UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Richard de Arruda Felix

Abordagem para implantação de produtos de dados de arquivo em lote em um ambiente de big data

Florianópolis
2026

Richard de Arruda Felix

Abordagem para implantação de produtos de dados de arquivo em lote em um ambiente de big data

Dissertação submetida ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Santa Catarina para a obtenção do título de mestre em Ciência da Computação.
Orientador: Profa. Patricia Della Méa Plentz, Dra.
Coorientador: Prof. Jean Carlo Rossa Hauck, Dr.

Florianópolis
2026

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Felix, Richard de Arruda
Abordagem para implantação de produtos de dados de arquivo em lote em um ambiente de big data / Richard de Arruda Felix ; orientadora, Patricia Della Méa Plentz, coorientador, Jean Carlo Rossa Hauck, 2026.
80 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Ciência da Computação, Florianópolis, 2026.

Inclui referências.

1. Ciência da Computação. 2. Processamento em lote. 3. Big data. 4. Engenharia de Software. 5. Pesquisa-Ação. I. Plentz, Patricia Della Méa. II. Hauck, Jean Carlo Rossa. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Computação. IV. Título.

Richard de Arruda Felix

Abordagem para implantação de produtos de dados de arquivo em lote em um ambiente de big data

O presente trabalho em nível de mestrado foi avaliado e aprovado, em 11 de dezembro de 2025, pela banca examinadora composta pelos seguintes membros:

Profa. Patricia Della Méa Plentz, Dra.
Universidade Federal de Santa Catarina

Prof. Davi Viana dos Santos, Dr.
Universidade Federal do Maranhão

Prof. Frank Siqueira, Dr.
Universidade Federal de Santa Catarina

Prof. Renato de Freitas Bulcão Neto, Dr.
Universidade Federal de Goiás

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de mestre em Ciência da Computação.

Documento assinado digitalmente
CARINA FRIEDRICH DORNELES
Data: 03/03/2026 11:23:07-0300
CPF: ***.834.250-**
Verifique as assinaturas em https://v.ufsc.br

Coordenação do Programa de
Pós-Graduação

Documento assinado digitalmente
PATRICIA DELLA MEA PLENTZ
Data: 18/02/2026 16:44:20-0300
CPF: ***.243.700-**
Verifique as assinaturas em https://v.ufsc.br

Profa. Patricia Della Méa Plentz, Dra.
Orientador

Florianópolis, 2026.

AGRADECIMENTOS

Inúmeras pessoas contribuíram, direta ou indiretamente, para o desenvolvimento do presente trabalho, tornando este período do mestrado extremamente enriquecedor para mim.

Agradeço, em especial, aos meus orientadores, Professora Patricia Della Méa Plentz e Professor Jean Carlo Rossa Hauck, pela orientação dedicada, pelas contribuições valiosas e, sobretudo, pela confiança em meu trabalho. A generosidade com que compartilharam seu conhecimento e a precisão com que nortearam cada etapa deste percurso foram fundamentais para que este projeto se concretizasse.

Por fim, dedico este trabalho à minha esposa, Marina, pelo apoio e paciência em todos os momentos desta jornada. Sua presença constante, incentivo e compreensão foram essenciais para que eu superasse os desafios do mestrado. Cada conquista aqui também é sua. E claro, Maggie a shitzu que mais entende de *big data* que eu conheço, por me fazer rir e relaxar nos momentos mais tensos.

RESUMO

A ciência de dados consolidou-se como pilar para a tomada de decisões em diversos setores, impulsionada pela análise de grandes volumes de dados. A fase de implantação de produtos de dados, que são artefatos tangíveis e reutilizáveis gerados a partir de dados e algoritmos, como modelos preditivos ou relatórios, frequentemente apresenta desafios. Em particular, a implantação de produtos de dados em lote (*batch*), como o cálculo periódico de *scores* de crédito, que opera sobre grandes volumes em intervalos fixos, tende a receber pouca atenção metodológica, resultando em ineficiências e retrabalho. Este estudo propõe a Abordagem Integrada de Implantação de Produtos de Dados (AIIPD), que integra práticas de engenharia de software e métodos ágeis para padronizar e otimizar a implantação de produtos de dados em lote em ambientes de *big data*. A AIIPD foi aplicada e avaliada por meio de pesquisa-ação conduzida em um *bureau* de crédito brasileiro, com implantação em produção. Os resultados quantitativos e qualitativos indicam redução significativa no tempo de implantação e melhoria na colaboração entre as equipes de Produto, Dados & Analytics, Tecnologia e Operação & Delivery. Observou-se ainda maior clareza de responsabilidades, rigor em validação e testes, e aprimoramento da documentação. Conclui-se que a integração de práticas estruturadas de engenharia de software com a flexibilidade dos métodos ágeis oferece uma solução escalável e robusta aos desafios do processamento em lote, contribuindo com um modelo prático que pode servir de referência para organizações que buscam otimizar a fase de implantação de projetos de ciência de dados.

Palavras-chave: Processamento em lote. *Big data*. Engenharia de Software. Metodologias Ágeis. Pesquisa-Ação.

ABSTRACT

Data science has become a cornerstone of decision making across sectors, driven by the analysis of large data volumes. The deployment phase of data products, defined as tangible and reusable artifacts generated from data and algorithms such as predictive models or reports, often presents challenges. In particular, the deployment of batch data products (for example, periodic credit score computation), which operate on large datasets at fixed intervals, tends to receive limited methodological attention, leading to inefficiencies and rework. This study proposes the Integrated Approach for Data Product Deployment (AIIPD), which integrates software engineering practices and agile methods to standardize and optimize the deployment of batch data products in big data environments. The AIIPD was applied and evaluated through action research conducted at a Brazilian credit bureau, culminating in a production deployment. Quantitative and qualitative results indicate a significant reduction in deployment time and improved collaboration among Product, Data & Analytics, Technology, and Operations & Delivery teams. Additional outcomes include clearer responsibilities, stronger validation and testing, and improved documentation. It is concluded that combining structured software engineering practices with agile flexibility provides a scalable and robust response to the challenges of batch processing, offering a practical model for organizations seeking to optimize the deployment phase of data science projects.

Keywords: Batch Processing. Big data. Software Engineering. Agile Methodologies. Action Research.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABBA	*Architecture-centric Agile Big data Analytics*
BDD	*Behavior Driven Development*
BPMN	*Business Process Model and Notation*
CRISP-DM	*Cross Industry Standard Process for Data Mining*
DDS	*Data Driven Scrum*
DevOps	*Development* (Desenvolvimento) e *Operations* (Operações)
ECL	*Enterprise Control Language*
ETL	*Extract, Transform, Load*
GQM	*Goal Question Metric*
HDFS	*Hadoop Distributed File System*
HPC	*High Performance Computing*
LDTM	*Lean Design Thinking Methodology*
MPSBR	Melhoria de Processo do Software Brasileiro
OpenUP	*Open Unified Process*
RUP	*Rational Unified Process*
SDLC	*Software Development Life Cycle*
SPEM	*Software & Systems Process Engineering Metamodel*
TDD	*Test Driven Development*
TDSP	*Team Data Science Process*
UP	*Unified Process*
XP	*Extreme Programming*

SUMÁRIO

1 INTRODUÇÃO

A ciência de dados tem se consolidado como uma prática essencial em diversos setores, como governo, saúde e finanças, ao viabilizar decisões informadas baseadas na análise de grandes volumes de dados. O avanço das ferramentas analíticas, aliado à crescente digitalização dos serviços, ampliou a capacidade das organizações de transformar dados em ativos estratégicos capazes de influenciar diretamente indicadores operacionais, táticos e estratégicos. Nesse cenário, a habilidade de extrair, interpretar e operacionalizar informações de forma eficiente tornou-se não apenas desejável, mas fundamental para organizações públicas e privadas que buscam competitividade em mercados cada vez mais dinâmicos e orientados por dados (Marz; Warren, 2015). Além disso, a consolidação de práticas de governança e qualidade da informação contribuiu para que a ciência de dados se tornasse um componente estrutural no ciclo de tomada de decisão em larga escala.

A evolução das tecnologias de *High Performance Computing* (HPC) e das arquiteturas de *big data* possibilitou o surgimento de projetos capazes de transformar dados brutos em dados acionáveis, influenciando desde estratégias de marketing até inovações em produtos e decisões financeiras (Grolinger et al., 2013). Ambientes distribuídos de processamento viabilizaram o uso de datasets massivos, operações paralelas e pipelines altamente complexos, ampliando de forma significativa a capacidade das instituições de realizar análises mais abrangentes e sofisticadas. Para orientar esses processos, diferentes frameworks foram desenvolvidos ao longo dos anos com o intuito de guiar o ciclo de vida da ciência de dados, como o *Cross Industry Standard Process for Data Mining* (CRISP-DM) (Chapman et al., 2000), considerado um dos modelos mais consolidados na indústria, e o *Lean Design Thinking Methodology* (LDTM) (Ahmed; Dannhauser; Philip, 2018), que propõe uma organização mais moderna e alinhada a práticas ágeis.

Apesar de sua ampla adoção, essas metodologias são reconhecidamente focadas nas etapas iniciais do processo, como compreensão do negócio, preparação dos dados e modelagem, e fornecem pouca orientação prática para a fase de implantação em produção (Cunha et al., 2021; Bender-Salazar, 2023). Essa lacuna metodológica torna-se crítica justamente no momento em que produtos de dados como *scores* de crédito, classificadores e modelos preditivos precisam ser integrados a sistemas corporativos em escala, atendendo requisitos rigorosos de confiabilidade, rastreabilidade, segurança e desempenho. A falta de diretrizes claras para essa etapa gera variações significativas na qualidade das entregas, favorece soluções ad hoc e aumenta a dependência de conhecimento tácito, o que compromete a continuidade operacional e dificulta a evolução dos produtos ao longo do tempo.

Nesse contexto, emerge um problema central que orienta esta pesquisa: a inexistência de processos estruturados e padronizados, fundamentados em boas práticas de engenharia, para a implantação de produtos de dados em lote em ambientes de *big data*.

Embora frameworks tradicionais auxiliem nas fases exploratórias e de modelagem, eles não fornecem orientações sobre como operacionalizar rotinas analíticas em ambientes distribuídos, nem sobre como lidar com desafios específicos da implantação, como controle de versões, integração entre times distintos, formalização de artefatos ou mecanismos de validação em larga escala. Assim, este trabalho busca responder à seguinte pergunta: *como estruturar um processo reprodutível, colaborativo e eficiente para a implantação de produtos de dados em lote em ambientes de big data?*

Em ambientes de big data, a implantação de produtos de dados em lote, também conhecidos como *batch files*, representa um conjunto específico de desafios. Esses produtos precisam ser escaláveis, executados com desempenho aceitável, e integrados com múltiplos sistemas e equipes, exigindo um alto grau de coordenação e governança operacional. Além disso, a diversidade de formatos, regras de negócio e etapas de transformação presentes em pipelines de dados aumenta a probabilidade de inconsistências quando não há um processo formal de implantação. Nesses contextos, a ausência de diretrizes claras e padronizadas para implantação em produção pode resultar em retrabalho, falhas de comunicação, atrasos na entrega e ineficiências técnicas (Stonebraker et al., 2010; Gorton; Bener; Mockus, 2015). Em organizações de grande porte, tais ineficiências podem afetar diretamente indicadores de nível de serviço e comprometer resultados de áreas inteiras.

Estudos recentes destacam que a engenharia de software pode fornecer soluções estruturadas para esses desafios, especialmente quando combinada com métodos ágeis (Grady, 2017; Amershi et al., 2019). A aplicação dessas práticas ao contexto de ciência de dados tem o potencial de melhorar a governança, a colaboração entre equipes e a rastreabilidade durante a entrega dos produtos analíticos. Ainda, segundo (Silva; Nicolau, 2023), práticas de engenharia e ciência de dados passam a influenciar-se mutuamente, exigindo novos modelos que contemplem não apenas a experimentação, mas também a implantação, manutenção e evolução dos produtos analíticos. Essa interação reforça a necessidade de abordagens que considerem tanto a natureza exploratória da ciência de dados quanto os requisitos formais de engenharia presentes em ambientes corporativos.

Um exemplo representativo da necessidade de uma abordagem estruturada para implantação de produtos de dados em lote pode ser observado nos *bureaus* de crédito. Essas instituições desempenham um papel central no sistema financeiro, processando milhões de registros para gerar *scores* de crédito e relatórios financeiros com base em dados cadastrais, transacionais, comportamentais e jurídicos (Kiviat, 2017). Os ambientes de processamento utilizados pelos bureaus normalmente operam com janelas de execução rígidas e exigem alta disponibilidade. Nesse cenário, a confiabilidade e a escalabilidade dos produtos de dados impactam diretamente a qualidade dos serviços prestados e, consequentemente, a competitividade da instituição. Assim, mesmo pequenas inconsistências em processos de implantação podem gerar impactos significativos para clientes corporativos e consumidores finais.

Com o objetivo de preencher a lacuna metodológica existente nas abordagens tradicionais, este trabalho propõe a Abordagem Integrada de Implantação de Produtos de Dados (AIIPD). Essa abordagem combina práticas da engenharia de software com métodos ágeis, estruturando etapas, artefatos e responsabilidades que apoiam equipes multidisciplinares na implantação de produtos de dados em lote. A AIIPD é avaliada por meio de uma pesquisa-ação conduzida em um *bureau* de crédito brasileiro, permitindo a realização de intervenções reais e colaborativas em um ambiente produtivo, com ciclos iterativos de diagnóstico, planejamento, ação e reflexão. Essa escolha metodológica possibilita observar problemas reais, ajustar o processo de forma incremental e verificar sua efetividade ao longo de múltiplos ciclos de implantação.

Além disso, parte dos resultados produzidos neste trabalho foi publicada no IoTBDS 2025, reforçando a relevância acadêmica da pesquisa e possibilitando sua disseminação para outros pesquisadores e profissionais da área. A publicação contribui para o debate sobre métodos de implantação e evidencia a importância de se estruturar processos formais para lidar com desafios ainda pouco explorados nos frameworks tradicionais.

São duas as principais contribuições desta dissertação. Para a comunidade acadêmica, é apresentada uma abordagem estruturada e iterativa, focada especificamente na fase de implantação, suprindo uma limitação evidente nos frameworks tradicionais de ciência de dados. Para os profissionais da área, são oferecidas diretrizes operacionais e boas práticas que podem servir de referência na implantação de projetos de ciência de dados em ambientes de computação de alto desempenho e *big data*, contribuindo para maior previsibilidade, qualidade e governança nos processos de entrega.

1.1 OBJETIVOS

Nas seções abaixo estão descritos o objetivo geral e os objetivos específicos.

1.1.1 Objetivo Geral

O objetivo geral deste trabalho é desenvolver e avaliar a AIIPD, uma abordagem específica para a implantação de produtos de dados em lote em ambientes de big data, visando melhorar a eficiência dos processos de implantação. A abordagem proposta será aplicada em um estudo de pesquisa-ação em uma organização que atua como *bureau* de crédito, integrando técnicas de engenharia de software e práticas ágeis.

1.1.2 Objetivos Específicos

Para alcançar o objetivo geral mencionado, os seguintes objetivos específicos foram estabelecidos:

a) Identificar e analisar os desafios atuais na implantação de produtos de dados em lote em ambientes de big data;

b) Desenvolver uma abordagem específica para a implantação de produtos de dados em lote;

c) Implementar a abordagem proposta em um estudo de pesquisa-ação;

d) Avaliar a eficácia da abordagem proposta;

e) Documentar e disseminar os resultados da pesquisa.

1.2 ESTRUTURA DO TRABALHO

O trabalho está organizado da seguinte forma: neste Capítulo, apresentamos a introdução e os objetivos. No Capítulo 2, discutimos os principais conceitos relacionados a *big data* e Engenharia de Software. O Capítulo 3 apresenta os trabalhos relacionados à proposta. O Capítulo 4 detalha a pesquisa-ação realizada, incluindo o diagnóstico do processo atual. O Capítulo 5 apresenta a abordagem proposta para a implantação de produtos de dados em lote, além de descrever a aplicação prática dessa abordagem em um estudo de caso real. No Capítulo 6, é apresentada a avaliação da abordagem, com a análise dos resultados obtidos. Por fim, o Capítulo 7 mostra a conclusão e os trabalhos futuros, com base nos resultados e nas lições aprendidas.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 COMPUTAÇÃO DE ALTO DESEMPENHO

A Computação de Alto Desempenho, ou HPC, é essencial para o processamento e a análise de grandes conjuntos de dados, especialmente em aplicações de *big data*. Ela desempenha um papel crítico em áreas como modelagem climática, bioinformática e simulação de terremotos. O advento de GPUs e sistemas de cluster escaláveis, como os clusters Beowulf (Sterling et al., 1995), permitiu que a HPC atendesse aos requisitos de velocidade e escalabilidade de ambientes de *big data*.

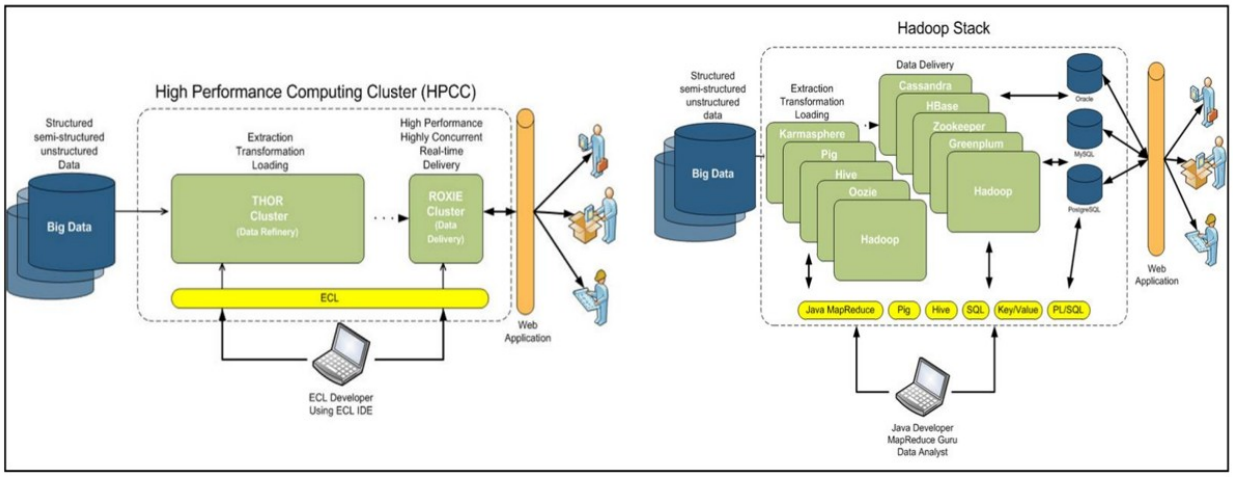
Diversas plataformas utilizam os recursos da HPC para processamento em lote em *big data*. O Hadoop (Shvachko et al., 2010), uma estrutura de processamento distribuído amplamente utilizada, inclui componentes como o *Hadoop Distributed File System* (HDFS), o MapReduce para execução paralela de tarefas e o YARN para gerenciamento de recursos. Embora o Hadoop se destaque no processamento em lote, frameworks mais recentes como Apache Spark e Apache Flink estendem seus recursos para incluir processamento de fluxo contínuo e em tempo real.

O Apache Spark soluciona as limitações do Hadoop processando dados na memória, acelerando tarefas iterativas, como aprendizado de máquina (Zaharia et al., 2010). Seu ecossistema oferece suporte a análises em lote e streaming por meio de componentes como Spark SQL para consultas e MLlib para aprendizado de máquina (Meng, X. et al., 2016). O Spark é particularmente eficaz para análise exploratória de dados e modelagem preditiva.

O Apache Flink é especializado em processamento de fluxo em tempo real, oferecendo recursos como tempo de evento e janelamento para análises baseadas em tempo. Com baixa latência e alta resiliência, o Flink é ideal para aplicações que exigem respostas imediatas, como detecção de fraudes e monitoramento de rede.

