Approach to Deploying Batch File Data Products in a Big Data Environment

Richard de Arruda Felix[a], Patricia Della Méa Plentz[b] and Jean Carlo Rossa Hauck[c]
Federal University of Santa Catarina, Florianópolis, Brazil

Keywords: Batch Processing, Big Data, Data Science, Agility.

Abstract: Data science has become essential across industries such as government, healthcare, and finance, driving decision-making through large-scale data analysis. Deploying batch data products, like the periodic calculation of credit scores for millions, presents significant challenges, including integration with existing big data architectures and ensuring scalability and efficiency. This study proposes an optimized approach that leverages software engineering and agile methodologies to streamline the deployment of such products. Validated through action research conducted at a Brazilian credit bureau, the approach demonstrated a substantial reduction in deployment time by improving documentation, development, and testing processes, offering a scalable solution to modern batch data processing challenges.

1 INTRODUCTION

Data science has become a cornerstone for various industries, enabling informed decision-making through the analysis of vast volumes of data. As data emerges as one of the most valuable organizational assets, the ability to extract meaningful insights is critical for maintaining competitive advantage (Marz and Warren, 2015). High-performance computing technologies facilitate the collection, storage, and analysis of raw data, transforming it into strategic information that drives decisions across marketing, innovation, and finance (Grolinger et al., 2013).

To ensure the effectiveness of data analysis and implementation, data science projects follow structured phases, including business understanding, data preparation, modeling, evaluation, and deployment (J. Gao and Selle, 2015). Frameworks such as CRISP-DM (Cross-Industry Standard Process for Data Mining) (Cunha et al., 2021) and LDTM (Lean Design Thinking Methodology) (Ahmed et al., 2018) have been developed to guide these stages. CRISP-DM, widely used in the field, emphasizes data preparation and analysis but provides limited guidance for the deployment phase (Chapman et al., 2000). Similarly, LDTM integrates concepts from CRISP-DM, Lean Startup, and Design Thinking to focus on early project stages (Bender-Salazar, 2023).

Deploying batch data products in big data environments introduces significant challenges. These products, often processing millions of data points (Gorton and Klein, 2015), require robust infrastructure to ensure scalability and performance. The complexity of integrating new solutions into existing architectures and coordinating across teams, systems, and technologies poses additional hurdles (Hashem et al., 2015). Effective deployment relies on overcoming these obstacles through improved processes and tools.

Software engineering, combined with agile methods, offers promising solutions to these challenges (Grady et al., 2017)(Amershi et al., 2019). Integrating these approaches can enhance team collaboration and streamline the deployment of complex data products.

This study addresses gaps in traditional data science frameworks, particularly regarding the deployment phase, by proposing a novel approach for deploying batch data products in big data environments. Using action research (Avison et al., 1999), this work employs iterative cycles of diagnosis, planning, action, evaluation, and learning (Baskerville, 1999), enabling continuous refinement based on team feedback and outcomes.

The contribution of this work is twofold. For researchers, it introduces a structured methodology that emphasizes deployment and iterative improvement in high-performance environments. For practitioners, it provides actionable insights for overcoming deployment challenges in real-world scenarios.

The paper is structured as follows: Section 2 discusses key concepts related to the research topic. Sec-

[a] https://orcid.org/0009-0009-9621-839X
[b] https://orcid.org/0000-0001-8029-9468
[c] https://orcid.org/0000-0001-6550-9092

Felix, R. A., Plentz, P. D. M. and Hauck, J. C. R.
Approach to Deploying Batch File Data Products in a Big Data Environment.
DOI: 10.5220/0013293100003944

In *Proceedings of the 10th International Conference on Internet of Things, Big Data and Security (IoTBDS 2025)*, pages 93-104
ISBN: 978-989-758-750-4; ISSN: 2184-4976

tion 3 reviews related work. Section 4 details the action research conducted. Section 5 presents the proposed approach for deploying batch data products. Section 6 evaluates the collected data, and Section 7 concludes the research.

2 BACKGROUND

2.1 Big Data and Software Engineering

Big data refers to the collection, storage, and analysis of large volumes of data from various sources, with the goal of extracting valuable insights for informed decision-making (Laigner et al., 2018). This concept has driven innovations across different sectors by integrating industry-specific characteristics to develop new products and services. Big data projects pose unique requirements compared to traditional projects, addressing challenges such as data distribution, intensive write workloads, variability in request loads, computationally intensive analyses, and the need for high availability (Gorton et al., 2016)(Hummel et al., 2018). These challenges have established big data as a significant subdiscipline within software engineering.

In response to the dynamic nature of big data projects, agile software development methods have gained prominence. Since the 2000s, these methods have provided a flexible and iterative framework that allows teams to quickly adapt to changing requirements (Hoda et al., 2018). Grounded in the four core values and twelve principles of the Agile Manifesto (Fowler and Highsmith, 2000), agile methodologies remain widely adopted due to their ability to enhance efficiency and foster team collaboration. Practical implementations, such as Scrum (Schwaber and Beedle, 2002) and eXtreme Programming (Beck and Andres, 2004), have been particularly successful, emphasizing individual competencies and structured ceremonies like iteration planning and daily meetings to maintain project agility (Sharma et al., 2021).

Data science, as an interdisciplinary field, has also benefited from agile practices. Combining statistical techniques, computational methods, and domain-specific expertise, data science aims to extract meaningful insights from large volumes of both structured and unstructured data (Irizarry, 2020). Its applications span diverse sectors, including government and healthcare, where it addresses complex problems and fosters innovation. The evolving definition of data science reflects its dynamic nature, integrating areas such as statistics, informatics, computer science, and specialized domain knowledge (Zhu and Xiong, 2015).

