NOVA IMS Information Management School

MGI

Master Degree Program in

Information Management

AFTER THE SUCCESS OF DEVOPS INTRODUCE DATAOPS IN ENTERPRISE CULTURE

Nuno Filipe Paulo da Silva

Dissertation

presented as partial requirement for obtaining the Master Degree Program in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

AFTER THE SUCCESS OF DEVOPS INTRODUCE DATAOPS IN ENTERPRISE CULTURE

by

Nuno Filipe Paulo da Silva

Master Thesis presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Information system and technology management.

Supervised by

Prof. Vitor Duarte dos Santos, PhD

July, 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

[Lisbon, July 2023]

ABSTRACT

A lot of organizations implemented DevOps processes with success. This allowed different areas like development, operations, security and quality work together. This cooperation, and processes associated to the work with these areas are producing excellent results. The organizations are developing many applications that support operation and are producing a lot of data. This data has a significant value for organizations because must be used in analysis, reporting and more recently data science projects to support decisions.

It is time to take decisions supported in data and for this is necessary to transform organizations in a data-driven organizations and for this we need processes to deal with this data across all teams.

This dissertation follows a design science research approach to apply multiple analytical methods and perspectives to create an artifact. The type of evidence within this methodology is a systematic literature review, with the goal to attain insights into the current state-of-the-art research of DataOps implementation. Additionally, proven best practices from the industry are examined in depth to further strengthen the credibility. Thereby, the systematic literature review shall be used to pinpoint, analyze, and comprehend the obtainable empirical studies and research questions. This methodology supports the main goal of this dissertation, to develop and propose evidence-based practice guidelines for the DataOps implementation that can be followed by organizations.

KEYWORDS

DataOps; DevOps; Agile; Data Management; Data Quality;

Sustainable Development Goals (SGD):

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

API	Application Programming Interface
CI/CD	Continuous integration / continuous delivery
DataOps	Data Operations
DevOps	Development Operations
DSR	Design Science Research
ETL	Extract, Transform, Load
GDPR	General Data Protection Regulation
ITIL	Information Technology Infrastructure Library
KPI	Key Performance Indicators
IoT	Internet of things

1. INTRODUCTION

1.1. BACKGROUND AND PROBLEM IDENTIFICATION

In the last years many organizations have introduced DevOps in their culture with success, promoting collaboration between development and operation teams. This culture brought a better alignment of all teams allowing the development of better solutions with improved processes. These solutions are generating many data, that can be used to bring more value. The increasing adoption of DevOps, the growing availability of data concerning data development processes gives rise to the need for a systematic process for collecting, processing and using data into companies (Rodriguez et al., 2020) After these successful investments, enterprises are making significant investments in data science applications, but at the same time, need processes for collecting, processing, and using these data.

Organizations need more than the latest AI algorithms, hottest tools, and best people to turn data into insight-driven action and useful analytical data products. Processes and thinking employed to manage and use data in the 20th century are a bottleneck for working effectively with the variety of data and advanced analytical use cases that organizations have today (Atwal, 2019).

The companies with high maturity in DevOps are producing many solutions, features, and processes quickly to meet business needs. These solutions are generating a lot of data and business areas have more requirements for analytics teams to use these data. However, Data analytic teams are facing many challenges:

- Requirements change frequently;
- Data live in silos;
- Data formats are not optimized (the best data structure for development is different for analytics);
- Data errors;
- Bad data destroy the best reports;
- Data pipeline maintenance never ends;
- Manual process fatigue.

The gap between user needs and what Data analytics teams can provide can be a source of conflict and frustration. Data analytics teams need to address these points to provide value to business areas. Comparing with development teams who are working with agile methodologies with cycles of development – sprints, Data analytic teams need similar processes to work with development, business, and operations teams.

A DataOps strategy can help companies to address these questions. DataOps is a collaborative data management practice focused on improving the communication, integration, and automation of data flows between data managers and data consumers across an organization.

The goal of DataOps is to deliver value faster by creating predictable delivery and change management of data, data models and related artifacts. DataOps uses technology to automate the design, deployment and management of data delivery with appropriate levels of governance and it uses metadata to improve the usability and value of data in a dynamic environment (Definition of DataOps - Gartner Information Technology Glossary, n.d.).

In the process of establishing DataOps as a data analytics methodology, people and organizations supporting the concept derived 18 principles of DataOps in the manifesto. The DataOps principle summarizes the best practices, goals, philosophies, mission, and values for DataOps practitioners. The manifesto sets team communication over tools and processes. Experimentation, interaction and feedback are more important than designing and developing the whole pipeline upfront. Sense of responsibility and cross-functional collaboration is advocated to increase the project efficiency reducing individual soiled responsibilities and heroism *(Mainali, 2021).*

A successful strategy for DataOps must follow principles of DataOps Manifesto (*The DataOps Manifesto - Read The 18 DataOps Principles*, 2021):

- Continually satisfy the customer;
- Value working analytics;
- Embrace change;
- It is a team sport;
- Daily interactions;
- Self-organization;
- Reduce heroism reflect;
- Analytics in code;
- Orchestrate;
- Make it reproductible;
- Disposable environments;
- Simplicity;
- Analytics in manufacturing;
- Quality is a paramount;
- Monitor quality and performance;
- Re-use;
- Improve Cycle times;

With the increase of data available, the requirements of data science solutions and regulations to manage data are a challenge for all organizations. Applying DataOps principles is complex because there is no single formula to implement this.

DataOps promises a remedy by combining an integrated and process-oriented perspective on data with automation and methods from agile software engineering, like DevOps, to improve

quality, speed, and collaboration and promote a culture of continuous improvement *(Ereth, 2018).*

The management of data becomes a business enabler: while users can get the data, they need to unlock their innovative capacity. In the past, we have changed the culture of development and operations with DevOps. It is time to introduce similar concepts in teams who work data like Data Analytics or Data Science. Like Development and Operations, Data needs methodology and processes to scale and bring value and DataOps can bring this to organizations.

Most of the companies have legacy, namely established processes and tools. Due to this, these companies need to deal with change, choosing the time to make changes and manage those changes without disrupt the business. On the other side, there are start-ups where DataOps implementation could be more facilitated because it is possible to design processes from scratch and choose the best tools for the new processes.

Most of organizations are collecting many data but are not ready to explore those data at the same velocity. Organizations can increase productivity if they obtain more value from data quicker. To achieve this, Data Analytics teams must be organized to:

- Work with Agile methodologies which will allow planning each sprint, prioritizing the most important things and deliver value in each release. Each release can collect feedback and deliver improvements in future releases. This is very important nowadays where requirements change constantly.
- Like DevOps, DataOps needs processes and automation to break the barriers between IT, software development and quality teams. Data Analytics teams can use cloud services that allow manage environments, replicate environments for each stage of development and create similar production environments for development and quality. This will reduce errors, misunderstandings and increase confidence in value provided.
- Use a methodology like DataOps that promotes collaboration between teams like data scientists, data analytics, data engineers, IT, quality assurance and business teams. Data Analytics teams work closer to other teams and have many dependencies to provide value from data. For this and to keep the process sustainable, it needs processes to access and transform data, build models and provide reports/dashboards. With DataOps this pipeline is highly automated and is recommended to include automated tests, control metrics, to approve changes and control deployment of new releases.

A successful DataOps implementation can increase the alignment between Data Analytics teams and other teams adding value to organization and address the following points:

- More agility answering to requirement changes;
- Provide value on each interation and collect feedback soon to improve;
- Increase user satisfaction;
- Give more flexibility to the business;

- Increase data quality;
- Increase return of investment (ROI) in data teams;
- Data teams works on priorities for the organization;
- Improve data governance;

However, even knowing these expected advantages, it is still not clear which is the best approach to introduce DataOps in an organization. This issue leads to the formulation of following research questions:

RQ1: which will be the best approach to introduce DataOps in a specific organization?

RQ2: How to measure the success/outcome of the DataOps implementation?

1.2. OBJECTIVES

The research objective is to help answer the research questions by proposing a comprehensive DataOps framework to successful introduce DataOps in enterprise culture.

To reach this goal, the following intermediate objectives were defined:

- Make a comprehensive study of DataOps Methodology from literature;
- How to start a DataOps implementation in organizations;
- Give visibility to the challenges that exist in DataOps implementation;
- Provide guidelines for the companies to establish DataOps in organization culture;
- Measure the success of DataOps in organization culture.

1.3. EXPECTED CONTRIBUTIONS AND OUTCOMES

DataOps implementation in organizations comes from similar challenges leading with large amounts of data and questions about how to bring more value from this data.

Firstly, this study will present DataOps methodology and how this can help organizations.

Secondly, this study discusses the challenges that organizations have when deciding to implement a DataOps strategy. Having more knowledge about this will avoid mistakes during the process.

Thirdly, this study provides the best practices to introduce DataOps, clarifying how organizations can start this process.

This study provides some strategies to include DataOps in whole organization culture promoting collaboration between all areas to explore data and extract value in all stages of enterprise development.

This study will discuss some metrics that can be taken into consideration to monitor the development of DataOps strategy in the organizations.

The research outcome will contribute to enterprise accelerate DataOps adoption to increase sharply data-driven decision making and take benefits sooner choosing the best way to start implementation and avoid mistakes knowing the main challenges that will be faced.

2. LITERATURE REVIEW

The increasing adoption of DevOps, the growing availability of data concerning data development processes gives rise to the need for a systematic process for collecting, processing and using data into companies (Rodriguez et al., 2020). Data preparation has become a crucial step to be conducted before any data can be consumed and analyzed. Data scientists typically need to gather available data from various sources and then prepare it to meet the requirements of a specific use case before actual data analytics work can begin. Data preparation activities include a range of data curation tasks such as data exploration, data cleaning, data transformation, data integration, etc. (Yu et al., 2023)

The possibilities of utilizing a wide variety of data force organizations to react quickly. One way to tackle issues of organizational data utilization is implementing DataOps practices, which assist organizations in setting up the data management structure agilely. (Gür et al., 2022).

To acquire a generic understanding, this chapter is divided into six subchapters and covers the concept of DataOps. The following part intends to explain the definition of DataOps, the relationship with DevOps and other methodologies, the necessary roles for implementation and present some case studies of implementation in different industries. This provides the baseline for initial development of respective implementation guidelines.

2.1. PROBLEMS IN THE DATA & ANALYTICS INDUSTRY

The data and analytics industry, like any other sector, faces its own set of challenges and issues. Data is the new oil for organizations (“Clive Humby,” 2006) and the amount of data is increasing fast. Currently, Organizations are dealing with problems to manage their data and some examples of this are:

- 87% of data science projects never go to production (“Why Do 87% of Data Science Projects Never Make It into Production?,” 2019);
- data-driven had declined over the past three years, from 37% in 2017 to 31% in 2019 (*Creating a Data-Driven Culture*, 2020);
- 85% of all big data projects fail (Henrion & Gatos, n.d.)

