.

1.1.2 Objetivos Específicos

I. Identificar os fatores que levam ao sucesso no processo de implementação de uma data lakehouse;
II. Verificar quais os impactos após a implementação do data lakehouse; e
III. Elaborar um guia de implementação de um data lakehouse.

1.2 JUSTIFICATIVA

Decisões orientadas por dados já são responsáveis por resultados e progresso nas organizações. Ao dar um passo além, em busca da mensuração da dimensão do retorno do investimento em tecnologias de análise e armazenamento de dados, surge a oportunidade de gerenciar dados de maneira eficiente. Na pesquisa “CIO Imperative” de 2022, a consultoria EY relata que investimento em análise de dados é uma prioridade para 53% dos CIOs (Chief Information Officer). Seguindo esta tendência, os gestores de times de dados devem adquirir

conhecimento dos métodos, arquiteturas e ferramentas que podem auxiliar no alcance de seus objetivos operacionais e financeiros.

Aprofundar os estudos sobre o “Lakehouse” é essencial primeiramente por ser uma tecnologia emergente, revelada em 2021, que promete superar as arquiteturas anteriores. Além disso, este estudo contribui para que empresas interessadas em aderir a nova tecnologia possam utilizar os resultados aqui disponibilizados como guia, antecipando requerimentos e desafios que foram enfrentados na implementação da mesma. Ressalta-se também a possibilidade de comparar os resultados alcançados pela empresa deste estudo com àquela que ainda passará pelo processo de implementação.

Por fim, o gerenciamento de dados tende a ser uma temática atípica na escola da administração empresarial, se limitando às mãos de engenheiros de dados, engenheiros de software e cientistas da computação. Porém, é importante para formação de futuros gestores e analistas de negócios que o conhecimento sobre o gerenciamento de dados seja difundido. Já é uma exigência de mercado o conhecimento de linguagens de consulta em banco de dados (SQL), modelagem de banco de dados e linguagens de análise (Python). Assim, este estudo também possui interesse em contribuir para a formação de gestores e analistas qualificados com os desafios do mercado.

2 FUNDAMENTAÇÃO TEÓRICA

2.1. Big Data e Big Data Analytics, crescimento sem precedentes

Tudo começa com o imperativo “você não consegue gerenciar aquilo que não consegue medir”, famoso ditado atribuído a Peter Drucker. A chegada do Big Data (BD) aumentou radicalmente a capacidade de coletar, mensurar e, consequentemente, analisar informações para transformá-las em decisões qualificadas em todas as áreas de trabalho, desde o combate de epidemias como a da Covid-19 ao aumento de vendas para lojas regionais. Hoje “podemos direcionar intervenções mais eficazes e fazê-lo em áreas que até agora foram dominadas pelo instinto e pela intuição, e não por dados e rigor” como citado por Mcafee e Brynjolfsson (2012).

Disciplinas emergentes, tal qual BD, geralmente sofrem com a falta de concordância sobre os conceitos principais. No entanto, há três características amplamente aceitas pela comunidade científica que fundamentam o que diferencia BD da mera prática de análise, elas definem características relacionadas exclusivamente a informação e são conhecidas como os 3V’s: **v**olume exuberante de dados, **v**elocidade de coleta e análise em tempo real (ou quase), e

grande variedade de tipos de dados (MCAFEE; BRYNJOLFSSON, 2012). Com a intenção de separar BD de Big Data Analytics (BDA), quando De Mauro, Greco e Grimaldi (2016) propõem que a definição de Big Data é "o ativo de informação caracterizado por tamanho, volume, velocidade e variedade que requer tecnologias e métodos de análise específicos para gerar valor", eles nos conferem a possibilidade de separar o ativo do método e aproximar a tecnologia da perspectiva de gestão.

Big Data Analytics (BDA) é uma temática que cai sobre a intersecção de ferramentas e métodos de aplicação orientada a Big Data. BDA torna-se necessário quando os métodos tradicionais de análise de dados não suportam mais a larga escala e complexidade de dados inerente ao BD. Os métodos tradicionais de análise de dados possuem limitações para além dos 3V's, tendo em vista que o BD trouxe novos problemas, novas questões de segurança, novas estruturas de dados, exigências profissionais, etc. É necessária a separação para que não se confundam áreas de estudo correlatas, mas substancialmente distintas (TSAI et al., 2015).

BDA é a aplicação de técnicas estatísticas, de processamento, armazenamento e de análise sobre BD a fim de aprimorar a tomada de decisão. Pode ser aplicado na geração de recomendações personalizadas, na identificação e previsão de potenciais falhas, compreensão profunda de comportamento, detectar anomalias, ajustar e analisar processos operacionais, e muito mais. Como dito por Grover et al. (2018, p. 390), "os conhecimentos obtidos com a análise de fluxos de dados estruturados e não estruturados podem responder a perguntas que as empresas nem sequer haviam considerado antes.".

2.2. Métodos de BDA: Business Intelligence, Artificial Intelligence e Machine Learning

De acordo com Grover et al. (2018), BDA é a tendência de negócios que teve maior impacto nos investimentos em TI das empresas na última década. Para aprofundar sobre o valor gerado dessas práticas, é necessário apresentar aquelas que se conectam com o objeto de estudo deste trabalho. Assumindo que a finalidade é desenvolver os meios para tomar a melhor decisão possível, existem três métodos de análise potencializados pelo BD: Business Intelligence (BI), Machine Learning (ML), e Artificial Intelligence (AI).

As definições sobre os métodos serão breves, considerando que o objetivo é compreender o que eles requerem para serem exercidos. Para BI, consideramos:

> O Business Intelligence (BI) reúne os dados coletados para descrever o estado passado e atual de uma empresa. O BI requer o processamento de dados brutos usando lógica de negócios. [...] Um sistema de BI mantém um repositório de definições e lógica de negócios (REIS; HOUSLEY, 2022).

Como introduzido pela citação anterior, BI faz referência tanto ao método de análise quanto ao conjunto de ferramentas que suportam esse tipo de análise. A geração de valor para negócios dá-se ao habilitar consultas a banco de dados, análises estatísticas, geração de relatórios, e visualização de dados por dashboards (EARLEY, 2017).

Artificial Intelligence (AI) ou inteligência artificial, é compreendida como a ciência e a engenharia de criação de máquinas inteligentes, especialmente softwares inteligentes (MCCARTHY, 2007). E Machine Learning (ML) ou aprendizado de máquina, é uma das disciplinas de AI entendida como a capacidade das máquinas imitarem a inteligência humana (MOHRI et al., 2018). ML trata dos métodos computacionais que usam da experiência para aprimorar decisões e fazer previsões mais precisas (BROWN, 2021).

O Business Intelligence (BI), a Inteligência Artificial (AI) e o Aprendizado de Máquina (ML) oferecem diversos benefícios de negócio. O BI proporciona *insights* valiosos sobre o desempenho passado e atual da empresa, facilitando a tomada de decisão. A AI capacita as máquinas a realizar tarefas complexas, automatizando processos e melhorando a eficiência operacional. O ML permite que as empresas extraiam padrões e tendências ocultas nos dados, possibilitando previsões precisas e aprimoramento contínuo. Sob a realidade do Big Data, infere-se que o gerenciamento e a engenharia de dados são fundamentais para explorar o potencial dessas metodologias e alcançar oportunidades de negócio significativas, visto que apenas os métodos e tecnologias mais modernas habilitam o uso de dados de diferentes tipos (estruturados, semi-estruturados e não estruturados), em larga escala, e em tempo real se necessário.

2.3. Engenharia de Dados como disciplina capacitadora da análise de big data

Precedente aos métodos de análise de dados, está a percepção de dados como um ativo organizacional estratégico. Um ativo é um recurso econômico, que pode ser controlado, protegido, guardado, bem ou mal administrado, que possui valor e também pode gerar valor (EARLEY, 2017). A fim de gerenciar este ativo com eficiência, a Engenharia de Dados (ED) protagoniza ao assumir o papel de gerenciar o ciclo de vida dos dados e viabilizar a extração de valor por métodos de BDA como BI, ML e AI. A Engenharia de Dados aborda:

> O desenvolvimento, implementação e manutenção de sistemas e processos que usam dados brutos para produzir informação de alta qualidade e consistentes que suportam casos de utilização posterior, como análises e aprendizado de máquina. Engenharia de dados é a intersecção de segurança, gerenciamento de dados, DataOps, arquitetura de dados, orquestração e engenharia de software (REIS; HOUSLEY, 2022).

Percebe-se que não se trata puramente de tecnologia, o objetivo é produzir dados confiáveis que atendam à empresa com qualidade e significado previsíveis. Sendo assim, há entre as disciplinas adjacentes à ED campos que contribuem para o alcance do objetivo citado acima. Para este trabalho é necessário compreender o gerenciamento e a arquitetura de dados, e como se relacionam com ED. Começar por gerenciamento de dados é adequado visto que:

> A gestão de dados é o desenvolvimento, a execução e a supervisão de planos, políticas, programas e práticas que fornecem, controlam, protegem e aumentam o valor dos dados e dos ativos de informação ao longo dos seus ciclos de vida (EARLEY, 2017).

O gerenciamento tem caráter promotor sobre a gestão do ciclo de vida dos dados realizada pela ED, e faz isso estabelecendo um conjunto de práticas recomendadas, promovendo o entendimento da utilidade estratégica dos dados e instaurando uma estrutura coesa de utilização dos dados em todos os níveis hierárquicos (REIS; HOUSLEY, 2022).