However, managing data science projects presents specific challenges, particularly in the deployment phase. Process models like CRISP-DM, which emphasize preparation, analysis, reflection, and dissemination stages, have been widely recognized in the field (Saltz and Shamshurin, 2016). Yet, their adoption has declined due to limitations in handling communication, knowledge sharing, and project management processes (Nagashima and Kato, 2019)(Schröer et al., 2021).

To address these limitations, new frameworks have emerged, incorporating agile principles to enhance data science project management. For instance, the Team Data Science Process (TDSP) model (Microsoft, 2024) integrates CRISP-DM with Scrum practices, modernizing the traditional approach by focusing on team collaboration and aligning better with current project needs (Saltz and Krasteva, 2022). Similarly, Data-Driven Scrum (Saltz et al., 2022) adapts Scrum for data science by emphasizing flexibility and experimentation, essential in projects characterized by frequent uncertainties and evolving requirements.

Additionally, hybrid frameworks that combine agile and traditional methodologies are increasingly popular. These approaches balance the detailed planning of traditional methods with the adaptability of agile practices, providing a structured yet flexible framework that accommodates emerging data and insights while maintaining organized project execution (Fleckenstein and Fellows, 2018).

2.2 High-Performance Computing

High-Performance Computing (HPC) is essential for processing and analyzing massive datasets, particularly in big data applications. It plays a critical role in fields such as climate modeling, bioinformatics, and earthquake simulation. The advent of GPUs and scalable cluster systems, such as Beowulf clusters (Sterling et al., 1995), has enabled HPC to meet the speed and scalability requirements of big data environments.

Several platforms harness HPC capabilities for batch processing in big data. Hadoop (Shvachko et al., 2010), a widely used distributed processing framework, includes components such as the Hadoop Distributed File System (HDFS), MapReduce for parallel task execution (Dean and Ghemawat, 2008), and YARN for resource management. While Hadoop excels in batch processing, newer frameworks like Apache Spark (Spark, 2024) and Apache Flink (Flink, 2024) extend capabilities to include real-time and

continuous stream processing.

Apache Spark addresses Hadoop's limitations by processing data in memory, accelerating iterative tasks such as machine learning (Zaharia et al., 2010). Its ecosystem supports batch and streaming analytics through components like Spark SQL for queries and MLlib for machine learning (Meng et al., 2016). Spark is particularly effective for exploratory data analysis and predictive modeling.

Apache Flink specializes in real-time stream processing, offering features like event time and windowing for time-based analysis. With low latency and high resilience, Flink is ideal for applications requiring immediate responses, such as fraud detection and network monitoring (Carbone et al., 2015).

Another notable platform is HPCC Systems (HPCC, 2024), which provides an integrated solution for data-intensive computing. Its architecture includes Thor for batch processing, Roxie for real-time data delivery, and ESP for integrating data services. Tasks and queries in HPCC are written using the Enterprise Control Language (ECL), facilitating efficient large-scale data analysis.

Table 1 highlights key characteristics of the discussed frameworks, comparing their types of processing, ideal use cases, and support for batch and real-time processing.

With the continuous evolution of processing needs in big data, HPC continues to adapt and expand. The combination of next-generation processors, GPUs, and high-speed networks ensures that HPC systems can handle increasingly larger volumes of data while maintaining the efficiency and speed required for real-time applications. Platforms such as Hadoop and HPCC Systems exemplify this capability, offering scalable and integrated solutions.

3 RELATED WORK

In (Saltz and Krasteva, 2022), a systematic review is conducted on the adoption of process frameworks in data science projects, highlighting a significant increase in research on the organization, management, and execution of these projects in recent years. The review identified 68 primary studies, categorized into six main themes related to data science project execution. CRISP-DM was the most commonly discussed workflow. However, the study found no standardized approaches specifically designed for the data science context, particularly in the deployment phase, indicating a gap in research on current practices. It is suggested that future research explore the combination of workflows with agile approaches to create a more comprehensive framework that covers different aspects of project execution. The novelty of this work lies in addressing this gap by proposing an approach that explicitly targets the deployment phase.

In (Dipti Kumar and Alencar, 2016), a study investigates how the application of software development principles at various stages of the project development cycle can contribute to the design of big data applications. The findings helped identify data project initiatives with significant potential for success. However, the researchers highlighted deficiencies in the development cycle of big data-related projects, underscoring the need for special attention. In comparison, this study complements these findings by presenting a structured approach that integrates development and deployment practices, enhancing coordination and reducing inefficiencies.

In (Chen et al., 2016), the authors conducted research aimed at understanding current architectural methodologies for big data, as well as the integration of architectural design with techniques to orchestrate technological tools in a unified and effective approach. The objective was to establish correlations between Agile Manifesto practices and an architecture-centered perspective. The study culminated in the proposal of a methodology named ABBA (Architecture-centric Agile Big data Analytics), which assigns software architecture a central role as an enabler of agility. While ABBA emphasizes architecture, the approach proposed in this study extends beyond architectural design by incorporating iterative feedback loops and collaborative practices specific to batch data product deployment.

The study in (Hummel et al., 2018) offers a detailed approach to twenty-six relevant challenges in developing big data systems. The authors carefully analyze and classify these challenges through a collaborative and systematic process, organizing them according to the various development phases. They highlight that critical issues influencing project success may not be fully addressed during the planning phase, making the development process highly exploratory. Similarly, this study acknowledges these exploratory aspects and mitigates them by introducing clear phases—such as initiation, elaboration, construction, and delivery—aimed at improving predictability and reducing ambiguity in deployment.