There are several problems in the data analytics industry that contribute for these results:

- Data quality issues: One of the major problems in the data analytics industry is the quality of the data being used. Data quality problems, such as missing values, inconsistent formatting, and incorrect data can result in incorrect insights and decisions. The impact of data quality is directly seen in lower revenue and higher

operational costs, both resulting in financial loss ("The 7 Most Common Data Quality Issues," n.d.);

- Data silos: Data silos are data available only to certain departments of an organization. When organizations use multiple apps and systems to manage various workflows, it leads to the development of data silos that make it difficult for teams to find the information they need (*Eight Factors That Can Impact Data Quality*, 2023). Many organizations struggle with data silos, which are separate, disconnected data sources that can make it difficult to access and analyze data in a comprehensive and integrated way;
- Lack of data governance: Data governance refers to the exercise of authority and control over the management of data. The purpose of data governance is to increase the value of data and minimize data-related cost and risk. Despite data governance gaining in importance in recent years, a holistic view on data governance, which could guide both practitioners and researchers, is missing (Abraham et al., 2019). Many of organizations deal with the lack of formal data governance processes. Without clear data governance policies, organizations can struggle to manage the quality, security, and privacy of their data;
- Difficulty integrating and managing multiple data sources: Many organizations struggle with integrating and managing multiple data sources, such as databases, cloud services, and IoT devices. This can result in data duplication, errors, and inconsistencies, making it difficult to get a complete and accurate view of data;
- Skills gap: There is a growing skills gap in the data and analytics industry, with a shortage of professionals who have the skills and knowledge to manage, process, and analyze data effectively;
- Data security and privacy: As more data is collected and processed, there is a growing concern about data security and privacy. Organizations need to ensure that they have the right security and privacy controls in place to protect sensitive data, namely personal data and prevent breaches.

Addressing these challenges requires a combination of technical expertise, organizational support, and a strong data-driven culture. By addressing these challenges, organizations can improve the quality, reliability, and security of their data and analytics systems. Organizations need to invest in data management practices, prioritize data governance, and build multidisciplinary teams with diverse skills to manage the complexities of the data and analytics industry.

2.2. SCOPE AND OBJECTIVES OF DATAOPS

Rapidly accelerating technology advances, the recognized value of data, and increasing data literacy are changing what it means to be "data driven." *(The Data-Driven Enterprise of 2025 | McKinsey, n.d.).*

DataOps is a set of practices and processes aimed at improving the speed, quality, and reliability of data-driven decision-making. The goal of DataOps is to automate and streamline the entire data pipeline, from data collection and storage to analysis and visualization.

2.2.1. Scope of DataOps

The scope of DataOps encompasses the entire data lifecycle and includes various processes, technologies, and practices aimed at optimizing and streamlining data operations within an organization. The scope of DataOps typically includes the following areas:

- Data Collection: This involves acquiring data from various sources, including databases, APIs, and log files, and storing it in a central repository;
- Data Integration: This involves integrating different data sources into a single, unified view, so that data analysts can work with a complete and accurate picture of the data;
- Data Processing: This involves transforming and cleaning the data to make it usable for analysis. This includes tasks such as data normalization, data enrichment, and data reconciliation;
- Data Analytics: This involves using data to gain insights and make decisions. This includes tasks such as data visualization, data mining, and machine learning;
- Data Delivery: This involves delivering the results of the data analysis to stakeholders in a way that is easily consumable and actionable.
- Data Governance and Compliance: DataOps incorporates data governance practices to ensure data integrity, security, privacy, and compliance with regulatory requirements.

2.2.2. Objectives of DataOps

The overall objective of DataOps is to create a data-driven culture and optimize the end-to-end data lifecycle, enabling organizations to derive maximum value from their data assets. The main goals of DataOps include:

- Accelerating Time-to-Value: DataOps aims to shorten the time it takes to deliver valuable insights and outcomes from data. By streamlining and automating data processes, organizations can reduce the time required for data preparation, analysis, and deployment. This enables faster decision-making and enhances agility in responding to business needs.

- Improve Data Quality: The goal of DataOps is to ensure that data is accurate, consistent, and up to date. This is achieved by automating data validation and reconciliation processes and implementing strict data governance policies.
- Increase Data Availability: DataOps aims to make data available to stakeholders as quickly as possible, so that they can make timely decisions based on the latest data;
- Enhance Data Collaboration: DataOps encourages collaboration between different teams and departments, so that they can work together to achieve a common goal. This is achieved using centralized data repositories, data pipelines, and shared data visualizations;
- Reduce Data Costs: DataOps aims to reduce the costs associated with managing and processing data, by automating manual tasks and reducing the time required to prepare data for analysis;
- Improve Data Security: DataOps is concerned with protecting sensitive data and ensuring that it is not misused or leaked. This is achieved using encryption, access controls, and data masking techniques.

2.3. Relationship between DataOps, DevOps and other methodologies

DataOps is a set of practices and principles aimed at improving the speed, quality, and reliability of data processing and management. DevOps became a major technology, approach and culture in applications and services development for modern agile datadriven companies, both information technology and business oriented. It is essential that DevOps implies the adoption of the DevOps culture by organisations that aligns people, processes, and tools toward a more unified customer focus (Demchenko et al., 2019). DevOps is a set of practices aimed at improving collaboration and communication between development and operations teams, with the goal of delivering software quickly and reliably.

The relationship between DataOps and DevOps can be seen as complementary, as both methodologies share similar goals and principles, such as:

- Collaboration and communication: Both DataOps and DevOps emphasize the importance of collaboration and communication between different teams within an organization.
- Continuous improvement: Both methodologies focus on continuous improvement and adaptation to changing business requirements.
- Automation: Both DataOps and DevOps rely heavily on automation to improve efficiency, reduce errors, and speed up processes.
- Quality assurance: Both methodologies place a strong emphasis on quality assurance and testing, to ensure that the final product meets business requirements and is reliable.

In addition to these similarities, DataOps and DevOps can also complement other methodologies such as Agile, Lean, and ITIL. For example, Agile provides a flexible and adaptive approach to software development, while DevOps focuses on delivering software quickly and reliably. DataOps can be used to manage the data aspect of software development, ensuring that data is processed and managed efficiently and effectively.

Overall, the relationship between DataOps, DevOps, and other methodologies is one of complementarity, where each methodology adds value to the others, and helps organizations to achieve their goals more effectively and efficiently.

2.4. THE ROLE OF DATA GOVERNANCE, DATA QUALITY AND DATA MANAGEMENT IN DATAOPS

Data Quality, Data Governance, and Data Management play crucial roles in DataOps and are interdependent and essential components of the overall DataOps process.

Data governance is a collection of processes, roles, policies, standards, and metrics that ensure the effective and efficient use of information in enabling an organization to achieve its goals. It establishes the processes and responsibilities that ensure the quality and security of the data used across a business or organization. Data governance defines who can take what action, upon what data, in what situations, using what methods (*What Is Data Governance (and Do I Need It)?*, n.d.). An effective data governance strategy provides many benefits to an organization like a common understanding of data, improved data quality and consistent compliance for meeting the demands of government regulations, such as the EU General Data Protection Regulation (GDPR). Data governance can guide all other data management activities managed properly at all levels according to policies and best practices (Phuong, 2021).

In DataOps, data governance is used to ensure that the data is used in a responsible and ethical way, and that it is protected from unauthorized access, misuse, and leakage. Data Governance is also used to ensure that the data is consistent, accurate, and up-to-date, and that data quality is maintained throughout the data pipeline.

The quality of the data is of utmost importance in DataOps as it directly affects the accuracy and reliability of the data-driven insights and decisions. Data Quality is concerned with ensuring that the data is accurate, consistent, and up-to-date, and it is achieved by implementing data validation and reconciliation processes, as well as data standardization and normalization. In DataOps, data quality is monitored in real-time, and any issues are addressed immediately to prevent them from affecting the quality of the data-driven insights and decisions.

The scope of data management is broader than data governance. It can be defined as the practice of ingesting, processing, securing and storing an organization’s data, where it is then

utilized for strategic decision-making to improve business outcomes. While this is inclusive of data governance, it also includes other areas of the data management lifecycle, such as data processing, data storage and data security. Since these other areas of data management can also impact data governance, these teams need to work together to execute against a data governance strategy (*What Is Data Governance?*, n.d.).

Data Management is concerned with the processes and technologies used to manage the lifecycle of data, from collection and storage to analysis and visualization. In DataOps, data management is used to automate and streamline the data pipeline, so that data is available to stakeholders as quickly as possible. Data Management is also used to ensure that the data is properly stored, secured, and backed up, and that data quality is maintained throughout the data pipeline.

In conclusion, Data Quality, Data Governance, and Data Management are all critical components of DataOps, and they work together to ensure that the data is accurate, consistent, secure, and available to stakeholders in a timely way so that they can make informed decisions based on the latest data.

2.5. THE BENEFITS AND CHALLENGES OF IMPLEMENTING DATAOPS IN ORGANIZATIONS

By 2025, smart workflows and seamless interactions among humans and machines will likely be as standard as the corporate balance sheet, and most employees will use data to optimize nearly every aspect of their work (The Data-Driven Enterprise of 2025 | McKinsey, n.d.).

Adopting DataOps requires a combination of technical investment, organizational restructuring and change management. It requires people to fundamentally change the way they work, and this change isn't going to happen overnight ("Key DataOps Challenges Enterprises Must Overcome," 2021).

DataOps implementation brings a lot of benefits to the organizations that include:

- Improved data quality - by automating and streamlining data processing and management, DataOps can help organizations improve the quality of their data. This can lead to more accurate insights and better decision-making.
- Increased speed and efficiency - DataOps can help organizations process and manage data more quickly and efficiently, which can lead to faster time-to-market for new products and services, and improved customer satisfaction.
- Enhanced collaboration and communication - DataOps emphasizes the importance of collaboration and communication between different teams within an organization. This can help to break down silos and improve overall efficiency and effectiveness. Data teams working closer to business teams will allow a better alignment between data projects and business goals and objectives.

- Reduced risk - DataOps mitigates various risks associated with data management by emphasizing data quality, security, compliance, operational efficiency, collaboration, scalability, and change management. By integrating these practices, organizations can build a more robust and resilient data infrastructure, enabling them to make data-driven decisions with confidence while minimizing potential risks.

Despite these benefits, there are also challenges associated with DataOps implementation in organizations, including:

- Organizational culture - Changing organizational culture can be a significant challenge, as it requires buy-in and support from all levels of the organization.
- Technical complexity - Implementing DataOps can require a significant investment in technology and infrastructure, which can be challenging for organizations with limited resources.
- Data privacy and security - DataOps can involve processing and managing large amounts of sensitive data, which can raise privacy and security concerns. Organizations must have strong data governance and security processes in place to ensure that data is protected.
- Skill and resource constraints – Implementing DataOps can require specialized skills and resources, which can be challenging for organizations with limited resources.
- Integration with existing systems - Integrating DataOps with existing systems and processes can be challenging and requires careful planning and execution.

Despite these challenges, the benefits of implementing DataOps can be significant and organizations that invest in this approach can reap significant rewards in terms of improved data quality, increased speed, efficiency and enhanced collaboration and communication. With DataOps implementation, organizations are more prepared to answer to changes in the market, customer needs and emerging trends. This means faster innovation and more decisions based on data.

2.6. THE TOOLS, TECHNIQUES AND BEST PRACTICES USED IN DATAOPS

DataOps is a methodology that combines data engineering, data integration, and DevOps principles to streamline and automate data operations processes. Therefore, building a data eco-system based on dataOps which is resilient and scalable, high performing and highly available requires a scientific and disciplined approach (Sahoo, 2019). In this chapter is presented some tools, techniques, and best practices commonly used in DataOps implementations.

Like DevOps, DataOps promotes collaboration and communication. Utilizing collaboration tools, such as Jira, Confluence, or Slack, fosters effective communication and knowledge sharing among team members. Centralized documentation, shared dashboards, and chat

channels enable efficient coordination and facilitate cross-functional collaboration. Because DataOps builds on DevOps, cross-functional teams that cut across “skill guilds” such as operations, software engineering, architecture and planning, product management, data analysis, data development, and data engineering are essential, and DataOps teams should be managed in ways that ensure increased collaboration and communication among developers, operations professionals, and data experts (Thor Olavsrud, 2022).