Já a arquitetura de dados define o plano para o gerenciamento de ativos de dados, que pode ser feito utilizando *Data warehouses, Data Lakes* ou até mesmo uma combinação desses componentes. A arquitetura deve selecionar componentes e projetos alinhados à estratégia organizacional (EARLEY, 2017).

> Armazenamento e processamento são os objetos de trabalho da arquitetura de dados, mas não se deve pensar arquitetura apenas como um conjunto de ferramentas, "a arquitetura é o design de nível superior, roteiro e modelo de sistemas de dados que satisfazem os objetivos estratégicos para o negócio", ou seja, é tudo aquilo que determina se um conjunto de tecnologias vai ou não gerar valor para o negócio (REIS; HOUSLEY, 2022).

As disciplinas de gerenciamento e arquitetura de dados já existem há décadas e passam por constante evolução, mas foi o BD que acelerou e adicionou novas variáveis, trazendo ED como um sistema de práticas adequadas para manejá-lo. No próximo tópico expõe-se o declínio dos Data Warehouses (DW) - arquitetura de armazenamento que será caracterizada posteriormente - pela inabilidade de suportar tipos variados (não estruturados) de dados em alta escala e velocidade, bem como a evolução histórica das arquiteturas de dados até o momento. Ao longo deste estudo, será explorada a noção de que a conformidade do esforço de Engenharia de Dados (ED) com o escopo determinado pelo gerenciamento e arquitetura pode resultar na redução dos custos de infraestrutura e na geração de retornos financeiros.

2.4. As práticas de BDA trazem novas exigências para arquiteturas de dados

É preciso conhecer o passado para compreender o presente e vislumbrar o futuro. DW e Data Lakes (DL) são passado e presente para as empresas que não fazem parte da vanguarda em arquitetura de dados. Após a compreensão desses métodos, avançaremos para uma proposta que supera as limitações das anteriores e se posiciona como a solução ideal para os desafios de BDA. De acordo com EARLEY (2017), Data Warehouse (DW) trata de dois componentes principais: o banco de dados integrado de suporte a decisões, e os programas de software usados para coletar, limpar, transformar e armazenar dados de diversas fontes operacionais e externas. E seu valor pode ser visto da seguinte forma:

> O data warehouse pode ser considerado uma tecnologia facilitadora da mineração de dados. [...] as empresas que decidem investir em DW geralmente podem aplicar a mineração de dados de forma mais ampla e profunda na organização. Por exemplo, se um DW integrar registros de vendas e faturamento, bem como de recursos humanos, ele poderá ser usado para encontrar padrões característicos de vendedores eficazes (PROVOST; FAWCETT, 2013).

O Data Lake (DL) representa o ambiente em que uma abundância de dados de vários tipos e estruturas pode passar pelo ciclo de vida de ingestão, armazenamento, avaliação e análise. É um conceito estritamente ligado ao surgimento do BD (EARLEY, 2017). O DL precisa habilitar o armazenamento do dado em sua forma bruta, facilitando a exploração, a automação das atividades de gerenciamento de dados e permitir uma grande variedade de métodos analíticos, entre eles aplicações de BI, ML e AI (ZAGAN; DANUBIANU, 2021).

A arquitetura dominante é atualmente a união de DW com DL, uma solução incompleta visto que ambas carecem separadamente de um conjunto de requisitos para aplicar métodos de BDA, e quando juntas criam ainda mais dificuldades, como o Quadro 1 apresenta:

Quadro 1 - Limitações relacionadas a Data Warehouse, Data Lake, e combinados.

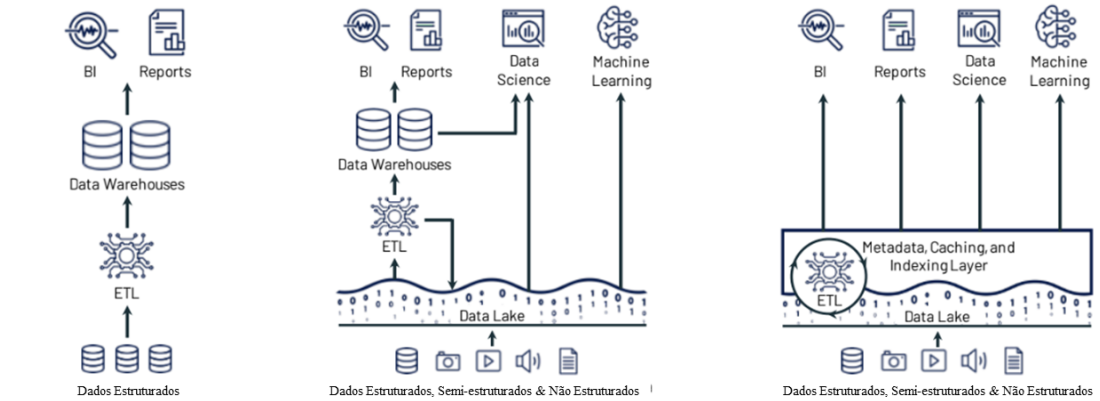
Data Warehouse (DW)	**Data Lake (DL)**	**DW + DL**
• Foco em dados estruturados de natureza transacional. Não suporta dados semi-estruturados ou não estruturados; • Não habilita ML e AI; • Acesso a dados baseado apenas em SQL; • Apresenta dificuldades com picos de volume e velocidade; • Esquematização de dados inflexível.	• Não suporta dados transacionais • Não pode garantir a confiabilidade dos dados • Processamento de consultas é lento - consequentemente, arrasta o DW da geração anterior para realizar ETL em porções de dados menores;	• Suporte não integrado e incompleto para casos de BI, ML e AI; • Métodos de governança e segurança incompatíveis entre os modelos; • Requer engenharia contínua e repetitiva para ETL de dados entre os sistemas;

Fonte: Elaboração Própria.

A dupla camada de armazenamento aumenta os custos de operações nas vias de armazenamento e processamento de ETL (ETL é a sigla para o processo de extrair, transformar e carregar. É uma forma tradicional aceita para que as organizações combinem dados de vários sistemas em um único banco de dados, repositório de dados, armazenamento de dados ou data lake) duplicados, não alcança integração dos dados e abre brechas para ferir a qualidade dos dados - dados desatualizados, confiabilidade questionável, indisponibilidade. O suporte para ML é limitado visto que, sob o ambiente de maior confiabilidade (DW) não há suporte para tal operação, e no ambiente com suporte (DL) a qualidade do dado é questionável e a capacidade de processamento é lenta. Data Lakes (lagos de dados) rapidamente se tornaram Data Swamps (pântano de dados), um local de difícil acesso, impróprio para exploração, com informações abandonadas (INMON et al., 2021).

O armazenamento em nuvem altamente escalável, barato e flexível é uma solução alinhada com as exigências dos métodos de BDA. É também o que coloca a arquitetura atual de dois níveis em questão, exigindo uma arquitetura de dados adequada e orientada ao atendimento dos métodos de BDA. Na Figura 1 é possível verificar a evolução das arquiteturas até a proposta mais recente, o Data Lakehouse (LH):

Figura 1 - Evolução das arquiteturas de plataforma de dados até o modelo atual de duas camadas (a-b) e o novo modelo Lakehouse (c).



Fonte: Adaptado de ARMBRUST et al. (2021).

Data Lakehouse (LH) é uma arquitetura de dados que combina os principais benefícios dos DL e DW: armazenamento de baixo custo em um formato aberto acessível por uma variedade de sistemas do primeiro, e recursos avançados de gerenciamento e otimização do segundo. É especialmente adequado para ambientes de nuvem com computação e armazenamento separados, mas pode ser aplicado *on-premise* (ARMBRUST et al., 2021).

Em tom complementar, para SCHNEIDER et al. (2023), LH é uma plataforma de dados integrada que utiliza o mesmo tipo de armazenamento e formato de dados para relatórios e OLAP (*On–line Analytical Processing*, ou processamento analítico on-line), mineração de dados e aprendizado de máquina, bem como cargas de trabalho de streaming. Lakehouse é a resposta para BDA com armazenamento na nuvem, mas como toda tecnologia, reserva pontos de atenção (JANSSEN et al., 2022):

- A tecnologia ainda não alcançou maturidade, foi introduzida ao mercado em 2020, enquanto DW existem há 40 anos e o data lake por mais de 10 anos.
- Não existe muita pesquisa e ciência sobre estudos de caso de aplicação da tecnologia, o que aponta para uma sub-exploração de suas capacidades e pouco desenvolvimento.
- Por ser nova, exige o treinamento de novas habilidades e convencimento da mudança, logo, esforços de capacitação também são uma desvantagem.
- Finalmente, a latência dependente do armazenamento em nuvem pode se tornar uma desvantagem.

2.5. Data Lakehouse sob um olhar comparativo

Para chegar na definição de Lakehouse, evidenciar as diferenças das outras arquiteturas, e destacar as vantagens que ele traz para o negócio é preciso compreender suas fundações. Primeiramente, destaca-se que o Delta Lake se refere a:

> Uma camada de armazenamento de banco de dados ACID sobre armazenamento de objetos na nuvem que permite uma gama mais ampla de recursos de desempenho e gerenciamento semelhantes aos do banco de dados, em um armazenamento na nuvem de baixo custo (ARMBRUST et al., 2020).

É um diretório em um armazenamento de objetos na nuvem ou sistema de arquivos que contém objetos de dados com o conteúdo da tabela e um registro de operações de transação (com pontos de verificação ocasionais). O Delta Lake é a base de uma arquitetura Lakehouse econômica e altamente escalável. Para Schneider et al. (2023), Delta Lake está entre as poucas ferramentas que alcançam os requisitos necessários e permitem a construção de uma arquitetura integrada como o LH.