In (Saltz et al., 2022), a new team process framework, Data Driven Scrum (DDS), is proposed to enhance the execution of data science projects. A case study conducted at a consultancy in Mexico explored the team's understanding and adaptation to the agile concepts of Lean. After transitioning from a waterfall approach, the team adopted DDS, refining their

Table 1: Comparison of data processing frameworks.

Technology	Type of processing	Ideal use	Batch	Real-time
Hadoop	Batch	Large volume processing	Yes	No
Spark	Batch and Real-Time	Iterative analysis, machine learning	Yes	Yes
Flink	Batch and Real-Time	Stream processing and real-time analytics	Yes	Yes
HPCC Thor	Batch	Intensive ETL and batch processing on big data	Yes	No
HPCC Roxie	Real-Time	Fast query and real-time data delivery	No	Yes

process for agile and lean development. The case study concluded that the organization understood and adapted to Lean agile concepts, validating the research questions. However, the main limitation was the application of DDS in only one organization. While DDS emphasizes team agility, this study focuses on improving both team collaboration and technical aspects, such as documentation and testing, to ensure scalability and efficiency in deploying batch data products.

The studies discussed in this chapter present various approaches, practices, and design patterns for big data and data science platforms, projects, and applications. However, none address the specific challenges of the production deployment phase in data science projects with the level of detail provided in this work. By focusing on deployment, this study fills a critical gap, offering a practical and replicable approach that integrates agile methods with software engineering principles tailored to big data environments.

4 ACTION RESEARCH

This chapter describes the methodological approach of this study. First, action research is introduced, highlighting its relevance and applicability within the context of this work. Then, the context of the research is presented, detailing the organization and the participants involved. Then, the data collection methods used are explained, with a detailed description of the data collection for the diagnostic phase that takes place before the execution of the action research. Next, the diagnostic based on the interviews is discussed, identifying the main problems and challenges that must be addressed in the approach to deploying batch data products within the company. Finally, the planning for data collection after the execution of the action research is presented, in order to identify that the proposed approach is effective.

Action research is a methodological approach that combines research and practical action, aiming to solve real-world problems and contribute to scientific knowledge. As previously mentioned, it is characterized by iterative cycles of diagnosis, planning, action-taking, evaluation, and learning. This approach is particularly suitable for contexts in which the researcher actively participates in the change process, works collaboratively, and intervenes consciously. In the context of this study, one of the authors is an employee of the company and played a dual role as both a facilitator of the interventions and a participant observer.

Although conducted in a credit bureau, the principles of action research, emphasizing collaboration, adaptability, and iterative improvement, are applicable to other industries like healthcare, manufacturing, and government. For instance, it has been used in healthcare to enhance patient care workflows through similar iterative cycles (Avison et al., 1999), demonstrating its potential to drive organizational change across diverse contexts.

4.1 Research Context

The research is being conducted at a Brazilian company that operates as a credit bureau. As a datatech, the company integrates various data sources and utilizes advanced technologies to provide analytical intelligence solutions, acting as an important intermediary between consumers, businesses, and financial institutions. The company undertakes data science projects to create comprehensive reports reflecting consumers' financial health and payment capacity based on their behavior (financial spending, geographic information, registration details, purchases, legal actions, online presence, among others). In addition to reports, the company provides credit scoring products, which are calculated using statistical models to indicate the probability of an individual or company meeting their financial obligations. For massive data processing, the open-source data lake platform HPCC Systems is used. At the beginning of the action research, the organization had approximately 60

employees. The team involved in the batch data product development flow consisted of 8 people.

In the Product team, there is 1 employee who serves as a product analyst with extensive experience in the area of credit score products, having worked at the organization for one year. In the Data & Analytics team, there are 3 employees with backgrounds in statistics: a junior analyst with less than a year in the organization, a senior analyst with two years of experience, and an analytics coordinator, also with two years in the organization. In the Technology team, the team consists of 2 employees with technology backgrounds, an intern and a specialist, both with less than a year in the organization. In the delivery area, there are 2 employees, also with technology backgrounds and less than a year in the organization. A legal representative of the organization signed the informed consent form to clarify the research procedures.

4.2 Data Collection for Diagnosis of the Current Process

The primary initial data collection technique used in this research was conducting interviews. These interviews combined elements of semi-structured and convergent interviews (Kallio et al., 2016), allowing for predefined themes and questions to guide the conversation while leaving room for open discussions. This approach was chosen because it offers flexibility, adapting to participants' responses and enabling the exploration of new directions during the interviews.

The interview results were examined through thematic analysis (Cruzes and Dyba, 2011), a popular approach in qualitative data analysis. The purpose of thematic analysis was to identify recurring themes or patterns in the interviews to diagnose the company's current process. This diagnosis was used to define the new approach that will be presented in this work.

4.3 Diagnosis

This stage involves understanding and defining the problem. An initial analysis of the organization's context and the interviews conducted was performed. Based on this analysis, we formalized the problem definition.

In the life cycle of a data science project, product development encompasses several phases before its deployment on the big data platform. The product development workflow involves interdisciplinary teams, beginning in the Product team. Next, the Data & Analytics team performs statistical modeling, defines business rules, and develops scripts using programming languages such as Python or R Language. For these activities, the organization already has a well-defined and mature process, requiring only minor adjustments, which will be detailed in the chapter on the proposed approach.