An important feature of DevOps is the automation of such a process: continuous delivery (CD) enables organizations to deliver new features quickly and incrementally by implementing a flow of changes into the production via an automated “assembly line” - the continuous delivery pipeline. This is coupled with continuous integration (CI) that aims at automating the software/product integration process of codes, modules and parts, thus identifying a CI/CD pipeline (Capizzi et al., 2020).

DataOps use DevOps principles that are established between development teams and operation teams, and one of that is the automated data pipelines that are used to streamline and automate data processing and management. These pipelines can be used to process data from various sources, such as databases, APIs, log files, and can be configured to perform various actions, such as data validation, data transformation, and data loading.

Version control systems, such as Git, are essential for managing and tracking changes to code, scripts, and configuration files related to data pipelines and workflows. It enables collaboration, rollback, and traceability of changes, ensuring reproducibility and reliability.

Continuous integration and continuous delivery practices enable automated building, testing, and deployment of data pipelines and workflows. These techniques can be applied to data processing and management systems, to ensure that changes are delivered quickly and reliably, and to minimize the risk of errors and downtime.

Another important practice imported from DevOps are automation, namely automatic tests. Implementing automated testing helps ensure the quality and reliability of data pipelines. Techniques such as unit testing, integration testing, and regression testing can be employed to validate the functionality and accuracy of data transformations, data quality checks, and ETL processes. This approach is used to ensure that data processing and management systems are functioning correctly and can help identify and resolve any issues before they become serious problems.

Monitoring logging and alerting are used to track the performance and health of data processing and management systems, and to identify and resolve any issues that may arise. Implementing robust monitoring and alerting mechanisms is crucial for proactive issue detection and troubleshooting. Monitoring tools, such as Prometheus or ELK stack, can be used to track the performance, availability, and health of data pipelines, while alerting systems can notify teams about anomalies or failures.

Data virtualization is an approach to data analysis that overcomes the challenges of drawing on data stores in various physical locations by creating a virtualized logical data layer that can integrate data sources of multiple types from many global sources without the need to draw in and manipulate data into an additional data store as in data warehouses. This approach beneficially eliminates the need for error prone data replication or data migrations that can lead to data corruption. For the end data users, data is presented in a single unified view, often with advanced visualizations (O Que é Virtualização de Dados? | Hitachi Vantara, n.d.). This can help reduce the complexity of data management and can be used to support real-time data processing and analysis.

Cloud-based data management are powerful tools that enable organizations to effectively manage and use their data providing a centralized, scalable, and flexible platform for managing and processing data. These platforms can be used to store, process, and analyze large amounts of data, and can support real-time data processing and analysis. Cloud-based data management platforms bring benefits like scalability, flexibility and real-time data processing enabling businesses to find insights and make decisions in near real-time. This capability is particularly useful in scenarios where, streaming data or Internet of Things (IoT) data processing is required.

Data governance promotes the availability, quality, and security of an organization's data through different policies and standards. These processes determine data owners, data security measures, and intended uses for the data. Overall, the goal of data governance is to maintain high-quality data that's both secure and easily accessible for deeper business insights (What Is Data Governance?, n.d.). Organizations must implement data governance practices that practices ensure the proper management, protection, and compliance of data assets. Establishing data policies, roles, and responsibilities, and utilizing tools like data lineage trackers, data access controls, and data encryption, helps organizations adhere to regulatory requirements and maintain data privacy and security.

These tools, techniques, and best practices promote collaboration, automation, repeatability, and reliability in data operations. By implementing DataOps principles, organizations can accelerate time-to-insights, improve data quality, and foster a culture of agility and innovation in data-driven decision-making.

2.7. USE CASES AND EXAMPLES OF DATAOPS IMPLEMENTATION IN DIFFERENT ORGANIZATIONS

There are several examples of successful DataOps implementation across different industries and organizations. By using these approaches, organizations can improve the speed, quality, and reliability of their data processing and management systems, leading to improved business outcomes. In this chapter, we show some use-cases of DataOps implementation.

Netflix implemented both DataOps and DevOps to manage its software, information on customer preferences, and viewing habits to improve its content recommendations, personalization, and streaming performance. Netflix uses DataOps to manage its huge data, which includes information on customer preferences, streaming performance, and viewing habits (Kitakabee, 2023). They adopted a culture of collaboration and automation, leveraging tools like Git for version control, Jenkins for continuous integration, and various monitoring and alerting systems. By implementing DataOps practices, Netflix reduced time spent on manual tasks, increased agility, and improved the accuracy of their data insights.

CERN, the European Organization for Nuclear Research, applied DataOps to manage and process the massive amounts of data generated from particle collisions in the Large Hadron Collider. They focused on automating data workflows, implementing real-time data monitoring and alerting, and utilizing containerization technologies like Docker. By embracing DataOps, CERN improved the efficiency of data processing, reduced downtime, and enabled faster analysis and discovery of scientific insights.

The experience of a global pharmaceutical company illustrates the impact of DataOps. The company was struggling with a range of challenges that included the failure to deploy data-science and data-engineering resources and a culture that didn't rely on data and analysis to inform decision making. Too often, data engineers and data scientists were focused on finding and modeling the data needed to run their models, meaning that it took three to six months for new algorithms to be incorporated into actual processes on the ground. The company saw an opportunity to significantly improve the performance of its drug-discovery process by accelerating the integration of advanced analytics into its operations. Specific use cases included the lead-finding process for new drugs as well as the actual operations of the plants. After adopting DataOps, the company was able to improve its development and deployment of analytics processes as well as the quality of the insights. It has now automated the generation of test data and developed improved methodologies to engineer the data for models. It can now catalog its inventory of models and algorithms, automate testing of a large part of the models aligned to the tool landscape, and increase access to various stakeholders (Młodziejewska & Soller, 2021).

General Electric: General Electric (GE) implemented DataOps to optimize their industrial operations and leverage data-driven insights. They established cross-functional teams comprising data engineers, data scientists, and domain experts to collaborate on data projects. General Eletric utilized agile methodologies, implemented automated testing, and adopted tools for data cataloging, data governance, and data quality assessment. The DataOps approach enabled GE to accelerate time-to-market for their data products, enhance data quality, and improve operational efficiency. When compared with data in other sectors (e.g., government, financial services, and retail), industrial data is different. Its creation and use are faster; safety considerations are more critical; and security environments are more restrictive. Computation requirements are also different. Industrial analytics need to be deployed on

machines (sometimes in remote locations) as well as run on massive cloud-based computing environments. As a result, the integration and synchronization of data and analytics, often in real time, are needed more than in other sectors (The Case for Industrial Big Data, n.d.).

These examples demonstrate how organizations across various industries have leveraged DataOps principles and practices to enhance their data operations, streamline processes, improve data quality, and drive better business outcomes. Each implementation is unique and tailored to the specific needs and challenges of the organization, showcasing the versatility and effectiveness of DataOps in different contexts.

3. METHODOLOGY

The main goal of this study is to present a framework that can help companies to introduce DataOps in the company with a strategy. This can be seen as being an artifact. In this way we will use Design Science Research to build this artifact.

Design Science Research (DSR) seeks to enhance technology and science knowledge bases via the creation of innovative artifacts that solve problems and improve the environment in which they are instantiated (Brocke et al., 2004). DSR is a problem-solving paradigm that is commonly used in information systems and is particularly well-suited to information systems research, as the object of our discipline is the study of information systems in organizations and "interest has shifted to organizational rather than technical issues" (Benbasat et al. 1987).

The main goal of DSR is to develop knowledge that professionals of the discipline in question can use to design solutions for their field problems. It is a methodology that intends to establish answers to the whys and how's of the phenomenon in question. This is particularly interesting for this study because the thesis output will be a framework to be used in the companies to implement a correct DataOps strategy.

3.1. DESIGN SCIENCE RESEARCH

This dissertation follows a DSR approach to apply multiple synthetic and analytical methods and perspectives to perform deep research in information systems. The goal of a DSR project is to extend the boundaries of human and organizational capabilities by designing new and innovative artifacts represented by constructs, models, methods, and instantiations (Gregor & Hevner, 2013). The artifacts of this dissertation are a framework and guidelines for the implementation of DataOps in organizational culture.

The DSR has different phases according to the figure 1.

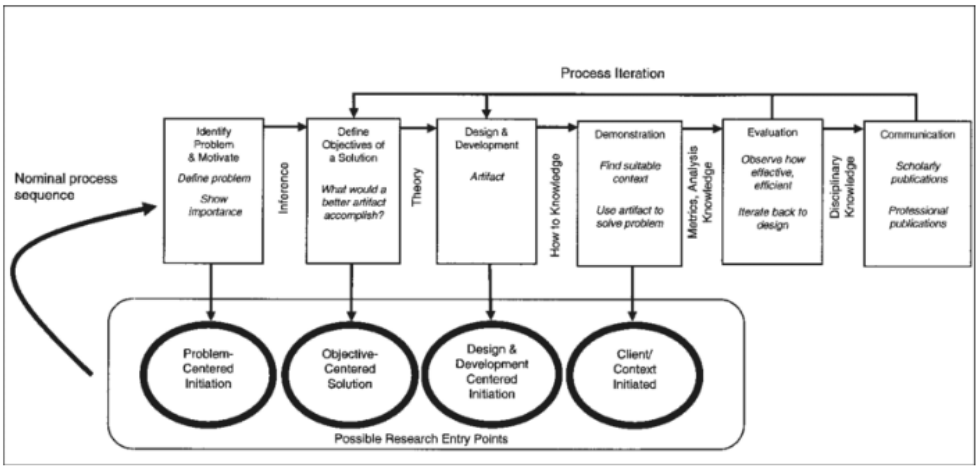


Figure 1: Design Science Research Methodology Process Model (Peffers et al., 2007)

The DSR process starts with problem identification, definition of the research problem and justifying the value of solution. Currently, with the number of applications growing and volume of data increasing, organizations are dealing with many challenges to manage data:

- Requirements change constantly.
- Data lives in silos;
- Data formats are not optimized (the best data structure for development is different for analytics);
- Data errors;
- Bad data destroy the best reports;
- Data pipeline maintenance never ends;
- Regulatory obligations (GDPR);
- Manual process fatigue.

Companies can apply DataOps as an enablement tool across the value chain, from data ingestion through processing, modeling, and insights for the end user. It empowers the provisioning of production data through automated data ingestion and loading from multiple sources. The use of automation for data transformation reduces time-consuming and error-prone steps in the pipeline, continuously improves analytics operations and performance and allows for faster deployment and releases. Last, it accelerates the time to value from data by enabling teams to access real-time data and adjust their business decisions based on the results (Młodziejewska & Soller, 2021).

Organizations need processes to manage this data to allow a sustainable growth.

Following DSR methodology, the second phase is defining the objectives of solution. After this study we expect:

- Build a framework that provides the guidelines to establish DataOps in organization culture. This framework shows how to start a DataOps implementation giving visibility of main challenges and how to deal with them.
- DataOps implementation can and should be implemented in small steps and in this way is important understand how organizations can measure DataOps adoption. The proposed framework intends give some guidance in how to measure the success of DataOps in organization culture.

In the third phase of DSR – design and development, the main objective of this step is to further improve and implement the previously proposed preliminary prototype design to create a specific artifact because of this process step. The approach in this phase may vary depending on the type of artifact to be created.

The fourth step of DSR is demonstration where the main goal is to demonstrate the use of the artifact to solve one or more instances of the problem. This could involve its use in experimentation, simulation, case study, proof, or other appropriate activity.