O Delta Lake apresenta para o ambiente de nuvem, característico do Data Lake, uma forte característica do Data Warehouse que é a propriedade ACID sobre as transações. No contexto de sistemas de armazenamento, uma transação é qualquer operação tratada como uma unidade de trabalho, que é totalmente concluída ou não, e deixa o sistema de

armazenamento em um estado consistente. ACID é um acrônimo que se refere ao conjunto de quatro propriedades principais que definem uma transação: atomicidade, consistência, isolamento e durabilidade. As transações ACID garantem a maior confiabilidade e integridade possíveis dos dados. Elas garantem que os dados nunca caiam em um estado inconsistente devido a uma operação concluída apenas parcialmente (DATABRICKS, [s.d.]).

Como estabelecido anteriormente neste trabalho, o Lakehouse é um sistema de gerenciamento de dados, potencializado pelo Delta Lake. A fim de posicionar o Lakehouse como solução ideal para uma arquitetura de dados moderna capaz de promover a exploração dos métodos de BDA, o Quadro 2 apresenta as principais diferenças entre as arquiteturas modernas.

Quadro 2 - Comparação entre DW, DL e LH.

Critérios	**Data Warehouse**	**Data Lake**	**Lakehouse**
Importância	Análise de dados e BI	ML e AI	Análise de dados, BI, ML e AI
Tipo de dado	Estruturado	Semi-estruturados e Não estruturados	Estruturado, Semi-estruturados e Não estruturados
Custo	Caro e demorado	Barato, rápido, e adaptável.	Barato, rápido, e adaptável.
Estrutura	Não configurável	Customizável	Customizável
Esquema	Definido antes do armazenamento	Desenvolvido depois que o dado é salvo	Desenvolvido depois que o dado é salvo
Usabilidade	Os usuários podem acessar e relatar dados facilmente	Analisar grandes quantidades de dados brutos sem ferramentas que classifiquem e cataloguem os dados pode ser árduo	Combina a estrutura e simplicidade de um DW com os casos de uso mais amplos de um DL
Conformidade ACID	Garante os mais altos níveis de integridade, os dados são registrados de forma compatível com ACID	Atualizações e exclusões são procedimentos difíceis que precisam de não conformidade com ACID	Compatível com ACID para garantir a consistência quando várias partes lerem ou gravarem dados ao mesmo momento
Qualidade	Alta	Baixa (pântano de dados)	Alta

Fonte: Harby e Zulkernine (2022).

Com o quadro acima torna-se claro que apenas o LH é capaz de habilitar plenamente os métodos de BDA apresentados, impactando positivamente a governança de dados ao garantir confiança e qualidade sobre os dados, e operando de uma maneira mais econômica. Uma vez que se decide por seguir com o LH como solução de gerenciamento de dados, partimos para as seções seguintes deste trabalho onde trabalhamos para encontrar conceitos que respondam às perguntas: "É possível mensurar o valor gerado por essa tecnologia? Se sim, como?" e "Como implementar essa tecnologia de uma maneira eficiente? Quais desafios podemos encontrar no percurso?", questões estas que fazem parte do objetivo deste trabalho.

2.6. Investir em dados gera valor para organizações, mas como medir?

O investimento em BDA é cada vez mais importante para uma empresa ser competitiva e inovadora. No entanto, medir o valor derivado desses investimentos pode ser um desafio. Neste tópico, serão exploradas as áreas em que os investimentos em dados geram valor, bem como discutidas as dificuldades de perceber e medir dados como um ativo econômico. Além disso, serão destacados os estudos atuais sobre medição de BDA e os métodos de medição mais relevantes usados na comunidade científica (GROVER et al., 2018).

De acordo com Manyka et al. (2021), BDA gera valor para negócios ao: prover informação em tempo real, permitir a personalização de produtos e serviços, aprimorar a tomada de decisão e iniciativas de inovação. Com similaridade, Davenport (2014) declara três classes de oportunidades de geração de valor com BDA: redução de custos, aprimoramento de decisões, e aprimoramento de produtos e serviços.

Para a análise dos benefícios citados acima, é necessária a mensuração precisa dos investimentos realizados em BDA, o que não é uma tarefa fácil. Em 2016, a Gartner conduziu uma pesquisa com líderes de TI e de negócios e reportou que 81 porcento não saberia afirmar em nome da companhia se os esforços relacionados a BD teriam retorno sobre investimento (ROI) positivo ou negativo (GROVER et al., 2018).

Há também a incerteza de como inferir o que é resultante da qualidade dos dados e o que é do gerenciamento. Além do caráter confuso da matéria-prima que alimenta BDA, o próprio campo acadêmico de BDA carece de clareza sobre métodos de medida, sobre quais são os itens e em qual dimensão devem ser medidos. Isso se torna ainda mais obscuro porque BDA pode ser medida como um processo, um projeto ou até mesmo como uma tecnologia.

A abordagem comum, fundamental, avalia BDA ou pela medição do uso da tecnologia (quanto cada recurso é utilizado e quanto custa o uso), ou por meio de indicadores financeiros

(retorno sobre investimento, redução de custos, realização de receita, etc.). A aferição de valor neste caso dá-se pela diferença do custo (de aquisição, armazenamento, tratamento) com a receita gerada através do uso dos dados (EARLEY, 2017). No entanto, reduzir a essas duas esferas é uma abordagem simplista e ignora as outras capacidades já abordadas neste texto. Há ainda, sob a categoria de benefícios tangíveis, a mensuração sob a perspectiva de desempenho do processo de BDA.

Figura 2 - Processo de BDA.

Fonte: Adaptado de Mohamed Ali et al. (2017).

A Figura 2 apresenta uma definição do processo de BDA que baseia o modelo sugerido por Mohamed Ali et al. (2017) para mensurar a performance do processo, apresentado no Quadro 2:

Quadro 2 - Medidas, Métricas e Indicadores para mensurar a performance do processo de BDA

Medidas	**Métricas**	**Indicadores**
Eficiência	Capacidade, velocidade de transferência, tempo de resposta, latência, acurácia, utilização do recurso	Entrada de dados são convertidas em ações / por período, tempo de ciclo / sobre (execução de processos, utilização do tempo de CPU da memória, armazenamento), precisão durante processamento dos dados, pontualidade na aquisição de dados, velocidade de transferência de dados pela rede.
Efetividade	Satisfação, interpretação de resultados, pontualidade, facilidade de entendimento, adequação a necessidades de negócio, personalização, confiança	Nível de satisfação do usuário com os resultados, como os resultados são representados, alcance das necessidades de negócio com os resultados, pontualidade dos resultados, compreensibilidade dos resultados.
Flexibilidade	Volume de entrada, possibilidade de modificação e personalização	• Manuseio dos volumes variáveis de entrada de dados (e.g. complexidade do processo, adição de novos processos, incorporação de novas ferramentas) • Os usuários podem escolher sua maneira de visualizar informações em gráficos ou forma tabular, num ecrã de computador ou num dispositivo portátil.
Tecnologia	Disponibilidade, maturidade, volatilidade, adequabilidade	• Ferramentas disponíveis (e.g. analíticas, armazenamento, programação) • A tecnologia cumpre com requerimentos • Incidência de mudanças de tecnologia • Quanto tempo anos/meses ferramentas estão em uso

Competência	Conhecimento tecnológico, conhecimento de processo, conhecimento de negócio, habilidades de comunicação	• Qualificações e experiências com ferramentas • Conhecimento de processo (habilidade para gerenciar o processo) • Compreensão das necessidades de negócio • Habilidade de comunicação • Habilidades computacionais
Compliance	Políticas e padrões, implementação, missão e objetivos, boas práticas	• Conformidade de implementação com processo de BDA • Conformidade com políticas de segurança, privacidade e confidencialidade • Observação das boas práticas em BDA • Conformidade com missão organizacional
Condições de trabalho	Motivação, conforto do ambiente de trabalho, carga de trabalho por colaborador.	• Disposição para performar trabalho processual • Conforto do ambiente • Capacidade física e intelectual do colaborador em performar o trabalho

Fonte: Adaptado de Mohamed et al. (2019).

Os indicadores são exemplos do que pode ser analisado sobre os fatores descritos em métricas a fim de alcançar uma avaliação da medida. Neste trabalho, percebem-se dimensões de análise que colocam o cliente (beneficiário final do processo de BDA) e o usuário (staff) como agentes de interesse da avaliação, que consideram visões organizacionais superiores ao citar missão organizacional e *compliance*, e que aferem benefícios tangíveis como a conta de *insights* acionáveis sobre a entrada de dados.

A fim de abordar também os benefícios intangíveis (ou simbólicos), há a abordagem de Jensen et al. (2023) defendendo que a aferição deve ser sobre o resultado. As medições dos benefícios devem se concentrar nos resultados do que a BD e a BDA produzem em conjunto para uma organização, e previamente deve-se estabelecer o que se espera alcançar com tais esforços. Portanto, estabelecer medidas apenas para o BD ou a BDA como um processo ou tecnologia pode ser insuficiente. A fim de reconhecer os benefícios intangíveis tal qual aqueles declarados em termos financeiros, apresenta-se o gerenciamento de benefícios como um modelo capaz de organizar o projeto de BDA de modo a monitorar o alcance dos benefícios propostos (WARD; DANIEL, 2012, p. 8, apud JENSEN et al., 2023). :

> O alto nível de complexidade no estabelecimento de medidas de benefícios exige uma visão mais ampla e uma investigação rigorosa sobre como estabelecer uma boa medida para a realização dos benefícios da BDA. A gestão de benefícios, no entanto, pode servir como um meio razoável para definir medidas de benefícios e, portanto, foi aplicada nesta pesquisa (JENSEN et al., 2023).