Outra plataforma notável é a *HPCC Systems* (HPCC Systems, 2024), que fornece uma solução integrada para computação com uso intensivo de dados. Sua arquitetura inclui Thor para processamento em lote, Roxie para entrega de dados em tempo real e ESP para integração de serviços de dados. Tarefas e consultas no HPCC são escritas usando *Enterprise Control Language* (ECL), facilitando a análise eficiente de dados em larga escala. A figura 1 ilustra a arquitetura do HPCC e sua relação com o Hadoop.

Figura 1 – Plataforma de sistemas HPCC e a arquitetura Hadoop



Fonte: (Sagiroglu; Sinanc, 2013).

A Tabela 1 destaca as principais características das estruturas discutidas, comparando seus tipos de processamento, casos de uso ideais e suporte para processamento em lote e em tempo real.

Tabela 1 – Comparação entre frameworks de processamento de dados.

Tecnologia	**Uso ideal**	**Lote**	**Tempo real**
Hadoop	Processamento de grandes volumes de dados	Sim	Não
Spark	Análise iterativa, aprendizado de máquina	Sim	Sim
Flink	Processamento de fluxos e análises em tempo real	Sim	Sim
HPCC Thor	*Extract, Transform, Load* (ETL) intensivo e processamento em lote de *big data*	Sim	Não
HPCC Roxie	Consultas rápidas e entrega de dados em tempo real	Não	Sim

Com a evolução contínua das necessidades de processamento em *big data*, HPC continua a se adaptar e expandir. A combinação de processadores de última geração, GPUs e redes de alta velocidade garante que os sistemas de HPC possam lidar com volumes de dados cada vez maiores, mantendo a eficiência e a velocidade necessárias para aplicações em tempo real. Plataformas como Hadoop e HPCC Systems exemplificam essa capacidade, oferecendo soluções escaláveis e integradas.

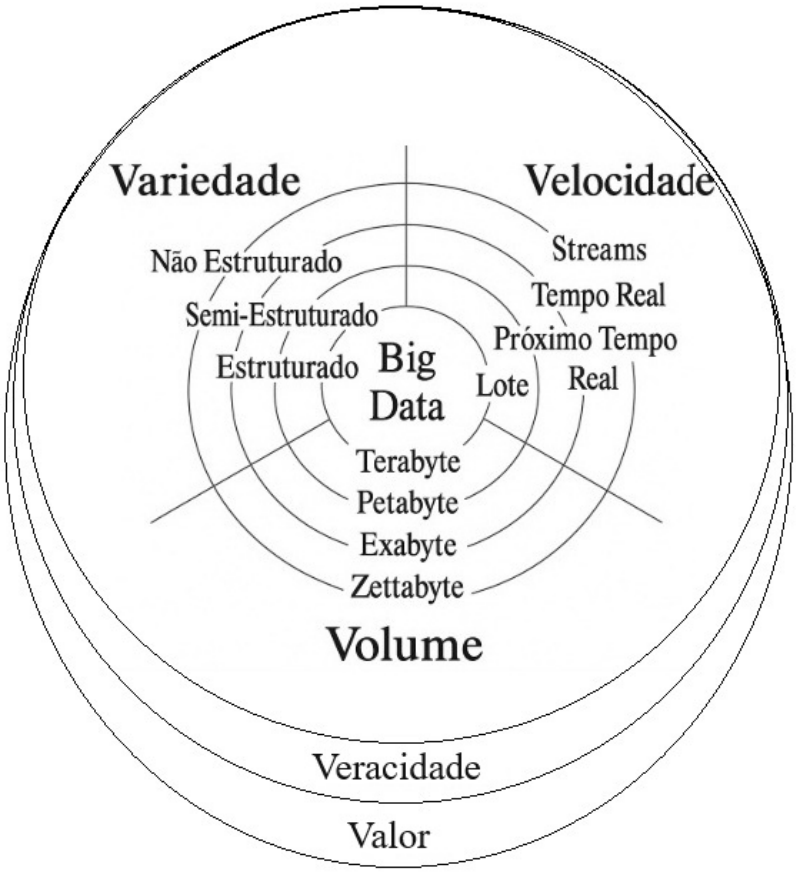
2.2 *BIG DATA*

O termo *big data* refere-se ao processo de coleta, armazenamento e análise de conjuntos de dados extremamente grandes e complexos que não podem ser processados por ferramentas tradicionais de gerenciamento de dados. Sua relevância aumentou significativamente devido à capacidade crescente de armazenamento e processamento computacional, permitindo que as organizações analisem padrões e tendências complexas, antes difíceis ou impossíveis de identificar (Chen, M.; Mao; Liu, 2014).

Os desafios e características de um ambiente de *big data* são frequentemente abordados sob o conceito dos cinco Vs: Volume, Velocidade, Variedade, Veracidade e Valor.

- O **Volume** refere-se à quantidade massiva de dados gerados e coletados, frequentemente alcançando petabytes, exabytes e até mesmo zettabytes. Esse volume massivo exige soluções de armazenamento e processamento distribuído (Sagiroglu; Sinanc, 2013).
- A **Velocidade** diz respeito à rapidez com que os dados são gerados, coletados e processados, demandando soluções que possam lidar com fluxos em tempo real, como dados de redes sociais ou sensores IoT (Gorton; Bener; Mockus, 2016).
- A **Variedade** aborda a diversidade das fontes e formatos dos dados, que podem variar de dados estruturados (bancos de dados relacionais), semi-estruturados (JSON, XML) a dados completamente não estruturados (áudios, vídeos, textos).
- A **Veracidade** diz respeito à confiabilidade e à qualidade dos dados. Em um ambiente *big data*, onde os dados vêm de inúmeras fontes, a garantia da veracidade é um desafio crucial (Taleb; Serhani; Dssouli, 2018).
- O **Valor** é a capacidade de transformar dados brutos em insights acionáveis e estratégicos. O real propósito de lidar com *big data* é extrair valor e apoiar a tomada de decisões informadas (Wang; Li; Zhang, 2019).

Figura 2 – Os cinco Vs do *big data*: Volume, Velocidade, Variedade, Veracidade e Valor.



Fonte: Adaptado de (Sagiroglu; Sinanc, 2013) pelo autor.

Os dados que compõem o *big data* são gerados de diversas fontes e em diferentes formatos, refletindo a crescente digitalização de processos e interações. Entre as principais formas de geração de dados, destacam-se:

- **Interações Humanas**: Dados provenientes de redes sociais, e-mails, mensagens instantâneas, blogs, fóruns e outras plataformas de comunicação online. Cada clique, curtida, comentário ou postagem contribui para um vasto volume de dados não estruturados e semi-estruturados.
- **Dispositivos e Sensores (IoT)**: A Internet das Coisas (IoT) é uma fonte massiva de dados, com bilhões de dispositivos conectados desde smartphones e wearables até sensores industriais, câmeras de segurança e veículos autônomos gerando informações contínuas sobre localização, temperatura, desempenho, uso e ambiente (Hashem et al., 2015).
- **Transações Comerciais**: Registros de vendas em varejo, transações bancárias, históricos de compras online, dados de programas de fidelidade e interações com serviços de atendimento ao cliente. Esses dados são geralmente estruturados e essenciais para análises financeiras e de comportamento do consumidor.

- **Dados de Máquina/Log**: Logs de servidores, aplicações, sistemas de rede e equipamentos de TI que registram eventos, erros e padrões de uso. Esses dados são cruciais para monitoramento de desempenho, segurança e diagnóstico de problemas.
- **Mídia e Entretenimento**: Conteúdo de vídeo, áudio, imagens e textos gerados por plataformas de streaming, jogos online, notícias e publicações digitais. A análise desses dados pode revelar tendências de consumo e preferências do público.
- **Dados Científicos e de Pesquisa**: Informações geradas por experimentos científicos, simulações complexas, pesquisas genômicas, dados climáticos e observações astronômicas, que frequentemente atingem volumes massivos e exigem processamento de alto desempenho (Grolinger et al., 2013).

A diversidade dessas fontes e a velocidade com que os dados são produzidos tornam o gerenciamento e a análise do *big data* um desafio complexo, exigindo infraestruturas e abordagens de engenharia de software especializadas.

Nesse contexto, a Ciência de Dados emerge como o campo interdisciplinar que transforma o potencial do *big data* em valor real. Ela combina técnicas estatísticas, métodos computacionais e conhecimentos de domínio específico para extrair conhecimento significativo de grandes volumes de dados (Irizarry, 2020; Provost; Fawcett, 2013). A Ciência de Dados utiliza técnicas como o aprendizado de máquina, a modelagem preditiva e a mineração de dados para resolver problemas complexos e identificar padrões ocultos. A integração de princípios da engenharia de software aos projetos de *big data* e Ciência de Dados é crucial para garantir não apenas a qualidade das soluções técnicas, mas também sua efetividade no atendimento às necessidades estratégicas das organizações modernas.

2.2.1 Tipos de Processamento de Dados em Ambientes de *big data*

Em ambientes de *big data*, a forma como os dados são processados é tão crucial quanto a sua geração e armazenamento. Existem diferentes paradigmas de processamento, cada um otimizado para lidar com características específicas de volume, velocidade e necessidade de resposta:

A escolha do tipo de processamento depende diretamente dos requisitos de negócio e das características dos dados. Muitas arquiteturas de *big data* modernas combinam esses diferentes paradigmas para atender a uma gama completa de necessidades analíticas e operacionais.

2.2.1.1 Processamento em Lote (*Batch Processing*)

Este tipo de processamento envolve a coleta e o processamento de grandes volumes de dados de uma só vez, em intervalos regulares (por exemplo, diário, semanal). É ideal

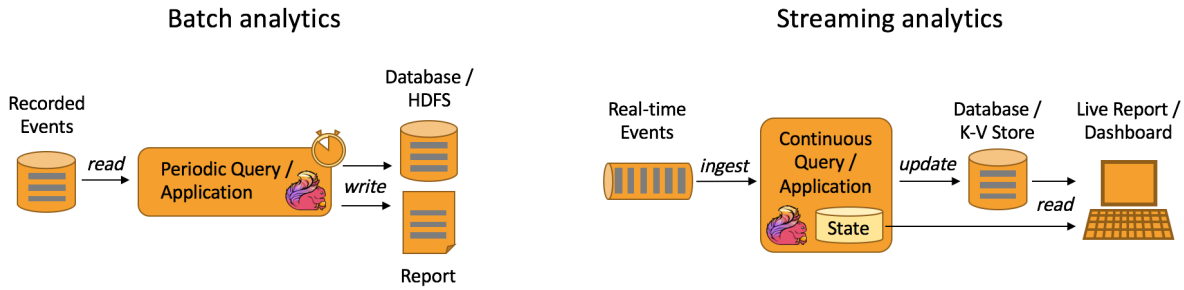
para tarefas que não exigem resultados em tempo real e que podem tolerar alguma latência. Exemplos incluem o cálculo de relatórios financeiros mensais, a análise de históricos de vendas para identificar tendências sazonais ou o processamento de grandes conjuntos de dados para treinamento de modelos de machine learning. Ferramentas como Apache Hadoop MapReduce e Apache Spark (Apache Spark Project, 2024) (para grandes volumes de dados estáticos) são comumente utilizadas para processamento em lote (Dean; Ghemawat, 2008; Apache Spark Project, 2024).

2.2.1.2 Processamento de Fluxo (*Stream Processing*)

Ao contrário do processamento em lote, o processamento de fluxo lida com dados contínuos e ilimitados à medida que são gerados, ou seja, em tempo real ou quase em tempo real. É essencial para aplicações que exigem respostas imediatas, como detecção de fraudes, monitoramento de sistemas em tempo real, análise de cliques em websites ou sistemas de recomendação. A baixa latência é uma característica fundamental. Plataformas como Apache Flink (Apache Flink Project, 2024) e Apache Spark Streaming são projetadas para este tipo de processamento, permitindo a análise de dados "em movimento"(Carbone et al., 2015; Tang; He; Yu, 2019).

A Figura 3 demonstra a principal distinção entre as duas abordagens. A análise em lote trabalha com dados estáticos e em intervalos de tempo, enquanto a análise de fluxo processa dados dinamicamente, à medida que chegam, para fornecer resultados em tempo real.

Figura 3 – Processamento em lote vs. Processamento de fluxo



Fonte: (Apache Flink Project, 2024)

2.2.1.3 Processamento Interativo (*Interactive Processing*)

Este paradigma permite que usuários e aplicações realizem consultas e análises exploratórias em grandes conjuntos de dados com baixa latência, obtendo resultados em segundos ou minutos, em vez de horas. É fundamental para cientistas de dados e analistas que precisam iterar rapidamente sobre os dados para descobrir insights. Embora possa operar sobre dados em lote, o foco está na velocidade de resposta às consultas ad-hoc. Ferramentas como Apache Impala, Presto e as capacidades interativas do Apache Spark

SQL são exemplos de tecnologias que suportam o processamento interativo (Meng, X. et al., 2016).

2.3 ENGENHARIA DE SOFTWARE

A Engenharia de Software é uma disciplina que aplica uma abordagem sistemática, disciplinada e quantificável ao desenvolvimento, operação e manutenção de software. Seu objetivo é produzir software de alta qualidade que seja econômico, confiável e adaptável. Em ambientes de *big data*, ela desempenha um papel crucial ao fornecer métodos, técnicas e ferramentas capazes de lidar com os desafios particulares impostos por esses tipos de aplicações, como a distribuição eficiente dos dados, a gestão de cargas de trabalho e a escalabilidade horizontal (Gorton; Bener; Mockus, 2016).

2.3.1 Abordagens e Práticas de Engenharia de Software

O desenvolvimento de software é orientado por diversos modelos de ciclo de vida de desenvolvimento de software, em inglês *Software Development Life Cycle* (SDLC).

O Modelo Cascata é um dos modelos mais tradicionais da Engenharia de Software, caracterizado por uma estrutura sequencial em que as fases de requisitos, projeto, implementação, testes e manutenção são executadas de forma linear. Embora tenha sido fundamental para consolidar a disciplina e introduzir maior rigor ao desenvolvimento de software, sua natureza rígida dificulta a adaptação a mudanças de requisitos e a retroalimentação entre fases, sendo atualmente mais referenciado como um marco histórico na evolução dos modelos de processo (Pressman, 2010).

Também surgiram modelos de ciclo de vida evolutivos e incrementais, que buscam mitigar as desvantagens da rigidez dos modelos sequenciais. Esses modelos permitem que o software seja desenvolvido em partes menores e entregues em iterações, possibilitando feedback contínuo e adaptação a mudanças.

Entre as abordagens iterativas e incrementais destaca-se o *Unified Process* (UP), amplamente difundido por meio do *Rational Unified Process* (RUP). O RUP organiza o desenvolvimento em quatro fases principais: *Inception* (concepção), *Elaboration* (elaboração), *Construction* (construção) e *Transition* (transição). Cada fase pode conter múltiplas iterações, produzindo incrementos executáveis do sistema. Na fase de Concepção são definidos escopo, visão e principais requisitos; na Elaboração ocorre a estabilização arquitetural e o refinamento dos requisitos; na Construção concentra-se a implementação incremental e os testes; e na Transição realiza-se a implantação e validação em ambiente real. O RUP também define papéis claros (como analista, arquiteto, desenvolvedor e gerente de projeto) e artefatos estruturados (modelos de caso de uso, modelos arquiteturais, planos de iteração e documentos de visão), promovendo rastreabilidade e governança do processo.

Como uma adaptação mais enxuta do UP, o *Open Unified Process* (OpenUP) foi

concebido para atender equipes menores, preservando os princípios iterativos e a ênfase arquitetural do RUP, porém com menor formalismo e quantidade reduzida de artefatos obrigatórios. O OpenUP mantém a divisão em fases semelhantes (*Inception*, *Elaboration*, *Construction* e *Transition*), mas privilegia ciclos curtos, colaboração contínua e foco em resultados incrementais, sendo particularmente adequado a contextos que exigem equilíbrio entre disciplina processual e agilidade.

O Modelo Espiral, proposto por Boehm (Boehm, 1988), é um modelo de ciclo de vida que combina elementos dos modelos sequenciais e iterativos, com forte foco na análise e mitigação de riscos. O desenvolvimento ocorre em espirais, onde cada volta representa um conjunto de atividades que incluem definição de objetivos, avaliação de alternativas, desenvolvimento e planejamento da próxima etapa. É particularmente adequado para projetos grandes, complexos e de alto risco.

Já o Modelo de Prototipagem envolve a criação de uma versão preliminar do sistema (protótipo) para coletar feedback dos usuários e refinar os requisitos. Este processo é iterativo, onde o protótipo é construído, avaliado e aprimorado até que os requisitos sejam bem compreendidos. Sua principal vantagem é a redução da incerteza e o aumento da satisfação do cliente (Pressman, 2010).

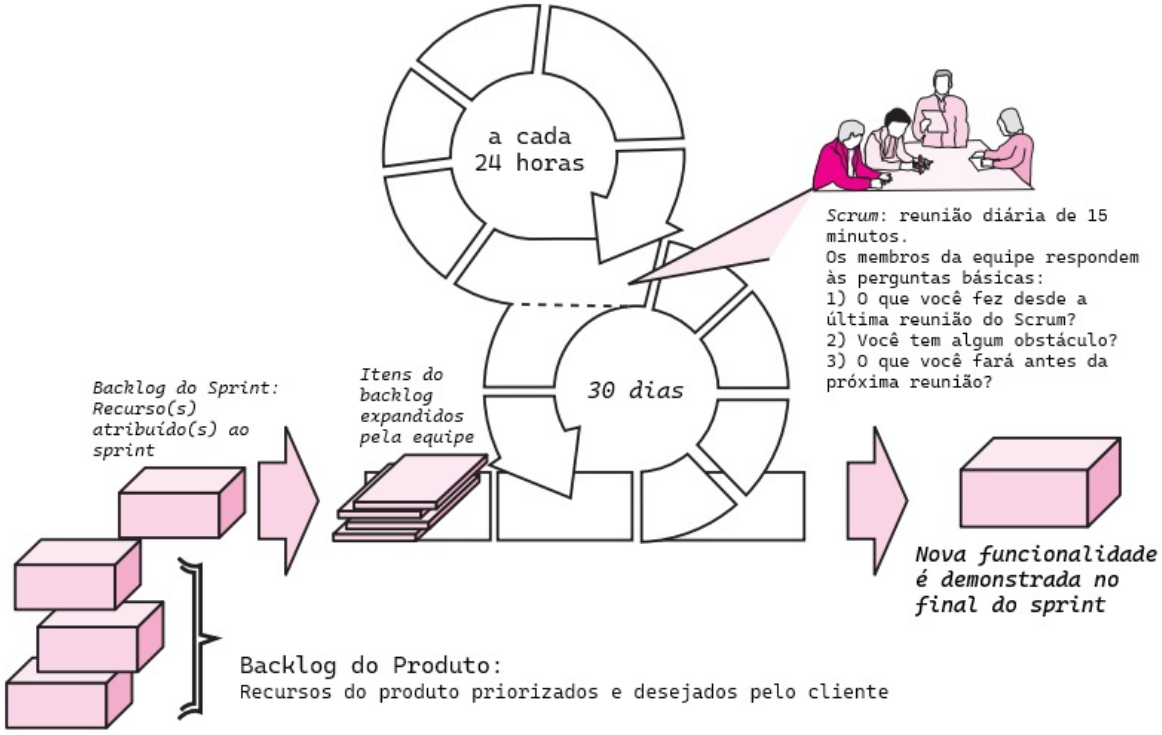
A relevância do RUP e do OpenUP para este trabalho reside na ênfase que essas abordagens atribuem à definição explícita de papéis, artefatos e fases bem delimitadas, aspectos que dialogam diretamente com os desafios identificados na fase de implantação de produtos de dados em lote. Em ambientes de *big data*, nos quais múltiplas equipes interagem e há necessidade de estabilização arquitetural antes da execução em larga escala, a combinação entre estrutura iterativa e disciplina processual mostra-se particularmente adequada, influenciando a concepção da abordagem proposta nesta dissertação.