The fifth step of DSR is evaluation. The main goal of this step is evaluating the quality of the artifact by using the captured feedback. The collected feedback can be used to refine, improve, or redesign artifacts. In the step we can compare the objectives and actual observed results from the use of artifact.

The last step of DSR methodology is to communicate the result of the entire research work. The aim is to ensure that the behavior of the proposed research solution is accepted as sufficient and that the knowledge and facts created are repeatedly applied. Appropriate forms of communication are employed depending upon the research goals and the audience, such as practicing professionals.

3.2. RESEARCH STRATEGY – DSR IMPLEMENTATION

In this study, we use qualitative research methods focusing on critical research. After literature review about DataOps and DevOps we will do a systematic review of the state of art. This section details how each process step of the DSR model was implemented and applied to the study of DataOps implementation.

1. Awareness: This study started with literature review, composed by analytical review of the latest literature in a different scholar and an examination of well-proven and best practices for DataOps implementation. The literature review evidence the main challenges in implementation, the necessary roles in organization, the benefits and present some case studies of DataOps implementation in different industries. This explains the problem definition and is used as a baseline for the suggestion and development of a framework for DataOps implementation.

2. Suggestion: After studying the area of DataOps implementation from a scientific point of view and analyze different case studies of implementation, it can be concluded that a correct strategy of DataOps in organizations brings value, but we need a guidance to the implementation. Moreover, another finding is the lack of conscience for these questions in top management. Besides that, the already presented frameworks and established best practices tend to ignore the different size of the companies and the relation with implemented processes in DevOps. It is necessary to introduce DataOps in systems development lifecycle. Therefore, it is necessary to propose sufficient guidelines that address the main issues and outline a framework for DataOps implementation in the overall domain including strategies for companies with different sizes and include the relation with DevOps processes. This framework illustrates the artifact.
3. Development: The guidelines are developed on the grounds of the respective learnings and key takeaways from the previously performed work, which will be described in detail after.
4. Evaluation: Since the subject of DataOps implementation and its underlying concept is not common matter, the validation of the artifacts will be performed qualitatively, by carrying out interviews with experts from relevant industries.
5. Communication: The last step pf DSR intends to share the outcomes of the study by publishing a scientific article in a relevant journal, whereas the fundamental learnings, limitations, and possible future work are of particular importance. By doing so, outcomes can easily be reached by organizations and DataOps can become a reality in organizations.

In the future, this primary framework can be improved by applying this DSR process model. It is important to note that there is typically a continuous sequence between the steps of development and evaluation. For further consideration and possible improvements, this interaction could be carried out more than once to ensure an even more extensive development phase with respective critical evaluation.

4. DEVELOPMENT AND PROPOSAL OF EVIDENCE-BASED PRACTICE GUIDELINES

This chapter presents a framework to design and facilitate the successful implementation of DataOps strategy in a company. The first section leverages the knowledge from the previously conducted literature review to define preliminary assumptions on which the model will be based on. The second section proposes an implementation framework, supported by diagrams explaining phases and decisions needed. Artifact validation is the subject of the third section. Thereby, the validation will be performed qualitatively, i.e., carrying out interviews with experts. In the last section, the proposed guidelines and respective validation will be critically assessed, discussed and a revised model introduced.

4.1. ASSUMPTIONS

Based on the insights and evidence gained from the extensive literature review on the research gaps, the different approaches and state-of-the-art implementations that serve as best practices within the industry, the assumptions (A) below serve as lessons learned and help to seek the full potential of the proposed implementation guidelines.

A1: The DataOps implementation is an interative process where it is important to learn and adapt the process to the organization.

A2: Build a great team. Find a mix between in-house employees and outsourcing / freelancers to bring more knowledge and discuss ideas. In-house employees know the organization and outsourcing / freelancers have knowledge about different DataOps implementations.

A3: DataOps implementation improve efficiency and collaboration by facilitating knowledge sharing across organizations.

A4: DataOps implementation reduce errors in processes and data with tests and automation.

A5: DataOps can scale data science in organizations accelerating data science projects, having more agility in data access, and creating the necessary environments.

A6: DataOps implementation is a continuous challenge because data sources are changing continuously, data lives in silos, data errors and data pipeline maintenance never ends.

A7: The correct strategy for DataOps implementation allow scale the number of decisions based on facts instead personal decisions.

4.2. FRAMEWORK FOR IMPLEMENTATION

With the growing number of software deliverables, the data available in organizations are increasing too and companies need extract value from this. The proposed framework for DataOps Implementation has two phases important phases as shown in diagram in figure 2.

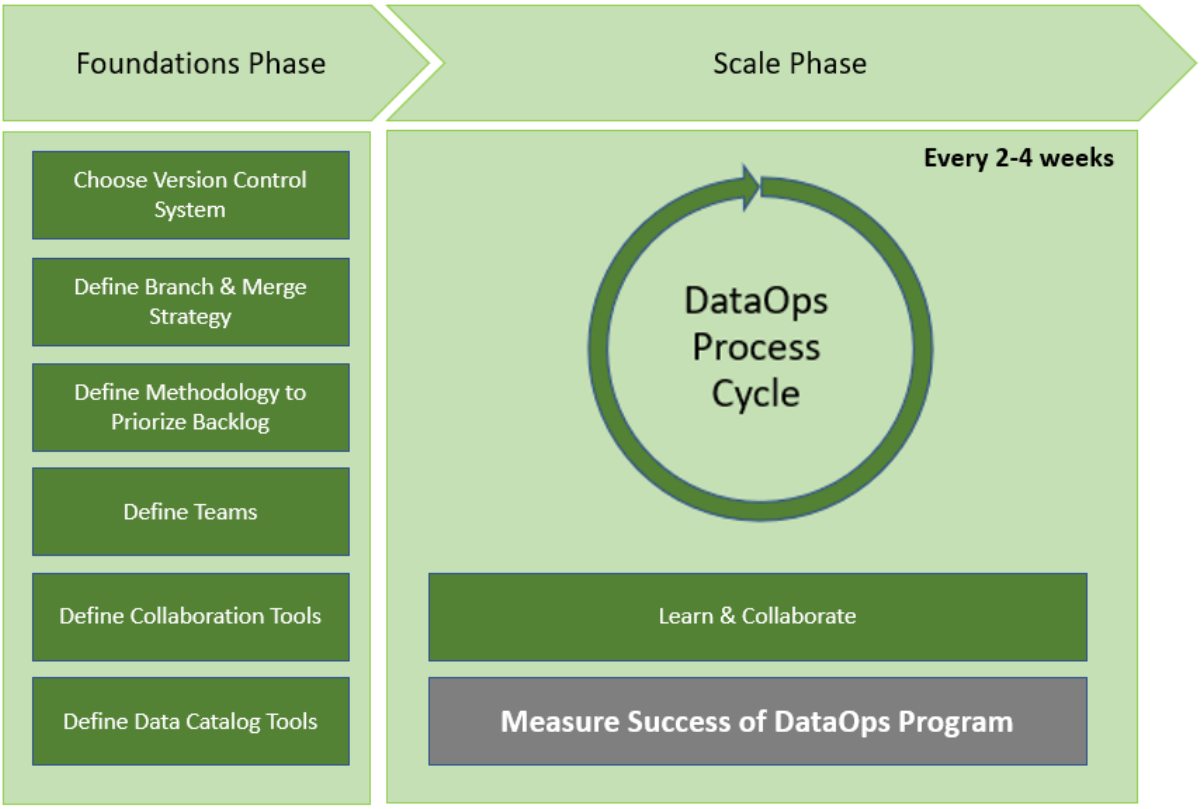


Figure 2: DataOps implementation phases

The first phase for DataOps implementation, called **Foundations Phase**, is crucial because the organization must build the first teams to start DataOps implementation and start the collaboration between them. It is highly recommended to choose a DataOps Program Manager to keep the alignment between data initiatives and business areas. DataOps Program Manager has the responsibility of measuring the impact of DataOps Program and showing the results to all organization. In this phase, teams must choose the systems that will support operations and prepare those systems.

One of the first systems that teams must choose is a **version control system** to record changes in a repository over time, so that it is possible to recall specific versions later. For version control system, if organization has implemented DevOps before, the organization has code repositories with control version system implemented. In this case, teams can use the same control version system that exists in organization considering that version control system did

not had limitations, exists knowledge in organization and this system fits with required processes. There are many excellent version control applications in the market. Git is a popular one, but there are other solutions in the market as shown in table 1.

Table 1: Samples of version control systems brands.

Brand tool	**Description**
Github	Git is an open-source distributed version control system that is available for free under the GNU General Public License version 2. The Git source code is hosted on GitHub, from where it can be downloaded or installed
Azure DevOps	Azure DevOps is a tool from Microsoft that allows manage code repositories and has a version control system.

The **branch and merge strategy** must be defined at Foundations Phase to enable people to safely work on their own tasks. To develop quality processes and be able to reuse them, we need to be able to track all changes and reverse them if necessary. Data teams must define a strategy for branch and merge. A merge is a process that unifies the work done in two branches. Branch and merge can be used in many ways. Currently, two of the most popular development styles that are possible to find are Git flow and Trunk-based development. Quite often, people are familiar with one of those styles and they might neglect the other one. Both have advantages and disadvantages.

Table 2: Git flow or feature-branched development

Git Flow or Feature-branched development
Git Flow is a classic approach to software engineering. Developing individual features is the focus. One of its primary differences from a trunk-based workflow is that it never pushes code changes to the main branch. Developers create feature branches from this main branch and work on them. Once they are done, they create pull requests. In pull requests, other developers comment on changes and may have discussions. Once it is agreed upon, the pull request is accepted and merged to the main branch. Once it is decided that the main branch has reached enough maturity to be released, a separate branch is created to prepare the final version. The application from this branch is tested and bug fixes are applied up to the moment that it is ready to be published to final users. Once that is done, we merge the final product to the master branch and tag it with the release version. In the meantime, new features can be developed on the develop branch.

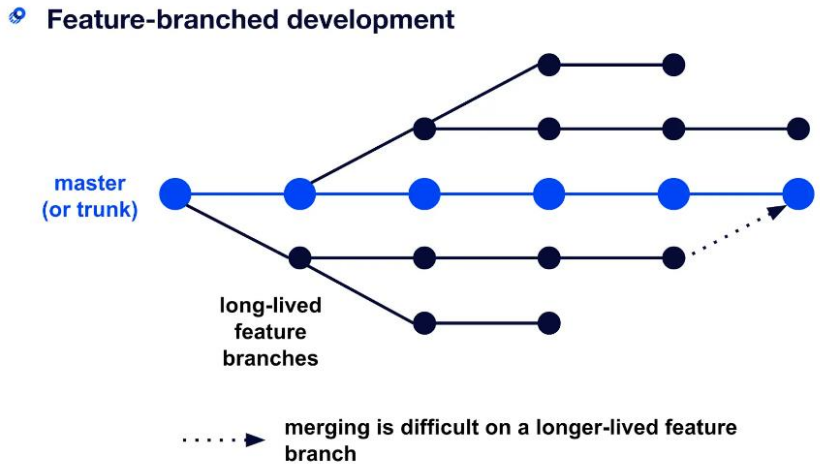


Figure 3: Git Flow or Feature-branched development (Optimezely, n.d.)

Advantages	**Disadvantages**
• Strict control. • Only authorized developers can approve changes. It ensures code quality and helps eliminate bugs early. • Better when the team have a lot of junior developers allows control the code of every feature. • When organization have established developments, allow control new releases.	• Features developed separately can create long-living branches that might be hard to combine with the main project. It delays approvals to production. • More micro-management • When teams need iterate quickly is not the best approach.