Entende-se que, para uma avaliação completa, é interessante enxergar BDA como um ativo (que custa e gera valor econômico a partir do acionamento), como um processo (cujo desempenho pode ser otimizado) e a partir dos benefícios esperados antes da sua aplicação.

Para compreender como alcançar sucesso com uma aplicação de BDA, resta abranger as boas práticas e desafios que podem ser acessados durante a implementação.

2.7. Recomendações para implementar uma tecnologia emergente

O claro entendimento do uso e das expectativas sobre a tecnologia são um desafio. É preciso que equipes técnicas e de negócios consigam alinhar as expectativas para alcançar a satisfação dos usuários finais.

Dessa forma, Tabesh et al. (2019) destacam que as dificuldades surgem tanto no âmbito técnico (que cobre a operação das ferramentas) quanto no gerencial - que trata da disponibilidade e motivação de talento, e perspicácia técnica sob níveis de gerência. Sobretudo, indica que as barreiras culturais são as que determinam o sucesso dos projetos de BDA. Uma organização sem a cultura orientada a dados tem altas probabilidades de falha. E modela um conjunto de recomendações para uma implementação bem sucedida:

Quadro 3 - Recomendações para implementação bem sucedida de estratégias de BD

Desafios que impedem os esforços de implementação	Responsabilidades gerenciais para uma implementação bem-sucedida	Recomendações
Tecnológico (T) • Ferramentas de gerenciamento de dados de alto custo • Carência de conhecimento gerencial sobre BDA • Desentendimento entre gerentes e técnicos • Desafios inerentes a natureza de BD • Cumprimento de questões de *compliance* e políticas de segurança da informação Cultural (C) • Ausência da compreensão disseminada sobre BD e seus objetivos • Grande dependência de abordagens de tomada de decisões intuitivas ou experimentais • Dominância da alta hierarquia no processo de decisão	Prover comprometimento e suporte	• Orientação de longo prazo sobre os investimentos em BDA (C, T) • Promover uma cultura de decisões orientadas a dados (C) • Atribuir times dedicados de dados providos de recursos, acesso e imunidade burocrática (C, T)
	Comunicação e coordenação efetiva	• Comunicar claramente o problema de negócio sendo endereçado pelo time técnico (C, T) • Criar entendimento comum dos objetivos de BD (C, T) • Incentivar a colaboração multifuncional (C, T) • Disseminar os *insights* gerados por dados e compartilhar os novos pontos de dados com a organização (C, T)
	Desenvolver perspicácia sobre gerenciamento de análise de negócios	• Obter e manter conhecimento relevante de *analytics* (T) • Ajudar as equipes a compreenderem de forma geral o ciclo de BDA e a agirem para contribuir com cultura orientada a dados (C, T) • Ajudar as equipes a desenvolver conhecimento estatístico (C, T) • Incentivar o ganho de conhecimento sobre BDA (T)

Fonte: Adaptado de Tabesh et al. (2019).

De acordo com Janusz Wielki (2013), são cinco os desafios mais significantes. Distribuir a informação em toda organização é o primeiro desafio pela existência dos silos organizacionais, que são equipes de pessoas que estão isoladas de outras partes da empresa devido a um fluxo mínimo de informações, operando separadamente uns dos outros, sem colaboração. O capital intelectual apropriado para explorar as oportunidades com dados configura o segundo obstáculo. A análise demorada de grandes conjuntos de dados aparece como um terceiro desafio. Em quarto, os autores mencionam as dificuldades de gerenciar e analisar dados não estruturados. E por fim, a percepção do valor de BD e BDA como ativo estratégico por parte da alta gerência descreve o último desafio.

2.8. Caracterização da empresa

Para a realização deste trabalho, a empresa solicitou que a marca não fosse revelada, e que as informações referentes a ela sejam utilizadas com descrição, portanto, a caracterização da empresa se dará com superficialidade para que ela não seja facilmente identificada. A empresa surgiu nos Estados Unidos, com o objetivo inicial de oferecer soluções de marketing digital para empresas de diferentes setores. Ao longo dos anos expandiu suas operações, adquirindo e desenvolvendo empresas em diversos segmentos, como tecnologia, telecomunicações, energia, serviços financeiros e muito mais.

No Brasil, iniciou suas atividades em 2015, estabelecendo-se como um player importante no mercado nacional. A empresa trouxe consigo uma abordagem inovadora, buscando combinar tecnologia, marketing digital e análise de dados para oferecer soluções personalizadas aos seus clientes. Ao longo do tempo, tem se destacado por sua capacidade de identificar oportunidades e adquirir empresas locais, integrando-as à sua estrutura e expandindo sua atuação no país.

Uma das principais características é sua atuação diversificada. A empresa busca estar presente em diferentes setores, explorando oportunidades de negócios e colaborando com o crescimento das empresas adquiridas. Essa estratégia tem possibilitado a expansão da companhia, tornando-a uma referência no mercado de soluções digitais no Brasil. Em relação às atividades, atua em diferentes áreas, como marketing digital, tecnologia da informação, análise de dados e serviços financeiros. Com uma equipe altamente qualificada e uma infraestrutura sólida, a empresa tem sido capaz de desenvolver soluções personalizadas para seus clientes, ajudando-os a alcançar seus objetivos de negócio.

3 PROCEDIMENTOS METODOLÓGICOS

3.1. Caracterização da pesquisa

A presente pesquisa adota uma abordagem que prioriza os elementos qualitativos, com suporte complementar, mas não fundamental de elementos quantitativos. A abordagem qualitativa é central neste trabalho, pois permitirá compreender e evidenciar os fenômenos e elementos que podem gerar dificuldades na implementação de um LH, assim como os que serão fatores de sucesso. Também evidenciará o padrão de comportamento e o sucesso da interação entre cargos de diferentes áreas de conhecimento e níveis hierárquicos. Assim como destaca Minayo (2009), a pesquisa permite ir além das questões materiais e acessar também o pensamento das pessoas envolvidas e a sua interpretação sobre o ocorrido.

Esta pesquisa assume também o caráter exploratório e descritivo. Exploratório porque trata de um tema de estudo muito recente com pouco material científico produzido ao seu respeito, onde as primeiras publicações surgiram em 2020 e, portanto, foi ainda pouco explorado (CASARIN, 2012). E descritivo decorrente do objetivo determinado que é de alcançar a produção de um guia de implementação que transcorre pelos fatores que determinam ou contribuem para o sucesso ou fracasso do projeto (GIL, 2002).

Faz parte também deste trabalho a pesquisa bibliográfica, que de acordo com Koche (2015), auxilia o investigador a tomar conhecimento sobre o tema e explorar o objeto de estudo com propriedade.

3.2. Sujeitos da pesquisa

Conforme Gil (2002), a população em uma pesquisa pode ser definida como o total de indivíduos de uma determinada classe que será avaliada, levando em consideração características específicas. Neste estudo, a população a ser investigada é composta pelos colaboradores da empresa que tiveram exposição tanto a arquitetura de dados legada (o conjunto de DW com DL) quanto a nova arquitetura implementada (LH). Foram escolhidos indivíduos que tiveram exposição em diferentes graus, com diferentes níveis de conhecimento sobre a nova tecnologia e que atuam em diferentes níveis hierárquicos na empresa. A amostragem utilizada pode ser classificada como amostragem não probabilística, na qual a seleção dos participantes depende de critérios definidos pelo pesquisador e consegue ser subdividida conforme tipicidade, acessibilidade e cota (GIL, 2008).

A seleção dos participantes foi baseada em critérios de conveniência e acessibilidade, permitindo entrevistas em locais e horários convenientes. Essa abordagem é comumente usada em pesquisas qualitativas e não requer precisão numérica (GIL, 2008). É importante reconhecer as limitações dessa abordagem e interpretar os resultados com cautela, considerando as perspectivas dos participantes selecionados.

3.3. Procedimentos e instrumentos de coleta de dados

Uma vez definido a dinâmica de seleção dos participantes, define-se o instrumento que viabilizará a realização do trabalho. Segundo Gil (1987), a entrevista consiste na aplicação de perguntas que podem ser pré-estruturadas ou fluídas conforme o objetivo da pesquisa, e permite interação social entre o pesquisador e o entrevistado, ocasionando numa coleta fluida e abrangente.

Neste estudo, utilizou-se a entrevista em profundidade como instrumento para coletar informações primárias. Essa abordagem envolve uma interação natural e semiestruturada com um pequeno grupo de entrevistados, conduzida por um entrevistador que possui conhecimento dos objetivos da pesquisa. O objetivo da entrevista em profundidade é obter um conhecimento mais aprofundado sobre o problema em questão, a partir da perspectiva do grupo-alvo (MALHOTRA, 2011).