Em resposta às limitações dos modelos tradicionais, os métodos ágeis transformaram significativamente a forma como o desenvolvimento de software é conduzido, estabelecendo-se como uma das práticas mais relevantes na engenharia de software desde sua popularização no início dos anos 2000 (Hoda; Salleh; Grundy, 2018). O Manifesto Ágil (Fowler et al., 2001) formalizou esta abordagem por meio de quatro valores fundamentais e doze princípios, destacando a importância da colaboração com o cliente, adaptabilidade às mudanças, entregas incrementais e comunicação eficaz nas equipes. Atualmente, a adoção dos métodos ágeis permanece em expansão, sendo uma prioridade estratégica para inúmeras empresas devido à sua capacidade de aumentar a eficiência, melhorar a qualidade do produto e garantir maior satisfação do cliente (VERSIONONE, 2022).

Entre os métodos ágeis mais adotados, destacam-se o Scrum (Schwaber; Beedle, 2001), com sua ênfase em entregas iterativas, ciclos curtos e cerimônias específicas que garantem transparência e inspeção contínua dos resultados (Figura 4), e o *Extreme Programming* (XP) (Beck, 2000), conhecido por priorizar qualidade de código, programação em pares e testes frequentes. Essas práticas têm se mostrado eficazes ao lidar com am-

bientes altamente dinâmicos e requisitos frequentemente mutáveis, típicos de projetos contemporâneos de software.

Figura 4 – Fluxo do processo Scrum



Fonte: Adaptado de (Pressman, 2010) pelo autor.

Um aspecto crítico do sucesso dos métodos ágeis é a dimensão humana e social envolvida nos projetos. A competência técnica individual, aliada à capacidade colaborativa e comunicacional dos membros da equipe, é essencial para o sucesso das iniciativas ágeis (Cockburn; Highsmith, 2001). Essa abordagem centrada nas pessoas promove um ambiente de trabalho saudável, produtivo e focado em resultados concretos.

Além disso, as cerimônias ágeis são fundamentais para a manutenção da agilidade operacional das equipes. No Scrum, dentre essas cerimônias estão o planejamento das iterações (*sprint planning*), as sessões de refinamento e priorização de requisitos (*backlog refinement*), as reuniões diárias de acompanhamento (*daily meetings*) e as reuniões retrospectivas, que buscam identificar melhorias contínuas e promover o aprendizado organizacional (Sharma; Kumar; Fayad, 2021). Essas práticas proporcionam um alinhamento constante das expectativas entre as partes interessadas e a equipe de desenvolvimento, garantindo entregas incrementais e um rápido feedback, cruciais para o êxito dos projetos ágeis.

A literatura acadêmica recente enfatiza a relevância das práticas ágeis não apenas em contextos tradicionais, mas também em ambientes altamente complexos como projetos de *big data* e Ciência de Dados. Begoli e Horey (Begoli; Horey, 2019) destacam que a

aplicação de métodos ágeis em projetos de *big data* facilita a gestão eficiente de riscos e melhora a adaptação às constantes mudanças técnicas e requisitos de negócio. Estudos também mostram que equipes ágeis são capazes de entregar produtos de dados em ciclos mais curtos e com maior qualidade, devido ao foco na comunicação constante e integração contínua de funcionalidades (Saltz, J. S.; Krasteva, 2022a).

Outro fator relevante para o sucesso dos métodos ágeis é a adoção de práticas de *Development* (Desenvolvimento) e *Operations* (Operações) (DevOps), que integram o desenvolvimento e as operações de TI para otimizar o ciclo de vida do software. DevOps, alinhado aos princípios ágeis, fortalece ainda mais a capacidade das equipes de entregar rapidamente valor ao cliente por meio de entregas frequentes e automatizadas, com alta qualidade e menor risco (Humble; Farley, 2010).

É fundamental reconhecer que a implementação eficaz dos métodos ágeis requer não apenas a adoção de práticas e cerimônias específicas, mas também mudanças culturais profundas nas organizações. Empresas que conseguem efetivamente incorporar uma mentalidade ágil, orientada à colaboração, experimentação e aprendizagem contínua, alcançam benefícios significativos em termos de desempenho do projeto, satisfação do cliente e engajamento das equipes (Dybå; Dingsøyr, 2014).

No contexto brasileiro, o programa Melhoria de Processo do Software Brasileiro (MPSBR) é coordenado pela SOFTEX que visa elevar a capacidade e a maturidade dos processos de desenvolvimento de software nas empresas do país. Seu principal componente, o MPS-SW, é um modelo de maturidade que estabelece sete níveis (de G a A), cada um representando uma evolução na forma como as organizações gerenciam seus projetos e processos, buscando maior previsibilidade, qualidade e eficiência (Menezes et al., 2017; SOFTEX, 2012). A adoção do MPSBR foi relevante para impulsionar a competitividade da indústria de software brasileira, alinhando as práticas locais com padrões de excelência e, em alguns casos, harmonizando-se com outros modelos e certificações, como demonstrado por estudos de caso na indústria (Hauck et al., 2015).

2.3.2 Requisitos de Software

A engenharia de requisitos é a disciplina que se preocupa em descobrir, analisar, especificar, verificar e gerenciar os requisitos de um sistema de software. A compreensão clara e precisa dos requisitos é fundamental para o sucesso de qualquer projeto (Wiegers; Beatty, 2013).

Requisitos funcionais descrevem as funções que o software deve executar. Eles definem o que o sistema deve fazer em termos de tarefas, serviços e comportamentos, como "o sistema deve calcular o score de crédito de um cliente".

Requisitos não funcionais descrevem as características de qualidade do software, como desempenho, segurança, usabilidade e escalabilidade. Em um ambiente de *big data*, requisitos não funcionais como desempenho (velocidade de processamento de dados) e

escalabilidade (capacidade de lidar com volumes crescentes de dados) são de suma importância. Outros exemplos incluem a disponibilidade, a manutenibilidade e a portabilidade do sistema (Sommerville, 2016).

2.3.3 Testes de Software

O teste de software é uma disciplina crítica para garantir a qualidade do produto final. Seu objetivo é identificar defeitos, garantir que o software funcione conforme os requisitos e aumentar a confiança no sistema. Em projetos de software, especialmente aqueles envolvendo *big data*, diferentes tipos de testes são aplicados para validar o sistema em vários níveis.

- **Testes de Unidade**: Validam o comportamento de componentes de software individuais, como uma função ou um método, isoladamente. Eles são tipicamente automatizados e realizados pelos desenvolvedores (Jorgensen, 2014).
- **Testes de Integração**: Verificam como os diferentes módulos ou componentes do software interagem entre si. O objetivo é detectar falhas de comunicação ou de interface entre as partes do sistema.
- **Testes de Sistema**: Avaliam o sistema completo e integrado, garantindo que ele atenda a todos os requisitos funcionais e não funcionais. Nesta fase, o sistema é testado em um ambiente que se assemelha ao ambiente de produção.
- **Testes de Aceitação**: Realizados para verificar se o software atende às necessidades e expectativas do cliente. É a última etapa de teste antes da implantação em produção e, muitas vezes, é conduzido pelos próprios clientes ou usuários finais.
- **Testes de Performance**: Avaliam o desempenho do sistema sob diferentes cargas de trabalho. Em projetos de *big data*, este tipo de teste é crucial para garantir que o sistema seja capaz de processar grandes volumes de dados dentro de prazos aceitáveis.

Além dos tipos de testes, existem abordagens que integram o desenvolvimento e o teste de forma mais coesa. Duas abordagens proeminentes são o Desenvolvimento Guiado por Testes ou *Test Driven Development* (TDD) e o Desenvolvimento Guiado por Comportamento ou *Behavior Driven Development* (BDD), que se destacam pela sua importância na engenharia de software moderna.

O TDD preconiza que os testes de software devem ser escritos antes do código de produção. O ciclo de trabalho é frequentemente descrito como "*Red-Green-Refactor*"(Beck, 2000):

1. ***Red***: O desenvolvedor escreve um teste para um novo recurso que falhará, pois o código para implementá-lo ainda não existe.

2. ***Green***: O desenvolvedor escreve o mínimo de código necessário para fazer o teste passar.
3. ***Refactor***: O código é então refatorado para melhorar sua qualidade, legibilidade e design, sem alterar sua funcionalidade.

A importância do TDD reside na sua capacidade de produzir um design de software mais limpo, modular e robusto, além de criar uma suíte de testes de regressão automatizada que protege o sistema de futuros defeitos.

O BDD é uma evolução do TDD que foca na colaboração entre desenvolvedores, analistas de negócio e clientes. Ele descreve o comportamento esperado do software a partir da perspectiva do usuário final, utilizando uma linguagem ubíqua e de fácil compreensão (North, 2006). Os cenários de teste são escritos em um formato estruturado, como "Dado-Quando-Então"(*Given-When-Then*), o que facilita a comunicação e garante que a equipe construa o produto certo. O BDD é importante para alinhar as expectativas de negócio com a implementação técnica, reduzindo mal-entendidos e retrabalho.

2.4 FRAMEWORKS E PROCESSOS PARA CIÊNCIA DE DADOS

Ciência de dados é um campo interdisciplinar que combina técnicas estatísticas, métodos computacionais e conhecimentos de domínio específico para extrair conhecimentos significativos e tomar decisões informadas a partir de grandes volumes de dados estruturados e não estruturados (Irizarry, 2020; Meng, X.-L., 2019). Este campo engloba várias etapas, incluindo coleta, processamento, análise e visualização de dados, frequentemente utilizando aprendizado de máquina para previsões e decisões baseadas em dados.

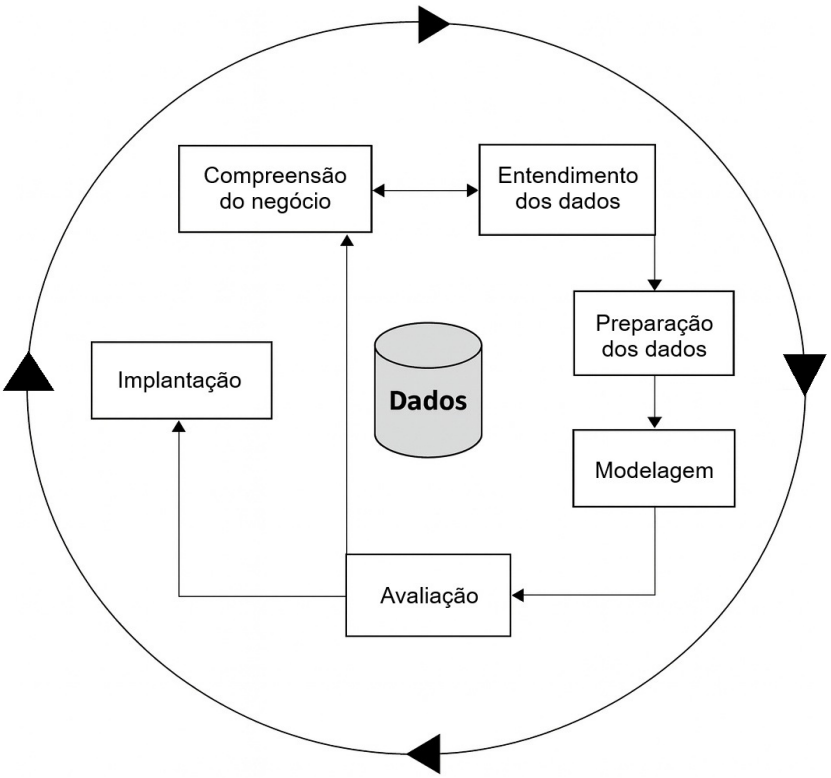
O termo ciência de dados tem sido amplamente adotado em diversos setores além da indústria tecnológica, como governo e saúde, devido à sua capacidade de resolver problemas complexos e facilitar a inovação. A definição de ciência de dados continua a evoluir, refletindo a natureza dinâmica e expansiva do campo, que integra áreas como estatística, informática, ciência da computação, e conhecimento de domínio (Zhu; Xiong, 2015).

Projetos de ciência de dados podem ser significativamente aprimorados por meio da aplicação de abordagens de gerenciamento de projetos e processos, que atuam como fatores críticos de sucesso. Modelos de processos como CRISP-DM se concentram em etapas como preparação, análise, reflexão e disseminação (Saltz, J. S.; Shamshurin, 2016), sendo amplamente reconhecidos na ciência de dados. No entanto, o uso desses modelos está em declínio, e publicações diversas destacam limitações nesses modelos padrão (Espinosa; Armour, 2016; Nabati; Thoben, 2016; Nagashima; Kato, 2019; Schröer; Kruse; Gómez, 2021), tais como não abordar adequadamente a gestão da comunicação, do conhecimento e de projetos.

O CRISP-DM (Figura 5) é o processo mais comumente utilizado em projetos de

ciência de dados, porém nenhuma atualização significativa foi feita ao modelo desde seu lançamento no início dos anos 2000 (Ahmed; Dannhauser; Philip, 2018). Além do CRISP-DM, outros frameworks e metodologias foram desenvolvidos para melhorar a gestão de projetos de ciência de dados. O modelo *Team Data Science Process* (TDSP) (Microsoft, 2024), por exemplo, combina práticas do CRISP-DM e do Scrum, incorporando conceitos ágeis e um foco forte na colaboração em equipe. O Data-Driven Scrum (Saltz, J. S.; Krasteva, 2022a) é outro exemplo, adaptando o framework Scrum para a ciência de dados, permitindo flexibilidade e foco em experimentação, o que é essencial para projetos de ciência de dados que frequentemente lidam com incertezas e mudanças nos requisitos.

Figura 5 – CRISP-DM



Fonte: Adaptado de (Wirth; Hipp, 2000) pelo autor.

Frameworks híbridos, como a combinação de métodos ágeis com métodos tradicionais, também estão ganhando popularidade. Estes frameworks balanceiam a rigidez e o planejamento detalhado dos métodos tradicionais com a flexibilidade e adaptabilidade dos métodos ágeis, criando um método único que atende melhor às necessidades específicas dos projetos de ciência de dados (Fleckenstein; Fellows, 2018).

2.5 CONSIDERAÇÕES DO CAPÍTULO

Neste capítulo, foi estabelecido o alicerce teórico do estudo, explorando os conceitos fundamentais de big data, com a análise de seus cinco Vs (Volume, Velocidade, Variedade,

Veracidade e Valor) e dos diferentes paradigmas de processamento, e de Engenharia de Software. A discussão sobre as abordagens e práticas da Engenharia de Software, incluindo modelos de ciclo de vida, métodos ágeis, e normas como o MPS.BR possibilitou a contextualização e a justificação da abordagem proposta. A interconexão desses campos é fundamental para a criação de soluções robustas para os desafios de implantação de produtos de dados em ambientes complexos.

3 TRABALHOS RELACIONADOS

3.1 PROTOCOLO DA REVISÃO DE LITERATURA

A revisão da literatura teve como objetivo identificar abordagens, frameworks e práticas metodológicas que abordem a implantação de produtos de dados em ambientes de *big data*, com foco na integração entre Engenharia de Software e Ciência de Dados. O protocolo de revisão foi estruturado conforme as diretrizes de revisões sistemáticas em Engenharia de Software (Kitchenham, 2007).

3.1.1 Objetivo e Perguntas de Pesquisa

O objetivo principal da revisão foi identificar e analisar trabalhos que tratam de metodologias, frameworks ou processos voltados à implantação de produtos de dados, especialmente em lote (*batch*), em ambientes de *big data*.

As seguintes perguntas de pesquisa (*Research Questions RQs*) nortearam a revisão:

- **RQ1:** Quais metodologias, frameworks ou processos têm sido propostos para a implantação de produtos de dados em ambientes de *big data*?
- **RQ2:** De que forma esses trabalhos integram práticas de Engenharia de Software e Ciência de Dados?
- **RQ3:** Quais limitações são identificadas nas abordagens existentes, especialmente na fase de implantação em produção?

3.1.2 Bases de Dados e Estratégia de Busca

A busca foi realizada nas bases *IEEE Xplore* e *Scopus*, por se tratarem de repositórios consolidados e revisados por pares na área de Computação. As estratégias de busca foram estruturadas em duas strings complementares, com o objetivo de delimitar com maior precisão o escopo da investigação.

A primeira string teve como foco principal a identificação de estudos relacionados à implantação de produtos de dados em ambientes de *big data*, com ênfase em aspectos de engenharia de software:

```
("Data Product" OR "Data Science" OR "Big Data")
AND ("Deployment" OR "Production" OR "Operationalization"
     OR "Batch Processing")
AND ("Software Engineering" OR "Process" OR "Methodology"
     OR "Framework")
```

A segunda string foi utilizada com o objetivo de capturar trabalhos que discutem frameworks e abordagens metodológicas aplicadas à ciência de dados, especialmente aquelas que envolvem práticas ágeis e integração com engenharia de software:

```
("Data Science" OR "Big Data Analytics")
AND ("Agile" OR "Scrum" OR "Process Model"
     OR "Lifecycle")
AND ("Framework" OR "Engineering Practices"
     OR "Project Management")
```

As buscas foram realizadas considerando títulos, resumos e palavras-chave dos trabalhos indexados. O recorte temporal abrangeu o período de 2020 a 2025, priorizando estudos recentes. Trabalhos anteriores considerados fundacionais foram incluídos quando relevantes ao embasamento teórico e à contextualização histórica do tema.

3.1.3 Critérios de Inclusão e Exclusão

- **Inclusão:** estudos revisados por pares que abordam frameworks, processos ou metodologias aplicadas à implantação de produtos de dados, engenharia de software ou ciência de dados em ambientes de *big data*.
- **Exclusão:** artigos não revisados por pares, duplicados, fora do escopo de implantação, ou voltados exclusivamente a modelagem ou visualização de dados.

3.1.4 Resultados da Busca e Seleção

A busca inicial retornou aproximadamente 130 artigos. Foi realizada uma triagem por título e resumo que resultou em 50 estudos. Após a eliminação dos artigos não diretamente relevantes para esse trabalho, foi empregado o método *Snowballing* (Juneja; Kaur, 2019), que compreende três etapas em cada ciclo de busca. Inicialmente foi realizada a fase de refinamento do conjunto inicial. Posteriormente, a fase *Backward Snowballing* envolveu a utilização da lista de referências para identificar artigos potencialmente pertinentes. Por último, a fase *Forward Snowballing* envolveu o estudo de artigos que citam o artigo que está sendo examinado.

Optou-se pela identificação da maioria dos documentos na fase de *Backward Snowballing*, dada a natureza relativamente recente do tema e a consequente necessidade de recorrer a trabalhos fundacionais. A leitura das seções de introdução e resultados permitiu reduzir o conjunto a 18 estudos que embasaram a abordagem proposta; dentre eles, cinco foram selecionados para análise detalhada na seção Trabalhos Relacionados por apresentarem maior aderência ao problema de implantação de produtos de dados em lote em ambientes de *big data*.

Tabela 2 – Trabalhos selecionados na revisão sistemática

Autores	**Título**	**Ano**	**Relevância para o tema**
Wirth & Hipp	CRISP-DM: Towards a Standard Process Model for Data Mining	2000	Processo de referência que estrutura o ciclo de *data mining*; base para comparar lacunas na etapa de implantação.
Stonebraker et al.	MapReduce and parallel DBMSs: Friends or foes?	2010	Debate arquitetural (MapReduce vs. SGBDs paralelos) útil para decisões de processamento em lote e suas implicações operacionais.
Grolinger et al.	Data management in cloud environments: NoSQL and NewSQL data stores	2013	Panorama de NoSQL/NewSQL para escalabilidade e integração; subsidia escolhas de armazenamento em pipelines *batch*.
Kumar & Alencar	Software engineering for big data projects: Domains, methodologies and gaps	2016	Mapeia domínios/metodologias e lacunas; evidencia falta de práticas maduras para a implantação em produção.
Chen et al.	Agile Big Data Analytics Development: An Architecture-Centric Approach	2016	Propõe abordagem arquitetural ágil (ABBA) focada no desenvolvimento; cobre pouco a fase de implantação.