After completing the phases described above, it is necessary to transform the data product into a marketable and scalable product. For this, the high-performance computing platform for big data, HPCC Systems, is essential for batch generation. This requires migrating the architecture used in the development phases to the scalable architecture of the big data platform. The development teams responsible for deploying the products on the organization's big data platform use the agile Scrum method as the basis for their software process. However, they have the flexibility to adapt the method according to their specific needs.

For each deployment, the Data & Analytics team provides detailed documentation of the statistical models used. However, these documents lack a defined standard, resulting in issues such as missing relevant information or logical errors. It is also important to note that the lack of centralized documentation is one of the main challenges, causing version loss and uninformed changes, leading to communication failures among different teams. In addition to documentation issues, the teams from different areas do not hold joint agile ceremonies and do not clearly define tasks and deliverables for each iteration in the deployment phase. This results in excessive time spent on deployments. Finally, another significant problem is product testing between the development architecture and the big data platform.

In this context, the importance of adopting a specific approach for the batch data product deployment phase within the organization was recognized. This need was driven by the urgency to deliver high-quality solutions in the big data environment at an accelerated pace, allowing the organization to remain competitive in the highly competitive Brazilian credit bureau market.

4.4 Planning for Data Collection after Action Research

The data collection phase following the execution of action research ensures that the information obtained is relevant and accurate for evaluating the interventions performed. We used the structured GQM (Goal Question Metric) approach to define and evaluate metrics based on specific objectives (R. Basili and Rombach, 1994).

The objectives and questions to evaluate the effectiveness of the proposed approach are described be-

low:

- O1: Assess the effort required to deploy a batch data product on the big data platform before and after using the approach within the target organization.

- Q1.1: What is the time in days for deploying batch data products to production?

- O2: Assess the effectiveness and acceptance of the batch data product deployment approach among employees within the target organization.

- Q2.1: How clear are the deployment process steps when using the approach?
- Q2.2: What is the level of employee satisfaction with the approach?

- O3: Assess the effectiveness of the batch data product deployment approach in facilitating team collaboration and clearly defining each team's responsibilities.

- Q3.1: Does the approach facilitate collaboration between different teams?
- Q3.2: Are the responsibilities of each team well-defined and understood?

- O4: Assess the overall effectiveness of the batch data product deployment approach, focusing on documentation efficiency, and the rigor and effectiveness of the validation and testing process.

- Q4.1: How is the efficiency of the documentation process evaluated?
- Q4.2: Is the validation and testing process rigorous and effective?

For each question, metrics were also defined following the GQM approach. For O1, the effort is recorded through the project tracking tool adopted by the organization. For data collection of metrics in O2, O3, and O4, a questionnaire containing all derived questions was developed to be administered at the end of the intervention.

5 PROPOSED APPROACH

To develop the approach, a detailed assessment was conducted through interviews with employees involved in the deployment process. These semi-structured interviews provided valuable insights into current practices, challenges faced, and the specific needs of the teams involved. Based on the initial diagnosis obtained from these interviews, we conducted an in-depth literature review to identify best practices and relevant approaches for executing big data and data science projects.

5.1 Teams and Responsibilities

Teams from the Product, Data & Analytics, Technology, and Delivery areas participate in the project life cycle for deploying a batch data product within the organization. Below, we describe the responsibilities of each team in the life cycle of the approach:

5.1.1 Product Team

Responsible for understanding the client and their needs, quantitatively evaluating the potential outcomes of these needs, prioritizing each proposed action, carrying out development, and closely monitoring the achievement of expected results.

5.1.2 Data & Analytics Team

Responsible for building statistical models (such as Credit Score, Behavior Score, and Churn), monitoring the effectiveness of these models, conducting analyses to improve existing procedures, and compiling detailed documentation with information about the models. This team also assists the technology team with troubleshooting and testing as needed.

5.1.3 Technology Team

Responsible for implementing the batch data product within the high-performance and big data architecture. The team thoroughly reviews the documentation provided by the Data & Analytics team and proceeds with coding in the big data environment. If the documentation is insufficient, the team notifies all involved areas so that adjustments can be made.

5.1.4 Delivery Team

Monitors the performance of the batch data product after deployment in production. This team oversees recurring deliveries, reports incidents, and requests improvements or fixes to the deployed product.

5.2 Batch Product Deployment Project Life Cycle Phases

The project life cycle for the approach to deploying a batch data product is divided into four main phases: 1) initiation, 2) elaboration, 3) construction, and 4) delivery (Figure 1). These phases should not be confused with the traditional development phases of a data science project. In addition to the preparatory and predictive model development stages, the proposed approach introduces these four phases specifically within the product deployment phase on the high-performance and big data platform.