Table 3: Trunck-based development

Trunk-based development
In the trunk-based development model, all developers work on a single branch with open access to it. In this strategy only exists the master branch. They commit code to it and run it. It is simpler. Trunk-based development gives programmers full autonomy and expresses more confidence. It provides excellent software development speed and reduces processes. On the other hand, is a shortcut to commit mistakes.

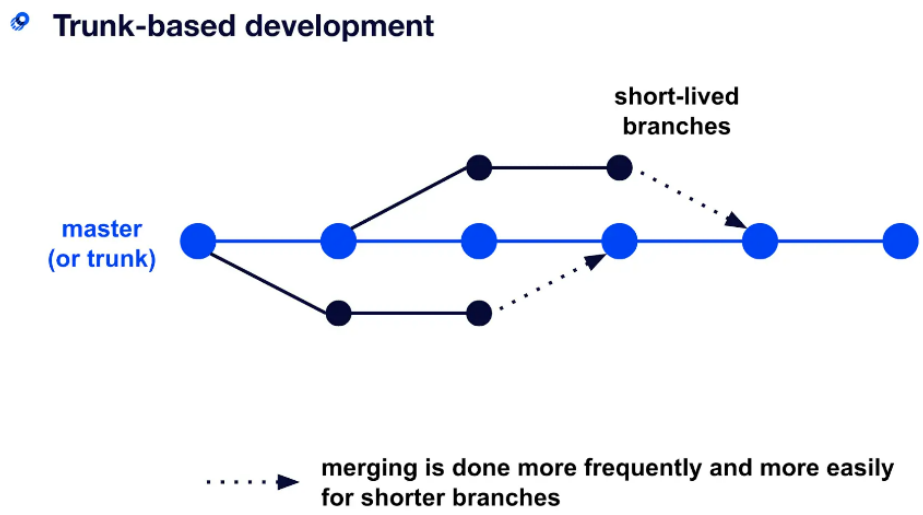


Figura 4: Trunck-based development (Optimezely, n.d.)

Advantages	**Disadvantages**
• Is more efficient when teams are starting a project. • Used when is necessary iterate quickly. • Only recommended when teams are composed by senior developers.	• Is not recommended when you have a lot of junior developers because you only have a master branch. • Is not recommended when you have an established product or manage large teams.

Another important step in foundations phase is defining the methodology to prioritize backlog. The teams are limited and in this way the companies must decide where they want to invest in each DataOps Process Cycle. The teams that work with DataOps, usually work for different teams or departments and for this a clear methodology will help business teams understand backlog prioritization. In this way, analytics use cases should be prioritizes based on feasibility and impact.

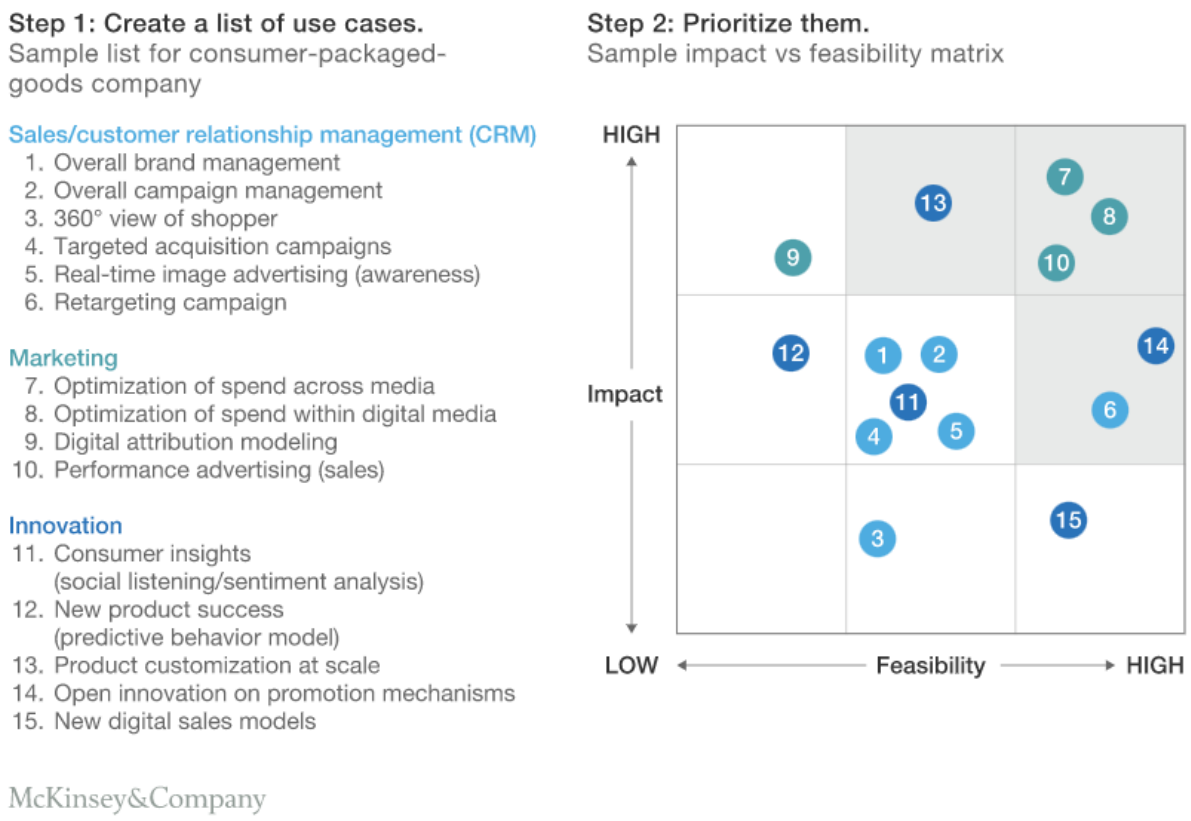


Figure 5: Prioritization based on feasibility and impact (Fleming et al., 2018)

When the team is planning and prioritizing backlog, they must think in the following questions that will help decide prioritization:

- Expected impact in organization?
- The necessary data for that use case is accessible and have enough quality?
- How many processes is necessary change to achieve results?
- Would the team involved in that process have to change?
- What Processes that can be changed with minimal disruption and huge impact?

This matrix shows transparency in organization about backlog prioritization.

In foundations phase, teams must define collaboration tools that will be used by all teams, it is where will be defined the features, user stories and tasks of each cycle. In this collaboration tools is expected that all teams have visibility about roadmaps of each initiative and what features will be developed in each sprint.

There are many applications in the market that organization can use for this, such as Jira or Azure DevOps tools, as shown in table 4.

Table 4: Examples of collaboration tools brands

Brand tool	Description
Jira	Jira is a software application developed by Atlassian that allows teams to track issues, manage projects and automate workflows. Jira is a suite of agile work management solutions that powers collaboration across all teams from concept to customer. Jira offers several products and deployment options that are purpose-built for Software, IT, Business, Ops teams, and more.
Azure DevOps	Azure DevOps is a tool from Microsoft that allows manage code repositories and has a version control system. Azure DevOps supports a collaborative culture and set of processes that bring together developers, project managers, and contributors to develop software. It allows organizations to create and improve products at a faster pace than they can with traditional software development approaches.

Build a great team. The organization must choose a program manager for DataOps program that have the responsibility of manage DataOps Program and guarantee the success of DataOps implementation measuring all KPI's defined during the process, namely IT KPI's, Data KPI's and Business KPI's. The organizations must find a mix between in-house employees and outsourcing / freelancers to bring more knowledge and discuss ideas. In-house employees know the organization and outsourcing / freelancers have knowledge about different DataOps implementations. Bring onboard data analysts, data architects, data engineers and other stakeholders. To start a DataOps implementation, is not required have all profiles in the beginning because organization can decide to grow the teams throughout implementation but to achieve success is mandatory bring onboard business skills, technological skills, and analytics skills. The figure 8 shows the recommend teams with respective profiles to start building a DataOps Program.

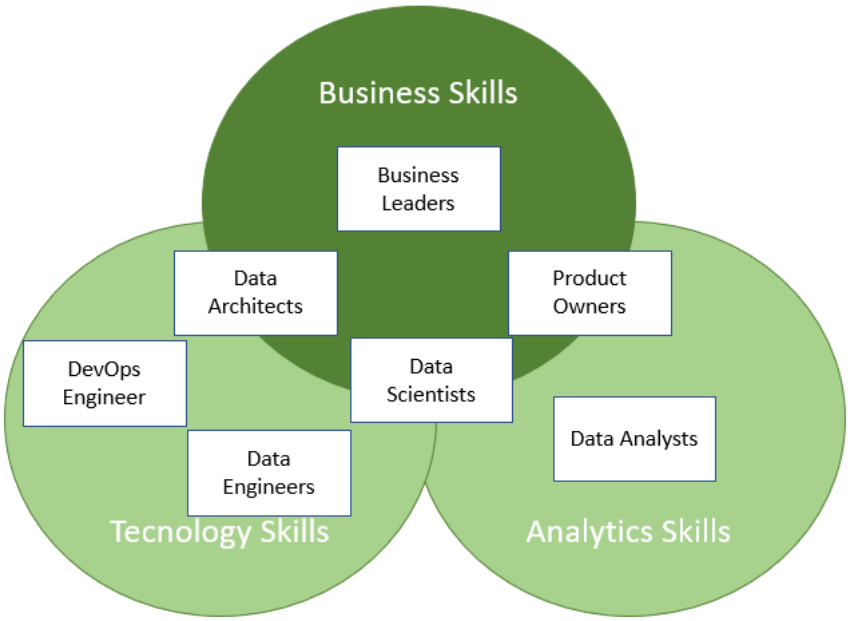


Figure 6: Recommended teams for DataOps implementation

During foundations phase, DataOps Program Manager and Data Teams must explain the strategy to stakeholders and bring them onboard. Give visibility that this strategy will produce Business KPI's that must be followed by stakeholders to know the value of data that have in your organization. Throughout DataOps implementation is common appear new Business KPI's and this can be a signal that this program has impact in organization and business areas are involved.

DataOps is an iterative implementation, and all teams must learn and collaborate during all cycles. It is crucial to collect the lessons learned to avoid repeated mistakes and establish better processes for the future. In the end of each cycle, the program manager or project manage can arrange a meeting to Cycle Review to collect and share this with all teams.

After foundations phase, it is time to enter in Scale Phase and start the implementation of the first use cases and show the results to the organization. This phase should not take longer than 3 or 4 weeks to introduce the idea of cycles where teams deliver new features / improvements to the organization. It is very important that teams learn with implementation and establish collaboration processes between them.

Scale Phase

After establish the foundations, the **Scale Phase** is where the business and technology teams are working together delivering features and improvements to the organization. In the beginning is not expected that all processes are implemented but is expected that the process efficiency increase cycle by cycle. Teams must learn and collaborate to improve processes according organization structure.

In the beginning of each cycle, like Agile Methodology, in the beginning of each cycle must be clear for all teams what is expected deliver in the cycle. This allow align expectations between all teams, namely technical teams, and stakeholders. During the cycle, the organization must not change the features or user stories that were accorded deliver in the cycle. If this happens, it will be difficult to have the commitment of all teams for next cycles.