Continua na próxima página

Tabela 2 – continuação da página anterior

Autores	**Título**	**Ano**	**Relevância para o tema**
Espinosa & Armour	The Big Data Analytics Gold Rush: A Research Framework for Coordination and Governance	2016	Enfatiza coordenação e governança; expõe riscos organizacionais que afetam entrega e operação de soluções analíticas.
Nabati & Thoben	On Applicability of Big Data Analytics in the Closed-Loop Product Lifecycle: Integration of CRISP-DM Standard	2016	Integra CRISP-DM ao ciclo de produto em *loop* fechado; aproxima análise e operação, mas com implantação pouco detalhada.
Saltz & Shamshurin	Big data team process methodologies: A literature review and the identification of key factors for a project's success	2016	Revisão de metodologias de times de *big data*; identifica fatores de sucesso e lacunas na etapa de *deployment*.
Grady	Challenges in Engineering for Big Data	2017	Discorre sobre desafios de engenharia para *big data*; reforça necessidade de práticas de SE na transição para produção.
Hummel et al.	A Collection of Software Engineering Challenges for Big Data System Development	2018	Compila 26 desafios de desenvolvimento; inclui obstáculos práticos para operacionalizar sistemas de *big data*.
Ahmed et al.	A Lean Design Thinking Methodology (LDTM) for Machine Learning and Modern Data Projects	2018	Metodologia LDTM como alternativa ao CRISP-DM; aborda pouco o detalhamento da implantação.

Continua na próxima página

Tabela 2 – continuação da página anterior

Autores	**Título**	**Ano**	**Relevância para o tema**
Fleckenstein & Fellows	Overview of Data Management Frameworks	2018	Levantamento de frameworks de gestão de dados; posiciona processos e limitações quanto à operacionalização.
Begoli & Horey	Agile Methodologies in Big Data Projects	2019	Relaciona métodos ágeis a projetos de *big data*; contribui para cadência/entrega, mas com foco limitado em implantação.
Amershi et al.	Software engineering for machine learning: A case study	2019	Práticas de engenharia para ML em escala; lições sobre pipeline, versionamento e liberação em produção.
Microsoft	Team Data Science Process (TDSP)	2020	Processo corporativo com artefatos e trilhas; inclui etapas de operacionalização e monitoramento.
Schröer et al.	A Systematic Literature Review on Applying CRISP-DM Process Model	2021	Revisão do uso do CRISP-DM; evidencia adoção e limitações, sobretudo na fase de *deployment*.
Saltz & Krasteva	Current approaches for executing big data science projects - a systematic literature review	2022	SLR que mapeia processos de *data science* e a carência de guias práticos para implantação.

Continua na próxima página

Tabela 2 – continuação da página anterior

Autores	**Título**	**Ano**	**Relevância para o tema**
Saltz, Sutherland & Hotz	Achieving Lean Data Science Agility Via Data Driven Scrum	2022	DDS adapta Scrum à *data science*; fornece artefatos/cerimônias com orientação parcial à implantação.

Os 18 estudos analisados evidenciam a evolução das práticas que integram Engenharia de Software e Ciência de Dados. Contudo, nota-se que a maior parte das propostas concentra-se nas etapas de preparação, modelagem e análise, dedicando menor atenção à implantação em produção. Essa lacuna fundamenta a proposta deste trabalho, voltada especificamente à implantação de produtos de dados em lote em ambientes de *big data*. Considerando esse objetivo, cinco estudos foram selecionados para análise aprofundada neste capítulo, por apresentarem maior aderência ao problema de implantação, descrições processuais mais completas e evidências práticas diretamente aplicáveis ao contexto desta abordagem.

3.1.5 Trabalhos Selecionados para Discussão Detalhada

Em (Dipti Kumar; Alencar, 2016), um estudo investiga como a aplicação de princípios de desenvolvimento de software em vários estágios do ciclo de desenvolvimento do projeto pode contribuir para o design de aplicações de *big data*. As descobertas ajudaram a identificar iniciativas de projetos de dados com potencial significativo de sucesso. No entanto, os pesquisadores destacaram deficiências no ciclo de desenvolvimento de projetos relacionados a big data, ressaltando a necessidade de atenção especial. Em comparação, este estudo complementa essas descobertas ao apresentar uma abordagem estruturada que integra práticas de desenvolvimento e implantação, aprimorando a coordenação e reduzindo ineficiências.

Em (Chen, H.-M.; Kazman; Haziyev, 2016), os autores conduziram uma pesquisa com o objetivo de compreender as metodologias arquitetônicas atuais para big data, bem como a integração do projeto arquitetônico com técnicas para orquestrar ferramentas tecnológicas em uma abordagem unificada e eficaz. O objetivo foi estabelecer correlações entre as práticas do Manifesto Ágil e uma perspectiva centrada na arquitetura. O estudo culminou na proposta de uma metodologia denominada *Architecture-centric Agile Big data Analytics* (ABBA), que atribui à arquitetura de software um papel central como facilitadora da agilidade. Embora o ABBA enfatize a arquitetura, a abordagem proposta neste estudo se estende além do projeto arquitetônico, incorporando ciclos de feedback

iterativos e práticas colaborativas específicas para a implantação de produtos de dados em lote.

O estudo publicado em (Hummel et al., 2018) oferece uma abordagem detalhada para vinte e seis desafios relevantes no desenvolvimento de sistemas de *big data*. Os autores analisam e classificam cuidadosamente esses desafios por meio de um processo colaborativo e sistemático, organizando-os de acordo com as diversas fases de desenvolvimento. Eles destacam que questões críticas que influenciam o sucesso do projeto podem não ser totalmente abordadas durante a fase de planejamento, tornando o processo de desenvolvimento altamente exploratório. Da mesma forma, este estudo reconhece esses aspectos exploratórios e os mitiga com a introdução de fases claras como iniciação, elaboração, construção e entrega com o objetivo de melhorar a previsibilidade e reduzir a ambiguidade na implantação.

Em (Saltz, J. S.; Krasteva, 2022b), uma revisão sistemática é conduzida sobre a adoção de frameworks de processos em projetos de ciência de dados, destacando um aumento significativo na pesquisa sobre a organização, gestão e execução desses projetos nos últimos anos. A revisão identificou 68 estudos primários, categorizados em seis temas principais relacionados à execução de projetos de ciência de dados. CRISP-DM foi o fluxo de trabalho mais discutido. No entanto, o estudo não encontrou abordagens padronizadas especificamente projetadas para o contexto da ciência de dados, particularmente na fase de implantação, indicando uma lacuna na pesquisa sobre as práticas atuais. Sugere-se que pesquisas futuras explorem a combinação de fluxos de trabalho com abordagens ágeis para criar um framework mais abrangente que abranja diferentes aspectos da execução do projeto. A novidade deste trabalho reside em abordar essa lacuna, propondo uma abordagem que visa explicitamente a fase de implantação.

Em (Saltz, J.; Sutherland; Hotz, 2022), uma nova estrutura de processos de equipe, *Data Driven Scrum* (DDS), é proposta para aprimorar a execução de projetos de ciência de dados. Um estudo de caso conduzido em uma consultoria no México explorou a compreensão e a adaptação da equipe aos conceitos ágeis do Lean. Após a transição de uma abordagem em cascata, a equipe adotou o DDS, refinando seu processo para desenvolvimento ágil e enxuto. O estudo de caso concluiu que a organização compreendeu e se adaptou aos conceitos ágeis do Lean, validando as questões de pesquisa. No entanto, a principal limitação foi a aplicação do DDS em apenas uma organização. Embora o DDS enfatize a agilidade da equipe, este estudo se concentra em aprimorar tanto a colaboração da equipe quanto os aspectos técnicos, como documentação e testes, para garantir escalabilidade e eficiência na implantação de produtos de dados em lote.

Os estudos discutidos neste capítulo apresentam diversas abordagens, práticas e padrões de design para projetos de ciência de dados e plataformas de *big data*. No entanto, mesmo com diferentes níveis de detalhamento nas fases dos projetos, a etapa de implantação, especialmente em contextos de produtos de dados em lote e ambientes de big data,

ainda recebe pouca atenção. Este trabalho busca preencher essa lacuna ao propor uma abordagem prática e replicável, que integra métodos ágeis com princípios de engenharia de software voltados à implantação em produção.

A Tabela 3 apresenta uma comparação entre os principais trabalhos relacionados e a abordagem proposta neste estudo. Foram considerados aspectos como o foco metodológico, o detalhamento da fase de implantação, o tipo de entrega de dados priorizado, a adoção de práticas ágeis e a aderência às demandas de ambientes de *big data*. Essa comparação evidencia as limitações das abordagens tradicionais e reforça a contribuição específica deste trabalho.

Tabela 3 – Trabalhos selecionados para análise detalhada

Autores	**Título**	**Ano**	**Foco principal**	**Limitações**
Kumar & Alencar	Software engineering for big data projects: Domains, methodologies and gaps	2016	Princípios/boas práticas de desenvolvimento em projetos de dados	Foco restrito ao design; implantação pouco detalhada; pouca ênfase em práticas ágeis
Chen et al.	Agile Big Data Analytics Development: An Architecture-Centric Approach	2016	Metodologia arquitetônica ágil para *big data*	Ênfase em arquitetura; implantação pouco ou não detalhada
Hummel et al.	A Collection of Software Engineering Challenges for Big Data System Development	2018	Desafios de desenvolvimento em sistemas de *big data*	Pouco detalhamento prático; implantação não é o foco central
Saltz & Krasteva	Current approaches for executing big data science projects - a systematic literature review	2022	Revisão de frameworks/processos em ciência de dados	Não detalha a fase de implantação; foco descritivo de frameworks

Continua na próxima página

Tabela 3 – continuação da página anterior

Autores	**Título**	**Ano**	**Foco principal**	**Limitações**
Saltz, Sutherland & Hotz	Achieving Lean Data Science Agility Via Data Driven Scrum	2022	Adaptação do Scrum à ciência de dados	Implantação parcialmente detalhada; validação limitada a uma organização

3.2 CONSIDERAÇÕES DO CAPÍTULO

A análise dos trabalhos relacionados neste capítulo revelou uma lacuna significativa na literatura: a ausência de abordagens metodológicas padronizadas e detalhadas focadas especificamente na fase de implantação de produtos de dados em lote em ambientes de *big data*. Embora existam frameworks que abordam o ciclo de vida de projetos de ciência de dados e o gerenciamento de sistemas de big data, a etapa de transição para produção, que é o ponto focal da nossa pesquisa, ainda recebe pouca atenção. Essa constatação justifica o desenvolvimento e a validação de uma nova abordagem, que será formalmente apresentada nos capítulos seguintes.

4 MÉTODO DE PESQUISA-AÇÃO

Este capítulo descreve a abordagem metodológica deste estudo. Primeiro, é apresentada a pesquisa-ação, destacando sua relevância e aplicabilidade dentro do contexto deste trabalho. Em seguida, é apresentado o contexto da pesquisa, detalhando a organização e os participantes envolvidos. Posteriormente, explicam-se os métodos de coleta de dados utilizados, com uma descrição detalhada da coleta de dados para a fase de diagnóstico, que ocorre antes da execução da pesquisa-ação. Na sequência, discute-se o diagnóstico baseado nas entrevistas, identificando os principais problemas e desafios a serem enfrentados na abordagem para implantação de produtos de dados em lote na empresa. Por fim, é apresentado o planejamento para a coleta de dados após a execução da pesquisa-ação, com o objetivo de avaliar a abordagem proposta.

4.1 PESQUISA-AÇÃO

A pesquisa-ação é uma abordagem metodológica que combina pesquisa com ação prática, visando resolver problemas reais e contribuir para o conhecimento científico (McNiff, 2013). Como mencionado anteriormente, ela se caracteriza por ciclos iterativos de diagnóstico, planejamento, ação, avaliação e aprendizagem (Figura 6). Essa abordagem é particularmente adequada para contextos em que o pesquisador participa ativamente do processo de mudança, trabalha de forma colaborativa e intervém conscientemente. No contexto deste estudo, o pesquisador atuou como colaborador da empresa, desempenhando papel duplo como facilitador das intervenções e observador participante.

Embora tenha sido conduzida em um bureau de crédito, os princípios da pesquisa-ação, com ênfase na colaboração, adaptabilidade e melhoria contínua, são aplicáveis a outros setores como saúde, indústria e governo. Por exemplo, tem sido utilizada na área da saúde para aprimorar fluxos de atendimento ao paciente por meio de ciclos iterativos semelhantes, demonstrando seu potencial para promover mudanças organizacionais em diferentes contextos.

4.1.1 Contexto da pesquisa

A pesquisa é realizada em uma empresa brasileira que atua como bureau de crédito[1]. O termo *datatech* refere-se a organizações cujo modelo de negócio é fundamentado no uso intensivo de dados e tecnologias analíticas para gerar valor, seja por meio de produtos, serviços ou insights estratégicos. Como uma datatech, a empresa integra diversas fontes de dados e utiliza tecnologias avançadas para fornecer soluções de inteligência analítica, atuando como um importante intermediário entre consumidores, empresas e instituições financeiras. A empresa conduz projetos de ciência de dados para criar relatórios abran-

[1] O nome da empresa não é apresentado por questões de confidencialidade.

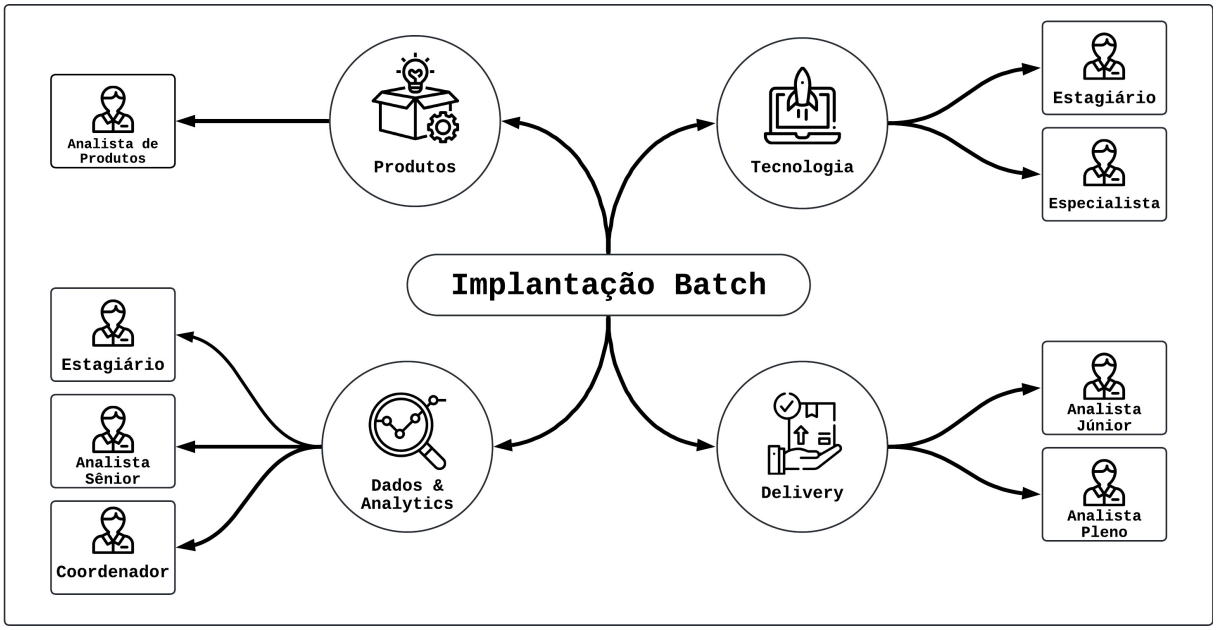
Figura 6 – Ciclo da pesquisa-ação.

Fonte: Adaptado de (McNiff, 2013) pelo autor.

gentes que refletem a saúde financeira e a capacidade de pagamento dos consumidores com base em seu comportamento (gastos financeiros, informações geográficas, dados cadastrais, compras, ações judiciais, presença online, entre outros). Além dos relatórios, a empresa oferece produtos de score de crédito, calculados por meio de modelos estatísticos que indicam a probabilidade de uma pessoa física ou jurídica cumprir com suas obrigações financeiras. Para o processamento massivo de dados, é utilizada a plataforma open-source de data lake HPCC Systems. No início da pesquisa-ação, a organização contava com aproximadamente 60 colaboradores. A equipe envolvida no fluxo de desenvolvimento de produtos de dados em lote era composta por 8 pessoas (Figura 7).

Na equipe de Produtos, há um colaborador atuando como analista de produto, com ampla experiência na área de score de crédito e um ano de atuação na organização. Na equipe de Data & Analytics, há três colaboradores com formação em estatística: um analista júnior com menos de um ano na organização, um analista sênior com dois anos de experiência e um coordenador de analytics, também com dois anos na organização. Na equipe de Tecnologia, há dois colaboradores com formação na área, sendo um estagiário e um especialista, ambos com menos de um ano na organização. Na área de Operação & Delivery, há dois colaboradores, também com formação em tecnologia e com menos de um ano de atuação. Um representante legal da organização assinou o termo de consentimento livre e esclarecido, formalizando a ciência sobre os procedimentos da pesquisa.

Figura 7 – Colaboradores que participam da pesquisa-ação



Fonte: Elaborado pelo autor.

4.1.2 Inserção do pesquisador e motivação organizacional

A motivação para a realização da pesquisa emergiu de um problema organizacional concreto relacionado à fase de implantação de produtos de dados em lote na plataforma de *big data*. Embora as etapas de modelagem estatística e definição de regras de negócio estivessem relativamente consolidadas na organização, a transição para o ambiente de produção apresentava inconsistências, ausência de padronização documental, dificuldades de versionamento e falhas de comunicação entre equipes.

Nesse contexto, o pesquisador já atuava na organização e foi diretamente envolvido nas discussões sobre a necessidade de estruturar essa fase do processo. A partir da identificação dessas dificuldades, o problema organizacional foi sistematizado como problema de pesquisa, permitindo que a intervenção prática fosse conduzida sob uma perspectiva metodológica estruturada.

A organização demonstrou apoio institucional à condução da pesquisa, autorizando formalmente sua execução e permitindo a aplicação das intervenções propostas no ambiente produtivo. Assim, a pesquisa-ação não se limitou a uma observação externa, mas constituiu um processo colaborativo de construção e avaliação de uma nova abordagem para a implantação de produtos de dados.

4.1.3 Coleta de dados para diagnóstico do processo atual

A principal técnica de coleta de dados utilizada nesta etapa inicial da pesquisa foi a realização de entrevistas. As entrevistas combinaram elementos de entrevistas semiestruturadas e entrevistas convergentes (Kallio et al., 2016), permitindo que temas e

perguntas pré-definidas guiassem a conversa, mas mantendo espaço para discussões abertas. Essa abordagem foi escolhida por oferecer flexibilidade, adaptando-se às respostas dos participantes e possibilitando a exploração de novas direções ao longo das entrevistas.

Os resultados das entrevistas foram analisados por meio da análise temática (Cruzes; Dyba, 2011), uma abordagem amplamente utilizada em pesquisas qualitativas. O objetivo da análise temática foi identificar padrões ou temas recorrentes nas entrevistas, de modo a diagnosticar o processo atual da empresa. Esse diagnóstico foi utilizado para definir a nova abordagem que será apresentada neste trabalho.

4.1.4 Diagnóstico

Esta etapa visa compreender e definir o problema. Foi realizada uma análise inicial do contexto da organização e das entrevistas realizadas. Com base nessa análise, formalizou-se a definição do problema.