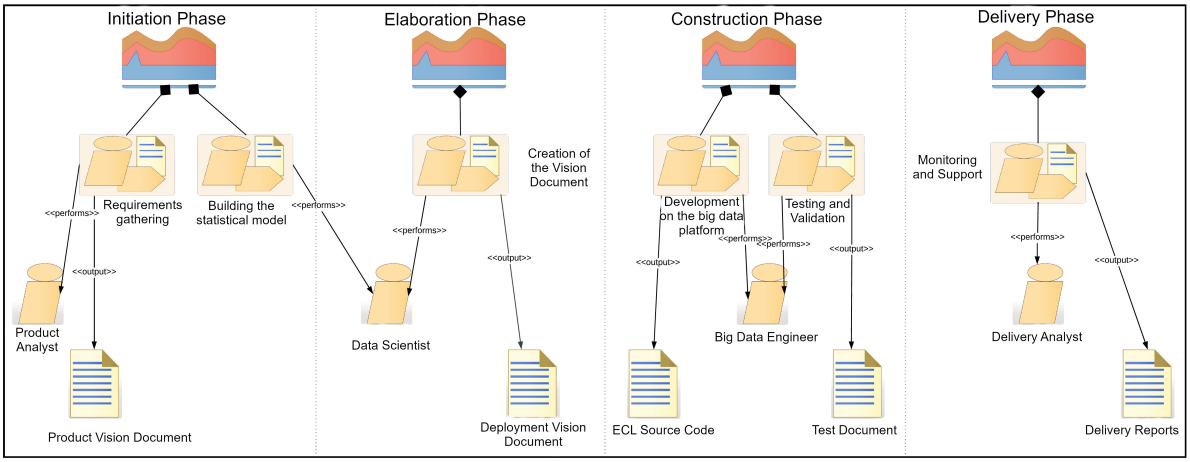

Figure 1: Project life cycle of the proposed approach.

Within each phase, weekly interactions are conducted among project stakeholders to monitor and evaluate progress, with each interaction culminating in the development of an incremental product, whether it be software or a project document. Upon the successful completion of all deliverables defined within a given phase, a brief meeting should be held to formally close the phase and transition to the next. Below, we will detail each phase of the project life cycle.

5.2.1 Initiation Phase

The main objective is to establish the product's scope, fostering an understanding of the needs of the client and stakeholders. A high-level understanding of project requirements is essential to mitigate potential risks and develop a robust business case. An initial vision document should be drafted, incorporating these requirements. Table 2 outlines the activities to be carried out during the Initiation phase and the respective responsibilities. At the end of this phase, a centralized vision document will be delivered.

5.2.2 Elaboration Phase

In this preparatory phase, critical risks are mitigated to allow for updates to cost estimates and schedules and to ensure stakeholder approval. The focus is on reducing key technical risks, ensuring that all necessary information is included in the vision document, and verifying platform compatibility. If needed, the technology team communicates with other teams to add information to the document. Tests to be conducted and acceptance criteria, which must be met during the construction phase, are also defined. Table 3 presents the activities of this phase and their responsibilities.

Table 2: Activities and responsible team of the Initiation phase.

Activity	Team
Alignment of product concept and response variables.	Product team
Alignment of development samples	Product team
Alignment of sample submission schedule and communication format for submission (e.g. Cloud, SFTP, Connect Direct).	Technology team
Alignment of date and consumption expectation and format.	Product team

5.2.3 Construction Phase

During this phase, technical development on the big data platform takes place to create the initial operational version of the product. Several internal versions ensure usability and alignment with client requirements. A functional beta version should be available for rigorous testing. All necessary validations and tests are conducted, focusing on compliance with acceptance criteria. The delivery team thoroughly validates the product, ensuring that acceptance criteria are met. Table 4 provides a description of the activities in this phase and their responsibilities.

5.2.4 Delivery Phase

The delivery phase begins once the product is aligned with the requirements defined in the elaboration phase. Preparations for the product's deployment into production are made, and minor adjustments may be implemented based on client feedback. Feedback at this stage focuses on fine-tuning, configuration, in-

Table 3: Activities and responsible team of the Elaboration phase.

Activity	Team	Artifact
Enrichment of the sample sent by the client with variables and indicators in the crop determined by the client.	Analytics Team	Samples enriched with the variables or samples for the client
Feedback from the client with formulas and databases used for the development of the data product.	Product Owner	Not Applicable
Completion of the vision document with the formulas, attributes used, and other important information regarding the analytical modeling.	Analytics Team	Partial deployment vision document
Completion of the vision document with other information necessary for the implementation of the data product	Projects and Implementation Team	Vision Document: This document provides a comprehensive overview of the intended product development.
Elaboration of the implementation schedule according to the prioritization	Projects and Implementation Team	Final deployment vision document: This document serves as a consolidated resource for developers, ensuring a smooth deployment process. It includes details on statistical models and is completed using macros to ensure that all essential information is included.

Table 4: Activities and responsible team of the Construction phase.

Activity	Team	Artifact
Use the code conversion tool if necessary.	Analytics Team	Code in the programming language of the Big Data platform.
Big Data and delivery reporting platform configurations.	Technology Team	Production-ready Big Data and communication platform.
Requesting product samples for approval.	Projects and Deployment Team and Analytics Team	Kit with databases containing the expected value.
Comparison of the product generated in the analytical environment with the product generated on the Big Data platform and preparation of an evidence document of the tests carried out.	Projects and Implementation Team	Document evidence: This document records the tests performed by the team responsible for deployment on the big data platform, following the requirements outlined in the deployment vision document. It includes screenshots of tests and detailed explanations of procedures. Involved areas are notified to formalize acceptance or rejection of the test cases.
Validation of test results performed by the Deployment Team.	Analytics Team	Acceptance or rejection of the document evidence of the comparison tests.

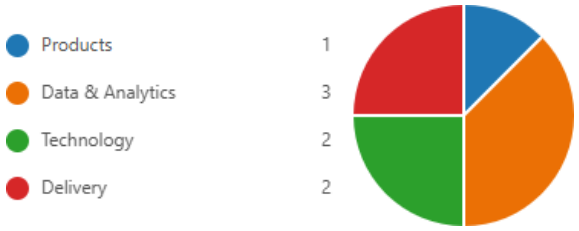

Figure 2: Teams that responded to the survey.

stallation, and usability.

6 EVALUATION

In this stage, the effects of the action are captured and analyzed. First, data was collected over six months of using the approach. Subsequently, each objective was analyzed based on the collected data.

6.0.1 Data Collection

Data collection was conducted throughout the duration of this study. For quantitative data, the activity management tool already used by the teams was utilized, along with a questionnaire administered at the end of the action research.

The questionnaire was sent at the conclusion of the study to all team members who participated in the deployment process using the proposed approach. Eight responses were received: one member from the product team, three from the Data & Analytics team, two from the technology team, and two from the delivery team (Figure 2).

6.0.2 Data Analysis

Following the data collection process, an analysis was conducted to assess whether the predefined objectives were achieved. The analysis is presented for each objective, with responses systematically grouped based on the collected data.

- Q1.1 - What is the time in days for deploying batch data products into production?

The comparative analysis of deployment times for batch data products, with and without the proposed approach, demonstrates significant improvements across all stages of the process. Figure 3 illustrates these differences. The time required for client approval of the model remained constant at five days for both approaches, indicating that external and administrative factors beyond the scope of the approach influenced this phase.

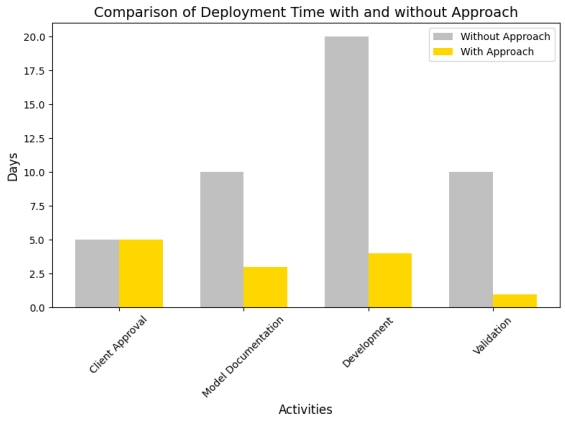

Figure 3: Comparison of deployment time with and without the deployment approach.

However, a notable reduction was observed in the time required for creating documentation, which decreased from 10 days to 3 days—a 70% improvement. This reduction can be attributed to the standardization of documentation processes and the introduction of more efficient templates. Similarly, the development phase on the big data platform experienced a significant time reduction, from 20 days to 4 days, representing an 80% improvement. This outcome reflects enhanced team coordination and the effective implementation of agile development practices. Lastly, the validation phase demonstrated the most substantial improvement, with the required time reduced from 10 days to 1 day (a 90% reduction). This improvement highlights the incorporation of efficient validation mechanisms and the establishment of a collaborative and integrated workflow between development and validation teams.

- Q2.1 - How clear are the deployment process steps when using the approach?

This data was obtained through a questionnaire. The majority of respondents rated the deployment process steps as clear or very clear, totaling seven positive responses out of eight. Specifically, 50% of respondents (4 out of 8) rated the steps as **very clear**, demonstrating excellent comprehension by half the participants. Additionally, 37.5% (3 out of 8) rated the steps as **clear**, indicating that nearly two-fifths of participants found the steps understandable. However, 12.5% (1 out of 8) rated the steps as **confusing**, highlighting a minor portion of participants who encountered challenges in understanding the process (Figure 4).

- Q2.2 - What is the level of satisfaction of team members with the approach?

The approach was well received by the majority of

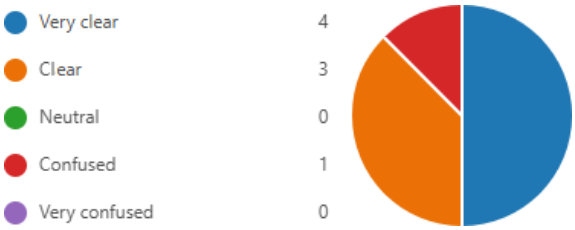

Figure 4: Clarity of the steps in the deployment process.

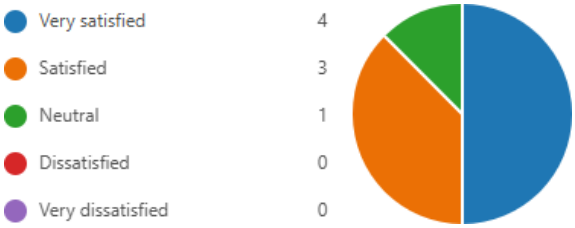

Figure 5: Level of satisfaction with the deployment approach.

respondents, with 50% (4 out of 8) rating their satisfaction level as **very satisfied** and 37.5% (3 out of 8) as **satisfied**. Together, these positive responses represent 87.5% of the feedback. Only 12.5% (1 out of 8) rated their satisfaction as **neutral**, indicating neither strong approval nor dissatisfaction (Figure 5).

- Q3.1 - Does the approach facilitate collaboration between different teams?

All respondents agreed that the approach facilitated collaboration between different teams. Specifically, 62.5% (5 out of 8) **strongly agreed**, reflecting a strong consensus regarding the approach's effectiveness in promoting teamwork. Additionally, 37.5% (3 out of 8) **agreed**, indicating a general positive sentiment toward the approach. No neutral or negative responses were recorded, underscoring the unanimous agreement on the collaborative benefits of the approach (Figure 6).

- Q3.2 - Are the responsibilities of each team well defined and understood?

Most respondents agreed that team responsibilities were clearly defined and well understood. Specifically, 50% (4 out of 8) **strongly agreed**, while 37.5% (3 out of 8) **agreed**. One respondent (12.5%)

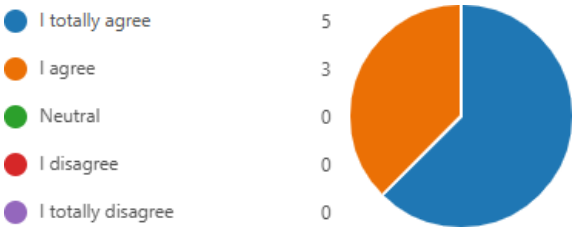

Figure 6: Perception of collaboration between teams.