A DataOps Cycle should have between 2 and 4 weeks (no more than this). During each cycle exist a lot of activities that are explained in the diagram in figure 9:

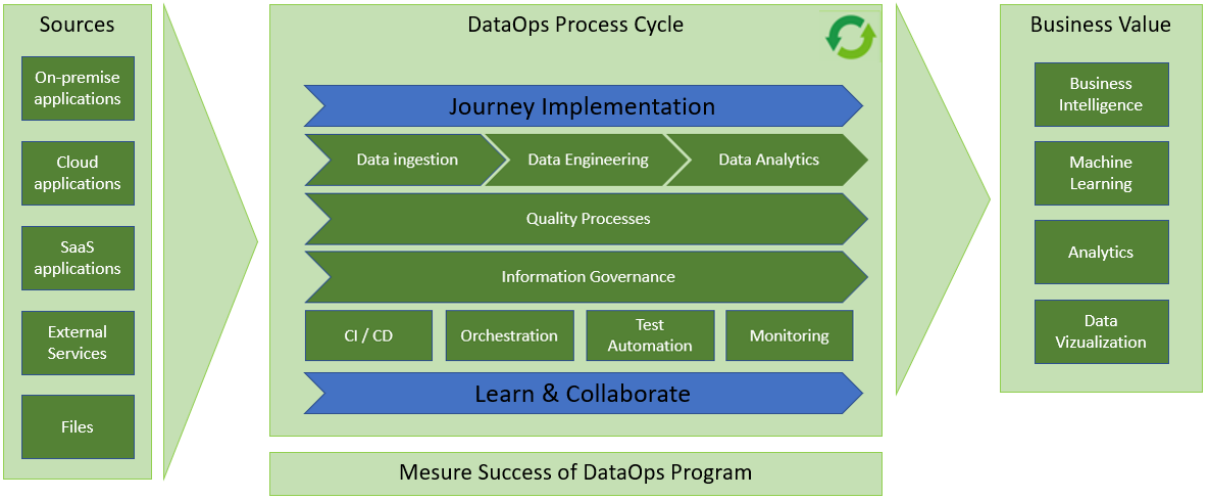


Figura 7: DataOps Process Cycle

Each cycle must start with planning and define the deliverables that the team must deliver in the end of each cycle. To deliver new features / improvements tasks of data ingestion, data engineering and data analytics are done by people of the team. To develop these tasks, teams need to deliver under some processes and according to rules explained in table 1.

Table 5: Rules that must be followed to deliver new features / improvements

CI/CD	The data pipeline architecture forms an integral part of a robust data platform. CI/CD (Continuous integration and continuous delivery) is a DevOps best practice for delivering code changes more frequently and reliably. Because of this, all teams must work in CI/CD.
Orchestration	Today data engineer and architect teams are stretched thin against an ever-growing list of data sources and data management tools. These teams need to keep vital information flowing to business end-users. However, most data pipelines are stitched together by a chaotic mess of point-to-point integrations and custom scripts that connect their data sources, ingestion tools, storage tools, and analytics solutions. A DataOps orchestration solution does not replace existing data pipeline and analytics tools. Instead, it serves as a platform that integrates with these tools. The orchestrator is a software entity that is responsible for managing the processes, control the execution of the steps and handles the exceptions. It is the only one who knows when a pipeline should be executed at a given moment. Therefore, it should be the only one who can trigger that execution (Rodriguez et al., 2020). Once integrated, teams may control and automate the actions and processes within each application or platform from the DataOps orchestration solution.
Test Automation	Every feature or improvement delivered to the organization must be accomplished by automated tests to reduce errors in production environments. These tests must be automated to be executed all days or when is needed.
Monitoring	All KPI's defined for every feature or improved must be monitored in production to understand if exist errors in production and detect future improvements to the organization.

Quality Processes

Every organization, when implementing DataOps must have Quality Processes that must be executed by different teams for auditing questions. Following this framework, we recommend two different types of quality processes:

- Data Quality Processes – Data quality processes are essential for ensuring the accuracy, reliability, and usability of data within an organization. These processes involve various activities and techniques aimed at improving the quality of data, identifying and correcting errors, and maintaining data integrity throughout its lifecycle. Data Quality

team must define KPI's with business areas to measure improvements in data quality cycle by cycle. Organization can establish KPI's based on Data Quality Metrics:

Table 6: Examples of KPI's for Data Quality Processes

Data Quality Metrics	**Description**	**KPI example**
Accuracy	Measure the number and types of errors in a dataset. Data can be compared with reference dataset.	Customer entered addresses and ZIP codes can be compared with address registry of CTT. The company must define the rules and can provide KPI based on that rule.
Completeness	It's important that all critical fields in a record be fully populated.	A Customer record missing phone number is incomplete. A KPI could be number of customers with completed data.
Consistency	Consistency measures how individual data points pulled from two or more sets of data synchronize with one another. If two data points are in conflict, it indicates that one of both records are inaccurate	A KPI could be the number of duplicated customers.
Timeliness	Measures the age of data in a database	Customers with updated data last year.
Uniqueness	The uniqueness metric tracks duplicate data.	The number of customers.
Validity	Validity measures how well data conforms to standards	Number of customers with wrong data format.

- Continuous monitoring of data quality is crucial to identify any emerging issues or deviations from predefined quality standards. This can be done through automated tools or manual checks to ensure that data remains accurate and reliable over time.

Monitor KPI's defined with business areas to verify if exist improvements in organization cycle by cycle. This monitorization can help find new features or improvements that can bring value to organization.

Data quality tools and processes are essential for ensuring that data is accurate, consistent, and compliant with regulations.

Information Governance

The organization must adopt Data Governance tools to manage data complexity, ensure data quality, comply with regulations, increase data trust, and reduce risk. The policies for Data Governance must be defined by Data Governance team and in each cycle of DataOps implementation, teams must comply with their policies.

Data Governance team must ensure that policies and responsibilities are clearly understood by all teams. Is this way, Data Governance team must prepare regular training sessions across all organization to spread this knowledge and increase organization maturity at this level.

Data catalog tools are essential for managing the increasing volume, variety, and complexity of data in today's data-driven world. It allows managing data assets, enabling self-service data access, and supporting data analytics. In figure 10, there are some examples of products in the market that organizations can adopt for data catalog.

Figura 8: Examples of Data catalog tools brands

There are several motivations to catalog and govern enterprise data. Governance motivates regulated enterprises, but observability and manageability are more common drivers. Continuous metadata processes could benefit enterprises in governance, security, data quality, compliance, Human / IT resource optimization, transparency, decision support and AI / Machine learning. (Underwood, 2023).

Measure success of DataOps Program

To achieve the goals proposed by this framework is mandatory to measure the success of this program. This job can start with a few KPI's and grow throughout the implementation. The Program manager must evolve all teams in KPI definition of the program and grow during implementation because some KPI's make sense in the beginning but not with growth. The Program manager, business leaders and IT leaders must develop strategies with the teams to transform this KPI's in "Keep Persons Involved" or "Keep Persons Informed".

The KPI's must be defined based on their specific objectives, challenges, and requirements. By tracking these metrics, organizations can continuously improve their data management processes, ensure data quality and security, and drive data-driven decision-making across the business.

DataOps implementation program must have different types of KPI's:

- Business KPI's – focus on the business goals and can measure for example number of customers, sales, revenue, customer satisfaction and can measure data culture promoting a data-driven decisions.
- IT KPI's - focus on the performance, reliability, and efficiency of data-related systems and technologies. These can include data availability, data processing time, number of errors in pipelines among others.
- Operational KPI's – focus on operation and could measure data integration time, compliance tracking adherence to data governance policies and regulations, number of data-driven decision making and data adoption rate. When teams deliver a feature or a report, must be measured if this feature are being used or not.
- Quality KPI's – focus on the quality of data and assesses the accuracy, completeness, and consistency of data to ensure high-quality information.

By following the KPI's defined throughout the program, the organization will gain greater knowledge about the value that data brings to the organization. The maturity of the organization about this will grow and more decisions will be taken based on data and not only in perceptions.

The program manager, project managers and business leaders must give visibility the KPI's for all organization.

4.3. EVALUATION

According to the DSR methodology, we have proceeded to the validation phase, where this framework for DataOps implementation was validated.

In this phase four quality assessment questions will be executed, as stated in the sixth guideline of (Hevner et al., 2004), where the artefacts, which adhere to this methodology, were presented to three different specialists who daily have the responsibility to manage data and responsibility on data analytics, data engineering and data governance. These specialists

have also answered the four question to validate the frameworks. These questions were regarding quality and utility assessment.

For the presentation of the quality assessment questions, a presentation in PowerPoint (Annex 1) format was prepared where the following points were mentioned: objective of this thesis; problem statement; main lessons learned from literature review, present proposed framework for DataOps implementation, the required elements of this framework; detailed explanation of each phase and detail some use-cases where this framework can be used. The tree people selected to answer the questions were people with responsibilities in data management and people who face the adversities of this area daily.

4.3.1. Interviews description

Presentation of the interviewees, education and current professional situation and presentation of the questions that were performed in the presentation and PowerPoint (Annex 1).

- **Interviewee #1** – PhD in systems and information technology. Specialist in enterprise architecture and digital transformation. Senior researcher in information systems. More than 25 years of experience in Information Systems.

- **Interviewee #2** – Business Intelligence Lead and Data Governance at the major football club in Portugal. More than 13 years of experience working with data in business intelligence projects.

- **Interviewee #3** – Business intelligence architect in the major airline company in Portugal. Previously he worked in different business intelligence projects as MicroStrategy Specialist in airline companies and other industries. More than 12 years of experience working with data in business intelligence projects.

Each interview was followed by artifacts presentation (Annex 1) and included four questions:

- **Question 1 (Q1):** Do you consider the proposed framework as useful and why? If not, why do you believe it is not?

- **Question 2 (Q2):** Do you have any criticism towards the proposed framework? Please explain.

- **Question 3 (Q3):** Would you consider implementing the proposed framework? Please clarify why/ why not.

- **Question 4 (Q4):** Do you have any recommendation or suggestions for further improvements of the proposed framework?

4.3.2. Discussion

On the 29th May it was arranged a meeting with three experts with large experience in Enterprise Architecture, Data Management, and experience in Business Intelligence projects. A focus group was set in a trial to answer the four key questions. These questions led to a very interesting debate around this issue which is a daily challenge.

The interviews for the evaluation and validation of framework proposal to introduce DataOps in enterprise culture, allowed a discussion about its utility, possible criticisms to the structure, technologies involved, its functioning, the challenges, among other important factors. The interviews also permitted to perceive if there are recommendations that improve this framework.

- **Question 1 (Q1):** Do you consider the proposed framework as useful and why? If not, why do you believe it is not?

The interviewees agree with the framework presented and considered very useful for DataOps implementation. All organization are facing challenges in data management and need processes to manage data. In many organizations, data engineering is disconnected from data analytics and this framework can address this point, by trying to bond them together.

The interviewees considered that the DataOps implementation become more facilitated when organization has DevOps established. The DevOps brings processes to application development and increase the data generated by these applications causing a disruption in data. The DataOps implementation is more than useful, it is vital for any organization and as soon as we realize this, easier will be to lead to its implementation.

This framework is very interesting and may help all areas, namely IT, Data, and business to establish processes and grow together which will make the Organization grow as well and increase its gains.

- **Question 2 (Q2):** Do you have any criticism towards the proposed framework? Please explain.

The interviewees considered the architecture of this framework very complete, and it details exactly the different activities that any organization need when work with data. In DataOps Process cycle, Interviewer #1 suggested change the name Information Governance to Compliance because Information Governance is above of operational activity and this change gives more visibility that in this step must be implemented the policies defined by team of information governance. This suggestion was subject of discussion, and we considered that could be an improvement to this framework and detail more these two important activities in organizations.

- **Question 3 (Q3):** Would you consider implementing the proposed framework? Please clarify why/ why not.

All interviewees consider implement this framework and agree that this framework could help to minimize inefficiencies between data management and information technology and will help to establish processes across all teams. The volume of data is increasing a lot in many organizations and is consensual that companies need processes to manage this.