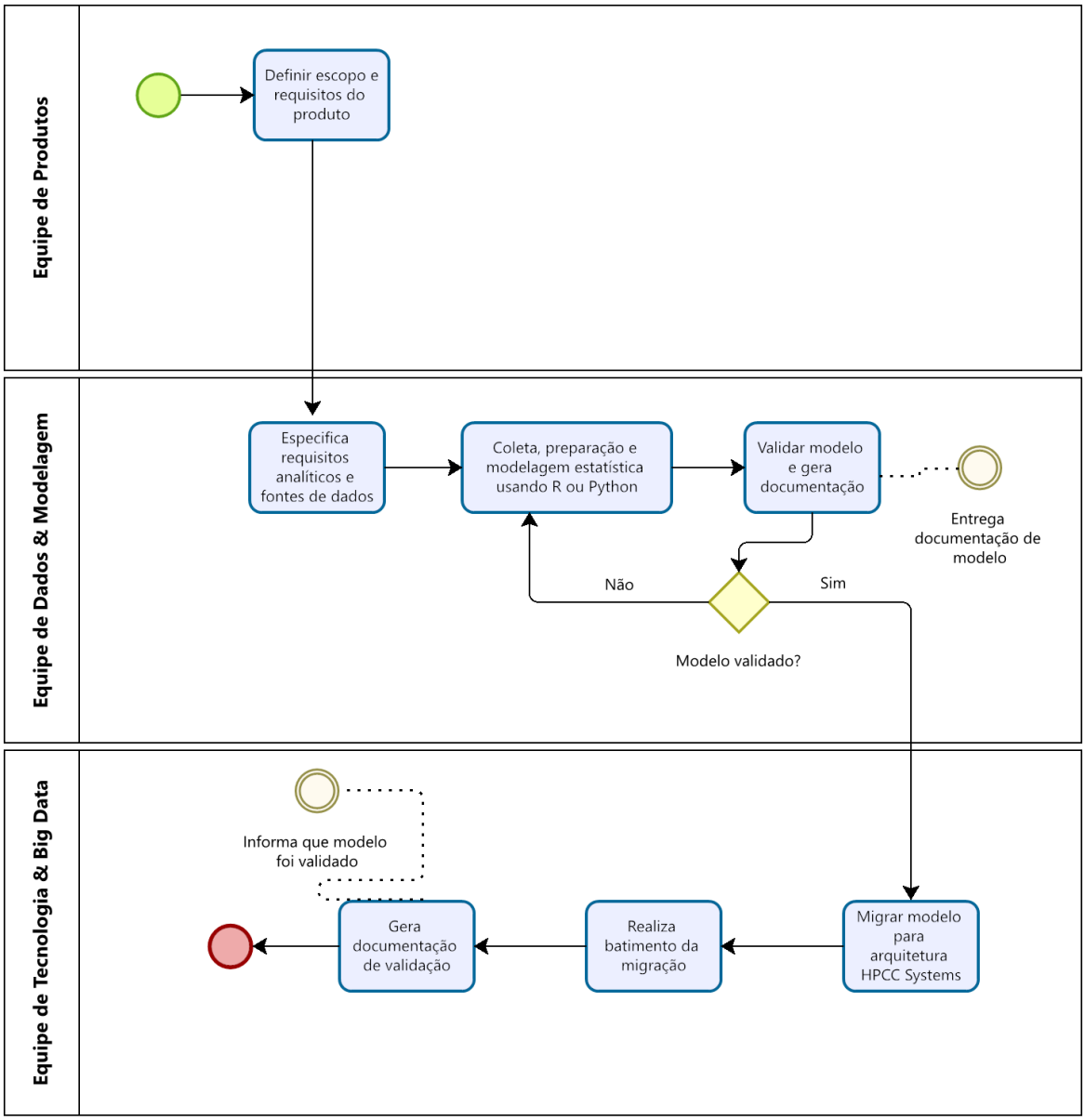
No ciclo de vida de um projeto de ciência de dados, o desenvolvimento do produto passa por várias etapas antes de sua implantação na plataforma de *big data*. O fluxo de desenvolvimento do produto envolve equipes multidisciplinares, iniciando na equipe de Produtos. Em seguida, a equipe de Dados & Modelagem realiza a modelagem estatística, define regras de negócio e desenvolve scripts utilizando linguagens como Python ou R. Para essas atividades, a organização já possui um processo bem definido e maduro, que exige apenas ajustes pontuais, os quais serão detalhados no capítulo que detalha a abordagem proposta.

Após a conclusão das etapas descritas anteriormente, é necessário transformar o produto de dados em um produto escalável e comercializável. Para isso, a plataforma de alto desempenho HPCC Systems é essencial para a geração *batch* do produto. Isso requer a migração da arquitetura utilizada na fase de modelagem para a arquitetura escalável da plataforma de *big data*. A equipe de desenvolvimento de software responsável pela implantação dos produtos na plataforma de *big data* da organização utiliza o método ágil Scrum como base para seu processo de software. No entanto, possui flexibilidade para adaptar o método conforme suas necessidades específicas.

A Figura 8 apresenta o fluxo do ciclo de vida de um produto de ciência de dados, utilizando a notação *Business Process Model and Notation* (BPMN) (Leopold; Mendling; Günther, 2016).

Para cada implantação, a equipe de Dados & Modelagem fornece documentação detalhada sobre os modelos estatísticos utilizados. Entretanto, esses documentos não seguem um padrão definido, o que resulta em problemas como ausência de informações relevantes ou erros lógicos. Além disso, a falta de uma documentação centralizada é um dos principais desafios, gerando perda de versionamento e alterações não comunicadas, o que acarreta falhas de comunicação entre as equipes. Outro problema identificado é que as equipes de diferentes áreas não realizam cerimônias ágeis em conjunto, nem definem

Figura 8 – Fluxo do ciclo de vida de um produto de ciência de dados utilizando notação BPMN



Fonte: Elaborado pelo autor.

claramente as tarefas e entregas de cada iteração na fase de implantação. Isso leva a um tempo excessivo gasto nas implantações. Por fim, outro problema significativo é a realização de testes do produto entre a arquitetura de desenvolvimento e a plataforma de *big data*.

Diante desse cenário, a empresa reconheceu a importância de adotar uma abordagem específica para a fase de implantação de produtos de dados em lote na organização. Essa necessidade foi impulsionada pela urgência de entregar soluções de alta qualidade no ambiente de *big data* em um ritmo acelerado, permitindo que a organização se mantenha competitiva no disputado mercado brasileiro de bureaus de crédito.

4.1.5 Ciclos executados da pesquisa-ação

A execução da pesquisa-ação ocorreu por meio de dois ciclos completos, cada um composto pelas etapas de diagnóstico, planejamento, ação, avaliação e aprendizagem.

No primeiro ciclo, realizou-se o diagnóstico inicial por meio das entrevistas e análise temática, identificando os principais gargalos na fase de implantação. Com base nesse diagnóstico, foi planejada uma primeira versão da abordagem proposta, incluindo definição inicial de artefatos, padronização documental e organização de responsabilidades entre equipes. A ação consistiu na aplicação piloto da abordagem em implantações reais. Ao final, foram coletadas métricas e percepções dos colaboradores, permitindo avaliar pontos fortes e limitações da primeira versão.

No segundo ciclo, as lições aprendidas foram incorporadas ao processo. Ajustes foram realizados na definição de artefatos, na formalização das etapas e na integração entre equipes. A abordagem revisada foi novamente aplicada, possibilitando uma avaliação comparativa entre o cenário anterior, o primeiro ciclo e o segundo ciclo. Esse processo evidenciou não apenas melhorias operacionais, mas também evolução na colaboração e clareza de responsabilidades.

Assim, além de dois ciclos de ação, houve dois ciclos explícitos de reflexão e aprendizagem, consolidando a abordagem apresentada no Capítulo seguinte.

4.1.6 Planejamento para coleta de dados na etapa de avaliação da pesquisa-ação

A fase de coleta de dados após a aplicação da pesquisa-ação visa garantir que as informações obtidas sejam relevantes para avaliar as intervenções realizadas. Foi utilizada a abordagem estruturada *Goal Question Metric* (GQM) para definir e avaliar métricas com base em objetivos específicos (R. Basili; Rombach, 1994), conforme detalhado no Apêndice A.

Os objetivos e perguntas para avaliar a efetividade da abordagem proposta são descritos a seguir:

- O1: Avaliar o esforço necessário para implantar um produto de dados em lote na plataforma de *big data*antes e depois da aplicação da abordagem na organização-alvo.
    - Q1.1: Qual o tempo, em dias, para implantar produtos de dados em lote em produção?
- O2: Avaliar a efetividade e aceitação da abordagem de implantação de produtos de dados em lote entre os colaboradores da organização-alvo.
    - Q2.1: Qual o nível de clareza das etapas do processo de implantação com o uso da abordagem?
    - Q2.2: Qual o nível de satisfação dos colaboradores com a abordagem?

- O3: Avaliar a efetividade da abordagem na facilitação da colaboração entre equipes e definição clara de responsabilidades de cada uma.
  - Q3.1: A abordagem facilita a colaboração entre as diferentes equipes?
  - Q3.2: As responsabilidades de cada equipe estão bem definidas e compreendidas?
- O4: Avaliar a efetividade geral da abordagem, com foco na eficiência da documentação e no rigor e efetividade do processo de validação e testes.
  - Q4.1: Como é avaliada a eficiência do processo de documentação?
  - Q4.2: O processo de validação e testes é rigoroso e efetivo?

Para cada pergunta, também foram definidas métricas seguindo a abordagem GQM. Para o objetivo O1, o esforço é registrado por meio da ferramenta de acompanhamento de projetos adotada pela organização. Para a coleta de métricas dos objetivos O2, O3 e O4, foi desenvolvido um questionário com todas as perguntas derivadas, aplicado ao final da intervenção.

4.1.7 Ameaças à validade

Como em toda pesquisa-ação, a atuação do pesquisador como membro ativo da organização pode introduzir vieses. Um possível viés refere-se à influência do pesquisador nas decisões tomadas durante a implantação, bem como à interpretação dos resultados obtidos.

Além disso, destaca-se o viés de aprendizagem ao longo dos ciclos. À medida que o pesquisador aprofundava sua compreensão do contexto organizacional e das dificuldades enfrentadas pelas equipes, sua capacidade de propor soluções também evoluía. Embora esse fenômeno seja inerente à pesquisa-ação, buscou-se mitigá-lo por meio do registro sistemático das intervenções, da utilização de métricas definidas via GQM e da validação dos resultados junto aos colaboradores envolvidos.

Outra limitação refere-se ao fato de a pesquisa ter sido conduzida em uma única organização do setor de bureau de crédito, o que pode restringir a generalização direta dos resultados para outros contextos. Ainda assim, os princípios estruturais da abordagem proposta são potencialmente aplicáveis a outras organizações que enfrentem desafios semelhantes na implantação de produtos de dados em lote.

4.2 CONSIDERAÇÕES DO CAPÍTULO

Neste capítulo, a metodologia de pesquisa-ação foi detalhada como o método escolhido para conduzir este estudo. A descrição do ciclo de diagnóstico, planejamento, ação e avaliação, em conjunto com a caracterização do contexto da pesquisa (o bureau de crédito), estabeleceu a base empírica e a validade do trabalho. A fase de diagnóstico, em

particular, identificou as principais ineficiências e desafios no processo de implantação existente, fornecendo os insumos necessários para a concepção da abordagem que será apresentada no próximo capítulo.

5 ABORDAGEM PROPOSTA

Para desenvolver a AIIPD, realizou-se coleta de dados por meio de entrevistas semiestruturadas com colaboradores envolvidos no processo de implantação. As entrevistas forneceram insumos sobre práticas vigentes, desafios recorrentes e necessidades específicas das equipes. Com base nesse diagnóstico, procedeu-se a uma revisão da literatura para identificar melhores práticas e abordagens relevantes à execução de projetos de *big data* e ciência de dados.

Na sequência, este capítulo apresenta as equipes participantes e suas responsabilidades no processo de implantação, detalha as fases do ciclo de vida do projeto que estruturam a AIIPD e, por fim, descreve a aplicação da abordagem em dois projetos distintos, discutindo decisões de engenharia e pequenos ajustes realizados para atender às necessidades específicas do ambiente de estudo.

5.1 CONCEPÇÃO DA ABORDAGEM

A AIIPD não foi concebida de forma teórica isolada, mas emergiu diretamente do diagnóstico organizacional realizado na etapa inicial da pesquisa-ação. As entrevistas conduzidas com as equipes de Produtos, Dados & Analytics, Tecnologia e Operação & Delivery revelaram gargalos recorrentes na fase de implantação de produtos de dados em lote, especialmente na transição entre a modelagem analítica e sua operacionalização na plataforma de big data.

Os principais problemas identificados foram: (i) ausência de padronização documental na passagem do modelo analítico para a equipe de tecnologia; (ii) retrabalho decorrente de informações incompletas ou ambíguas; (iii) falta de definição clara de responsabilidades na fase de implantação; (iv) inexistência de pontos formais de validação entre equipes; e (v) dificuldades no controle de versionamento e rastreabilidade das entregas.

A partir desses gargalos, iniciou-se um processo estruturado de concepção da abordagem. Primeiramente, identificou-se a necessidade de criar um artefato centralizador que servisse como fonte única de verdade do projeto resultando na formalização do Documento de Visão da Implantação. Em seguida, observou-se que os problemas estavam distribuídos ao longo de momentos distintos do ciclo de implantação, o que motivou a organização do método em fases claramente delimitadas.

A fase de Iniciação surgiu da necessidade de formalizar o alinhamento inicial entre cliente e equipe de Produtos. A fase de Elaboração foi estruturada para mitigar riscos técnicos e garantir completude das informações antes do desenvolvimento. A fase de Construção consolidou as práticas de codificação, testes comparativos e geração de evidências. Por fim, a fase de Entrega formalizou a transição para produção e o acompanhamento operacional.

Dessa forma, as fases da AIIPD não foram definidas arbitrariamente, mas derivadas

diretamente da análise dos pontos críticos identificados no processo existente. A estrutura final da abordagem representa, portanto, uma resposta sistematizada aos problemas observados na prática organizacional, incorporando princípios iterativos e disciplina processual inspirados na Engenharia de Software.

5.1.1 Evolução do processo: do modelo anterior à AIIPD

Antes da concepção da AIIPD, o processo de implantação de produtos de dados em lote na organização não possuía uma estrutura formalmente definida além da etapa inicial de levantamento de requisitos, conduzida pela equipe de Produtos. Na prática, o que existia era uma fase equivalente à Iniciação, na qual eram discutidos escopo e objetivos do produto. A partir desse ponto, as atividades subsequentes incluindo refinamento técnico, desenvolvimento na plataforma de big data, testes e entrega ocorriam de forma pouco estruturada, variando conforme a experiência individual dos envolvidos.

Não havia delimitação clara entre momentos de análise técnica, implementação e validação. Tampouco existiam artefatos padronizados para formalizar a transição entre equipes. As fases que hoje são denominadas Elaboração, Construção e Entrega não eram explicitamente reconhecidas como etapas distintas do processo, o que dificultava rastreabilidade, controle de responsabilidades e previsibilidade de prazos.

A AIIPD introduziu uma reorganização estrutural do fluxo de implantação, formalizando fases antes implícitas e estabelecendo critérios de entrada e saída para cada uma delas. A fase de Elaboração passou a consolidar documentação técnica e validação arquitetural antes do desenvolvimento. A fase de Construção estruturou a implementação incremental e os testes comparativos entre ambientes. Já a fase de Entrega passou a incluir validação formal, geração de evidências e acompanhamento pós-implantação. Dessa forma, o processo evoluiu de uma sequência informal de atividades para um fluxo semi-sequencial estruturado, com papéis, artefatos e responsabilidades claramente definidos.

5.2 EQUIPES E RESPONSABILIDADES

Equipes das áreas de Produtos, Dados & Analytics, Tecnologia e Operação & Delivery participam ativamente do ciclo de vida do projeto de implantação de um produto de dados em lote dentro da organização. Cada equipe desempenha um papel específico e essencial, com responsabilidades claramente definidas para garantir o sucesso das entregas e o alinhamento com os objetivos estratégicos do negócio. Abaixo, detalhamos as responsabilidades e interações entre as equipes no contexto da abordagem proposta.

5.2.1 Equipe de Produtos

Responsável por compreender profundamente as necessidades e expectativas do cliente, realizando entrevistas, reuniões e análises quantitativas para definir e validar re-

quisitos. A equipe avalia possíveis resultados esperados com base em dados e insights, priorizando ações de acordo com critérios estratégicos e impacto potencial. Além disso, atua diretamente no desenvolvimento das soluções propostas, garantindo que cada entrega incremental esteja alinhada com os objetivos definidos e acompanhando continuamente os resultados para garantir a satisfação do cliente e a geração de valor para a organização.

5.2.2 Equipe de Dados & Analytics

Responsável pela construção, validação e manutenção de modelos estatísticos como *Score* de Crédito, *Score* de Comportamento, Churn, entre outros. Esta equipe monitora continuamente a performance e a acurácia dos modelos, realizando ajustes e melhorias com base em novos dados e análises estatísticas detalhadas. Além disso, a equipe de Dados & Analytics é responsável pela produção de uma documentação detalhada, contendo todas as especificações técnicas e metodológicas relacionadas aos modelos. Colabora estreitamente com a equipe de Tecnologia no esclarecimento de dúvidas técnicas, apoio na resolução de problemas operacionais e execução de testes para garantir a qualidade e robustez dos modelos desenvolvidos.

5.2.3 Equipe de Tecnologia

Responsável pela implantação dos produtos *batch* dentro do ambiente de computação de alto desempenho e *big data*. A equipe realiza uma análise detalhada da documentação produzida pela equipe de Dados & Analytics, garantindo a clareza e a completude das informações necessárias para o desenvolvimento. Procede com a codificação, otimização e implementação das soluções no ambiente produtivo, assegurando que todas as especificações técnicas e requisitos de performance sejam atendidos. Caso identifique inconsistências ou insuficiências na documentação, notifica imediatamente as áreas responsáveis, coordenando os ajustes necessários. Além disso, realiza testes rigorosos para assegurar a integridade e a estabilidade das soluções implantadas.

5.2.4 Operação & Delivery

Responsável pelo monitoramento contínuo dos produtos após sua implantação em ambiente produtivo. Esta equipe acompanha o desempenho dos produtos entregues, assegurando sua operacionalidade e eficiência no uso contínuo. Atua proativamente na identificação e reporte de incidentes, coletando e registrando feedbacks dos usuários finais para garantir que as soluções estejam aderentes às necessidades reais do negócio. Adicionalmente, a equipe é responsável por coordenar solicitações de melhorias, correções e ajustes evolutivos, mantendo uma comunicação constante com as equipes de Produtos, Dados & Analytics e Tecnologia para garantir que todas as necessidades operacionais sejam prontamente atendidas.

5.3 FASES DO CICLO DE VIDA DO PROJETO DE IMPLANTAÇÃO DE PRODUTO EM LOTE

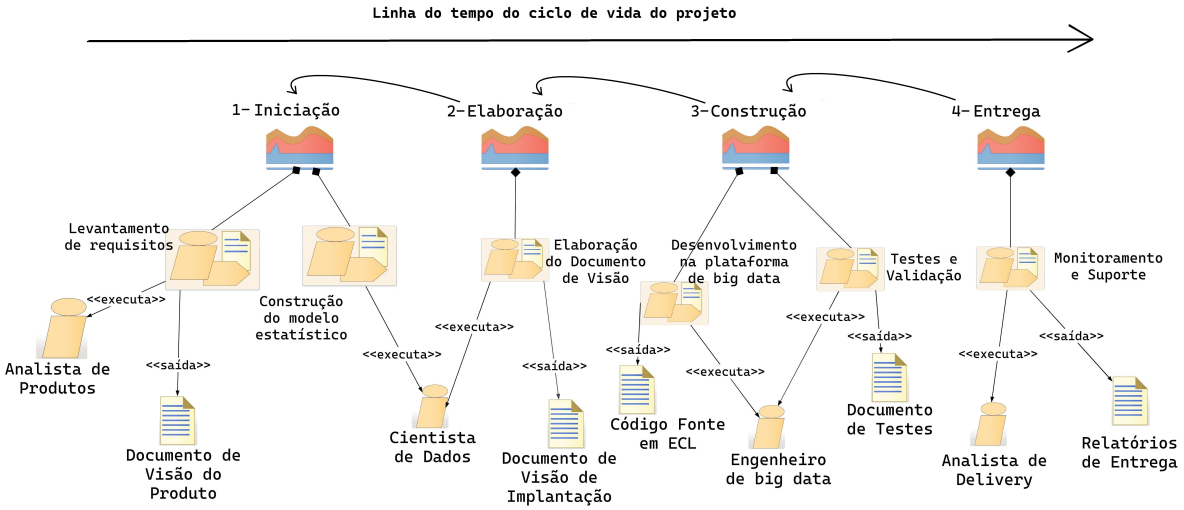
A Figura 10 ilustra o fluxo de trabalho da AIIPD em linguagem do modelo *Software & Systems Process Engineering Metamodel* (SPEM) (Sparx Systems, 2022), metamodelo da OMG que descreve papéis, tarefas, artefatos e fluxos de trabalho de processos de engenharia de software e sistemas, destacando a interação entre as equipes e os artefatos gerados em cada uma das quatro fases.

- **Atores** (representados por ícones de pessoas): Indicam qual equipe ou papel é responsável por uma determinada ação ou pela criação de um documento.
- **Ações/Documentos** (representados por ícones de documentos e caixas de processo): Simbolizam os entregáveis e as principais atividades de cada fase.
- **Linhas de Conexão**: As setas indicam a sequência e a dependência entre as atividades. Por exemplo, uma linha que sai de um ator e aponta para um documento significa que aquele ator é responsável por criar ou modificar aquele artefato. Uma linha que conecta dois documentos ou ações demonstra a transição e a evolução do trabalho ao longo do ciclo de vida.

O fluxo inicia-se na fase de Iniciação, onde a equipe de Produtos, a partir das necessidades do cliente, gera o Documento de Visão inicial. Na fase de Elaboração, este documento é enriquecido pela equipe de Dados & Analytics com as especificações do modelo e, em seguida, complementado pela equipe de Tecnologia com os requisitos técnicos para a implantação. A colaboração e a passagem de responsabilidade são representadas pelas linhas que ligam as equipes aos sucessivos aprimoramentos do documento.

Na fase de Construção, a equipe de Tecnologia utiliza o documento de visão finalizado para desenvolver os scripts e gerar a documentação de evidência dos testes. A linha que conecta este documento de volta à equipe de Dados & Analytics representa o processo de validação. Finalmente, na fase de Entrega, o produto implantado e processado é entregue ao cliente, finalizando o ciclo de implantação, com o acompanhamento da equipe de Operação & Delivery. Esse fluxo é iterativo, com pontos de passagem e validação claros, garantindo o alinhamento e a qualidade do produto final.

Figura 9 – Ciclo de vida do projeto da abordagem proposta, representado utilizando a notação SPEM



Fonte: Elaborado pelo autor.

Dentro de cada fase, são realizadas interações semanais entre os envolvidos no projeto para monitorar e avaliar o progresso, sendo que cada interação culmina no desenvolvimento de um produto incremental, seja ele um software ou um documento do projeto. Após a conclusão bem-sucedida de todos os entregáveis definidos em uma determinada fase, uma breve reunião deve ser realizada para encerrar formalmente a fase e avançar para a próxima. A seguir, detalharemos cada fase do ciclo de vida do projeto.

É importante destacar que, embora a abordagem apresente uma organização em fases claramente delimitadas, sua aplicação prática ocorre de maneira semi-sequencial e adaptativa. Essa escolha é intencional e está alinhada ao contexto organizacional no qual a pesquisa foi conduzida. A equipe envolvida na implantação de produtos de dados em lote é relativamente pequena e composta por profissionais com perfis multidisciplinares, que acumulam responsabilidades técnicas e de coordenação. Nesse cenário, um modelo excessivamente complexo ou altamente formalizado poderia gerar sobrecarga processual e reduzir a eficiência operacional.

A estrutura semi-sequencial adotada permite manter clareza quanto às responsabilidades, artefatos produzidos e pontos de validação, ao mesmo tempo em que possibilita revisões e ajustes entre fases sempre que necessário. Dessa forma, combina-se disciplina processual suficiente para garantir rastreabilidade e governança com flexibilidade prática compatível com equipes enxutas. Essa característica mostrou-se adequada ao ambiente estudado, no qual a proximidade entre os membros da equipe facilita comunicação direta e rápida tomada de decisão.

5.3.1 Fase de Iniciação

O principal objetivo é estabelecer o escopo do produto, promovendo o entendimento das necessidades do cliente e das partes interessadas. Uma compreensão de alto nível dos requisitos do projeto é essencial para mitigar riscos potenciais e desenvolver um caso de negócio robusto. Um documento de visão inicial deve ser elaborado, incorporando esses requisitos. A Tabela 4 apresenta as atividades a serem realizadas durante a fase de Iniciação e as respectivas responsabilidades. Ao final desta fase, será entregue um documento de visão centralizado.

Tabela 4 – Atividades e equipe responsável pela fase de Iniciação

Atividade (*Activity*)	**Equipe (*Role*)**
Alinhamento do conceito do produto e das variáveis de resposta.	Equipe de Produtos
Alinhamento das amostras de desenvolvimento.	Equipe de Produtos
Alinhamento do cronograma de envio das amostras e do formato de comunicação para o envio (ex: Cloud, SFTP, Connect Direct).	Equipe de Tecnologia
Alinhamento da data, expectativa de consumo e formato.	Equipe de Produtos

5.3.2 Fase de Elaboração

Nesta fase preparatória, riscos críticos são mitigados para permitir atualizações nas estimativas de custo e cronogramas, além de garantir a aprovação das partes interessadas. O foco está na redução de riscos técnicos chave, assegurando que todas as informações necessárias estejam incluídas no documento de visão e verificando a compatibilidade com a plataforma. Se necessário, a equipe de tecnologia se comunica com outras equipes para adicionar informações ao documento. Também são definidos os testes a serem realizados e os critérios de aceitação que devem ser atendidos durante a fase de construção. A Tabela 5 apresenta as atividades desta fase e suas respectivas responsabilidades.

Tabela 5 – Atividades e equipe responsável pela fase de Elaboração

Atividade (*Activity*)	**Equipe (*Role*)**	**Artefato (*Work Product*)**
Enriquecimento da amostra enviada pelo cliente com variáveis e indicadores no corte determinado pelo cliente.	Equipe de Dados & Analytics	Amostras enriquecidas com as variáveis ou amostras para o cliente
Feedback do cliente com fórmulas e bases de dados utilizadas para o desenvolvimento do produto de dados.	Equipe de Produtos	Não se aplica
Complementação do documento de visão com as fórmulas, atributos utilizados e demais informações importantes sobre a modelagem analítica.	Equipe de Dados & Analytics	Documento de visão parcial da implantação
Complementação do documento de visão com outras informações necessárias para a implementação do produto de dados.	Equipe de Tecnologia	Documento de Visão: Este documento fornece uma visão abrangente do desenvolvimento do produto pretendido.
Elaboração do cronograma de implantação conforme a priorização.	Equipe de Tecnologia	Documento de visão final da implantação: Este documento serve como um recurso consolidado para os desenvolvedores, garantindo um processo de implantação fluido. Inclui detalhes sobre os modelos estatísticos e é completado utilizando macros para assegurar que todas as informações essenciais estejam incluídas.

5.3.3 Fase de Construção

Durante esta fase, ocorre o desenvolvimento dos scripts da plataforma de *big data* para criar a versão operacional inicial do produto. Diversas versões internas garantem a usabilidade e o alinhamento com os requisitos do cliente. Uma versão beta funcional deve estar disponível para testes rigorosos. Todas as validações e testes necessários são realizados, com foco no cumprimento dos critérios de aceitação. A equipe de Operação & Delivery valida minuciosamente o produto, garantindo que os critérios de aceitação

sejam atendidos. A Tabela 6 apresenta uma descrição das atividades desta fase e suas respectivas responsabilidades.

Tabela 6 – Atividades e equipe responsável pela fase de Construção

Atividade (*Activity*)	**Equipe (*Role*)**	**Artefato (*Work Product*)**
Configurações da plataforma de *big data* e dos relatórios de entrega do produto.	Equipe de Tecnologia	Plataforma de *big data* com o produto implantado e pronto para produção.
Solicitação de amostras do produto para aprovação.	Equipe de Tecnologia e Equipe de Dados & Analytics	Kit com bases de dados contendo o valor esperado.
Comparação do produto gerado no ambiente analítico com o produto gerado na plataforma de *big data* e elaboração de documento de evidência dos testes realizados.	Equipe de Tecnologia	Documento de evidência: Este documento registra os testes realizados pela equipe responsável pela implantação na plataforma de *big data*, seguindo os requisitos descritos no documento de visão da implantação. Inclui capturas de tela dos testes e explicações detalhadas dos procedimentos. As áreas envolvidas são notificadas para formalizar a aceitação ou rejeição dos casos de teste.
Validação dos resultados dos testes realizados pela equipe de Implantação.	Equipe de Dados & Analytics	Aceite ou rejeição do documento de evidência dos testes comparativos.

5.3.4 Fase de Entrega

A fase de Entrega se inicia assim que o produto estiver alinhado com os requisitos definidos na fase de elaboração. São realizados os preparativos para a entrada do produto em produção, e pequenos ajustes podem ser implementados com base no feedback do cliente. O feedback nesta etapa foca em refinamentos, configuração, instalação e usabilidade.

5.4 EXECUÇÃO DA ABORDAGEM PROPOSTA

Conforme detalhado anteriormente, a AIIPD foi elaborada para integrar práticas de engenharia de software e metodologias ágeis, visando solucionar as ineficiências identificadas no processo tradicional. A seguir, será apresentada a aplicação prática da aborda-

gem, detalhando como as suas fases e princípios foram implementados no contexto de um estudo de caso real em um *bureau* de crédito. A execução de dois projetos, aqui denominados **Modelo A** e **Modelo B**, permitiu não apenas a avaliação da AIIPD, mas também sua evolução contínua. As observações e os resultados desta etapa serão posteriormente analisados no Capítulo 6.

5.4.1 Implantação dos Modelos A e B

5.4.1.1 Implantação do Modelo A: Estabelecimento de um Processo Base

O Modelo A consistia em um *score* de crédito que utiliza o algoritmo de *gradient boosting* XGBoost. O desenvolvimento e a modelagem do produto foram conduzidos pela equipe de Dados & Analytics, utilizando a linguagem de programação *R* no ambiente de desenvolvimento RStudio, em uma plataforma de nuvem própria da empresa. Essas atividades aconteceram na fase **Iniciação** do ciclo de vida do projeto na AIIPD. Após a aprovação do modelo pelo cliente, iniciou-se a fase de **Elaboração** da abordagem, focada na criação da documentação necessária para a implantação.

Ainda na fase de **Elaboração**, para formalizar e padronizar o processo, foi elaborado um documento de visão detalhado, conhecido internamente como **documentação de implantação**. Este artefato se tornou a fonte única de verdade para todas as informações técnicas e de negócio do projeto. Para o controle de versionamento e acessibilidade, foi estabelecido um sistema de gestão de documentos no Sharepoint, onde cada projeto recebia uma estrutura de pastas e um código de implantação padronizado, garantindo a rastreabilidade e a organização do conhecimento.

A convenção para o código de implantação foi definida como **IMP-CLI2025-CRD01-PF**, sendo cada parte do código explicada da seguinte forma:

- **IMP:** Sufixo para identificar que se trata de uma implantação de produto.
- **CLI:** Sigla do cliente, permitindo uma identificação clara do destinatário do projeto.
- **2025:** Ano de desenvolvimento do produto.
- **CRD:** Tipo do produto (neste caso, "crédito"), com outras opções como **FRD** (fraude) ou **COB** (cobrança).
- **01:** Um número sequencial para diferenciar produtos do mesmo tipo para o mesmo cliente no mesmo ano.
- **PF:** Tipo de pessoa ("pessoa física"), podendo ser também **PJ** ("pessoa jurídica").

O documento de visão foi elaborado como um arquivo no formato Word, com o uso de macros, para que o analista de modelagem pudesse preencher os campos com as informações necessárias de forma estruturada. A colaboração entre as equipes de Dados &

Analytics e Tecnologia foi fundamental nessa etapa, garantindo que o documento atendesse tanto aos requisitos do modelo quanto às necessidades de desenvolvimento.

Na fase de **Construção**, a equipe de Tecnologia utilizou o documento de visão para desenvolver os scripts necessários para a implantação do modelo na plataforma de *big data* HPCC. Após a conclusão do desenvolvimento, foi elaborado um documento de evidência dos testes realizados, comparando os resultados do modelo no ambiente analítico com os resultados obtidos na plataforma de *big data*. A equipe de Dados & Analytics validou esses resultados, assegurando que o modelo implantado atendesse aos critérios de aceitação definidos previamente.

Na fase de **Entrega**, o produto foi oficialmente implantado em ambiente produtivo e entregue ao cliente. A equipe de Operação & Delivery assumiu a responsabilidade pelo monitoramento contínuo do produto, garantindo sua operacionalidade e eficiência no uso diário. Para execuções pontuais, foi realizada uma passagem de conhecimento para a equipe de Operação & Delivery, assegurando que todos os aspectos técnicos e operacionais do produto fossem compreendidos e gerenciados adequadamente. Essa passagem de conhecimento foi gravada em vídeo e documentada para futuras referências.

5.4.1.2 Implantação do Modelo B: Aprimoramento e Automação

A implantação do Modelo B, que também utilizava o algoritmo *XGBoost*, representou uma oportunidade para aplicar as lições aprendidas durante o projeto anterior e aprimorar a abordagem. As equipes, com base na experiência do Modelo A, identificaram a necessidade de adicionar novos campos à documentação de implantação, tornando o documento ainda mais completo e reduzindo ambiguidades.

O principal avanço técnico desta fase foi a identificação de um padrão nos scripts de predição dos modelos *XGBoost*. Essa constatação levou ao desenvolvimento de uma ferramenta de automação: um conversor de script que traduzia o código desenvolvido em *R* diretamente para a linguagem ECL, utilizada pela plataforma HPCC. Essa inovação resultou em uma redução de 50% no tempo de desenvolvimento dos scripts em ECL, otimizando significativamente a fase de Construção.

Foi realizada uma reunião com todas as pessoas de Dados & Analytics, onde o analista de modelagem pode passar para os colegas de equipe as melhores práticas e os aprendizados obtidos durante a implantação do Modelo A. Essa troca de conhecimento foi essencial para alinhar as expectativas e garantir que todos estivessem cientes das melhorias implementadas na abordagem na fase de **Elaboração**.

5.4.2 Testes e Aceite Multifuncional

Para ambos os modelos (A e B), a fase de testes foi conduzida de forma colaborativa, envolvendo todas as equipes multidisciplinares. Ao final de cada etapa do processo de implantação, todas as áreas precisaram formalmente dar seu aceite, o que garantiu dois

pontos cruciais: a qualidade e a confiabilidade dos modelos implantados e o engajamento e alinhamento de todas as pessoas envolvidas. Este processo reforçou a responsabilidade compartilhada e assegurou que o produto final atendesse às expectativas de todas as partes interessadas.

5.5 EXEMPLO DIDÁTICO DE APLICAÇÃO DA AIIPD

Com o objetivo de ilustrar de forma concreta a aplicação da AIIPD, apresenta-se nesta seção um exemplo didático de implantação de um produto de dados em lote. O exemplo não representa código ou dados reais da organização estudada, sendo uma abstração construída para fins acadêmicos, preservando confidencialidade e aspectos estratégicos internos.

5.5.1 Contexto do Caso Fictício

Considera-se um produto denominado *Score de Risco Pessoa Física*, cujo objetivo é calcular, em ambiente de *big data*, um score baseado em variáveis como renda, histórico de atraso e indicadores comportamentais. O modelo estatístico foi previamente desenvolvido em ambiente de modelagem (RStudio) e, conforme a AIIPD, necessita ser implantado na plataforma de processamento distribuído para execução em lote.

O produto é segmentado por perfis de cliente, sendo cada segmento responsável por regras específicas de transformação. A implantação exige organização estruturada de artefatos, definição clara de responsabilidades e validação rigorosa dos resultados.

5.5.2 Estrutura Organizacional do Projeto

A organização modular do projeto segue o princípio de separação de responsabilidades adotado na AIIPD. A estrutura simplificada é apresentada a seguir:

```
root/
  Delivery/
       Scores/
            Cliente/
                 Produto/
                      BWR/
                          BWR_RunBatch.ecl
                      SEG_A/
                          fModelCalculate.ecl
                          fModelJSON.ecl
                      Testes/
                          BWR_Test.ecl
```

```
            fAttributes.ecl
            fCalculatedR.ecl
        macRunBatch.ecl
        modConstants.ecl
```

Nessa estrutura:

- A pasta `SEG_A` contém regras específicas do segmento, incluindo a árvore do modelo em formato JSON e as transformações aplicadas.
- O arquivo `modConstants.ecl` centraliza constantes, parâmetros e definições globais do modelo.
- A macro `macRunBatch.ecl` orquestra a execução consolidada dos segmentos.
- A pasta `Testes` contém funções auxiliares e rotinas de validação comparativa.

Essa organização favorece rastreabilidade, versionamento e clareza estrutural, alinhando-se à fase de Construção definida na AIIPD.

5.5.3 Pipeline de Execução

O pipeline de execução do produto segue as seguintes etapas:

1. Extração dos atributos previamente calculados;
2. Aplicação das regras de transformação do modelo por segmento;
3. Consolidação dos resultados por meio de macro orquestradora;
4. Execução de testes comparativos com base de referência;
5. Geração do dataset final para entrega.

A fase de Elaboração da AIIPD garante que atributos e regras estejam formalizados antes da implementação. A fase de Construção contempla o desenvolvimento incremental e execução dos testes. Por fim, a fase de Entrega formaliza a geração do produto final e seus relatórios de validação.

5.5.4 Trecho Fictício de Código ECL

A seguir apresenta-se um exemplo simplificado de função de cálculo do score:

```
EXPORT fModelCalculate(DATASET(Layout_Input) ds) := FUNCTION
    RETURN PROJECT(ds,
        TRANSFORM(Layout_Output,
            SELF.Score := (LEFT.Renda / 1000)
                        - (LEFT.AtrasoMeses * 5)
                        + LEFT.ScoreComportamental;
            SELF := LEFT;
```

```
        )
    );
END;
```

A macro responsável pela consolidação pode ser representada de forma simplificada:

```
EXPORT macRunBatch := MACRO
    OUTPUT(SEG_A::fModelCalculate(InputA));
    OUTPUT(SEG_B::fModelCalculate(InputB));
ENDMACRO;
```

Essa estrutura demonstra a separação entre lógica de negócio, organização por segmento e orquestração final.

5.5.5 Validação Comparativa e Prática Inspirada em TDD

A validação do modelo implantado é realizada por meio de comparação entre os resultados calculados na plataforma de *big data* e os valores previamente obtidos no ambiente de modelagem estatística (RStudio). Essa prática é inspirada nos princípios de *Test-Driven Development* (TDD), adaptados ao contexto de produtos analíticos.

Um exemplo simplificado de validação pode ser representado da seguinte forma:

```
Diff := JOIN(CalcECL, CalcR,
             LEFT.ID = RIGHT.ID,
             TRANSFORM({STRING ID; DECIMAL diff},
                 SELF.ID := LEFT.ID;
                 SELF.diff := LEFT.Score - RIGHT.Score;
             ));

ASSERT(COUNT(Diff(diff != 0)) = 0,
       'Erro de validação: divergência entre R e ECL');
```

Caso divergências sejam identificadas, a execução é interrompida, impedindo a promoção do produto para ambiente produtivo. Essa abordagem reduz riscos operacionais e garante consistência matemática entre ambientes.

5.5.6 Relação com as Fases da AIIPD

O exemplo apresentado evidencia como a AIIPD estrutura a implantação:

- **Iniciação:** definição do escopo do produto e requisitos do modelo;
- **Elaboração:** formalização dos atributos, constantes e regras de transformação;

- **Construção:** implementação modular por segmento e execução de testes comparativos;
- **Entrega:** consolidação do batch e geração de evidências de validação.

Assim, o exemplo didático demonstra que a abordagem proposta não se limita a uma organização conceitual de fases, mas orienta efetivamente a estrutura técnica, a governança de artefatos e a validação operacional do produto de dados em lote.

5.5.7 Reutilização da estrutura de implantação em outros contextos

Embora a AIIPD tenha sido concebida inicialmente para a implantação de produtos de dados em lote na plataforma de big data, observou-se posteriormente que a estrutura de artefatos e organização em fases passou a ser reutilizada em outros fluxos de desenvolvimento da organização.

Em especial, a equipe responsável pelo desenvolvimento de APIs passou a adotar elementos estruturais originalmente definidos para o contexto batch, como padronização documental, definição explícita de critérios de validação e formalização de responsabilidades entre equipes. Essa reutilização ocorreu de maneira orgânica, a partir do reconhecimento de que os artefatos e práticas definidos na AIIPD contribuíam para maior clareza técnica e redução de retrabalho.

Esse desdobramento evidencia que a abordagem proposta não apenas solucionou os gargalos identificados na implantação de produtos em lote, mas também influenciou positivamente outros processos internos, ampliando seu impacto organizacional.

5.6 CONSIDERAÇÕES DO CAPÍTULO

Este capítulo apresentou a AIIPD, detalhando as equipes envolvidas (Produtos, Dados & Analytics, Tecnologia e Operação & Delivery) e as fases do ciclo de vida (Iniciação, Elaboração, Construção e Entrega) concebidas para mitigar ineficiências recorrentes na implantação de produtos de dados em lote por meio da integração de práticas de engenharia de software e métodos ágeis. Essas definições fornecem a estrutura necessária para padronizar papéis, artefatos e fluxos, endereçando lacunas identificadas na literatura e na prática organizacional.

A execução da abordagem em dois projetos (Modelos A e B) consolidou um processo base com documentação de visão padronizada, versionamento e convenções de codificação e promoveu melhorias incrementais, incluindo a automação da tradução de scripts de predição de *R* para ECL, o que reduziu sensivelmente o tempo de desenvolvimento. O ciclo de testes e aceite multifuncional reforçou a qualidade, a confiabilidade e o alinhamento entre as áreas. Esses resultados práticos, quantitativos e qualitativos, constituem a base para a avaliação de eficácia da AIIPD apresentada no capítulo seguinte.

6 AVALIAÇÃO DA ABORDAGEM PROPOSTA

Nesta etapa, os efeitos da ação são capturados e analisados. Primeiro, os dados foram coletados ao longo de seis meses de uso da abordagem. Subsequentemente, cada objetivo foi analisado com base nos dados coletados.

6.1 COLETA DE DADOS

A coleta de dados foi realizada durante toda a duração deste estudo. Para dados quantitativos, utilizou-se a ferramenta de gestão de atividades Jira Software (Jira Software, 2025), já empregada pelas equipes. Para dados qualitativos, aplicou-se um questionário no Microsoft Forms ao final da pesquisa-ação.

O questionário, derivado da abordagem GQM utilizada, foi enviado ao término do estudo a todos os membros das equipes que participaram do processo de implantação da abordagem proposta (ver Apêndice B). Foram recebidas oito respostas: um membro da equipe de Produto, três da equipe de Dados & Analytics, dois da equipe de Tecnologia e dois da equipe de Entrega (Figura 10).

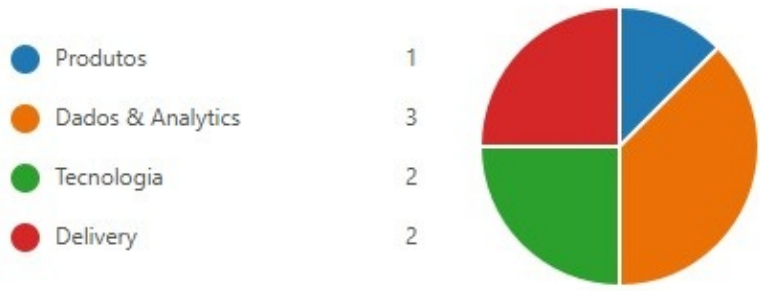
6.2 ANÁLISE DOS DADOS

Após o processo de coleta de dados, foi realizada uma análise para verificar se os objetivos previamente definidos foram alcançados. A análise é apresentada para cada objetivo, com respostas sistematicamente agrupadas com base nos dados coletados.

- Q1.1 - Qual é o tempo, em dias, para implantação de produtos de dados batch em produção?

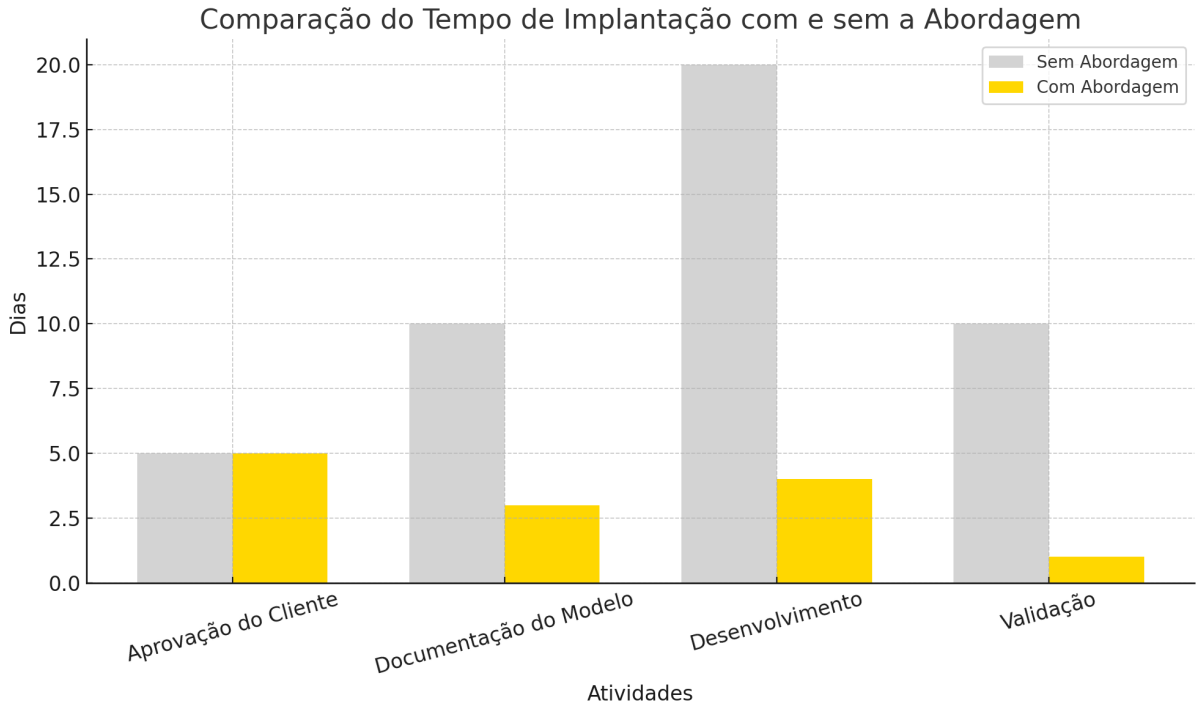
A análise comparativa dos tempos de implantação de produtos de dados batch, com e sem a abordagem proposta, demonstra melhorias significativas em todas as etapas do processo. A Figura 11 ilustra essas diferenças. O tempo requerido para aprovação do modelo pelo cliente permaneceu constante em cinco dias para ambas as abordagens, indi-

Figura 10 – Equipes que responderam ao questionário



Fonte: Elaborado pelo autor.

Figura 11 – Comparação do tempo de implantação com e sem a abordagem.



Fonte: Elaborado pelo autor.

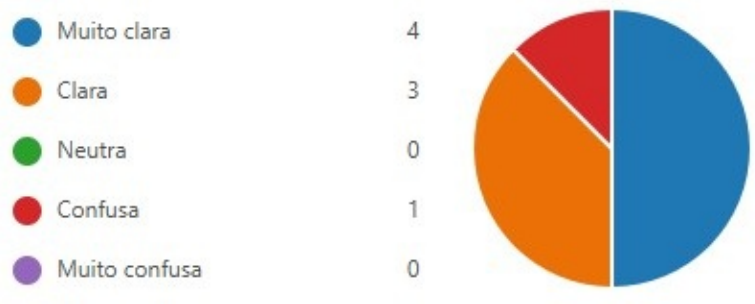
cando que fatores externos e administrativos, além do escopo da abordagem, influenciaram essa fase.

No entanto, observou-se uma redução notável no tempo necessário para a criação da documentação, que caiu de 10 dias para 3 dias uma melhoria de 70%. Essa redução pode ser atribuída à padronização dos processos de documentação e à introdução de modelos mais eficientes. De forma semelhante, a fase de desenvolvimento na plataforma de *big data*apresentou uma redução significativa no tempo, de 20 dias para 4 dias, representando uma melhoria de 80%. Esse resultado reflete uma melhor coordenação das equipes e a implementação eficaz de práticas de desenvolvimento ágil. Por fim, a fase de validação demonstrou a melhoria mais substancial, com o tempo requerido reduzido de 10 dias para 1 dia (uma redução de 90%). Essa melhoria destaca a incorporação de mecanismos de validação eficientes e o estabelecimento de um fluxo de trabalho colaborativo e integrado entre as equipes de desenvolvimento e validação.

- Q2.1 - Quão claras são as etapas do processo de implantação ao usar a abordagem?

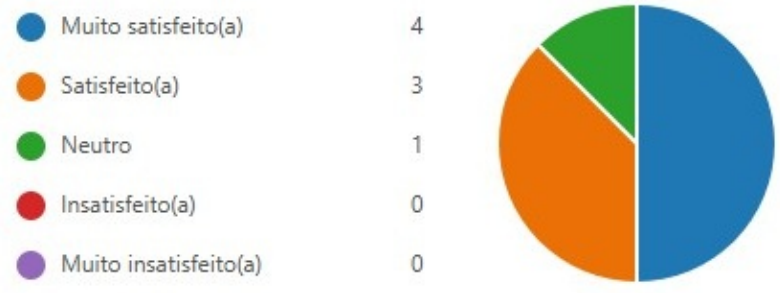
Esses dados foram obtidos por meio do questionário. A maioria dos respondentes avaliou as etapas do processo de implantação como claras ou muito claras, totalizando sete respostas positivas de oito. Especificamente, 50% dos respondentes (4 de 8) classificaram as etapas como **muito claras**, demonstrando excelente compreensão por metade dos participantes. Além disso, 37,5% (3 de 8) avaliaram as etapas como **claras**, indicando que

Figura 12 – Clareza das etapas do processo de implantação.



Fonte: Elaborado pelo autor.

Figura 13 – Nível de satisfação com a abordagem de implantação.



Fonte: Elaborado pelo autor.

quase dois quintos dos participantes consideraram as etapas compreensíveis. No entanto, 12,5% (1 de 8) consideraram as etapas **confusas**, destacando uma pequena parcela de participantes que encontrou dificuldades na compreensão do processo (Figura 12).

- Q2.2 - Qual é o nível de satisfação dos membros da equipe com a abordagem?

A abordagem foi bem recebida pela maioria dos respondentes, com 50% (4 de 8) classificando seu nível de satisfação como **muito satisfeito** e 37,5% (3 de 8) como **satisfeito**. Juntas, essas respostas positivas representam 87,5% do feedback. Apenas 12,5% (1 de 8) classificaram sua satisfação como **neutra**, indicando nem forte aprovação nem insatisfação (Figura 15).

- Q3.1 - A abordagem facilita a colaboração entre as diferentes equipes?

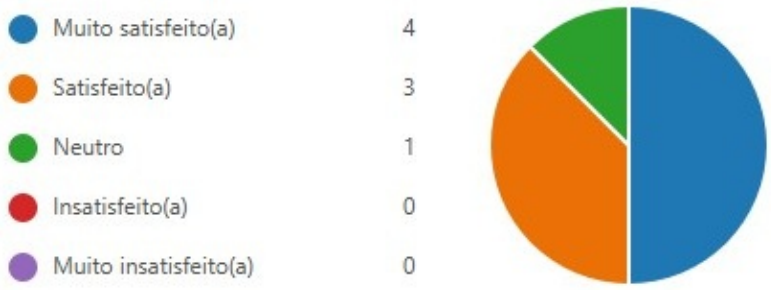
Todos os respondentes concordaram que a abordagem facilitou a colaboração entre as diferentes equipes. Especificamente, 62,5% (5 de 8) **concordaram totalmente**, refletindo um forte consenso quanto à eficácia da abordagem em promover o trabalho em equipe. Além disso, 37,5% (3 de 8) **concordaram**, indicando um sentimento geral positivo em relação à abordagem. Nenhuma resposta neutra ou negativa foi registrada, ressaltando a concordância unânime quanto aos benefícios colaborativos da abordagem (Figura 14).

- Q3.2 - As responsabilidades de cada equipe estão bem definidas e compreendi-

Figura 14 – Percepção sobre a colaboração entre as equipes.

Fonte: Elaborado pelo autor.

Figura 15 – Percepções sobre a clareza dos papéis e definição de responsabilidades.



Fonte: Elaborado pelo autor.

das?

A maioria dos respondentes concordou que as responsabilidades das equipes estavam claramente definidas e bem compreendidas. Especificamente, 50% (4 de 8) **concordaram totalmente**, enquanto 37,5% (3 de 8) **concordaram**. Um respondente (12,5%) expressou uma posição **neutra**, indicando margem para maior esclarecimento ou melhoria na comunicação (Figura 15).

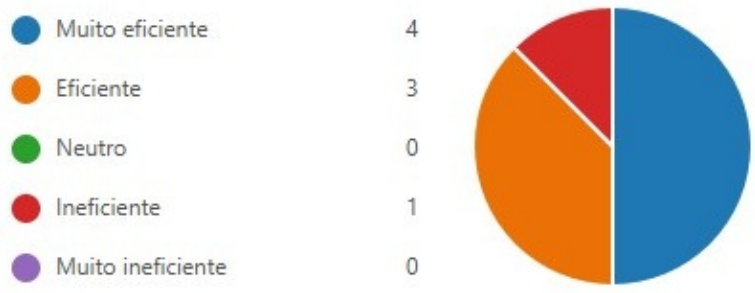
- Q4.1 - Quão eficiente é o processo de documentação?

O processo de documentação foi predominantemente avaliado como eficiente ou muito eficiente, com 87,5% dos respondentes fornecendo feedback positivo. Entre esses, 50% (4 de 8) classificaram o processo como **muito eficiente**, enquanto 37,5% (3 de 8) o classificaram como **eficiente**. No entanto, um respondente (12,5%) avaliou o processo como **ineficiente**, sugerindo uma área para possível melhoria (Figura 16).

- Q4.2 - O processo de validação e testes é rigoroso e eficaz?

O processo de validação e testes foi avaliado positivamente pela maioria dos respondentes, com 62,5% (5 de 8) **concordando totalmente** e 37,5% (3 de 8) expressando uma posição **neutra**. Embora nenhum feedback negativo tenha sido registrado, as respostas neutras destacam uma oportunidade de reforçar ainda mais a confiança nesse aspecto da abordagem (Figura 17).

Figura 16 – Avaliação da eficiência da documentação.



Fonte: Elaborado pelo autor.

Figura 17 – Avaliação do processo de validação e testes.

Fonte: Elaborado pelo autor.

6.3 DISCUSSÃO

A implementação da abordagem proposta demonstrou melhorias substanciais na eficiência da implantação de produtos de dados batch. As reduções nos tempos de documentação, desenvolvimento e validação evidenciam sua eficácia em agilizar processos e aprimorar a coordenação.

Embora a maioria dos participantes tenha considerado o processo de implantação claro e as responsabilidades bem definidas, um feedback isolado sugere margem para refinamento na comunicação e esclarecimento de papéis. De forma semelhante, a recepção positiva dos processos de documentação e validação destaca suas forças, embora melhorias adicionais possam abordar as poucas respostas neutras ou críticas.

A concordância unânime quanto aos benefícios colaborativos da abordagem potencialmente pode contribuir para o sucesso em promover o trabalho em equipe e quebrar silos, um fator crítico em projetos complexos de *big data*. No geral, os resultados levantam indícios iniciais sobre a utilidade da abordagem e fornecem uma base para refinamentos futuros visando garantir maior aplicabilidade e satisfação.

6.4 AMEAÇAS À VALIDADE

Embora este estudo ofereça insights valiosos sobre a implantação de produtos de dados batch em ambientes de big data, algumas limitações devem ser reconhecidas. Os resultados podem carecer de generalização, pois a pesquisa foi conduzida em um único bureau de crédito com infraestrutura e práticas específicas. Fatores organizacionais, como experiência das equipes e recursos disponíveis, também podem limitar a replicabilidade da abordagem em outros contextos. Potenciais vieses na coleta e análise dos dados, juntamente com avaliações subjetivas, podem afetar a validade interna. Por fim, a rápida evolução das tecnologias e práticas da indústria de *big data*exige atualizações constantes para manter a relevância da abordagem.

6.5 CONSIDERAÇÕES DO CAPÍTULO

A avaliação da abordagem apresentada neste capítulo revela indícios iniciais, refletidos nos resultados positivos da pesquisa-ação. Os dados coletados demonstraram melhorias substanciais nos tempos de implantação, na qualidade da documentação e na colaboração entre as equipes. Os resultados qualitativos, obtidos por meio de questionários, indicaram alta satisfação dos colaboradores e maior clareza nas responsabilidades. As conclusões gerais do estudo, bem como as propostas para trabalhos futuros, serão apresentadas no próximo e último capítulo.

7 CONCLUSÃO E TRABALHOS FUTUROS

7.1 CONCLUSÕES

A crescente adoção de soluções baseadas em Ciência de Dados e *big data* tem ampliado significativamente a complexidade do ciclo de vida de produtos analíticos nas organizações. Embora modelos consolidados, como CRISP-DM e abordagens ágeis, ofereçam diretrizes consistentes para as etapas iniciais de exploração, modelagem e validação, observa-se uma lacuna persistente na formalização da fase de implantação em produção, especialmente em ambientes de processamento em lote.

Esta dissertação partiu dessa lacuna, identificada tanto na literatura quanto na prática organizacional, e buscou responder à seguinte questão de pesquisa: como estruturar um processo sistemático, colaborativo e eficiente para a implantação de produtos de dados em lote em ambientes de *big data*? A investigação revelou que, apesar da maturidade técnica das equipes envolvidas, inexistia um processo estruturado que organizasse de forma explícita papéis, artefatos, responsabilidades e critérios de validação na transição entre modelagem analítica e execução em larga escala.

A partir do diagnóstico conduzido por meio de pesquisa-ação, foi concebida a Abordagem Integrada de Implantação de Produtos de Dados (AIIPD). Diferentemente de frameworks tradicionais que concentram-se predominantemente nas fases de exploração e modelagem, a AIIPD formaliza a implantação como um processo composto por quatro fases semi-sequenciais: Iniciação, Elaboração, Construção e Entrega. Essa estrutura estabelece mecanismos explícitos de coordenação entre áreas multidisciplinares, promovendo maior previsibilidade e governança técnica.

A pesquisa-ação permitiu não apenas observar o processo organizacional, mas intervir ativamente em sua estrutura, possibilitando a implementação prática da abordagem. O processo anteriormente concentrado na fase de iniciação foi reorganizado com a criação formal das etapas de Elaboração, Construção e Entrega, cada uma com artefatos definidos, critérios de entrada e saída e responsabilidades associadas.

Os resultados obtidos indicaram melhorias observáveis tanto em métricas quantitativas quanto qualitativas. Houve redução no tempo médio de implantação, maior clareza na definição de responsabilidades e melhoria na qualidade da documentação técnica. A introdução de práticas sistemáticas de validação e testes inspiradas em princípios consolidados da Engenharia de Software elevou o nível de confiabilidade dos produtos implantados, reduzindo retrabalho e falhas decorrentes de inconsistências entre ambientes de desenvolvimento e produção.

Do ponto de vista científico, esta dissertação contribui ao formalizar a implantação como etapa estruturada no ciclo de vida de produtos de dados, ampliando a discussão sobre operacionalização analítica em ambientes de *big data*. A integração entre fundamentos da Engenharia de Software, práticas ágeis e necessidades específicas do processamento em lote

oferece uma perspectiva interdisciplinar que aproxima a Ciência de Dados da engenharia de processos.

A contribuição metodológica reside na aplicação rigorosa da pesquisa-ação como instrumento de transformação organizacional, permitindo validar a abordagem em ambiente real e analisar seus efeitos de forma sistemática. Essa característica fortalece o caráter aplicado da pesquisa e amplia sua relevância prática.

Sob a perspectiva organizacional, a abordagem impactou a dinâmica entre as equipes de Produto, Dados, Tecnologia e Operação. A formalização das fases reduziu ambiguidades, melhorou a comunicação interequipes e promoveu maior alinhamento entre objetivos de negócio e execução técnica. Observou-se ainda que artefatos estruturados para o processamento em lote foram posteriormente reutilizados em outros contextos internos, como integrações via APIs, indicando efeito multiplicador da intervenção.

Dessa forma, conclui-se que a AIIPD contribui para reduzir a distância entre experimentação analítica e operacionalização em larga escala, promovendo maior previsibilidade, governança e qualidade na entrega de produtos de dados em ambientes batch.

Reflexões sobre Maturidade Processual

Um dos aspectos mais relevantes evidenciados nesta pesquisa foi a necessidade de tratar a implantação como disciplina própria dentro do ciclo de vida de Ciência de Dados. A transição de modelos estatísticos para produtos executáveis em plataformas de alto desempenho exige não apenas competência técnica, mas coordenação processual. A formalização proposta neste estudo demonstra que a maturidade na implantação depende tanto da clareza estrutural quanto da colaboração entre áreas especializadas.

Limitações do Estudo

Apesar dos resultados consistentes, algumas limitações devem ser reconhecidas. A pesquisa foi conduzida em uma única organização, inserida no domínio financeiro, o que pode restringir a generalização imediata dos achados. A atuação do autor como pesquisador-praticante, embora fundamental para a efetividade da intervenção, pode ter influenciado determinadas dinâmicas organizacionais.

Além disso, a aplicação concentrou-se em processamento em lote utilizando infraestrutura tecnológica específica. A adaptação da abordagem para outros contextos tecnológicos ou arquiteturas orientadas a eventos não foi explorada neste estudo.

7.2 TRABALHOS FUTUROS

A pesquisa abre múltiplas possibilidades de aprofundamento:

- **Replicação intersetorial:** Aplicar a AIIPD em organizações de diferentes domínios para avaliar sua robustez e adaptabilidade.

- **Comparação experimental:** Conduzir estudos comparativos entre a AIIPD e outras abordagens de operacionalização analítica, utilizando métricas padronizadas.
- **Automação do ciclo de implantação:** Integrar a abordagem com ferramentas de CI/CD, MLOps e testes automatizados, ampliando o nível de automação e rastreabilidade.
- **Extensão para arquiteturas em tempo real:** Investigar adaptações da abordagem para cenários de processamento em fluxo e arquiteturas orientadas a eventos.
- **Avaliação longitudinal:** Analisar impactos de médio e longo prazo na manutenção evolutiva, incidência de falhas e indicadores organizacionais.
- **Formalização como framework replicável:** Estruturar a AIIPD como metamodelo formal baseado em SPEM, permitindo sua institucionalização como guia de referência.

Em síntese, esta dissertação reforça que a implantação de produtos de dados não deve ser tratada como etapa meramente operacional, mas como componente estruturante do ciclo de vida analítico. Ao propor, aplicar e avaliar uma abordagem formal em ambiente real, o estudo contribui para o amadurecimento da interface entre Engenharia de Software e Ciência de Dados, fortalecendo a governança e a qualidade na operacionalização de soluções baseadas em dados.

REFERÊNCIAS

AHMED, B.; DANNHAUSER, T.; PHILIP, N. A Lean Design Thinking Methodology (LDTM) for Machine Learning and Modern Data Projects, 2018.

AMERSHI, S.; BEGEL, A.; BIRD, C.; DELINE, R.; GALL, H.; KAMAR, E.; NAGAPPAN, N.; NUSHI, B.; ZIMMERMANN, T. Software engineering for machine learning: A case study, p. 291–300, 2019.

APACHE FLINK PROJECT. **Apache Flink Project**. [S.l.*:* s.n.], 2024. `https://flink.apache.org/`. Accessed: 2024.

APACHE SPARK PROJECT. **Apache Spark Project**. [S.l.*:* s.n.], 2024. `https://spark.apache.org/`. Accessed: 2024.

BECK, Kent. **Extreme Programming Explained: Embrace Change**. [S.l.]: Addison-Wesley Professional, 2000.

BEGOLI, Edmon; HOREY, James. Agile Methodologies in Big Data Projects. **IEEE Software**, v. 36, n. 3, p. 10–13, 2019.

BENDER-SALAZAR, R. Design thinking as an effective method for problem-setting and needfinding for entrepreneurial teams addressing wicked problems. **Journal of Innovation and Entrepreneurship**, v. 12, n. 1, 2023.

BOEHM, Barry W. A Spiral Model of Software Development and Enhancement. **Computer**, v. 21, n. 5, p. 61–72, 1988.

CARBONE, Paris; KATSIFODIMOS, Asterios; EWEN, Stephan; MARKL, Volker; HARIDI, Seif; TZOUMAS, Kostas. Apache Flink: Stream and Batch Processing in a Single Engine. **IEEE Data Engineering Bulletin**, v. 38, jan. 2015.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. CRISP-DM 1.0: Step-by-step data mining guide, 2000.

CHEN, Hong-Mei; KAZMAN, Rick; HAZIYEV, Serge. Agile Big Data Analytics Development: An Architecture-Centric Approach. In*:* 2016 49th Hawaii International Conference on System Sciences (HICSS). [S.l.*:* s.n.], 2016. p. 5378–5387.

CHEN, Min; MAO, Shiwen; LIU, Yunhao. Big data: A survey. **Mobile Networks and Applications**, v. 19, n. 2, p. 171–209, 2014.

COCKBURN, Alistair; HIGHSMITH, Jim. Agile software development: The people factor. **Computer**, v. 34, n. 11, p. 131–133, 2001.

CRUZES, Daniela S.; DYBA, Tore. Recommended Steps for Thematic Synthesis in Software Engineering. In*:* 2011 International Symposium on Empirical Software Engineering and Measurement. [S.l.*:* s.n.], 2011. p. 275–284.

CUNHA, A. F.; FERREIRA, D.; NETO, C.; ABELHA, A.; MACHADO, J. A CRISP-DM Approach for Predicting Liver Failure Cases: An Indian Case Study, 2021.

DEAN, Jeffrey; GHEMAWAT, Sanjay. MapReduce: simplified data processing on large clusters. **Commun. ACM**, v. 51, n. 1, p. 107–113, jan. 2008.

DIPTI KUMAR, Vijay; ALENCAR, Paulo. Software engineering for big data projects: Domains, methodologies and gaps. In*:* 2016 IEEE International Conference on Big Data (Big Data). [S.l.*:* s.n.], 2016. p. 2886–2895.

DYBÅ, Tore; DINGSØYR, Torgeir. Agile Project Management: From Self-Managing Teams to Large-Scale Development. **IEEE Software**, v. 31, n. 5, p. 49–55, 2014.

ESPINOSA, J. Alberto; ARMOUR, Frank. The Big Data Analytics Gold Rush: A Research Framework for Coordination and Governance, p. 1112–1121, 2016.

FLECKENSTEIN, Mike; FELLOWS, Lorraine. Overview of Data Management Frameworks. In*:* MODERN Data Strategy. [S.l.]: Springer International Publishing, Cham, 2018. p. 55–59.

FOWLER, Martin et al. **Manifesto for Agile Software Development**. [S.l.*:* s.n.], 2001. `https://agilemanifesto.org/`. Acessado em: 31 de julho de 2025.

GORTON, I.; BENER, A. B.; MOCKUS, A. Software Engineering for Big Data Systems. **IEEE Software**, v. 33, n. 2, p. 32–35, 2015.

GORTON, I.; BENER, A. B.; MOCKUS, A. Software Engineering for Big Data Systems. **IEEE Software**, v. 33, n. 2, p. 32–35, 2016.

GRADY, N. Challenges in Engineering for Big Data. In*:* ADVANCES in Computers. [S.l.]: Elsevier, 2017. v. 105. p. 1–35.

GROLINGER, K.; HIGASHINO, W. A.; TIWARI, A.; CAPRETZ, M. A. Data management in cloud environments: NoSQL and NewSQL data stores. **Journal of Cloud Computing: Advances, Systems and Applications**, v. 2, n. 1, p. 22, 2013.

HASHEM, I. A. T.; YAQOOB, I.; ANUAR, N. B.; MOKHTAR, S.; GANI, A.; KHAN, S. U. The rise of big data on cloud computing: Review and open research issues. **Information Systems**, v. 47, n. 1, p. 98–115, 2015.

HAUCK, Jean Carlo Rossa et al. Harmonizing MPS.BR and CERTICS: A Case Study in a Maturity Level F Organization, p. 61–70, 2015.

HODA, R.; SALLEH, N.; GRUNDY, J. The Rise and Evolution of Agile Software Development. **IEEE Software**, v. 35, n. 5, p. 58–63, 2018.

HPCC SYSTEMS. **HPCC Systems**. [S.l.*:* s.n.], 2024. `https://www.hpccsystems.com`. Accessed: 2024.

HUMBLE, Jez; FARLEY, David. **Continuous Delivery: Reliable Software Releases through Build, Test, and Deployment Automation**. [S.l.]: Addison-Wesley Professional, 2010.

HUMMEL, Oliver; EICHELBERGER, Holger; GILOJ, Andreas; WERLE, Dominik; SCHMID, Klaus. A Collection of Software Engineering Challenges for Big Data System Development. In*:* 2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA). [S.l.*:* s.n.], 2018. p. 362–369.

IRIZARRY, Rafael A. The Role of Academia in Data Science Education. **Harvard Data Science Review**, v. 2, n. 1, jan. 2020.

JIRA SOFTWARE. **Jira Software**. [S.l.*:* s.n.], 2025. `https://www.atlassian.com/software/jira`. Accessed: 2025.

JORGENSEN, Paul C. **Software Testing: A Craftsman's Approach**. 4th. [S.l.]: CRC Press, 2014.

JUNEJA, Prateek; KAUR, Preeti. Software Engineering for Big Data Application Development: Systematic Literature Survey Using Snowballing. In*:* 2019 International Conference on Computing, Power and Communication Technologies (GUCON). [S.l.*:* s.n.], 2019. p. 492–496.

KALLIO, Hanna; PIETILÄ, Anna-Maija; JOHNSON, Martin; KANGASNIEMI, Mari. Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. **Journal of Advanced Nursing**, v. 72, n. 12, p. 2954–2965, 2016. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1111/jan.13031`.

KITCHENHAM, Barbara. **Guidelines for Performing Systematic Literature Reviews in Software Engineering**. UK, 2007. Disponível em: `https://www.elsevier.com/__data/promis_misc/525444systematicreviewsguide.pdf`.

KIVIAT, B. The art of deciding with data: evidence from how employers translate credit reports into hiring decisions. **Socio-Economic Review**, 2017.

LEOPOLD, Henrik; MENDLING, Jan; GÜNTHER, Oliver. Learning from Quality Issues of BPMN Models from Industry. **IEEE Software**, v. 33, n. 4, p. 26–33, 2016.

MARZ, Nathan; WARREN, James. **Big Data: Principles and Best Practices of Scalable Real-Time Data Systems**. [S.l.]: Manning Publications, 2015.

MCNIFF, Jean. **Action Research: Principles and Practice**. 3. ed. [S.l.]: Routledge, 2013.

MENEZES, G. S. et al. Adoção e Avaliação de um Processo de Software Baseado no MPS.BR: um Estudo de Caso, 2017.

MENG, Xiangrui et al. MLlib: machine learning in apache spark. **J. Mach. Learn. Res.**, v. 17, n. 1, p. 1235–1241, jan. 2016.

MENG, Xiao-Li. Data Science: An Artificial Ecosystem. **Harvard Data Science Review**, v. 1, n. 1, jul. 2019.

MICROSOFT. **What is the Team Data Science Process? - Azure Architecture Center**. [S.l.*:* s.n.], 2024. `https://learn.microsoft.com/en-us/azure/architecture/data-science-process/overview`.

NABATI, Elaheh; THOBEN, Klaus-Dieter. On Applicability of Big Data Analytics in the Closed-Loop Product Lifecycle: Integration of CRISP-DM Standard, p. 457–467, 2016.

NAGASHIMA, Hiroko; KATO, Yuka. APREP-DM: a Framework for Automating the Pre-Processing of a Sensor Data Analysis based on CRISP-DM, p. 555–560, 2019.

NORTH, Dan. **Introduction to Behavior Driven Development**. [S.l.*:* s.n.], 2006. Dan North and Associates.

PRESSMAN, Roger S. **Software Engineering: A Practitioner's Approach**. 7th. [S.l.]: McGraw-Hill, 2010.

PROVOST, Foster; FAWCETT, Tom. Data Science and its Relationship to Big Data and Data-Driven Decision Making. **Big Data**, v. 1, n. 1, p. 51–59, 2013.

R. BASILI, G. Caldiera; ROMBACH, H. D. The Goal Question Metric Approach. **Encyclopedia of Software Engineering**, v. 1, p. 528–532, 1994.

SAGIROGLU, Seref; SINANC, Duygu. Big Data: A Review, p. 42–47, 2013.

SALTZ, J. S.; KRASTEVA, I. Current approaches for executing big data science projects - a systematic literature review. **PeerJ Computer Science**, v. 8, 2022.

SALTZ, J. S.; KRASTEVA, I. Current approaches for executing big data science projects - a systematic literature review. **PeerJ Computer Science**, v. 8, 2022.

SALTZ, J. S.; SHAMSHURIN, I. Big data team process methodologies: A literature review and the identification of key factors for a project's success, p. 2872–2879, 2016.

SALTZ, Jeffrey; SUTHERLAND, Alex; HOTZ, Nicholas. Achieving Lean Data Science Agility Via Data Driven Scrum. In*:*

SCHRÖER, Christoph; KRUSE, Felix; GÓMEZ, Jorge Marx. A Systematic Literature Review on Applying CRISP-DM Process Model. **Procedia Computer Science**, v. 181, p. 526–534, 2021.

SCHWABER, Ken; BEEDLE, Mike. **Agile Software Development with Scrum**. [S.l.]: Prentice Hall PTR, 2001.

SHARMA, Sandeep; KUMAR, Dinesh; FAYAD, Mohamed E. An Impact Assessment of Agile Ceremonies on Sprint Velocity Under Agile Software Development, p. 1–5, 2021.

SHVACHKO, Konstantin; KUANG, Hairong; RADIA, Sanjay; CHANSLER, Robert. The Hadoop Distributed File System, p. 1–10, 2010.

SILVA, S.; NICOLAU, B. A Case Study on Data Science Processes in an Academia-Industry Collaboration, 2023.

SOFTEX. **MPS.BR-Melhoria de Processo do Software Brasileiro - Guia Geral MPS de Software**. [S.l.*:* s.n.], 2012. SOFTEX.

SOMMERVILLE, Ian. **Software Engineering**. 10th. [S.l.]: Pearson Education, 2016.

SPARX SYSTEMS. **Software & Systems Process Engineering Meta-Model (SPEM)**. Versão 16.0. [S.l.], out. 2022. Enterprise Architect User Guide Series. Disponível em: `https://sparxsystems.com/resources/user-guides/16.0/model-domains/languages/spem.pdf`. Acesso em: 1 nov. 2025.

STERLING, Thomas L.; SAVARESE, Daniel; BECKER, Donald J.; DORBAND, John E.; RANAWAKE, Udaya A.; PACKER, Charles V. BEOWULF: A Parallel Workstation for Scientific Computation, 1995.

STONEBRAKER, M.; ABADI, D. J.; DEWITT, D. J.; MADDEN, S.; PAULSON, E.; PAVLO, A.; RASIN, A. MapReduce and parallel DBMSs: Friends or foes? **Communications of the ACM**, v. 53, n. 1, p. 64–71, 2010.

TALEB, Ibrar; SERHANI, Mohamed A.; DSSOULI, Rachida. Big Data Quality: A Survey. **IEEE Transactions on Services Computing**, v. 11, n. 1, p. 98–115, 2018.

TANG, Shuhao; HE, Bingsheng; YU, Haibo. A Survey on Spark Ecosystem: Big Data Processing Infrastructure. **IEEE Transactions on Knowledge and Data Engineering**, v. 31, n. 2, p. 293–311, 2019.

VERSIONONE. **16th Annual State of Agile Report**. [S.l.*:* s.n.], 2022. `https://stateofagile.com`. Acessado em: 31 de julho de 2025.

WANG, K.; LI, X.; ZHANG, L. From Data to Value: A Systematic Framework for Data-Driven Decision Making, 2019.

WIEGERS, Karl; BEATTY, Joy. **Software Requirements**. 3rd. [S.l.]: Microsoft Press, 2013.

WIRTH, Rüdiger; HIPP, Jochen. CRISP-DM: Towards a Standard Process Model for Data Mining, p. 29–39, 2000.

ZAHARIA, Matei et al. Spark: cluster computing with working sets, p. 10, 2010.

ZHU, Yangyong; XIONG, Yun. Defining Data Science. **arXiv:1501.05039 [cs.DB]**, 2015.

APÊNDICE A – ESTRUTURA DAS ENTREVISTAS

Processo Atual: Os entrevistados foram convidados a descrever o processo atual de implantação de produtos de dados em lote em suas equipes e a identificar os pontos críticos envolvidos.

Desafios e Problemas: Foram coletadas informações sobre os principais desafios enfrentados durante a implantação mais recente, incluindo como questões relacionadas à comunicação entre diferentes equipes foram tratadas.

Papéis e Responsabilidades: Foram exploradas as responsabilidades específicas de cada participante no processo de implantação, juntamente com suas percepções sobre se essas responsabilidades estavam adequadamente distribuídas entre as diferentes áreas.

Ferramentas e Tecnologias: As ferramentas e tecnologias utilizadas durante a implantação de produtos de dados foram investigadas, bem como sua eficácia percebida de acordo com os entrevistados.

Documentação: A qualidade da documentação fornecida para os modelos a serem implantados foi avaliada, com atenção a potenciais problemas como informações ausentes ou erros na documentação.

Métodos de Trabalho: Foi analisada a implementação do Scrum dentro das equipes durante a fase de implantação, incluindo se ocorreram cerimônias ágeis conjuntas com outras equipes e a percepção de utilidade dessas cerimônias.

Sugestões de Melhoria: Foram solicitadas sugestões para aprimorar o processo de implantação de produtos de dados, incluindo ajustes específicos que poderiam beneficiar potencialmente as equipes envolvidas.

Voltar ao texto principal na Seção 4.1.6.

APÊNDICE B – QUESTIONÁRIO DE AVALIAÇÃO DA ABORDAGEM DE IMPLANTAÇÃO

Este apêndice apresenta o questionário aplicado aos colaboradores das equipes de Produto, Dados & Analytics, Tecnologia e Delivery, utilizado para avaliar a abordagem proposta para implantação de produtos de dados em lote.

1. **Departamento**
    - Produtos
    - Dados & Analytics
    - Tecnologia
    - Delivery
2. **Tempo de trabalho na organização**
    - Menos de 1 ano
    - 1–3 anos
    - 3–5 anos
    - Mais de 5 anos
3. **Avalie a clareza das etapas do processo de implantação**
    - Muito clara
    - Clara
    - Neutra
    - Confusa
    - Muito confusa
4. **A abordagem facilita a colaboração entre as diferentes equipes?**
    - Totalmente concordo
    - Concordo
    - Neutro
    - Discordo
    - Totalmente discordo
5. **As responsabilidades de cada equipe são bem definidas e compreendidas?**
    - Totalmente concordo
    - Concordo
    - Neutro
    - Discordo
    - Totalmente discordo

6. **Como você avalia a eficiência do processo de documentação?**
    - Muito eficiente
    - Eficiente
    - Neutro
    - Ineficiente
    - Muito ineficiente
7. **O processo de validação e teste é rigoroso e eficaz?**
    - Totalmente concordo
    - Concordo
    - Neutro
    - Discordo
    - Totalmente discordo
8. **A abordagem ajuda a identificar e mitigar riscos de forma eficaz?**
    - Totalmente concordo
    - Concordo
    - Neutro
    - Discordo
    - Totalmente discordo
9. **Você está satisfeito(a) com a abordagem de implantação no Thor utilizada?**
    - Muito satisfeito
    - Satisfeito
    - Neutro
    - Insatisfeito
    - Muito insatisfeito
10. **Você recomendaria essa abordagem para outros projetos na organização?**
    - Sim
    - Talvez
    - Não

Voltar ao texto principal na Seção 6.1.