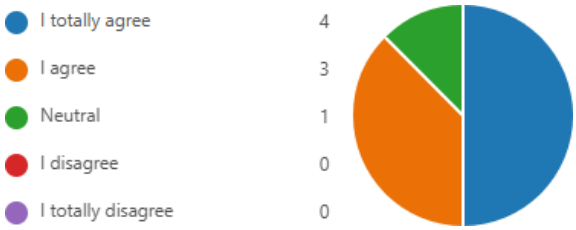

Figure 7: Perceptions of role clarity and responsibility definition.

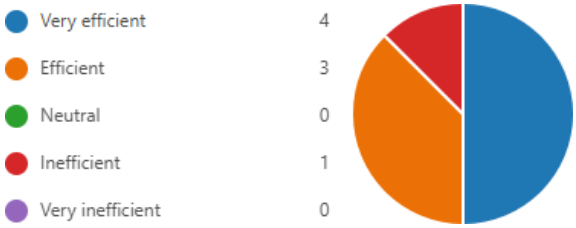

Figure 8: Evaluation of documentation efficiency.

expressed a **neutral** position, indicating room for further clarification or communication improvement (Figure 7).

- Q4.1 - How efficient is the documentation process?

The documentation process was predominantly rated as efficient or very efficient, with 87.5% of respondents providing positive feedback. Among these, 50% (4 out of 8) rated the process as **very efficient**, while 37.5% (3 out of 8) rated it as **efficient**. However, one respondent (12.5%) rated the process as **inefficient**, suggesting an area for potential improvement (Figure 8).

- Q4.2 - Is the validation and testing process rigorous and effective?

The validation and testing process was rated positively by the majority of respondents, with 62.5% (5 out of 8) **strongly agreeing** and 37.5% (3 out of 8) expressing a **neutral** position. While no negative feedback was recorded, the neutral responses highlight an opportunity to further strengthen confidence in this aspect of the approach (Figure 9).

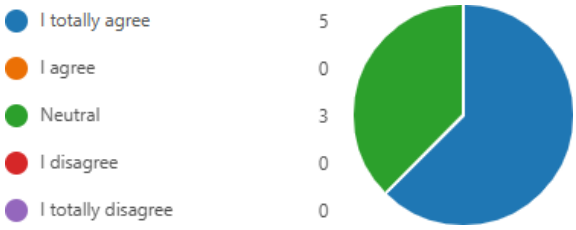

Figure 9: Evaluation of the validation and testing process.

6.0.3 Discussion

The implementation of the proposed approach demonstrated substantial improvements in the efficiency of deploying batch data products. Reductions in documentation, development, and validation times underscore its effectiveness in streamlining processes and enhancing coordination.

While most participants found the deployment process clear and the responsibilities well defined, isolated feedback suggests room for refinement in communication and role clarification. Similarly, the positive reception of the documentation and validation processes highlights their strengths, although further enhancements could address the few neutral or critical responses.

The unanimous agreement on the approach's collaborative benefits reflects its success in fostering teamwork and breaking down silos, a critical factor in complex big data projects. Overall, the findings validate the approach's utility and provide a foundation for further refinements to ensure broader applicability and satisfaction.

6.0.4 Threats to Validity

While this study offers valuable insights into deploying batch data products in big data environments, some limitations should be acknowledged. The results may lack generalizability as the research was conducted in a single credit bureau with specific infrastructure and practices. Organizational factors, such as team expertise and available resources, may also limit the replicability of the approach in other contexts. Potential biases in data collection and analysis, along with subjective evaluations, may affect internal validity. Lastly, the fast-paced evolution of big data technologies and industry practices necessitates ongoing updates to maintain the approach's relevance.

7 CONCLUSION

This study proposed and validated a novel approach for deploying batch data products in big data environments, achieving significant improvements in efficiency and team coordination within a credit bureau. Key results included notable reductions in documentation, development, and validation times, alongside high employee satisfaction and acceptance. The structured documentation guidelines and rigorous validation processes ensured quality and accuracy while identifying areas for further refinement. These findings demonstrate the approach's potential to enhance deployment processes, offering valuable insights for both researchers and practitioners in data science and big data projects.

REFERENCES

Ahmed, B., Dannhauser, T., and Philip, N. (2018). A lean design thinking methodology (ldtm) for machine learning and modern data projects. In *2018 10th Computer Science and Electronic Engineering (CEEC)*, pages 11–14.

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., and Zimmermann, T. (2019). Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 291–300.

Avison, D. E., Lau, F., Myers, M. D., and Nielsen, P. A. (1999). Action research. *Commun. ACM*, 42(1):94–97.

Baskerville, R. L. (1999). Investigating information systems with action research. *Communications of the Association for Information Systems*, 2.

Beck, K. and Andres, C. (2004). *Extreme programming explained : embrace change*. Addison-Wesley, Boston, MA.

Bender-Salazar, R. (2023). Design thinking as an effective method for problem-setting and needfinding for entrepreneurial teams addressing wicked problems. *Journal of Innovation and Entrepreneurship*, 12(1).

Carbone, P., Katsifodimos, A., Ewen, S., Markl, V., Haridi, S., and Tzoumas, K. (2015). Apache flink™: Stream and batch processing in a single engine. *IEEE Data Engineering Bulletin*, 38.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. Technical report.

Chen, H.-M., Kazman, R., and Haziyev, S. (2016). Agile big data analytics development: An architecture-centric approach. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 5378–5387.

Cruzes, D. S. and Dyba, T. (2011). Recommended steps for thematic synthesis in software engineering. In *2011 International Symposium on Empirical Software Engineering and Measurement*, pages 275–284.

Cunha, A. F., Ferreira, D., Neto, C., Abelha, A., and Machado, J. (2021). A crisp-dm approach for predicting liver failure cases: An indian case study. In Ahram, T. Z., Karwowski, W., and Kalra, J., editors, *Advances in Artificial Intelligence, Software and Systems Engineering*, pages 156–164. Springer International Publishing, Cham.

Dean, J. and Ghemawat, S. (2008). Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113.

Dipti Kumar, V. and Alencar, P. (2016). Software engineering for big data projects: Domains, methodologies and gaps. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 2886–2895.

Fleckenstein, M. and Fellows, L. (2018). Overview of data management frameworks. In *Modern Data Strategy*, pages 55–59. Springer International Publishing, Cham.

Flink (2024). Apache flink project. https://flink.apache.org/ [Accessed: 2024].

Fowler, M. and Highsmith, J. (2000). The agile manifesto. 9.

Gorton, I., Bener, A. B., and Mockus, A. (2016). Software engineering for big data systems. *IEEE Software*, 33(2):32–35.

Gorton, I. and Klein, J. (2015). Distribution, data, deployment: Software architecture convergence in big data systems. *IEEE Software*, 32(3):78–85.

Grady, N. W., Payne, J. A., and Parker, H. (2017). Agile big data analytics: Analyticsops for data science. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 2331–2339.

Grolinger, K., Higashino, W. A., Tiwari, A., and Capretz, M. A. (2013). Data management in cloud environments: Nosql and newsql data stores. *Journal of Cloud Computing: Advances, Systems and Applications*, 2(1).

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., and Ullah Khan, S. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, 47:98–115.

Hoda, R., Salleh, N., and Grundy, J. (2018). The rise and evolution of agile software development. *IEEE Software*, 35(5):58–63.

HPCC (2024). Hpcc systems. https://www.hpccsystems.com [Accessed: 2024].

Hummel, O., Eichelberger, H., Giloj, A., Werle, D., and Schmid, K. (2018). A collection of software engineering challenges for big data system development. In *2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pages 362–369.

Irizarry, R. A. (2020). The Role of Academia in Data Science Education. *Harvard Data Science Review*, 2(1). https://hdsr.mitpress.mit.edu/pub/gg6swfqh.

J. Gao, A. K. and Selle, S. (2015). Towards a process view on critical success factors in big data analytics projects. Core.ac.uk.

Kallio, H., Pietilä, A.-M., Johnson, M., and Kangasniemi, M. (2016). Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. *Journal of Advanced Nursing*, 72(12):2954–2965.

Laigner, R., Kalinowski, M., Lifschitz, S., Salvador Monteiro, R., and de Oliveira, D. (2018). A systematic mapping of software engineering approaches to develop big data systems. In *2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pages 446–453.

Marz, N. and Warren, J. (2015). *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications Co., USA, 1st edition.

Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D., Amde, M., Owen, S., Xin, D., Xin, R., Franklin, M. J., Zadeh, R., Zaharia, M., and Talwalkar, A. (2016). Mllib: machine learning in apache spark. *J. Mach. Learn. Res.*, 17(1):1235–1241.

Microsoft (2024). What is the team data science process? - azure architecture center. Available at: https://learn.microsoft.com/en-us/azure/architecture/data-science-process/overview.

Nagashima, H. and Kato, Y. (2019). Aprep-dm: a framework for automating the pre-processing of a sensor data analysis based on crisp-dm. In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 555–560.

R. Basili, G. C. and Rombach, H. D. (1994). The goal question metric approach. *Encyclopedia of Software Engineering*, 1:528–532.

Saltz, J., Sutherland, A., and Hotz, N. (2022). Achieving lean data science agility via data driven scrum.

Saltz, J. S. and Krasteva, I. (2022). Current approaches for executing big data science projects - a systematic literature review. *PeerJ Computer Science*, 8.

Saltz, J. S. and Shamshurin, I. (2016). Big data team process methodologies: A literature review and the identification of key factors for a project's success. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 2872–2879.

Schröer, C., Kruse, F., and Gómez, J. M. (2021). A systematic literature review on applying crisp-dm process model. *Procedia Computer Science*, 181:526–534. CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020.

Schwaber, K. and Beedle, M. (2002). *Agile Software Development with Scrum*. Prentice Hall, Upper Saddle River, New Jersey.

Sharma, S., Kumar, D., and Fayad, M. (2021). An impact assessment of agile ceremonies on sprint velocity under agile software development. In *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pages 1–5.

Shvachko, K., Kuang, H., Radia, S., and Chansler, R. (2010). The hadoop distributed file system. In *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, pages 1–10.

Spark (2024). Apache spark project. https://spark.apache.org/ [Accessed: 2024].

Sterling, T. L., Savarese, D., Becker, D. J., Dorband, J. E., Ranawake, U. A., and Packer, C. V. (1995). Beowulf: A parallel workstation for scientific computation. In *International Conference on Parallel Processing*.

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., and Stoica, I. (2010). Spark: cluster computing with working sets. In *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing*, HotCloud'10, page 10, USA. USENIX Association.

Zhu, Y. and Xiong, Y. (2015). Defining data science.