The interviewees considered that a DataOps implementations become more facilitated when DevOps is established because some processes are similar and the way to implement DataOps is faster.

- **Question 4 (Q4):** Do you have any recommendation or suggestions for further improvements of the proposed framework?

The interviewees agreed with the adoption of this framework and methodology. One of the interviewees considered that Foundations Phase is fundamental to process definition and bring all areas onboard. It's crucial implement DataOps using this framework or other similar because many organizations work a lot in an old-fashioned way, and we need improve, establishing processes across all teams and grow together, making good use on new technologies.

For future work, this project could be complemented with a use-case and could detail the starting point to DataOps implementation, because organizations need some maturity to implement this.

According to the interviewees' answers, all organizations need processes to deal with data and the proposed framework for DataOps implementation is very complete, allowing collaboration across all areas that will generate value for the company.

The interviewees congratulate for the excellent work presented.

4.3.3. Revised framework for implementation

The foundation for the revised model is provided by the previously conducted expert interviews. As a result of the constructive discussion above, there was an improvement in the proposed framework: all agree that it brings value.

After the interviews to validate the proposed framework, it was proposed to better clarify the activity - Information Governance. Information Governance is a discipline above of operational activities where data policies are defined. In DataOps Process Cycle, Information Governance is changed to Information Governance (compliance) because during each DataOps Process Cycle, teams, when working with data, must comply with Information Governance Policies and responsibilities which are defined in the organizations.

This change increases the understanding of activities that are expected during DataOps Process Cycle and evidences the responsibilities of teams that participate in these cycles. This keeps the independence of the teams who define Information Governance policies in the organization. The DataOps Process Cycle was changed, as shown in the figure 11.

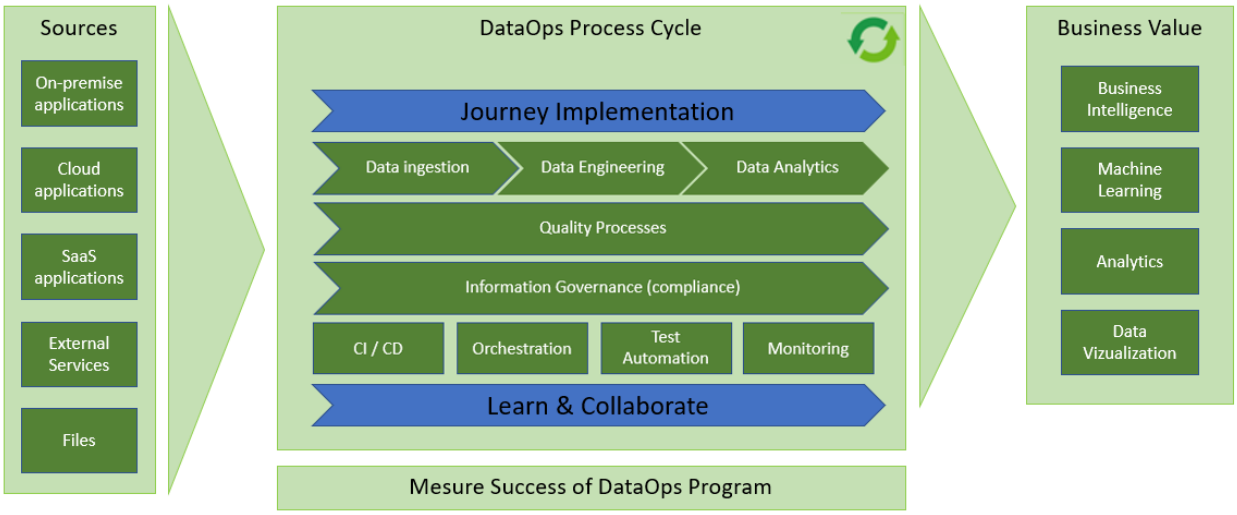


Figura 9: DataOps Process Cycle Revised

5. CONCLUSIONS AND FUTURE WORKS

Many organizations are dealing with problems in data management and sooner or later need to implement DataOps to deal with complexity of data. According to interviews in validation phase, this is a common problem in many organizations and in many cases, data engineering is disconnected from data analytics and the proposed framework can address this point, by trying to bond them together.

The DSR was used to build this framework. In the Evaluation phase, the results of the architecture were evaluated and interpreted from interviews with experts in this sector with a lot of experience in data management. The interviewees responded to four questions after a brief presentation about the architecture. According to expert interviews, is very complete and could provide the guidance to this implementation. With this framework organizations can establish DataOps and measure improvements during this way.

Finally, the research objective was achieved, and the research questions were answered. In the literature review, this study presented DataOps methodology and show the challenges that organizations have when decide to implement a DataOps strategy. The proposed framework provides the guidance and best practices to follow and show how can start this process.

5.1. SYNTESIS OF THE RESEARCH

DataOps is an approach that combines agile methodologies, DevOps practices, and data management principles to improve the efficiency and quality of data operations. The key principles of DataOps include collaboration, automation, continuous integration, version control, and monitoring and observability to establish monitoring systems to track the health, performance, and reliability of data pipelines, allowing for proactive issue detection and resolution.

A lot of organizations are facing challenges in data management by several reasons like lack of data governance, data lives in different silos, poor data quality, difficulty for integrate and manage multiple data sources, manage data security and privacy among others. DataOps implementation can bring benefits like Improved data quality through automated processes, Faster time-to-insight, enhanced collaboration between data engineers, data scientists, and other stakeholders. This brings Increased agility to respond quickly to changing business needs and adapt their data operations accordingly.

The implementation of DataOps represents a cultural change because requires a shift in mindset and collaboration between different teams. It is crucial to promote a data-driven culture in organizations. Several organizations have successfully implemented DataOps and reported significant improvements in data delivery, collaboration, and data quality.

5.2. RESEARCH LIMITATIONS

The main limitations of this thesis dissertation were mainly the time and scope. This research was designed to complete a master's thesis that features deadlines. The scope of DataOps implementation is very wide and with more time is possible explore in detail each topic of proposed framework.

A literature review was done on DataOps and DevOps implementation that require high knowledge and experience to understand and build the proposed framework. Each phase of this framework was explained, however each activity can be detailed more and complemented with samples if there was more time to do the research.

Another limitation is noticeable in the process of validation. Due to the time restriction, it is not possible to re-validate the revised model, which has been improved with additional feedback from expert interviews. The DSR approach is applied, but not fully exploited.

5.3. FUTURE WORKS

The subject of DataOps implementation is very wide and, in this way the proposed framework is very complete but also offers some opportunities for future work.

As future work, this framework can be used to implement DataOps in different scenarios / industries. All organization are implementing or will need to implement in near future DataOps and the adoption of this framework allows collect feedback for future improvements.

In a future revision of this framework, it can be detailed what organizations need in terms of maturity level to start the DataOps implementation. In this dissertation, it is clear that companies that have DevOps established, the DataOps implementation become more facilitated, but the starting point could be more detailed and maybe defined as pre-requirements to start.

The limited approach to the topic of data privacy, data ownership, and security within this work represents an opportunity to improve this framework to meet regulatory requirements of organizations, namely GDPR in European Union.

BIBLIOGRAPHICAL REFERENCES

Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management*, *49*, 424–438. https://doi.org/10.1016/j.ijinfomgt.2019.07.008

Atwal, H. (2019). *Practical DataOps: Delivering agile data science at scale* (p. 275). Scopus. https://doi.org/10.1007/9781484251041

Brocke, J. vom, Hevner, A., & Maedche, A. (2004). *Introduction to Design Science Research* (pp. 1–13). https://doi.org/10.1007/978-3-030-46781-4_1

Capizzi, A., Distefano, S., & Mazzara, M. (2020). *From DevOps to DevDataOps: Data Management in DevOps Processes* (pp. 52–62). https://doi.org/10.1007/978-3-030-39306-9_4

Clive Humby. (2023). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Clive_Humby&oldid=1160630515

Creating a data-driven culture. (2020, June 2). IDG. https://thelivingenterprise.cio.com/collection/building-data-driven-business/article/thelivingenterprise.cio.com/collection/building-data-driven-business/article/creating-a-data-driven-culture

Definition of DataOps—Gartner Information Technology Glossary. (n.d.). Gartner. Retrieved November 20, 2022, from https://www.gartner.com/en/information-technology/glossary/dataops

Demchenko, Y., Zhao, Z., Surbiryala, J., Koulouzis, S., Shi, Z., Liao, X., & Gordiyenko, J. (2019). Teaching DevOps and Cloud Based Software Engineering in University Curricula. *2019 15th International Conference on EScience (EScience)*, 548–552. https://doi.org/10.1109/eScience.2019.00075

Eight Factors that can impact Data Quality. (2023, January 20). https://www.managedoutsource.com/blog/factors-that-can-affect-data-quality/

Ereth, J. (2018). *DataOps – Towards a definition*. *2191*, 104–112. Scopus.

Fleming, O., Fountaine, T., Henke, N., & Saleh, T. (2018, May 14). *Getting your organization's advanced analytics program right | McKinsey*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/ten-red-flags-signaling-your-analytics-program-will-fail

Gregor, S., & Hevner, A. (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly*, *37*, 337–356. https://doi.org/10.25300/MISQ/2013/37.2.01

Gür, I., Möller, F., Hupperz, M., Uzun, D., & Otto, B. (2022). Requirements for DataOps to foster Dynamic Capabilities in Organizations—A mixed methods approach. *2022 IEEE 24th Conference on Business Informatics (CBI)*, *01*, 166–175. https://doi.org/10.1109/CBI54897.2022.00025

Henrion, M., & Gatos, L. (n.d.). *Why most big data analytics projects fail: How to succeed by engaging with your clients*.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (n.d.). *Design Science in Information Systems Research*.

Key DataOps Challenges Enterprises Must Overcome. (2021, July 12). *Business of Data*. https://business-of-data.com/articles/dataops-challenges/

Kitakabee. (2023). *DataOps vs DevOps: Key Differences*. BrowserStack. https://browserstack.wpengine.com/guide/dataops-vs-devops/

Mainali, K. (2021). *Discovering DataOps: A Comprehensive Review of Definitions, Use Cases, and Tools*.

Młodziejewska, M., & Soller, H. (2021, May 8). *How companies can use DataOps to jump-start advanced analytics | McKinsey & Company*. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/tech-forward/how-companies-can-use-dataops-to-jump-start-advanced-analytics

O que é virtualização de dados? | Hitachi Vantara. (n.d.). Retrieved June 25, 2023, from https://www.hitachivantara.com/pt-br/insights/faq/what-is-data-virtualization.html

Optimezely. (n.d.). *Trunk-based development*. Optimizely. Retrieved April 12, 2023, from https://www.optimizely.com/optimization-glossary/trunk-based-development/

Peffers, K., Tuunanen, T., Rothenberger, M., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*, 45–77.

Phuong, N. T. T. (2021). *DataOps for Product Information Management: A study of adoption readiness*.

Rodriguez, M., Araújo, L. J. P. de, & Mazzara, M. (2020). Good practices for the adoption of DataOps in the software industry. *Journal of Physics: Conference Series*, *1694*(1), 012032. https://doi.org/10.1088/1742-6596/1694/1/012032

Sahoo, P. R. (2019). DataOps in Manufacturing and Utilities Industries. *International Journal of Applied Information Systems*, *12*.