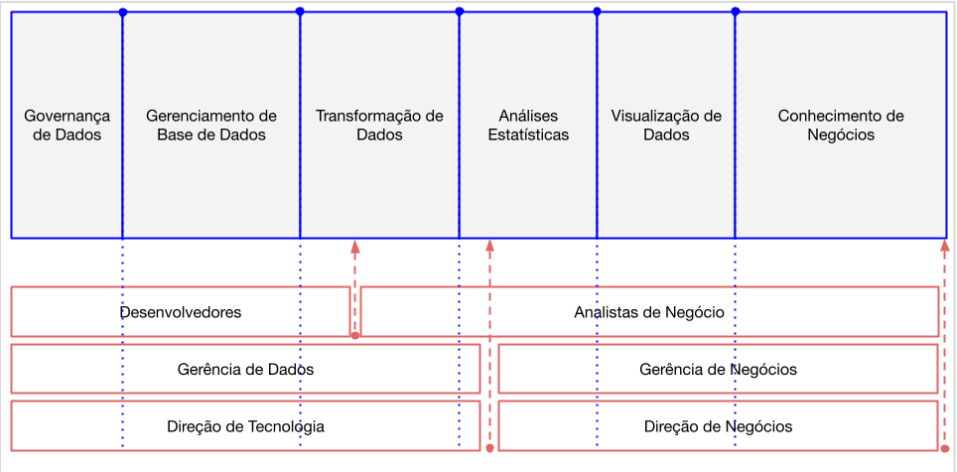
Em resumo, o questionário foi dividido em quatro partes: (1) perfil dos entrevistados, (2) percepção sobre a implementação, (3) identificação dos efeitos pós-implementação, e (4) oportunidades de melhoria sobre o processo. Visando conhecer as características dos participantes, a primeira parte investigou idade, gênero, cargo, tempo no cargo, nível hierárquico do cargo. As demais questões foram elaboradas pelo autor, baseando-se na fundamentação teórica deste trabalho, a fim de direcionar respostas completas e amplas sobre as dimensões avaliadas.

As entrevistas foram realizadas por meio da plataforma de videochamada Zoom e agendadas previamente conforme a disponibilidade dos entrevistados. Durante esse processo, as entrevistas foram gravadas e posteriormente transcritas, o que facilitou a análise das respostas pelo autor e também possibilitou a utilização de trechos específicos no desenvolvimento do trabalho. Os participantes foram designados para um ambiente confortável, sem distrações externas. Para o estudo foram convidados vários candidatos, mas apenas 7 candidatos preenchiam os pré-requisitos estabelecidos. Os critérios de seleção são os seguintes:

- Ter atuado direta ou indiretamente no projeto de implementação do lakehouse.
- Ter tido exposição a arquitetura anterior.
- Ser analista de negócio, gerente de negócio ou de tecnologia, ou desenvolvedor(a).

Na seleção de candidatos, foram selecionados os analistas de negócio por serem consumidores da infraestrutura de dados, realizando análises sobre a atuação das unidades de negócio e reportando erros ao time técnico. Os desenvolvedores por serem os responsáveis pela instalação, manutenção e criação de aplicações de dados. E os gerentes, tanto técnicos quanto de negócios, por serem estes os profissionais que olham para o resultado desses esforços, avaliando o funcionamento das aplicações e a geração de *insights*. Abaixo, na figura 3, consta uma simplificação do processo de BDA realizado na companhia, assim como os cargos atuantes por etapa. Na sequência, no quadro 4, a identificação dos profissionais entrevistados.

Figura 3 - Simplificação da operação de dados e atuação de agentes por etapa



Elaborado pelo autor.

Quadro 4 - Identificação dos entrevistados

Identificação numérica	Área de atuação	Cargo
Entrevistado 1, 2 e 3	Negócios	Analista
Entrevistado 4	Negócios	Gerência
Entrevistado 5 e 6	Técnica	Gerência
Entrevistado 7 e 8	Técnica	Desenvolvedor

Elaborado pelo autor.

3.4. Tratamento e análise dos dados

Após a coleta dos dados, inicia-se a análise para encontrar respostas para as perguntas da pesquisa. Ao contrário da análise quantitativa, a análise qualitativa depende da habilidade e do estilo do pesquisador, pois não segue uma abordagem sequencial predefinida (GIL, 2008). Nesse processo, entretanto, existem três etapas que podem auxiliar o pesquisador: redução, apresentação e conclusão.

A etapa de redução envolve selecionar e reduzir os dados, codificando-os conforme os temas definidos nos objetivos do estudo. Isso permite que as conclusões sejam construídas e validadas. Na fase de apresentação, os dados selecionados e codificados são sistematicamente organizados visando identificar suas diferenças, semelhanças e inter-relações. Por fim, a fase de conclusão, ou validação, requer um exame cuidadoso dos dados, buscando identificar significados, regularidades, padrões e explicações para os fenômenos observados (GIL, 2008).

Neste trabalho, a fim de alcançar os objetivos propostos, a análise será realizada sobre as dimensões de cargos entrevistadas: analistas (A), desenvolvedores (D) e gerentes (G). As questões, seus respectivos objetivos, e a qual cargo se destinam são demonstrados nos quadros abaixo

3.4.1. Roteiro de entrevista

Quadro 4 - Questões por objetivo específico e cargos questionados

Objetivo específico	
1. Identificar os fatores que levam ao sucesso no processo de implementação de uma data lakehouse	
Questão	**Cargos questionados**
Como foi o processo de transição para Lakehouse (Delta Lake) para você? Quais dificuldades você encontrou?	A/D/G
Quais foram as principais melhorias na capacidade de análise e geração de *insights* a partir dos dados após a migração para o data lakehouse? Você consegue apontar a melhora em alguma métrica?	A/G
Como você avalia a usabilidade da tecnologia e a adesão dos times a ela?	D/G
Houve alguma barreira técnica para realizar a migração?	D
Quais foram os principais passos e etapas seguidos durante o processo de implementação?	G
Quais são os desafios em equilibrar os investimentos em tecnologia de dados com a redução de custos?	G
2. Verificar quais os impactos após a implementação do data lakehouse	
Questão	**Cargos questionados**

Sobre o direcionamento da migração e a superação de desafios relacionados ao LH (Delta Lake): Qual sua percepção do papel da liderança, o que voce acha que foi bom, o que poderia ter sido melhor?	A/D
Quais os principais benefícios ou vantagens alcançadas com a adoção do Lakehouse (Delta Lake)?	D
Quais os pontos que você percebe que ainda podem ser melhorados em relação ao Lakehouse (Delta Lake)?	A
Qual o impacto dessa mudança no dia a dia da sua função?	A
Como você avalia a dependência do time de negócios para com o time de dados/técnico antes e depois da migração?	A/D
Qual a sua avaliação sobre o LH em comparação ao DW?	D
Que diferenças você percebeu no dia a dia após migração?	D
A transição para o lakehouse levou a alguma inovação de produtos ou serviços na empresa?	G
3. Elaborar um guia de implementação de um data lakehouse	
Questão	**Cargos questionados**
Me fale sobre a motivação da mudança e Como a empresa promoveu a migração para o Lakehouse (Delta Lake)?	A/D/G
Com base na experiência da empresa, quais recomendações você daria a quem deseja implementar um Data lakehouse?	A/D/G
O objetivo de negócio a ser alcançado com a migração para o LH esteve claro para você durante a implementação? Qual era?	D
Quais são os principais aspectos a considerar ao planejar a arquitetura e a infraestrutura para um Data lakehouse?	G
Quais são os principais desafios ou armadilhas que as lideranças devem estar cientes ao implementar um Data lakehouse?	G
Quais são os desafios enfrentados ao tentar monetizar os dados?	G

Fonte: Elaborado pelo autor.

A entrevista foi conduzida com flexibilidade, utilizando a ordem das perguntas para dar foco, mas permitindo desvios conforme a contribuição do entrevistado direcionasse para oportunidades de aprofundamento ou expansão do tema. O tratamento dos dados coletados foi realizado a partir da análise das entrevistas gravadas em vídeo, em busca de padrões, depoimentos e fenômenos que pudessem estruturar a visão sobre o processo de implementação, impactos e aprendizados para orientar um novo processo otimizado.

4. RESULTADOS

Com as organizações ganhando escala por meio de operações digitais, os custos para armazenar e analisar (processar) esses dados vem se tornando um grande obstáculo em termos

de custos. Encontrar alternativas escaláveis de arquitetura de dados, que não comprometam os custos, é o que pode desbloquear o potencial analítico e o ganho de eficiência em times de dados e de negócios. O entrevistado 6 demonstrou justamente essa preocupação com os custos, conforme é possível verificar:

> "Depois de algum tempo, você tem tantos dados, e tantas tabelas diferentes, que a performance costuma cair. E para melhorar, ou você paga mais, ou você cria mais trabalho. E do ponto de vista pragmático, você precisa tentar chegar numa estrutura mais eficiente, mais escalável, que não te obrigue a gastar mais só para otimizar consultas". (Entrevistado 6)

Além disso, a entrada do lakehouse no mercado não torna as arquiteturas anteriores obsoletas, o que ela faz, na verdade, é trazer uma nova abordagem para problemas de operação de grande escala. Como diz Schneider (2023), "parece ser do senso comum que a motivação fundamental é simplificar as arquiteturas analíticas empresariais existentes e reduzir a sua complexidade."

> "A gente se perguntou assim: será que não tem uma solução melhor para isso? Alguma que o processamento escale conforme a gente precisa? E aí surgiu a ideia do Lakehouse integrado com o Databricks. O Databricks já separa o armazenamento do processamento, e a gente consegue ser cobrado por uso, aumentando o processamento só nos momentos de pico. Foi essa motivação". (Entrevistado 4)

4.1. Processo de Implementação

Muitos desafios foram observados no processo de implementação, mas é possível resumi-los em duas dimensões: compreensão da tecnologia e engajamento de stakeholders. Ao ser indagado sobre sua impressão sobre o processo de implementação, o entrevistado 8 comentou:

> "Foi um pouco conturbada, tá? Ainda mais quando se fala de uma tecnologia nova que não tem tanto material, tanto suporte, tantas conexões com terceiros. E tivemos alguns desafios de cultura também, teve resistência das equipes de negócios". (Entrevistado 8)

O planejamento de implementação deve considerar essas duas dimensões e colocar o usuário final no centro do planejamento para realizar o processo com sucesso. No entanto, por tratar-se de uma tecnologia muito nova, fazer um planejamento assertivo também é uma questão que testa as lideranças.

> "Tivemos alguns tropeços no meio do caminho, o primeiro foi o pior, nos fez rever planejamento e tudo mais. Depois deu muito certo e foi muito bom em termos técnicos. Já na área de negócio tudo fica em adesão, muitas pessoas foram resistentes e continuaram criando coisas na *stack* antiga, ou seja, geraram mais trabalho para a migração. Tivemos que fazer um grande trabalho de conversar e nos aproximar das pessoas para garantir que elas estivessem trabalhando na nova *stack*". (Entrevistado 5)

Por outro lado, a resistência e barreiras de adesão não são encontradas do lado técnico da operação, como dito pelo entrevistado 7: “Barreiras? Não, praticamente nenhuma”. Reforçando uma percepção da liderança sobre o processo:

> "O time de tecnologia é apaixonado por inovação e gosta de mudanças. O time de negócios já apresentou mais resistência, a curva de aprendizado é maior, demorou para entender e abraçar a mudança". (Entrevistado 5)

Em seguida, para realizar a implementação, a empresa começou (i) desenhando uma abstração da arquitetura de dados ideal, e seguiu com as etapas de (ii) começar a migração dos principais processos sem desligar a operação original, (iii) documentar os aprendizados dos primeiros testes, (iv) validar a escrita de dados e por último (v) desligar a operação antiga. Como é possível visualizar na fala do entrevistado 6:

> "Em resumo é: planejar, migrar, e matar o legado. E tudo que surgir de novidade neste período, já começa a fazer na nova arquitetura". (Entrevistado 6)

Gerenciar projetos de dados, especialmente de tecnologias inovadoras, é um grande desafio para as lideranças. Segundo Tabesh et al. (2019) “As barreiras culturais, assim como as barreiras tecnológicas, podem prejudicar a aplicação efetiva de estratégias de big data.”, o que pode ser confirmado pela fala do entrevistado 4: "Teve resistência. Fomos muito perguntados se realmente precisava mudar e coisas do tipo, mas começamos a fazer os *office hours* para tirar a dúvida do pessoal. Eles começaram a entender mais e a resistência foi diminuindo. Não adianta só o time de dados saber a motivação e o resto não, sabe?".

A compreensão completa sobre a mudança é determinante para trazer agilidade à adesão da nova estrutura. Essa infraestrutura funciona melhor com a adesão absoluta das partes envolvidas. Uma vez que a equipe entende o funcionamento da estrutura e como se mensura a performance, torna-se possível mensurar e tirar conclusões sobre um uso ideal da aplicação.

> "O principal problema que tivemos foi a adaptação ao dialeto de SQL do Delta. Depois foi a questão do cluster, demorava para iniciar e desligava depois de um tempo em desuso, mas conforme o time foi entendendo e aderindo à ferramenta esse problema foi resolvido". (Entrevistado 6)

E ao tratar da compreensão da mudança, há aquilo que diz respeito à motivação da mudança, e àquilo que trata do que está sendo alterado propriamente. Na organização, a motivação esteve clara entre todos os agentes de diferentes perfis, mas a compreensão conceitual chegou com mais superficialidade na ponta da operação entre os analistas.

> "Custo era sempre o que mais se falava. Outra coisa interessante é poder trabalhar com dados organizados em diferentes níveis de refinamento, isso ajuda muito na

> hora de fazer uma análise porque fica tudo muito mais rápido. E na parte de governança temos mais oportunidades agora para explorar." (Entrevistado 7)
>
> "O grande objetivo era reduzir custos, e do lado mais técnico, melhorar a disponibilidade e a operação de dados" (Entrevistado 8)

Segundo o entrevistado 1,

> "No começo eu precisei entender o que é o Delta, né? E no começo isso era muito confuso. Então teve todo um primeiro passo de entender o que é, para depois entender como usa, e aí sim entender como isso gera valor para o negócio".

Os desafios observados no processo de implementação não comprometeram o alcance dos objetivos desejados tampouco a aplicação bem sucedida da tecnologia. É observado pelo entrevistado 4 que "O principal desafio é a curva de aprendizado. No caminho você encontra vários problemas e situações que você não conhecia antes e isso inevitavelmente vai aumentar os custos do projeto", que aponta para a importância do planejamento. Segundo Earley (2017):

> "O planejamento de melhores dados requer uma abordagem estratégica à arquitetura, modelação e outras funções. Depende também de uma colaboração estratégica entre a direção comercial e a direção de TI. E, claro, depende da capacidade de execução eficaz de projetos individuais. O desafio é que existem normalmente pressões organizacionais, bem como as eternas pressões de tempo e dinheiro, que impedem um melhor planejamento."

4.2. Impactos após Implementação

O principal benefício que se busca alcançar com o Lakehouse é uma redução de custos significativa em comparação às arquiteturas de dados anteriores. A companhia conseguiu alcançar e medir o sucesso sobre essa métrica.

> "A mais fácil de mensurar é o custo, que já se provou real, mas os outros aspectos são mais difíceis. O quanto a gente melhorou em processos não é algo mensurado e também não conseguimos comparar com o passado, apesar de sabermos que hoje é melhor" (Entrevistado 5)

Segundo o entrevistado 4,

> "A principal métrica que acompanhamos é o custo, e tem sido um sucesso. Também tem uma questão de uso, com o lakehouse usando databricks nós conseguimos separar processamento de armazenamento, usar straming e batch ao mesmo tempo, e consegue criar e testar tabelas muito mais rápido do que no Redshift".

Os benefícios ligados à operação foram percebidos entre os analistas e desenvolvedores, destacando a centralização dos recursos de análise numa única ferramenta, e a possibilidade de utilizar Python para análises mais refinadas no mesmo ambiente das consultas ao banco de dados feitas em SQL.

> "É tudo uma questão de performance. Quando você roda queries mais rápidas, você faz mais análises, você não perde o ritmo de trabalho." (Entrevistado 3)

Por outro lado, a ausência de uma abordagem científica sobre a operação impossibilitou uma análise quantitativa da mudança, restando apenas a impressão qualitativa sobre a praticidade advinda da nova tecnologia. Segundo o entrevistado 1:

> "Em questão de performance eu não medi né? Não consigo falar em números, mas ter um lugar centralizado para as análises é ótimo. Antes tinha que consultar num lugar, extrair uma tabela, ou então criar uma tabela nova, para conectar com a ferramenta de BI. Agora não, agora eu gero uma conexão e pronto, faço todas as análises e visualizações ali."

Segundo o entrevistado 8: "O Lakehouse é muito superior (em comparação ao Warehouse), principalmente na questão de custos. Trazendo a parte de processamento, você tem uma série de funcionalidades para otimizar processamento, várias features para o trabalho analítico, tem versionamento dos dados e tabelas. No geral, você tem um controle de como opera a ferramenta muito melhor", uma arquitetura com essas atribuições possibilita que o gerenciamento de dados atinja níveis mais maduros. Quando numa infraestrutura de dados simplificada, times de negócios são potencializados por terem maior liberdade e autonomia para realizar análises, levando também a uma dedicação mais especializada da equipe técnica.

> "Eu senti que as equipes se tornaram mais independentes, uma vez que aprenderam como fazer algumas coisas, os pedidos de ajuda diminuíram." (Entrevistado 7)

Essa percepção, no entanto, não é homogênea, como afirma o entrevistado 8:

> "Não houve mudança. A dependência varia e está relacionada com quão interessada a pessoa de negócios é pelo entendimento da ferramenta, da linguagem, etc."

Com uma infraestrutura simplificada, a escalabilidade fica ao alcance da companhia sem comprometer os custos, como a fala do entrevistado 7 reforça: "Outra coisa muito legal é a escalabilidade, o negócio vai crescer e com isso vai aumentar o volume de dados, mas o custo vai continuar baixo". E isso se aplica também a escala do trabalho individual, analistas conseguem produzir mais neste formato.

> "Eu tive que fazer análises exaustivas esse semestre que se fosse no ambiente antigo seria muito mais demorado." (Entrevistado 3)

Outro aspecto a se investigar é o quanto essa mudança é capaz de promover inovação. A percepção da inovação causada pela implementação do Lakehouse não é homogênea. Por um lado, a nova estrutura facilitou a aplicação de práticas mais avançadas como o uso de Machine Learning, mas esta prática não estava impossibilitada pela stack anterior, apenas sob a necessidade de mais trabalho para ser aplicado. Como se pode conferir pela fala do

entrevistado 5: "Essa nova tecnologia abre um pouco a nossa cabeça. Já temos alguns exemplos de uso de machine learning, ainda embrionários, mas já começamos". Segundo o entrevistado 6,

> "A transição em si não levou a nenhuma inovação de negócios, mas o fato de termos aproveitado esse momento para organizar a stack, criar padrões, novos recursos, etc. Isso beneficiou bastante o negócio, e poderia ter ocorrido com ou sem a implementação"

Por fim, há ainda um benefício sob a perspectiva comercial. Estar a frente de uma tecnologia inovadora abre portas para oportunidades de novos negócios, como destacado na fala do entrevistado 4: "Por ser uma tecnologia nova, é muito interessante para o fornecedor provar as suas vantagens, então nós ganhamos muita visibilidade com isso".

4.3. Aprendizados para uma implementação mais eficiente

Os relatos dos entrevistados apontam que a espinha dorsal de uma boa implementação é o planejamento, e para bem orientá-lo é preciso colocar o usuário no centro. A boa execução do que foi planejado será fruto de práticas apresentadas por Tabesh et al. (2019)

> "A comunicação eficaz dos objetivos de *big data* pode atenuar os obstáculos culturais e tecnológicos à implementação bem sucedida de estratégias de *big data"*.

O engajamento do usuário, seja consumidor final ou agente da implementação, deve ser tomado em consideração desde o começo. Fazê-lo se envolver com o processo e compreender a mudança torna tudo mais fluido. Segundo o entrevistado 2,

> "O primeiro e mais importante é garantir que as pessoas que vão consumir disso, utilizar o produto dessa mudança, tenham tempo para se envolver com a migração e estejam alinhadas com prazos. Fazer uma mudança de dados sem estar claro para todo mundo não rola"

Além dos esforços de comunicação, destaca-se a elaboração de um bom cronograma que respeite também a etapa de validação de dados, como visto na fala do entrevistado 8:

> "Acho que estruturar um roadmap bem claro do que será feito e quais são os prazos, é o melhor que se pode fazer." (Entrevistado 8)

Além disso, segundo o entrevistado 4,

> "A principal dica é dar bastante tempo para validar se os dados estão certos. Não dá pra confiar que vai ser acertado tudo de primeira, também é bom colocar uma data de corte para diminuir o processamento, por exemplo: os dados serão migrados de uma data tal pra frente, e daí partir para validação"

Ainda em etapa de planejamento devem surgir as considerações técnicas, segundo o entrevistado 6:

> "O primeiro é avaliar se a mudança para um Lakehouse vai exigir que você mude a sua stack ou não. E tem que ter uma operação com escala suficiente para justificar fazer essa mudança. E no fim do dia, dá pra tentar estimar os ganhos com a mudança, mas vale mais a pena mensurar depois que você migrou"

Na sequência, o estudo da ferramenta deve ser focado na aplicação de boas práticas. As funções de *optimize*, *vacuum*, e de otimização de consultas devem ser estudadas ainda na etapa de planejamento do processo para fazer uma migração alinhada visando redução de custos.

> "O Lakehouse tem que ser mais barato, mas para isso é preciso seguir algumas boas práticas, se ele não for gerenciado com atenção, pode até sair mais caro." (Entrevistado 4)

Havia um ponto focal capaz de coordenar e transmitir segurança sobre o assunto, e isso foi reconhecido como um ponto positivo por analistas:

> "Foi muito bom ter alguém que você possa aprofundar a conversa sobre um determinado assunto" (Entrevistado 2)

Entretanto, a empresa poderia ter se beneficiado de um planejamento mais refinado que considerasse a atribuição do projeto a mais pessoas com escopo limitado. Foi notado desvio de foco entre todos os perfis durante a execução, resultando em atrasos e demandas ostensivas sobre desenvolvedores, como a fala do entrevistado 5 afirma: "Olhar sempre para o objetivo e ter as pessoas compradas na ideia. É comum em projetos longos ocorrer afastamentos do objetivo, mas é papel da liderança trazer o time para o curso certo".

> "Com a motivação de reduzir custos, e todas as outras mudanças que a empresa estava passando, nós ficamos com poucas pessoas e uma esfera muito grande de prioridades quando deveria ser uma esfera pequena."(Entrevistado 8)

Por fim, trabalhar sobre um projeto de inovação agrupa os desafios de gerenciamento de projetos e gerenciamento de mudanças, como diz Earley (2017) “Este nível de mudança não é conseguido através da tecnologia, embora a utilização adequada de ferramentas de software possa apoiar a sua realização. Em vez disso, é alcançado mediante uma abordagem cuidadosa e estruturada da gestão da mudança na organização”. Ou seja, não basta a correta operação da parte ferramental, tampouco se pode ignorar o fator humano e realizar uma mudança estrutural de forma abrupta.

> "É uma coisa inovadora né, não temos muito claro o retorno real daquilo. Temos que trabalhar com muitas projeções para provar que aquele investimento não é tecnologia por tecnologia, que é a tecnologia utilizada para um benefício de negócio." (Entrevistado 5)

Segundo o entrevistado 6,

> "Uma coisa que falta muito em empresa de tecnologia é uma noção de gerenciamento de projeto. Você pode fazer uma solução perfeita, que resolve todos os problemas da empresa, mas se ela for muito cara, ou mal implementada, não vai adiantar nada." (Entrevistado 6)

É de interesse da companhia, já no momento de idealização, formular maneiras de mensurar de forma quantitativa o valor agregado do projeto. Dados geram valor, pois geram conhecimento através da informação, todavia, a monetização desse valor não é muito clara dada a natureza intangível do ativo, e da cultura de uma operação de dados vista como área de suporte em empresas de tecnologia.

> "Esse é um super desafio. Dados é uma área de muitos custos, e quando você auxilia com a geração de receita em alguma área, isso é atribuído àquela área e não a dados. No final, nós que tentamos mensurar quanto trouxemos de resultado, não tem um modelo estabelecido." (Entrevistado 4)

Há um espaço em branco na produção de conhecimento sobre como atribuir o valor gerado por operações de dados aos resultados de negócio. É difícil porque a operação de dados sustenta as decisões de negócio, sustenta a operação comercial e promove o conhecimento, mas a atribuição direta disso aos resultados não é clara e também não é um conhecimento bem distribuído, como afirma o entrevistado 5: "Acho que os desafios que temos hoje é o desafio de mensurar algo que é novo, temos que criar muitas coisas da nossa cabeça. Não tem muitos cases para buscar ou pessoas que trabalharam com isso. Nós temos que criar a nossa maneira de resolver."

5. CONSIDERAÇÕES FINAIS

O presente trabalho teve como objetivo geral analisar os impactos e o processo de implementação de um data lakehouse em uma empresa de tecnologia, considerando a disciplina de gerenciamento de dados sinergética com a formação de um administrador e observando a compreensão de arquiteturas de dados como um conhecimento capaz de potencializar a produção de análises em operações com alto volume de dados sem comprometer os custos da empresa. Neste capítulo, a fim de alcançar os objetivos específicos, apresenta-se uma análise sobre os fatores que levam ao sucesso no processo de implementação do Lakehouse, quais os impactos observados após a implementação, e a elaboração de um guia de implementação de um data lakehouse.

A partir do relato dos entrevistados é possível concluir que os desafios de implementação se encontrarão sob três principais dimensões: compreensão da tecnologia, engajamento de stakeholders, e gerenciamento do projeto. As primeiras dimensões são enfatizadas entre os analistas, neste caso, os consumidores do produto final desta mudança.

Quando uma liderança tornou-se um ponto focal da comunicação, estabelecendo rotinas de comunicação e suporte, os obstáculos ocasionados pela mudança foram superados.

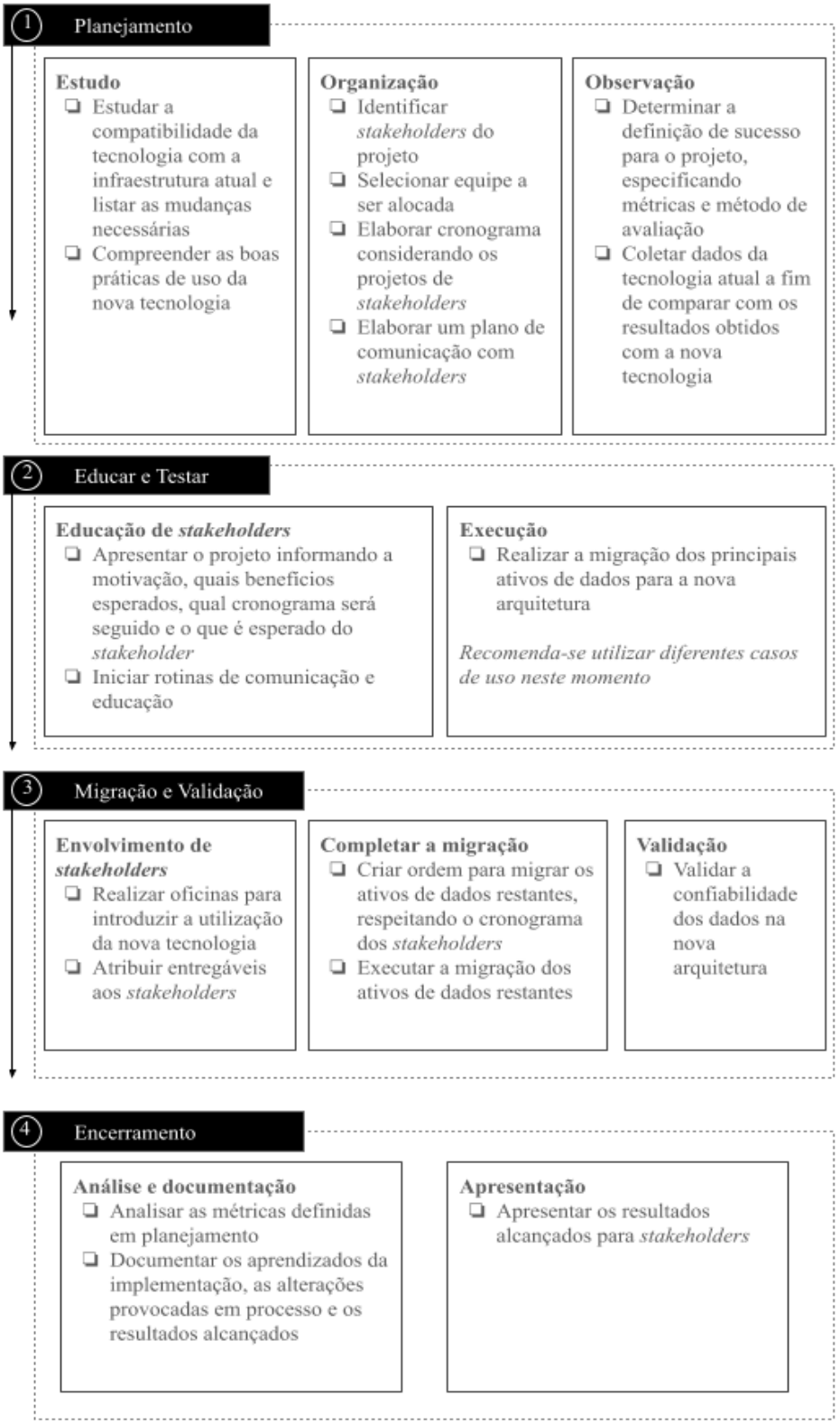
A participação de todos os *stakeholders* neste tipo de projeto é determinante para alcançar sucesso, apenas a exposição da motivação originada pela liderança é insuficiente para uma implementação ágil e assertiva, é necessário envolvimento regular e atenção à educação dos stakeholders. Considerando que uma tecnologia recentemente lançada ainda estará produzindo documentações e exemplos da sua aplicação, cai sobre a liderança um papel importante de facilitar a transmissão do conhecimento técnico para perfis de negócios, auxiliando-os a enxergar e utilizar os novos recursos.

Quanto a dimensão de gerenciamento de projetos, houve um grande foco na redução de custos da infraestrutura de dados, objetivo alcançado, mas pouco se pode aferir sobre como esses custos reduziram, quais as melhorias quantitativas ao nível de processo, e como as funcionalidades mais refinadas de análise estavam atribuídas a geração de valor. Empresas de tecnologia podem se beneficiar do método científico, observando a infraestrutura e os processos que serão impactados pela mudança, declarando métricas de sucesso e planejando a coleta de dados que comprovem o atingimento dos objetivos do projeto.

Os benefícios, ao nível de custo e operação, prometidos pelo Lakehouse foram observados com sucesso por todos os perfis. A gama de possibilidades é tamanha que em algumas funcionalidades mais avançadas nem mesmo eram de conhecimento de todos. Muitas portas se abriram para a organização ao trabalhar com uma tecnologia de vanguarda.

Por fim, podemos concluir que uma implementação bem sucedida passa por um (i) planejamento refinado, que deve incluir um estudo sobre as mudanças necessárias da infraestrutura de tecnologia, a definição de sucesso que orientará o projeto, um plano de comunicação e educação dos *stakeholders*, um cronograma alinhado com o cronograma das operações dependentes, e a seleção de uma equipe dedicada para o projeto. Em seguida, (ii) realiza a migração dos principais ativos de dados, realizando testes e colocando diferentes casos de uso em validação, para então (iii) realizar a migração dos ativos restantes e iniciar a (iv) validação dos dados, a fim de garantir confiabilidade sobre a nova operação. Por último (v) deve-se coletar os dados a fim de analisar se as definições de sucesso idealizadas na etapa de planejamento foram alcançadas, e encerrar a operação da tecnologia anterior. A fim de guiar gestores na implementação do Lakehouse, o infográfico abaixo apresenta fatores-chave a serem endereçados em cada etapa.

Figura 4 - Guia para implementação de um Data Lakehouse



Elaborado pelo autor.

REFERÊNCIAS

ARMBRUST, M. et al. Delta Lake: High-Performance ACID Table Storage over Cloud Object Stores. **Proceedings of the VLDB Endowment**, v. 13, n. 12, 2020.

ARMBRUST, M. et al. **Lakehouse: a New Generation of Open Platforms That Unify Data Warehousing and Advanced Analytics.** 11th Annual Conference on Innovative Data Systems Research (CIDR'21). **Anais**… 1 jan. 2021.

BROWN, S. **Machine Learning Explained**. Disponível em: <https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained>.

DATABRICKS. **What Is a Lakehouse?** Disponível em: <https://www.databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html#:~:text=A%20lakehouse%20gives%20you%20data>. Acesso em: 28 maio. 2023.

DATABRICKS. **What Are ACID Transactions?** Disponível em: <https://www.databricks.com/glossary/acid-transactions>. Acesso em: 30 maio. 2023.

DAVENPORT, T. H. **Big Data @ Work : Dispelling the myths, Uncovering the Opportunities**. Boston, Mass.: Harvard Business Review Press, 2014.

DE MAURO, A.; GRECO, M.; GRIMALDI, M. A Formal Definition of Big Data Based on Its Essential Features. 4 abr. 2016.

EARLEY, S. **DAMA-DMBOK : Data Management Body of knowledge**. 2. ed. Basket Ridge, New Jersey: Technics Publications, 2017.

GROVER, V. et al. Creating Strategic Business Value from Big Data Analytics: a Research Framework. **Journal of Management Information Systems**, v. 35, n. 2, p. 388–423, 3 abr. 2018.

HARBY, A. A.; ZULKERNINE, F. **From Data Warehouse to Lakehouse: a Comparative Review**. IEEE Xplore. **Anais**...1 dez. 2022. Disponível em: <https://ieeexplore.ieee.org/document/10020719>. Acesso em: 1 fev. 2023

INMON, B.; LEVINS, M.; SRIVASTAVA, R. **Building the Data Lakehouse**. [s.l.] Technics Publications, 2021.

JAIN, P. et al. **Analyzing and Comparing Lakehouse Storage Systems**. [s.l: s.n.]. Disponível em: <https://www.cidrdb.org/cidr2023/papers/p92-jain.pdf>. Acesso em: 30 maio. 2023.

JANSSEN, N. et al. **The Evolution of Data Storage Architectures: Examining the Value of the Data Lakehouse**. [s.l: s.n.]. Acesso em: 31 maio. 2023.

JENSEN, M. H.; PERSSON, J. S.; NIELSEN, P. A. Measuring Benefits from Big Data Analytics projects: an Action Research Study. **Information Systems and e-Business Management**, v. 21, n. 323-352, 28 fev. 2023.

MANYIKA, J. et al. **Big data: the next Frontier for innovation, competition, and Productivity | McKinsey**. Disponível em: <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation>.

MCAFEE, A.; BRYNJOLFSSON, E. **Big Data: The Management Revolution**. Disponível em: <https://hbr.org/2012/10/big-data-the-management-revolution>.

MCCARTHY, J. **What is artificial intelligence?** Disponível em: <https://www-formal.stanford.edu/jmc/whatisai.html>. Acesso em: 27 maio. 2023.

MIT TECHNOLOGY REVIEW. **Building a high-performance Data and AI Organization**. Disponível em: <https://www.technologyreview.com/2021/04/15/1022754/building-a-high-performance-data-and-ai-organization/>. Acesso em: 1 jun. 2023.

MOHAMED ALI, I. et al. A Conceptual Framework for Measuring the Performance of Big Data Analytics Process. **Acta Informatica Malaysia**, v. 1, n. 2, p. 13–14, 7 dez. 2017.

MOHAMED, I. et al. Measuring the Performance of Big Data Analytics Process. **Article in Journal of Theoretical and Applied Information Technology**, v. 31, n. 14, 2019.

MOHRI, MEHRYAR; AFSHIN ROSTAMIZADEH; AMEET TALWALKAR. **Foundations of Machine Learning**. Cambridge, Massachusetts: The Mit Press, 2018.

PROVOST, F.; FAWCETT, T. **Data science for business**. Beijing: O'reilly, 2013.

REIS, J.; HOUSLEY, M. **Fundamentals of Data Engineering**. [s.l.] "O'Reilly Media, Inc.", 2022.

RUSSEL, S.; NORVIG, P. **Artificial Intelligence : A Modern approach.** 4. ed. [s.l.] Prentice Hall, 2020.

SCHNEIDER, J. et al. **Assessing the Lakehouse: Analysis, Requirements and Definition:** SCITEPRESS - Science and Technology Publications, 2023. Disponível em: <https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0011840500003467>. Acesso em: maio. 2023

TABESH, P.; MOUSAVIDIN, E.; HASANI, S. Implementing Big Data strategies: a Managerial Perspective. **Business Horizons**, v. 62, n. 3, p. 347–358, maio 2019.

TSAI, C.-W. et al. Big data analytics: a survey. **Journal of Big Data**, v. 2, n. 1, 1 out. 2015.

ZAGAN, E.; DANUBIANU, M. **Cloud DATA LAKE: the New Trend of Data Storage**. 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). **Anais**...11 jun. 2021.

CASARIN, Helen de Castro S.; CASARIN, Samuel S.Pesquisa científica: da teoria à prática. Curitiba: Ed. Intersaberes, 2012.

GIL, Antônio Carlos.Métodos e técnicas de pesquisa social. 2. ed. São Paulo: Atlas, 1989.

GIL, Antônio Carlos. Como elaborar projetos de pesquisa. 4. ed. - São Paulo : Atlas, 2002.

KÖCHE, José Carlos. Fundamentos de Metodologia Científica: teoria da ciência e iniciação à pesquisa. 34. ed. Petrópolis, RJ: Vozes, 2015.

MINAYO, Maria Cecília de S. O desafio da pesquisa social. In: ____ (org.); DESLANDES, Suely F.; GOMES, Romeu. Pesquisa social: teoria, método e criatividade. 28. ed. Petrópolis, RJ: Vozes, 2009, p. 9-29.