The 7 most common data quality issues. (n.d.). *Collibra*. Retrieved June 25, 2023, from https://www.collibra.com/us/en/blog/the-7-most-common-data-quality-issues

The Case for Industrial Big Data. (n.d.). Retrieved June 26, 2023, from https://www.ge.com/digital/blog/case-industrial-big-data

The data-driven enterprise of 2025 | McKinsey. (n.d.). Retrieved February 12, 2023, from https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-data-driven-enterprise-of-2025

The DataOps Manifesto—Read The 18 DataOps Principles. (2021, May 10). https://dataopsmanifesto.org/en/

Thor Olavsrud, N. (2022, December 22). What is DataOps? Collaborative, cross-functional analytics. *CIO*. https://www.cio.com/article/227979/what-is-dataops-data-operations-analytics.html

Underwood, M. (2023). Continuous Metadata in Continuous Integration, Stream Processing and Enterprise DataOps. *Data Intelligence*, *5*(1), 275–288. https://doi.org/10.1162/dint_a_00193

What is data governance? | IBM. (n.d.). Retrieved June 25, 2023, from https://www.ibm.com/topics/data-governance

What is Data Governance (and Do I Need It)? (n.d.). Talend - A Leader in Data Integration & Data Integrity. Retrieved June 25, 2023, from https://www.talend.com/resources/what-is-data-governance/

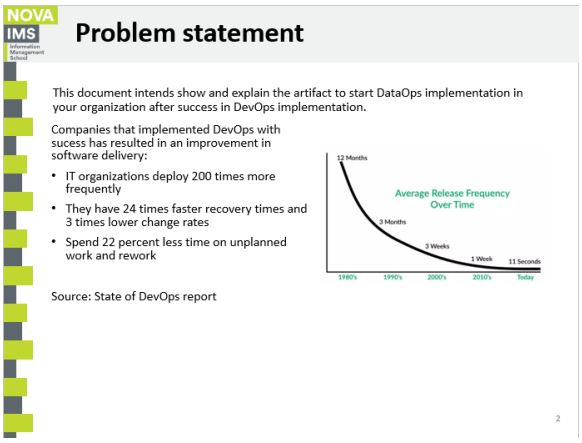
Why do 87% of data science projects never make it into production? (2019, July 19). *VentureBeat*. https://venturebeat.com/ai/why-do-87-of-data-science-projects-never-make-it-into-production/

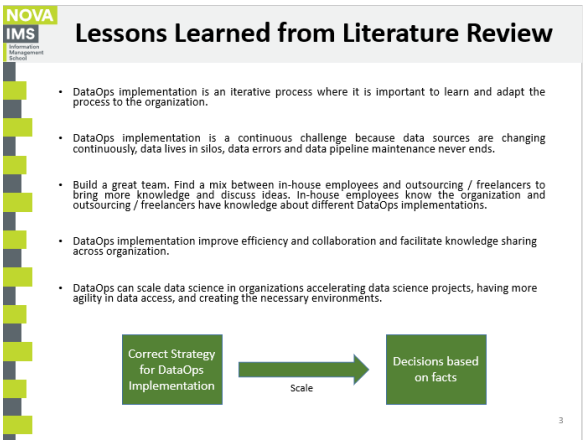
Yu, S., Chen, T., Han, L., Demartini, G., & Sadiq, S. (2023). DataOps-4G: On Supporting Generalists in Data Quality Discovery. *IEEE Transactions on Knowledge and Data Engineering*, *35*(5), 4668–4681. https://doi.org/10.1109/TKDE.2022.3151605

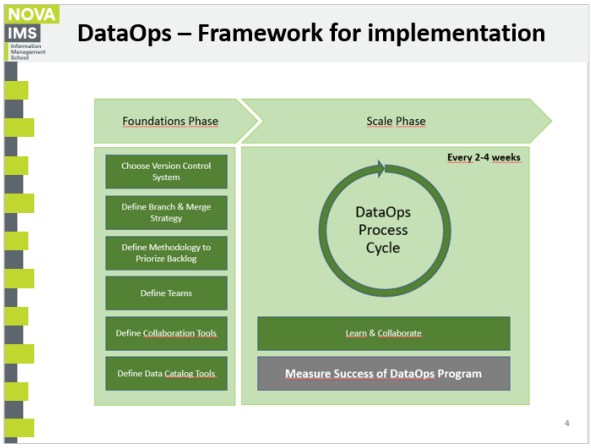
ANNEXES

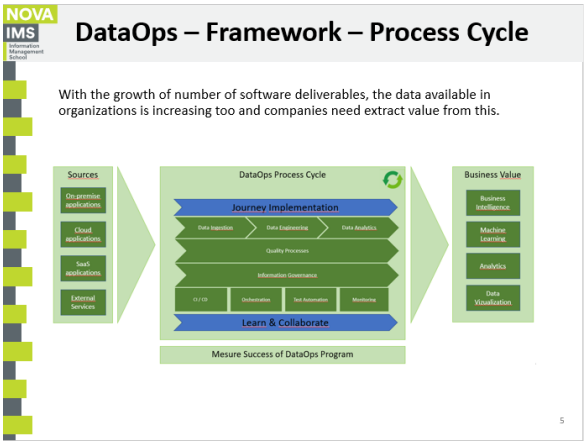
ANNEX 1 – Interviews Presentation

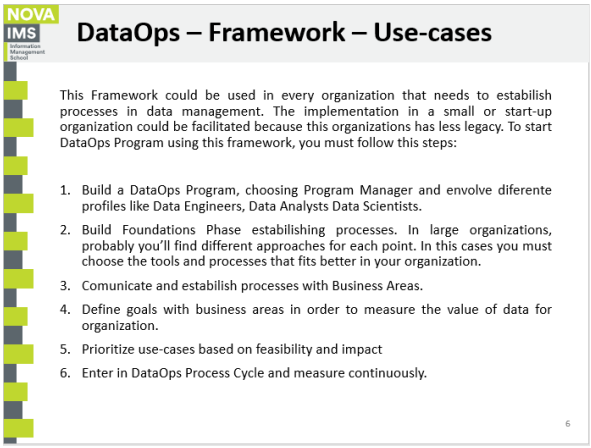
On the 29th May a meeting was arranged with three experts in Data Management and Developing Projects with Data where the following presentation about how DataOps can be introduced in Enterprise Culture was shown.

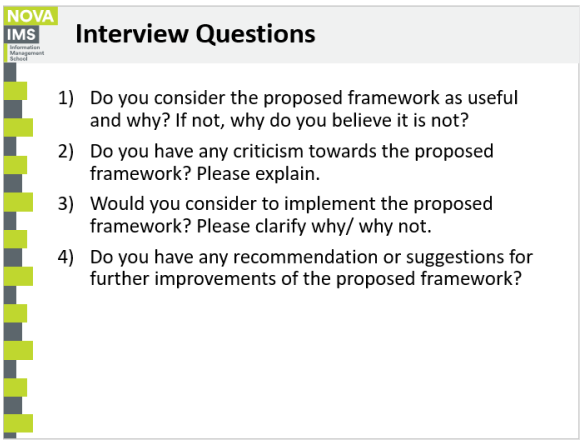












ANNEX 2 – Interviews Transcription

On the 29th May a meeting was arranged with three experts in Data Management and Developing Projects with Data. A focus group was set in a trial to answer the four key questions. These questions led to a very interesting debate around this issue which is a daily challenge.

The three specialists have different profiles as it is summed in the next paragrapher.

- **Interviewee #1** – PhD in systems and information technology. Specialist in enterprise architecture and digital transformation. Senior researcher in information systems. More than 25 years of experience in Information Systems.

- **Interviewee #2** – Business Intelligence Lead and Data Governance at the major football club in Portugal. More than 13 years of experience working with data in business intelligence projects.

- **Interviewee #3** – Business intelligence architect in the major airline company in Portugal. Previously he has worked in different business intelligence projects as MicroStrategy Specialist in airline companies and other industries. More than 12 years of experience working with data in business intelligence projects.

Interviews #1: Date: 29/05/2023

Q1: Do you consider the proposed framework useful? why? If not, why do you believe it is not?

I would say the framework is useful because it is a subject that is missing in organizations. The use of different technologies and data exploration in different clouds is growing and that brings new challenges for the organizations. All organizations need to answer to the data

management challenges that lie ahead. The data engineering is disconnected from data analytics and this framework can address this point, by trying to bond them together.

Q2: Do you have any critics of the proposed framework? Please explain.

In DataOps Process Cycle, I suggest change Information Governance to compliance. Information Governance is above of this and must be used to define policies. In this step, if you change this to compliance, it could be the implementation of these policies to guarantee the correct implementation of policies defined by Information Governance.

Q3: Would you consider implementing the proposed framework? Please clarify why / why not?

Yes, If I could, I would implement this framework or other similar as this framework could help to minimize inefficiencies between data management and information technology. This framework can help to establish collaboration processes across all teams. We need understand if an organization must start DataOps implementation if DevOps is not established in the organization. Some processes can be made more efficient just by implementing this framework and as so become more valuable.

Q4: Do you have any recommendations or suggestions for further improvements of the proposed framework?

For future work, this project could be complemented with a use-case and could detail the starting point to DataOps implementation, because organizations need some maturity to implement this.

A more detailed explanations and set up point with close follow up, may help to get them more efficient.

Interviewee #2: Date: 29/05/2023

Q1: Do you consider the proposed framework useful? why? If not, why do you believe it is not?

I consider that DataOps implementation is easier when DevOps is established in the organization. The DevOps brings processes to application development and elevates the data generated by these applications causing a disruption in data. After this, we need processes to manage data and make them liable and secure and this framework is very useful for that. The DataOps implementation is more than useful, it is vital for any organization and as soon as we realize this, easier will be to lead to its implementation.

Q2: Do you have any critics of the proposed framework? Please explain.

The architecture of this framework is very complete, and it details exactly the different activities that any organization needs when work with data. To improve this framework, I agree with Interviewee #1 when he mentions it is important to clarify the activities related with Information Governance. The Information Governance is above all of operation and operation teams must implement the policies defined by Information Governance to achieve the final goal. I agree that you can change this activity to compliance, as it will make clearer what you intend to do.

Q3: Would you consider implementing the proposed framework? Please clarify why / why not?

Yes, all data teams will need to implement this framework in the future. We need establish collaboration processes across all teams to manage data correctly and with the help of this tool, that goal will be easily and securely achieved.

Q4: Do you have any recommendations or suggestions for further improvements of the proposed framework?

I agree with the adoption of this framework and methodology. I consider that Foundations Phase is fundamental to process definition and bring all areas onboard. So, if you manage to convince the Board of an Organization that this will make Data more secure and easily accessed, fo sure they will wish to implement it and they will see its value.

Interviewee #3: Date: 29/05/2023

Q1: Do you consider the proposed framework useful? why? If not, why do you believe it is not?

In the organization where I work, is still implementing DevOps and I consider that we need processes to manage data in any organization. Now a days, everything happens faster, and business areas have more requirements and want to be more involved in development process. This framework is very interesting and may help all areas, namely IT, Data, and business to establish processes and grow together which will make the Organization grow as well and increase its gains.

Q2: Do you have any critics of the proposed framework? Please explain.

This framework is very complete, and organizations must consider the implementation of DataOps processes. In terms of monitoring is very important receive and act in real-time to avoid mistakes for the users. So, my critic, in a positive way, would be, to be to have extra care with security of Data.

Q3: Would you consider implementing the proposed framework? Please clarify why / why not?

Yes, I would implement it. Nowadays is crucial. We work a lot in an old-fashioned way, and we need improve, establishing processes across all teams and grow together, making good use on new technologies.

Q4: Do you have any recommendations or suggestions for further improvements of the proposed framework?

Like Interviewee #1 and #2, I agree with the adoption of this framework and methodology and all organizations need DataOps implementation and this framework is very useful in that way to make Organizations achieve their goals.

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa