# GOVERN RELEVANT KEY PERFORMANCE INDICATORS FOR BUSINESS ALIGNMENT WHILE DEVELOPING DATA PRODUCTS

by

LEELA RAVI SHANKAR DHULIPALLA

DISSERTATION

Presented to the Swiss School of Business and Management Geneva

In Partial Fulfillment

Of the Requirements

For the Degree

DOCTOR OF BUSINESS ADMINISTRATION

SWISS SCHOOL OF BUSINESS AND MANAGEMENT GENEVA

JUNE, 2024

**GOVERN RELEVANT KEY PERFORMANCE INDICATORS FOR BUSINESS ALIGNMENT WHILE DEVELOPING DATA PRODUCTS**

by

LEELA RAVI SHANKAR DHULIPALLA

Supervised by

Dr. Anna Provodnikova

APPROVED BY

Dissertation chair

RECEIVED/APPROVED BY:

Admissions Director

## Dedication

This thesis report is dedicated to all the people who have supported me throughout my education.

ABSTRACT

**GOVERN RELEVANT KEY PERFORMANCE INDICATORS FOR BUSINESS ALIGNMENT WHILE DEVELOPING DATA PRODUCTS**

LEELA RAVI SHANKAR DHULIPALLA
2024

In today's world of data driven business landscape, data product development has become pivotal for all organizations for effectively generating insights and taking timely data driven decisions. Aligning business objectives and expectations versus the data product development is a challenge for any organization. The fields of Data Operations (DataOps) and Machine Learning Operations (MLOps) have paved paths for the accelerated delivery of data products by unlocking the potential of the data that is present in the organization data stores. While developments in the areas of DataOps and MLOps are providing new horizons to explore, simultaneously bringing more challenges to organizations. Organizations are increasingly relying on data powered data products for strategic decision making, the effective governance of Key Performance Indicators (KPIs) become pivotal, requiring attention towards measurable methods that can seamlessly align with the business objectives.

This paper presents the challenges that are being observed in the areas of DataOps and MLOps and the need for a measuring system that can measure the maturity of these systems. The research will enhance the understanding of governing KPIs to mature and

align the data products towards business maturity. With a measurable framework, organizations will be empowered to make decisions in an agile manner, thereby accelerating the developments of the data products that aid in decision making.

This study employed mixed method approach, it begins with a comprehensive literature review for understanding the need and significance of the KPIs that are aligned with data product development to achieve business goals. Subsequently, qualitative interviews were conducted with industry experts to understand the challenges in the current practices and also to select the KPIs that fit to the real-world problem statements.

The findings from the study revealed a multi-faceted problem where the organizations are facing diverse challenges in effectively utilizing the data insights through the data products. Throughout the data product life cycle, there are challenges at every stage and self-healing process are to be established to make the processes reliable and effective. This study in overall gives an insight into the complexities of data product development and offers actionable insights to optimize their data operations.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# CHAPTER 1 :
# INTRODUCTION

## 1.1 Introduction

The world recognizes the data as the most valuable resources and is often referred as the “new oil” (Manyika et al., 2011). In the dynamic landscape of modern data-driven business, the merging of data operations and machine learning operations has emerged as a major driving force by effectively utilizing their data repositories for data-driven decision-making (Marz and Warren, 2015). As Organizations are spending significant effort and money on the data product development for faster decision-making, the governance of the key performance indicators becomes very crucial and maturing in the areas of Data Operations (DataOps) and Machine Learning Operations (MLOps) has become a necessity than an option for ensuring a seamless alignment with business objective (Kaisler et al., 2013; Sivarajah et al., 2017).

Majority of the organizations have equipped themselves with the insights from the data by centralizing the data at enterprise level making it easily accessible across the company using cost effective strategies for storage and compute (Provost and Fawcett, 2013). Having centralized systems will help in reducing the redundancy, ensuring consistency and integrity across data applications. It will also help in following the same set of standards, technologies, optimized infra, reduces technical debt and is easy to maintain reducing the support costs as well. From personal experience, data science has taken a key role in improving the business decision-making in recent years by

understanding the underlying patterns in data using machine learning and advanced deep learning techniques. These machine learning techniques have shown promising results by enabling predictive analytics, helping business to foresee the market trends and demands with good accuracy leading to optimized efforts across all business units in the organizations. The recent advancements in areas of natural language processing and computer vision have brought artificial intelligence near to reality, the ChatGPT system is one breakthrough that is revolutionizing the industry in the recent time. Organizations are implementing these advanced systems into their day-to-day operations, example the AI powered chatbots and virtual agents are now handling the customer service tasks with human like efficiency. The data empowerment has facilitated businesses in improving products, services, marketing, and overall business strategy to meet the evolving demands of end customers (Davenport and Dyché, 2013). Moreover, this is enabling organizations to make data driven decisions and are becoming more agile and responsive to business environment. The integration of insights into business operations is achieving competitive advantage in the current marketplace.

Organizations are leveraging the new advancements in the areas of data engineering and data science to gain a competitive advantage by unravelling the hidden opportunities and optimizing the operations. Organizations with centrally organized data will enable real time data analytics and support data-based products to provide immediate insights that drive agile decision making and strategic pivots. Integration of AI based tools like ChatGPT into the various domains of the organization like support operations has improved the customer interactions, providing optimized performance, reducing the

response time and also provide rich personalized experiences. Adoption of these AI tools and technologies is not only confined to large organizations, but the small and medium scale organizations are also reaping the benefits of the AI based systems.

The advancements and convergence in the areas of big data, cloud computing and AI has democratized the access to powerful analytical tools allowing organizations to operate with unprecedent foresight. Democratization means that even the small and medium scale organizations are leveraging the advanced technologies to compete with larger organizations by being more creative and innovative. Data science teams are now closely working with the various internal teams to collaborate and demonstrating the data driven capabilities in each of their domains. This close collaboration is cultivating a culture of innovation and continuous improvement across business units in the organizations. Advanced models and analytics are being used to do forecasting, risk assessments, provide personalization and others to maximize the return on investments by the organizations. The machine learning models are empowering the organizations to have proactive market strategies for identifying potential business opportunities and threats in the business operations. Organizations are empowering themselves to have real time analytics to generate actionable insights almost in real time. The machine learning capabilities can help in refining the business process, enable predictive maintenance and improve customer experience, the synergy between the technologies is helping towards having a more strategic approach towards business growth and innovation irrespective of the organization industries.

Industries such as Healthcare, Ecommerce, Fincare and retail are seeing transformative changes with the help of AI to analyze complex data systems and have actionable insights. In healthcare AI is advancing in the areas of manufacturing, quality testing, supply chain and also providing personalized recommendations. In the areas of ecommerce, AI is progressing towards providing personalized recommendations and also influencing the customer buying patterns. In the areas of Fincare, AI is advancing in the areas of risk management, fraud detection and portfolio management. Advancements in the areas of AI and machine learning technologies is promising to unlock more sophisticated applications which will drive innovation and efficiencies across all sectors in the industry in the upcoming days.

Amidst the recent developments and advancements in the areas of data engineering and data science, the business are struggling with the challenges related to the streamlining and productionalization of data pipelines and data science applications (Gandomi and Haider, 2015; Marz and Warren, 2015). On a day-to-day basis, it is a normal scenario that the data pipelines failing due to various reasons varying from platform related issues, data related and manual errors. Productionalization of a data pipeline from its source to a decision-making application is not in line with the business expectations of it being delivered in time which is causing the data to be outdated and ineffective in decision-making for business even after spending significant amount of money on these systems. A study conducted in 2018 at Gartner (2018) highlights that business on average incur losses of $15 million due to bad-quality data. The pressing question to the organization leadership teams from experience is how to reduce these losses attributed to bad-quality data.

Development and Operations (DevOps) is a collaborative and multidisciplinary effort to automate continuous delivery of new software updates while guaranteeing their correctness and reliability (Leite et al., 2019). Based on personal experience, implementation of DevOps into any organization has its own challenges, it requires practitioners who has sound knowledge in architecting systems in continuous delivery, assessing existing systems and how to make it adaptable across projects. Successful implementation of DevOps principles can help teams in release software very frequently in production-like-environment and business can validate the changes before moving into production in short cycles effectively. The DevOps principles may not entirely apply to data science problems because of its nature of being data driven and developed applications guided by algorithms that require continuous evaluation metrics and retraining pipelines which are not seen in regular software applications. Also, the data applications are prone to data drift, concept drift and model drift which is not usual in software applications so applicability DevOps principles require a different approach in the fast-moving data landscapes.

Data and Operations (DataOps) is the new discipline that emerged in the recent years that combines an integrated and process-oriented perspective on data pipelines with automation and methods from agile software engineering to improve quality, speed, and collaboration and promote a culture of continuous improvement. Agility can be brought with an ability to react to volatile environments regarding the functionality or the content of the data products. It is not a particular method or tool, but it’s rather a collection of

principles and way of doing things on a cultural, organizational and technological level (Ereth, 2018).

Similar to DataOps we have Machine Learning and Operations (MLOps) principles that can be used for continuous delivery of machine learning models. Compared to DataOps or DevOps, MLOps requires different datasets used for training model and their versioning, model versioning, monitoring of the model to detect bias and drift problems (Granlund et al., 2021).

As organizations are aimed at achieving a state of maturity, addressing these challenging concerns becomes necessary for the data leaders to eliminate the dilemma of maintaining the data quality against the operational challenges. This research is aimed at exploring the areas of having governance on the Key Performance Indicators (KPIs) in measuring the maturity of DataOps and MLOps practices. As part of it the research is focused to explore the strategies and building metrics to measure the maturity of the practices and frameworks in the areas of data operations and machine learning operations. This research studies about the challenges the organizations are facing in implementing these Key Performance Indicators in measuring the maturity of the DataOps and MLOps practices. Establishing a mechanism and understanding the benefits of a successful DataOps and MLOps practices in any data critical companies is crucial for delivering quality and reliable applications. The findings from this study will help organizations to enhance their maturity in the areas of data operations and machine learning operations and have a competitive edge in strategizing the business operations.

## 1.2 Research Problem

Organizations are increasingly relying on advanced analytics and machine learning which are fueled by data, having an effective governance of Key Performance Indicators (KPIs) has become an important challenge. The absence of such a governance framework will not allow companies to scale themselves in achieving their business objectives in this fast-evolving data world. The lack of matured systems will lead to sub-optimal decisions causing operational inefficiencies, financial losses and re-design their systems again and again. The importance of the maturity frameworks has been recognized by the industry.

The efficient convergence of the DataOps and MLOps practices will help in effective data management with operationalization of ML models. Organizations are aware of the importance of these practices, however, integrating these practices into their workflows to improve the agility of data product development is of utmost importance in the current time. Maturity around these practices not only helps in achieving technical efficiency but also improves the business strategic alignment.

Major challenges arise in both technical and non-technical aspects during the implementation of the DataOps and MLOps principles during the data product development. Understanding these challenges and building best practices around them, improvising them, and continuously measuring them is an important aspect for any organization. Setting up a path through best practices and principles to achieve a state of maturity is what needs to be done and it is a journey and not an instant state that can be achieved.

DevOps over the period has shown some decent success in delivering software products following agile practices. This success has led to improvement in terms of development, deployment and operations of a data product development with increased efficiencies. Taking inspiration from these DevOps practices and adapting them to DataOps and MLOps practices can be an area that can provide an approach to mature the data product development. The integration of continuous integration and continuous delivery pipelines can enhance workflows and increase productivity. Identifying the key components of the DataOps and MLOps systems will be critical, including activities like data ingestion, data transformation, data processing. model training and model deployments. Establishing a metric for assessing the maturity of these components through measurable KPIs will be an important area of study for anyone to achieve an optimal state for data product development. By leveraging these metrics organization can measure their performance on the different critical areas of application development, ensure they are aligned with the organization goals, track progress and identify areas of improvement. It inculcates the culture of collaboration and continuous learning in the areas of DataOps and MLOps to drive innovation and operational excellence for every organization.

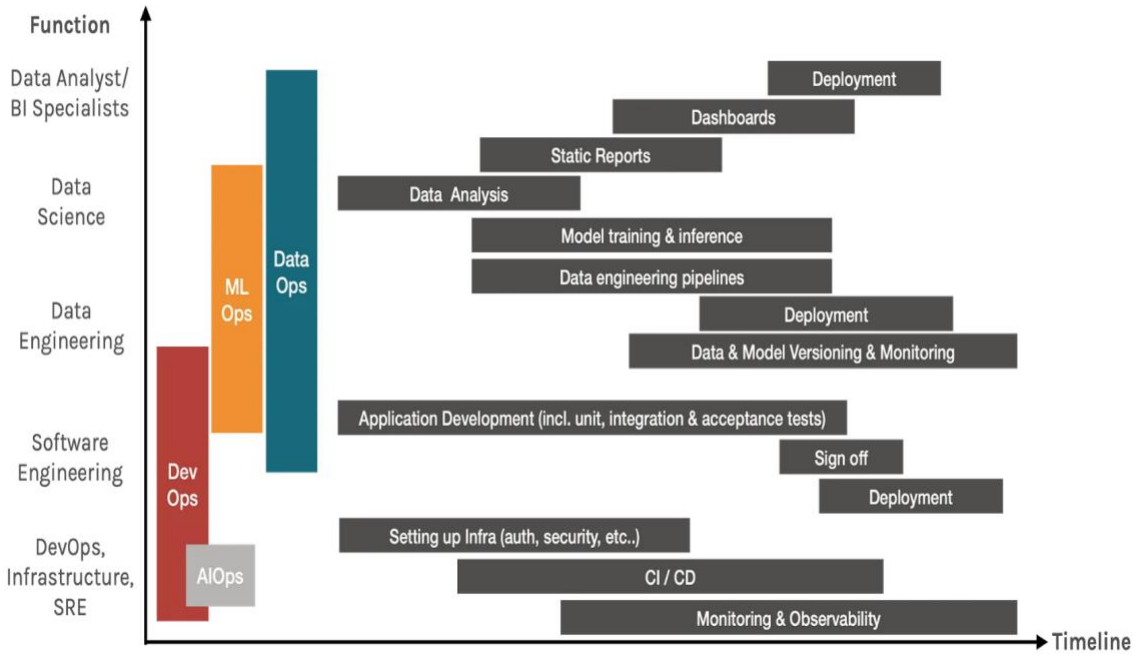


Figure 1.1 Graph of the tasks that each job function performs across the timeline (Source: Merelda Wu, 2021)

Integration of DataOps, MLOps, DevOps and AIOps practices in data centric organizations has led to the emergence of multiple job functions like data engineers, ml engineers, and devops engineers. These roles are responsible for development and maintenance of the data products which has challenges on its own in maintaining data quality, data pipelines and infrastructure. Fig 1.1 has the details of the various job functions and the associated tasks linked with the each job function role. DevOps involves people specialized in the roles of software engineers with experience in deployments, infrastructure, reliability and data engineering. Ensuring smooth coordination between the development and operations without any challenges related to deployments, infrastructure scaling and maintenance of the systems is a critical task and require a measurable framework to identify the gaps in the processes and address those challenges.

DataOps works includes data analysis, data integration and transformations. The primary challenge for the dataops team is to maintain data quality and consistency, complex data workflows and dependencies. Managing data version control and data privacy regulations and continuously monitir the data pipelines for the smooth operations are critical activities for DataOps team.

MLOps consists of data scientists, ml engineers, software and devOps engineers. The prinary challenge with the MLOps teams is to track the different versions of the models and their corresponding data used for training of the model, having the ability to re-produce the machine learning models, testing of the models, mitigating the biases within the models, detecting the degradation in the models and retraining of the models as and when required. Ensuring these activities are tracked on regular basis and necessary processes are established to ensure there are no failures or deviations which are critical for any organization to achieve the state of success. Data is core component required for development and building of machine learnining models. Data needs to be ensured to be accurate and with quality to ensure the machine learning models work as per the business needs. Inaccurate data will result in low accurate models and can cause the ineffective predictions which can result in wrong decision making and causing challenges and negative impact to business operations (Mylavarapu et al., 2019).

The key challenges for AIOps include integrating diverse data sources, accurately identifying the anomalies in the data, having automated capabilities, ensuring the scalability of the applications and maintaining compliance with the IT security and privacy regulations. With all the job functions, ensuring a collective collaboration between the

various teams and having processes and metrics established to track and optimize the processes and methods is important for any organization to ensure efficient and reliable operations.

### 1.3 Purpose of Research

The purpose of the research is to understand the industry challenges in the development of Data & AIML products and support the organizations in reducing the risk of not meeting their strategic goals because of the delays. Data products development are complicated as there are multiple stakeholders, platform teams, enterprise teams, data level complexities and complex business process. Data required for the data products can exist in various sources, extratcion and transformation of this data requires data engineering and data modeling capabilities to rightly organize the data as per the business reporting process. Understanding the business process and ensuring the data is aligned as per the business process is a complicated activity in my opinion and without clear guidelines on how the data needs to be stored, processed and utilized will result in a complicated can result in lot of reworks causing financial impacts to the organization and also losing critical time for the development of the data products. Developing solutions in this complex landscape is tough and would require a deeper undertsanding on the organization complexities and the importance of the strategic goals. A measuring framework that can help in organizations that can track and take course corrective decisions is required to help and track the progress of the development process. This research is targeted in understanding the challenges and help organizations with Key Performance Indicators that can support them for accelerated product development. This research will support the organizations in maturing themselves

with the correct appropriate trackers and metrics to ensure their progress is inline with the organization goals.

### 1.4 Significance of the Study

Starting from the way the data is extracted from source systems, establishing framework around extraction, automating this framework for accelerated development and deployment, an observability framework that can provide a realistic and futuristic state for effective monitoring, building metrics to measure around these components is critical for stable and efficient data products. Irene O'Callaghan et.al (2024) in their research about key performance indicators in “KPIs for Quality and Availability of Data in an Industrial Setting” stressed the importance of developing a framework capable of adapting to changing data sources and logging formats. They advocate for automated feature extraction methods to minimize dependencies on specific log sources and broaden the framework's applicability. This adaptability is crucial in maintaining the framework's relevance and effectiveness as data environments evolve. Absence of standardized governance mechanism for these KPIs can pose a significant roadblock for effective alignments of the strategic alignment of the data products (Marz and Warren, 2015; Sivarajah et al., 2017). So, employing effective governance around these metrics and baselining these metrics will be required for navigating and stabilizing the complex data-centric developments.

Implementation of robust data lineage systems and having systems that can have transparency and traceability of the data pipelines is crucial for organizations to detect any discrepancies or quality issues that may arise during the day to day operations. Integrating

advanced analytics and machine learning models into the business process will enhance the decision making process.

By focusing on the governance of the Ky Performance Indicators for business alignment for development of data products, this research helps in strategizing the data product development by making strategic decisions and adopting risk management practices to ensure the strategic goals are met in time.

### 1.5 Research Purpose and Questions

The purpose of this study is to understand the relevant Key Performance Indicators (KPIs) that effectively govern the alignment of business objectives with the development of data products. By understanding the data product development process and its challenges, the research aims to improve the understanding on how organizations can mesaure and mature their data product developments to achieve their strategic goals. There is a need for a better understanding the constraints and providing the governance to ensure the organizations can scale for scalable and manageable data products. This research aims in addressing the following questions

1. How can the DataOps and MLOps principles be effectively governed using Key Performance Indicators (KPIs) for the successful and seamless delivery of quality data products?
2. What are the diverse technical and non-technical challenges involved and the current practices followed for implementation of DataOps and MLOps principles in developing data products?

3. What best practices that are being currently used for successful implementation of DevOps can be adapted to DataOps and MLOps?
4. What can be the baseline that can be defined for various components involved in DataOps and MLOps pipelines that can measure the maturity of the implementation of these principles?

# CHAPTER 2 :

# REVIEW OF LITERATURE

## 2.1 Introduction

Organizations require fast and accurate analytics to be able to compete in the evolving markets. The ability to quickly generate insights from data is crucial for any organization to make informed decisions and be competitive in the current world. Almost every company has invested in their data engineering and analytical teams for building data products and it would be a setback for the companies if these teams are not properly aligned in delivering reliable results. Misalignment can lead to inconsistencies, errors thereby resulting in wrong decision making and will create a negative impact to business financially and reputation wise resulting in losses to organizations. Based on my experience, the high-level challenges that has been faced by the Organizations and how the decision making is delayed between team is shown in Figure 2.1. The challenges generally stem from silos between the teams, lack of collaboration, ineffective communications, different priorities and agendas, lack of processes and ineffective usage of tools and techniques that can help in addressing the challenges. To deliver value to a company it requires different functional groups to be collaboratively working towards the implementation of DataOps principles. This collaborative effort ensures the data is managed correctly and insights are generated efficiently. Establishing standardized processes and using integrated platforms can improve the collaboration between the cross functional teams and the streamline the workflows for smoother operations. Establishing standard processes for every line of

operations can remove the confusions and provide a clear understanding and accountability of the processes and reducing the conflicts on ownership and deliverables. DataOps principles can help in ensuring the data quality standards are maintained on a day-to-day process and ensure faster analytics and decision making thereby driving the business success of the organization.

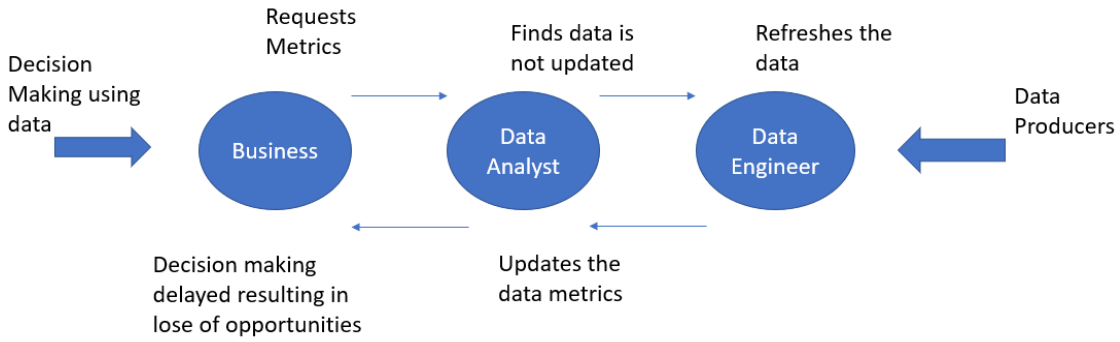


Figure 2.1 How delay happens in Decision making.

Initial literature review shows the evolution of DataOps and MLOps to address the challenges in building and delivering value of the data products like predictive models or prescriptive models that help in generating revenue, mitigating risks, improve compliance etc. These DataOps and MLOps methodologies are targeted to enhance the efficiency and effectiveness of data management, machine learning model management and machine learning workflows ensuring data-driven insights are regular, on time and accurate helping business to mature. The DevOps principles of automation, continuous integration, and continuous deployment are critical and helpful in reducing the time-to-market for software products and has delivered real time benefits for every organization. DevOps which evolved for streamlining software engineering to a continuous delivery model has gain a

lot of attention and companies adopted it to full extent based on the success it has generated. Ineta Bucena et.al (2017) in their study “Simplifying the DevOps Adoption Process”, focuses on the evolution of DevOps and its successful integration into software development processs and the challenges business had in adapting into the culture of the teams. It has attracted software developers, managers, stakeholders and experts from the domain to understand how it has fast paced the delivery activities across the areas of software development. Similar to DevOps, drawing parallels of DataOps and MLOps integration into the organizational workflows of product development requires cultural and proceedural changes. Companies should also inculcate the habits of environment of collaboration, continuous learning and innovation to realize the benefits of these methodologies. Alignment of Dataops and MLOps principles with busines objectives ensure the data products are not only technically viable but also strategically valuable for the business. As organizations are navigating the complexities of modern landscape the insights of DevOps can provide a valuable roadmap for the successful implementation of DataOps and MLOps.

## 2.2 Exploration of DevOps Practices: Insights from Key Studies

There has been valuable researches conducted on DevOps practices which describes the various challenges, drivers for adopting DevOps, engineering capabilities, technical enablers that can help in moving software developed to a production environment using agile practices (Bucena and Kirikova, 2017; Senapathi et al., 2019). These studies have higlighted the importance of DevOps and how organizations are leveraging the benefits of DevOps to enhance the delivery of software products, highlighting the

significant benefits of these approaches and potential pitfalls associated with them. Based on Huttermann (2012) in his study “DevOps for Developers”, the goals of operations team and the development team are not the same and sometimes they are opposite to each other and it can be noticed in Figure 2.2. Bring collaboration between these teams, shared ways of working and aligning to the goals require a lot of effort from the business to address the challenges.

Researchers have highlighted the technical and non-technical challenges that arise during the implementation of DevOps principles. Technical challenges include the integration of various tools and technolgies to implement the DevOps practices for faster integation testing and deployments, automating workflows involved at different stages of the system, and ensuring the reliability and security of the deployment processes and various other steps involved in the data product development. Non technical challenges includes the process gaps, cultural resistance, lack of collaborations between the teams, different priorities and expectations and misalignment of objectives among the teams. Bucena et.al (2017) in their research “Simplifying the DevOps Adoption Process” mentions the top three common challenges related to DevOps practices and which were the challenges that I also had noticed in my professional experience:

1. Missing the definition of maturity of the concept
2. Lack of awareness
3. Lack of coordination between the teams

Power et.al (2014) in their research “Impediments to Flow: Rethinking the Lean Concept of 'Waste' in Modern Software Development” have proposed a “Nine Impediment

Categories" framework which can be used to identify necessary DevOps practices a company could use for DevOps adoption to achieve a desired maturity level. The framework discussed about the nine key impediments that hinder the successful adoption of the DevOps practices in any organization. These impediments include organization silos, inadequate tooling, lack of process automation, incorrect tooling, insufficient skills and resistance towards cultural changes.

Power et.al (2014) in their research mentions to identify the nine impediments of the software development and prioritize them based on a survey questionnaire. This survey based approach will help the organization to understand the impediments and critical challenges that are involved in moving towards the transition of DevOps practices. Once the organizations identify the impediments then the organizations can follow a strctured approach in ensuring each of the impediment is priortized and resolved and move towards an establishment of DevOps maturity model. This model helps in setting a structured way to assess the current state of the organization and have clear goals for improvement. Selecting and setting the DevOps maturity goals involves in identifying and setting up realistic and measurable targets that align with organization strategic objectives. This process requires continuous monitoring and adjustment to ensure that the DevOps practices evolve in response to changing the business needs and technological advancements of the organization. By addressing both technical and non-technical challenges organization can have a more cultural, collaborative, agile and efficient software delivery process. Bucena and Kirikova ( 2017), Power and Conboy (2014) in their research mentions business can understand and navigate the complexities involved in Devops practices and work towards

establishment of a DevOps maturity model and select the desired maturity level to be achieved in a shorter period and in longer periods.

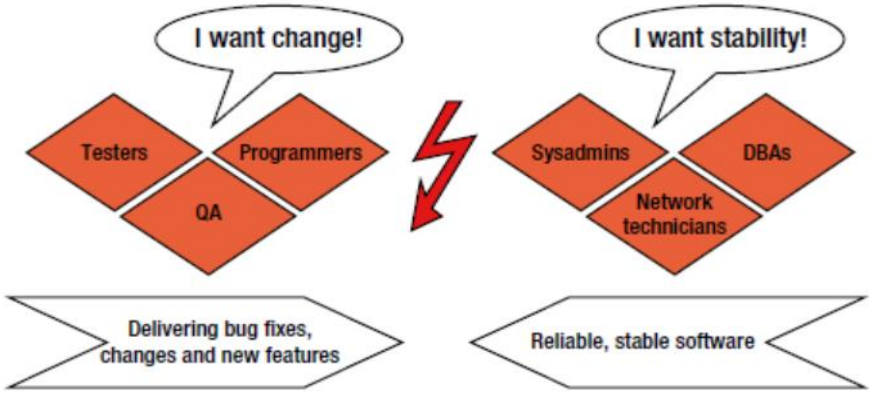


Figure 2.2 DevOps strives to improve collaboration between development and operations (Source: Michael Huttermann, 2012, p. 20)

Delays in detecting the downtime of the data issues can lead to financial loss, and also qualitative loss such as delayed, incorrect and redudant decisions could lead to reputation damage and slowness in repoting and compliance (Barr Moses, 2024). Data downtimes if not addressed promptly within the stipulated time can result in significant operational disruptions and loss of customer trust and business. Fig 2.3 displays how the time to detect the issues can lead to impact illustrating the correlation between the delays and severity of their consequences.

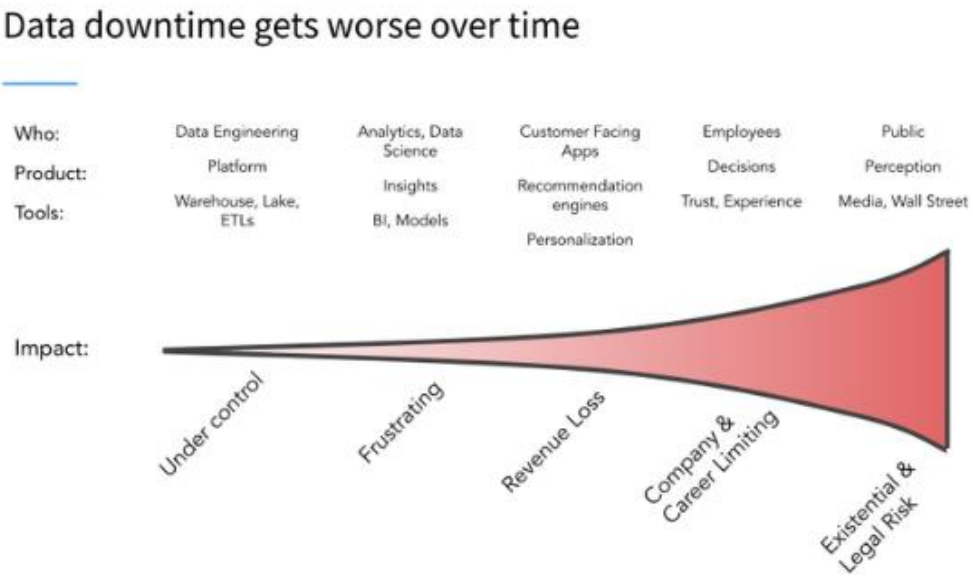


Figure 2.3 Data downtime gets worse over time (Source: Barr Moses, 2024)

Early detection of data issues and taking necessary steps in resolving these issues are crucial for maintaining data integrity and ensuring timely, accurate decision making for the business. Prolonged downtimes or not resolving the errors can lead into a state of piling up large volume of data issues and can be a state where the data would needs to reprocessed from the beginning due to not able to resolve the issue due to multiple causes. From personal experience, some of the data pipelines have not executed for some months and at the time when the data was required these pipelines where not operational and products and teams which were relying on this data had been impacted as the data was stale. Organizations would end up spending money and effort for these pipelines to re-execute and with an effect that these cannot be used imediately loosing advantage to business. For

institutes like financial organizations, delayed detection of data anamolies can result in incorrect decision making of trades, identifying frauds causing reputation and regulatory non-compliance risks and considerable financial penalties. In healthcare industries it can result in delayed drug discovery, incorrect diagrnosis of patient diseases, delayed treatment and compromising patient care and safety.

Implementing robust monitoring and alerting mechanisms for DataOps and MLOps practices can mitigate the risks associated with the data issues. Building observability platforms that can track and monitor the data quality issues, downtime issues, job failures, job executions, data freshness, infrastructure stability can help in promptly identifying the issues and resolve them before they impact the organizations. Establishing practices for day to day monitoring and improvement further helps in minimizing the downtime and maintaining high data quality standards.

Power et.al (2014) in their research “Impediments to Flow: Rethinking the Lean Concept of 'Waste' in Modern Software Development” mentions to study the list of existing tools available in the organization and adopt the DevOps principles in phased manner which can be applied across projects. This phase wise approach will be a decent applicable approach if it can be ensured and rightly collaborated with the various teams. Cross functional collaboration between the data science, data engineering and operations teams can enhance effectiveness of these practices thereby improving the quality of the data products. The implementation of data quality frameworks that can assess and check the data quality issues within the data can play a critical role in mitigating the risks associated with the data down time. Fig 2.3 shows and serves as a critical reminder on the importance

of these data quality issue mitigations and how un attended could lead to an impact to business operations and cause financial and reputation losses.

The research “KPIs for Quality and Availability of Data in an Industrial Setting” (Irene O’Callaghan, Andriy Hryshchenko, 2024) has discussed about the critical role of data quality and availability in data completion exercises. They suggest that organizations should focus on developing KPIs that reflect the unique characteristics and needs of their specific industrial scenarios. Tailored approach can help organizations prioritize their efforts and resources more effectively to achieve the required organizational goals. One of the key proposals of the study is the importance of considering feature utility coefficients in determining the success of data completion exercises. The authors illustrate how this method can help organizations in prioritizing features based on the relative importance of the data features thereby optimizing their data collection strategies.

### 2.3 Advancing from DevOps to DataOps and MLOps

Similar to DevOps, DataOps and MLOps are the domains which primarily focused on Data Science and Data Analytic processes which typically deals with data pipelines. Adopting these principles will definitely help the Data systems to resolve some of the long-standing issues that are causing delay in delivery of production grade applications. DataOps has gained significant attention with many organizations recognizing the potential to accelerate the production of high quality data insights. Real time monitoring of the data pipelines will help in improving the quality of the data along with reduction in the number of issues, thereby improving the quality of the data products. DataOps Manifesto (2021) has published 18 principles which can be used as a reference for implementing DataOps in

the organization. However, there are quite number of ambiguities like DataOps is usage of set of tools, expensive methodology, can only use on data analytics etc., in DataOps practices leads to certain challenges in its maturity and its rapid adoption in the organizations (Mainali et al., 2021). Some of the common challenges like continuous change in the requirements, data being not structured, unavoidable errors in the collection of the data, unavoidable manual intervention in the data collection process are creating additional pressure in maintaining the data quality and implementation of DataOps principles (Rodriguez et al., 2020). Also being an evolving field, this has caused additional challenges in its adoption as the definitions and usage are changing continuously.

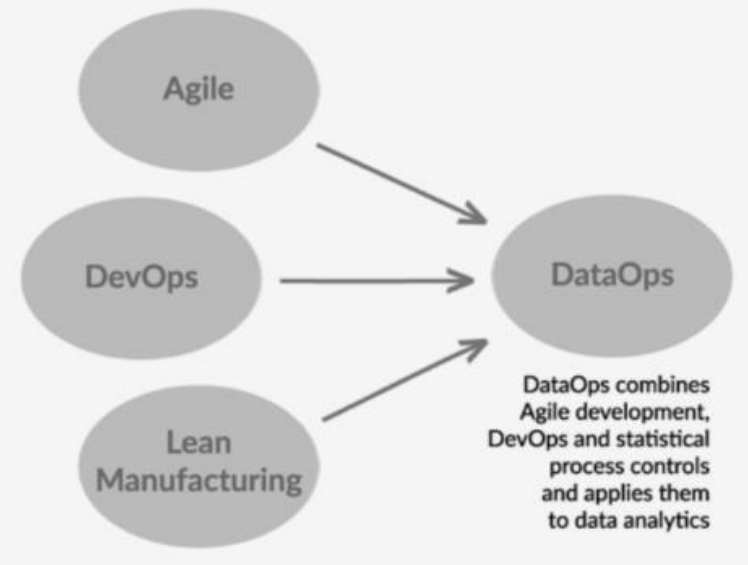


Figure 2.4 The intellectual heritage of DataOps (Source: DataOps Manifesto, 2021, p.18)

Data Observability platforms help in improving the real time monitoring of the data pipelines and timely detection of anomalies and failures. In the research “Real-Time

Monitoring of Data Pipelines: Exploring and Experimentally Proving that the Continuous Monitoring in Data Pipelines Reduces Cost and Elevates Quality" conducted by the Narayanan et al. (2024) has highlighted the importance of the data observability and continuous monitoring pipelines for effective improvement of the data pipelines and thereby improving the quality of the data itself. The research emphasizes the identification and early detection of issues using various practices related to real time monitoring which improves the data integrity, enhanced operational efficiency and increased trust in the organization. The study also discusses about how costs can be reduced with real time monitoring by minimizing the down time and mitigating the impacts on the quality. The researcher has focused on the improving the data quality and integrity of the systems but critical aspects such as security and compliance and also does not discuss about the scale of real-time monitoring of solutions. Real time data ingestion systems carry their own level of complexity due to their frequent executions and any breakage at any one pipeline could potentially result in data integrity challenges. So, organizations should be careful and have effective strategies for ensuring the observability systems are available to track the real time data pipelines. Data Quality, Anomaly detection rate, System Performance, Data Lineage and Operational Efficiency are the metrics that the research has discussed for the success of the DataOps and MLOps practices.

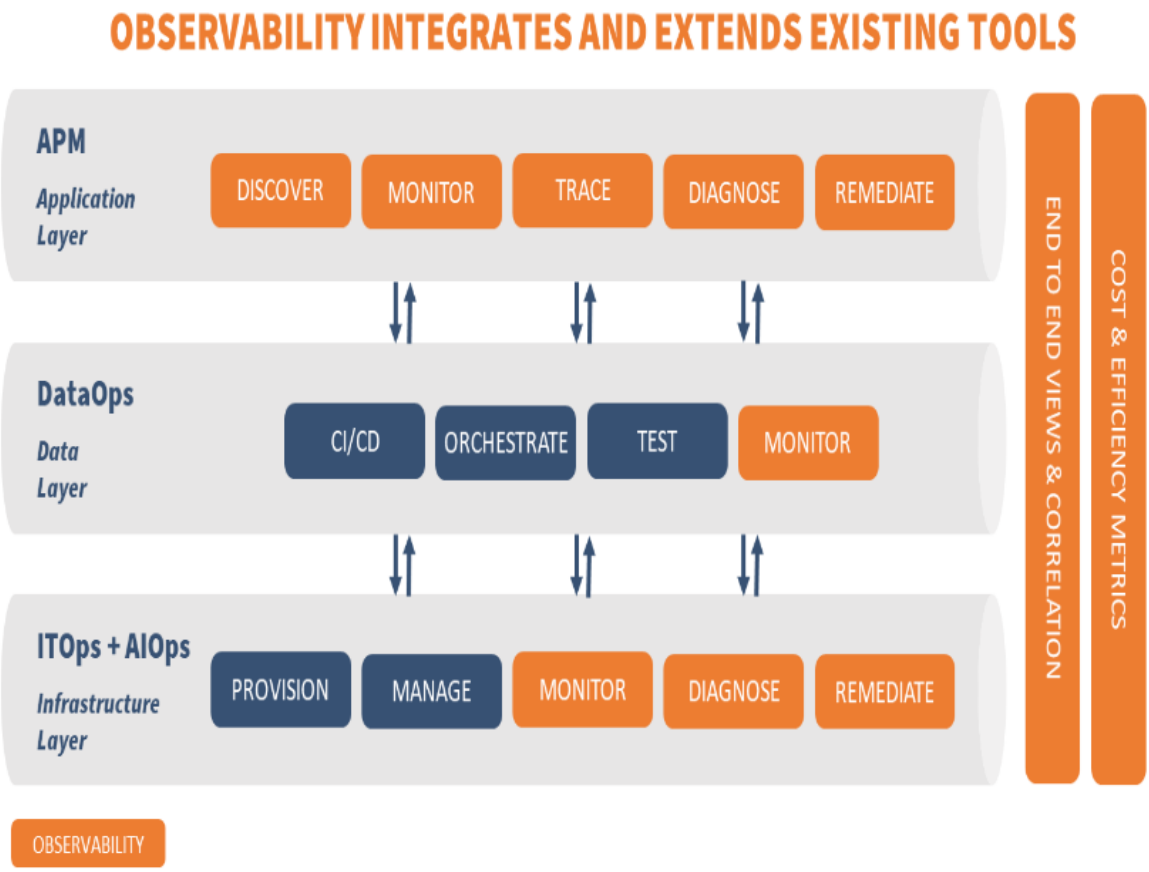


Figure 2.5 Data Pipelines integration with Observability

(Source: Narayanan et al., 2024, p. 4)

MLOps is an amalgamation of ML and Operations which refers to advocation and monitoring of all steps of ML development and deployment. There are various stages that are present starting from training, testing and cross validation tests. Each stage of it has its own complexity and each stage has to be strictly monitored and evaluated for any data leakages, inefficiencies in hyper parameter tunings etc. as suggested as antipatterns in MLOps (Muralidhar et al., 2021). Once the models are developed, they need to be productionalized to generate insights or predictions for decision making but due to the

complex nature of the ML model development and deployment. It has been surveyed that there are 87% of the models do not make into production (VB Staff, 2019). Algorithmia's report says the majority of the companies take 8 to 90 days to deploy a single model into production (2020 state of enterprise machine learning, 2019). Even though after having sophisticated infrastructure and MLOps principles laid it has become always a challenge to the business to fully utilize the benefits of these models due to challenges in implementation of these principles and also due to lack of clarity on the maturity levels of these principles. Also, in one of the research the researchers have compared various tools and it has been said that no single tool has the capability of realizing a fully automated MLOps workflow and different tools have different overlapping features increasing redundancy (Ruf et al., 2021).

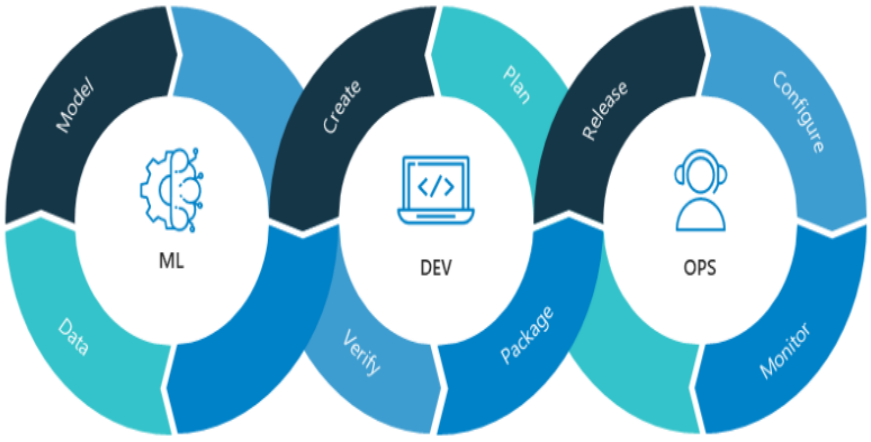


Figure 2.6 MLOps Lifecycle (Source: Murali, 2021)

Continuous delivery, test quality framework, monitoring are the primary component that has been defined. Not many sources have details on how these are to be implemented in various challenging scenarios and how business can measure the success of their DataOps teams (Ereth, 2018; Munappy et al., 2020; Rodriguez et al., 2020). Once when the DataOps team is set, how the DataOps team should navigate in the realm of various teams and what would be the factors that should be considered while prioritization is minimally known based on my experience. Best practices as part of DataOps has been explained in the researches, the practical benefits and what should be the output and measurement at each stage of these DataOps has been little studied (Ereth, 2018; Munappy et al., 2020; Rodriguez et al., 2020). So, in this research, a study of various approaches that the organization are following in implementing DataOps and MLOps will be studied and will also study on the metrics that can be measured for understanding the maturity of the DataOps and MLOps.

### 2.4 Harnessing Data Engineering and AI Engineering for Organizational Advancement

Enterprises are adopting Data Engineering and AI Engineering practices to build and manage decision making systems to outperform their peers and competitors. Data and Analytics capabilities within the Organization is going to be the key for anyone to accelerate in their respective areas. Organizations are focusing on building data economy in their organizations by gathering and organizing the data to drive value from the data. From a study done by Statista, the data creation is projected to grow more than 180 zettabytes by 2025 compared to 64.2 zettabytes in 2020 (Statista.com, 2022). Though there is a vast amount of data being captured it is very difficult to find patterns, information that

is present in the DNA of the data that helps in decision making for organizations. AI/ML applications are helping in this regard in analyzing these huge volumes of big data using various statistical techniques and algorithms to identify the patterns. However, many organizations are facing challenges having a comprehensive data management design, deployment of AI/ML applications in production and consume their benefits in real time.

Data in the organizations are vast and would be sourced from various sources having different formats, making the data heterogeneous and complex to process. Organizing these huge volumes of data and making these data available as a service for various consumptions needs is a huge challenge. The data needs to be managed effectively for faster decision making, but in majority of the cases it has been noticed that data being in silos and the teams spend a lot of time in connecting with teams to check the availability of the data and sorting the issues related to cross team data sharing policies.

Multiple researches have emphasized the necessity of evolving traditional approaches to handle large volume and complexity of data in organizations (Kai Hartzell, 2023; Irene O'Callaghan, Andriy Hryshchenko, 2024). Hartzell (2023) in his research "Comparison of Big Data SQL Engines in the Cloud" focused on the significance of Big Data SQL engines in modern organizations, exploring various methodologies for processing large-scale datasets using technologies like Hadoop, Spark, Presto, and Trino. These engines are vital for businesses that need to efficiently manage and analyze vast amounts of data. The authors highlight the need for effective and efficient data processing frameworks that are capable of handling complex queries and providing real-time insights for organizations. Traditional relational databases have limitations in managing big data,

making it necessary to adopt columnar storage and advanced query optimization techniques to enhance performance and scalability.

Data Fabric is a system that addresses this challenge by providing a unifying architecture for management and provisioning of data (Östberg et al., 2022). Data Fabric systems facilitate the integration of various systems across the organization boundaries with properties to scale, easy integration, distributed storage, and support for interfaces that can offer self-service to the end users. In total, Data Fabric is an information management platform that helps in data management and supports data integration and has API's and interface that do data communication with the source systems and support applications to consume for generation of insights.

Alvord et al. (2020) in his research "Big Data Fabric Architecture: How Big Data and Data Management Frameworks Converge to Bring a New Generation of Competitive Advantage for Enterprises", it was mentioned that the success of the Data Fabric architecture can be achieved through a combination of technical and non-technical factors like good data management and clear data strategies. However, implementing and development of these Enterprise Architectures (EA) are not an easy task, and organizations need to be aware of the critical success factors to reduce the risk of their failures. In systematic literature based research done by Ansyori (2018) in "A systematic literature review: Critical Success Factors to Implement Enterprise Architecture", the commonly used frameworks for Enterprise Architecture implementation are The Open Group Architecture Framework (TOGAF) and US Federal Enterprise Architecture (FEAF) (Ansyori et al., 2018). Despite of having many Enterprise Architecture Frameworks,

comprehensive guidelines, it has been difficult in implementing them due to inflexibility and complexity of the business and IT structures. It was revealed that 66% of the EA programs did not fulfil the expectations based on the study "Why Two Thirds of Enterprise Architecture Projects Fail: An Explanation for The Limited Success of Architecture Projects" (Roeleven, 2010). One of the key reasons for this failure is due to the lack of coodination between business strategy and IT architecture. The study also discussed the enterprise architecture should be guided by vision, strategy and objectives setting clear expectations within the organization. The most common reasons for disappointing EA results include difficulties in connecting EA to business elements, lack of support from C-level executives, limited commitment from interested parties, and financial and political issues that hinder EA projects. To ensure the success of EA projects, the author suggests three key principles: setting clear enterprise-wide EA objectives before starting a project, establishing EA governance, and involving the business in EA initiatives. In the whitepaper "Why Two Thirds of Enterprise Architecture Projects Fail: An Explanation for The Limited Success of Architecture Projects" published by Roeleven (2010) underscores the significance of involving the business in EA initiatives and establishing clear objectives. It also highlights the need for effective governance to ensure the success of EA projects. The authors views emphasize the importance of aligning EA with business strategy, making it a holistic, business-driven discipline. Gartner mentioned that the most consistent pattern for digital business is by focusing on technology enabled models and less by on independent technologies.

Irene O'Callaghan et al. (2024) in their research "KPIs for Quality and Availability of Data in an Industrial Setting", have delved into three key metrics that serve as Key Performance Indicators (KPIs) for evaluating the success of data transformation in industrial settings. The research has proposed the use of feature centric, usage based and data-centric metrics to measure the quality and availability of the data. These metrics are supposed to provide a comprehensive framework for organizations to assess and measure the data collection efforts of the organization effectively.

The authors also highlighted the need for considering context based utility coefficients of features in determining the relative importance of a feature among other features. This approach has helped to understand the significance of a particular feature is appropriately weighted according to its utility in a specific context. The research also discussed about the need for normalizing the utility coefficients to ensure and maintain a balanced evaluation system.

In evaluation of these metrics, the study demonstrates how different industrial sites can have varying levels of success in meeting data collection targets. For example, the results reveal that site A has achieved a high level of feature completeness, while site C struggles with no features having complete descriptions. This disparity underscores the variability in data collection success across different sites and highlights the need for tailored approaches giving the organizations a view of how the teams and systems are positioned in terms of their maturity levels.

Roeleven (2010) in his whitepaper "Why Two Thirds of Enterprise Architecture Projects Fail: An Explanation for The Limited Success of Architecture Projects" provided

valuable insights into the drivers, roles, and challenges of EA initiatives. The key metrics for the success of DataOps and MLOps include setting clear enterprise-wide objectives, establishing effective governance, involving the business in EA initiatives, and choosing the right EA tool. By focusing on these metrics, organizations can better align their data strategy with business goals and ensure the success of their EA projects. Future research should continue to explore the factors influencing EA success and develop strategies to address the challenges identified.

### 2.5 Streamlining Big Data with Agile DevOps

Demonstrating the capabilities with the big data has been difficult due to its complexity involved, compute required, cost involved and data quality factors. The effort involved in setting up and maintaining complex data sources is frequently increasing and it is limiting its usability and requires sophisticated computing tools to perform any analysis. The development associated in developing high processing data pipelines requires coordination between the various teams starting from the data sourcing team to data consumption team. An agile process needs to be followed for timely delivery of the data to ensure Organizations can get the benefits and insights of the data. DevOps process helps in this regard in bringing coordination between the cross functional teams and bring automation for the speedy development and deployment of the data products.

DevOps which is an agile movement that advocates continuous small developments releases with continuous end user reviews (Bou Ghantous et al., 2017). The processes are focused towards the agile development and automation of steps involved starting from development to deployment (Leite et al., 2019). From my experience, adoption of DevOps

practices to a Bigdata solution developments requires tuning in terms of introducing methods that can address the various aspects of the data (quality, pipeline reusability & performance etc.,). In the survey research done by Leite et al. the researchers have surveyed multiple researches and have highlighted the implications to the engineers on how one should architect the systems, the implications to the managers on how mangers should face the DevOps phenomenon and the implications to the researchers on what could be exploited for future research (Leite et al., 2019). The research though discusses the majority of the areas where there are open challenges, but it has not discussed on what are the success factors that you could use a metric to measure the organizational success in terms of your DevOps maturity in the Organization.

Each of the key principles related to the DevOps have been widely discussed in various research but majority of the research in our opinion are focused towards solving a particular problem in the DevOps area and have not discussed in detail about the success criteria for those areas. In the research "The Intersection of Continuous Deployment and Architecting Process: Practitioners' Perspectives" done by Shahin et al. (2016) they have discussed how DevOps impacts the architecture, DevOps tools that are to be considered in the rapid changing environment, learning & training for the employees and also how the continuous delivery can be adopted using microservices strategy. In the research "On the Impact of Mixing Responsibilities Between Devs and Ops " done by Nybom et al. (2016), they have discussed about the adoption of the DevOps processes in Organizations. They have mentioned about the common risks involved in the adoption like not having clarity

on the responsibilities of the various teams will result in friction in Organization during DevOps adoption.

One cannot manage what is not measured (Forsgren and Kersten, 2018), similarly if one cannot measure the DevOps practices of an Organization it would not be successful. Forsgren and Kresten mentions that the metrics should be mainly used to identify the areas of improvement rather than punishing the teams as it would result in unreliable data resulting in undesired behaviors (Forsgren and Kersten, 2018). Feijter et al., (2018) provided a maturity model for their organization that can help in measuring the current maturity and identify the areas of improvement. Adoption of such maturity models will give an excellent scope for Organizations to mature their DevOps capabilities. DevOps practices though have lots of benefits but there are some challenges as well. Ghantous et al. has discussed the common challenges faced in DevOps practices with some of them being difficulty in adopting the mindset of DevOps in the teams, the clashes in the Dev and Ops tools and how migration of ongoing projects become difficult in getting adopted to DevOps principles (Bou Ghantous, Gill and Bou, 2017).

## 2.6 Understanding Data Integration Complexities in Organizations

Data sources are heterogeneous in nature and the amount of vast data that is getting generated is tremendously increasing day by day. For any Organization to understand and generated meaningful insights it requires data to be sourced from all the heterogenous sources into one common location. Ingesting high volumes of data coming from various heterogeneous sources requires scalable storage systems and compute systems to support and build applications. For analysts to analyze these large volumes of data from operational

databases are difficult as the data is spread across multiple database system in most of the cases and it has been noticed that each of it have different standards being followed further complicates the problem of doing the analysis. Also, if the operations are done directly on the operational databases will have an impact on the performance of the DBMS and its related application and there are chance to get errors or wrong formats. Martin et al. in their research "Lakehouse architecture for simplifying data science pipelines: data engineering and graph data mining explorations in Trase.earth for the traceability of supply chains driving deforestation" (Martín et al., 2023) discussed about the data engineering lifecycle that is being used across organizations. Fig 2.7 & Fig 2.8 show clearly the underlying components and the active components that are involved in active data moveement between the storage layers and how the data is being consumed at data and ml applications.

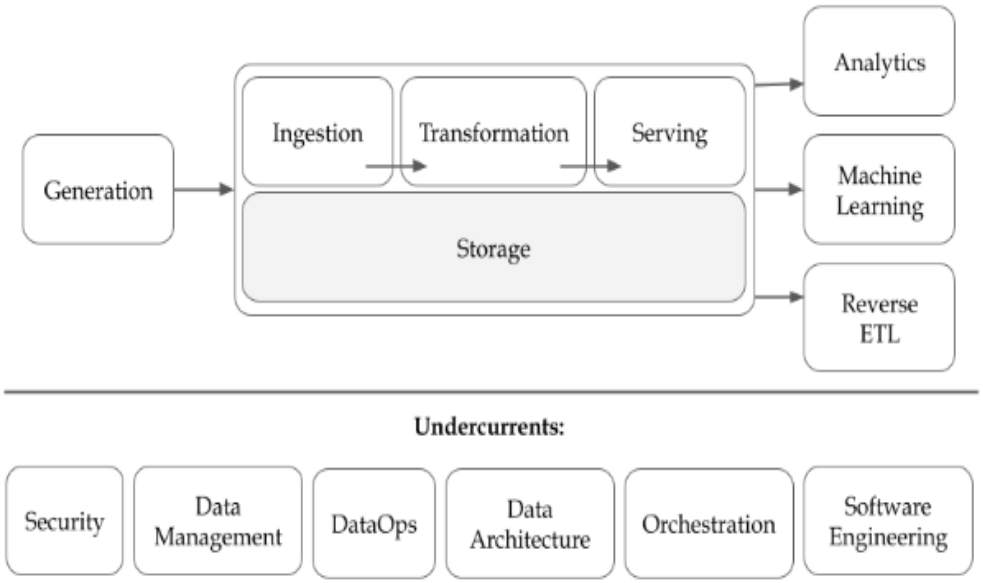


Figure 2.7 Data engineering lifecycle (Source: Martín et al., 2023, p. 17)

Fig 2.8 shows the high-level components involved in the data engineering undercurrents and the subcomponents involved in it. Access control, Data Management, DataOps, Data Architecture, Orchestration and Software Engineering are the underlying components and each of the areas have their own specific functionalities targeted at for a successful data architecture. Access Management is required to ensure the right level of access are in place for organizations to have control on who can access or not. Data Architecture helps in considering the components required for building a scalable and high-volume processing data application.

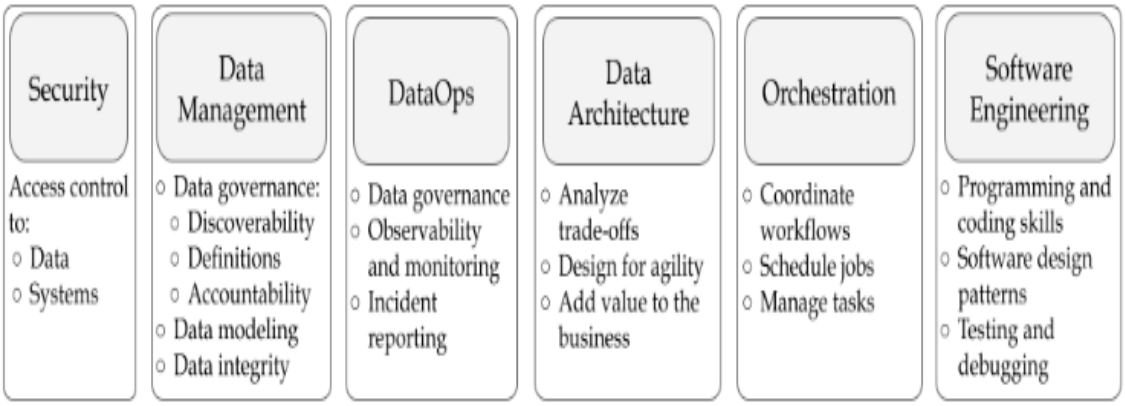


Figure 2.8 Data engineering undercurrents (Source: Martín et al., 2023, p. 17)

For the world of Data Analysis and insights generation, Online Analytics Transactional Systems (OLTP) are not suitable for processing the need of the analysts. These are operational systems which are primarily designed for the transactional operations. For Data Analysis one would need an Online Analytical Processing System (OLAP) which has the data aggregated and ready for consumption of analytical needs. The model used for the design of OLAP systems are well suited for the analytical needs and are not good fit for the OLTP systems as there is loss of data integrity and due to redundancy in the data. Poe et al. (1998) in his research "Building a data warehouse for decision support" shared his idea that the OLAP systems are built for comparisons and also for analyzing patterns and is difficult for such analysis using OLTP systems. The development lifecycle for the OLTP and OLAP systems differ and Inmon (2005) in his study "Building the data warehouse" favored towards data driven approach while Kimbal (2011) favored towards requirements driven approach for the Data warehousing systems in his research

"The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling". Sang-youl Kim et al. (2005) in their research have argued that in the absence of metadata management for a Datawarehouse system would make the decision support rely on the technical users.

Data warehousing is designed with the idea of collecting and modeling high-volume data to be stored in a managed database in which the data are extracted from the operational databases in 3 steps known as Extraction, Transformation and Loading (ETL). The data are extracted from variety of heterogeneous transaction sources and are transformed for analytical purposes based on a certain model. Data from all the sources needs to be transformed to a correct format for consumption at the BI layers. The metadata is used for describing the model and also relates it to the definition of the source model.

According to Inmon in his research Building the data warehouse (Inmon, 2005), the data modeling process can be divided into 3 levels. The data at a first step needs to be modeled for entity relationships and is called as Entity Relationship Modeling and it contains entities, features and relationships, no attributes and primary keys are specified at this stage. The mid-level modeling called as Logical modeling main objective is to document the business data structure, processes, relationships, and rules by a single view data model. At this stage, the attributes are defined, primary key of each entity is defined, and referential keys are specified. The third level modeling called as physical model is to optimize the performance. The purpose of the physical model is to map the logical model to the physical structure of the RDMS system. The data modeling is one of the crucial steps for designing an enterprise data warehouse system which can be used by reporting applications, if it is not designed correctly, it can become a dumping ground. Data being

conformed meaning the definitions of various fields should remain consistent irrespective from where they origin is one of the critical requirements of a big data system. Also, data should be historical meaning one should be able to retrieve data for any point of time, Data should be sharable meaning data should be accessible to all and Data should be comprehensive meaning it can be captured and consolidated from multiple systems. Ballared (1999) describes data modeling is the process of developing a model for effectively storing data in his research "Data Modeling Techniques for Data Warehousing". There are two data modeling techniques which are commonly used for the modeling data to data warehouses which are Entity Relation Modeling and Dimensional Modeling. Entry Relation model focus is primarily on the two concepts of entities and relationship between the entities while Dimensional Modeling focuses on measures, facts, and dimensions.

Schema-on-write approach like Extract, Transform and Load (ETL) that is followed in the Datawarehouse have limitation in handling the semi structured and unstructured data which has led to the development of NoSQL Databases. NoSQL management databases enables schema-on-read manner by storing the data in schema less manner. The authors in the study proposed framework aims to optimize ETL (Extract, Transform, Load) dataflows by classifying components based on their characteristics, partitioning the dataflow at various granularities, and utilizing shared caching schemes and parallelization techniques. This approach is intended to enhance the efficiency and performance of data warehousing systems. Improving the efficiency of ETL processes is crucial for enhancing the overall performance of data warehousing systems. Implementing partitioning techniques can significantly reduce the execution time of ETL workflows. Utilizing shared caching

schemes can minimize memory footprint and CPU consumption, leading to more efficient data processing. The optimization process may require significant computational resources and expertise, which can be a barrier for some organizations. Data Lake was proposed by Jame Dixon (2010) in his research "Hadoop and Data Lakes" as a solution which can raw data from more than on source. Data Lakes support the NoSQL formats and stores the raw data from various sources by storing their data in the source format but provides maintenance and query processing. Nargesian et.al (2020) in their research "Organizing Data Lakes for Navigation" has described the seven functions of data lakes discussing about the technologies and systems that can help in data lakes. Couto et al. in their research "A mapping study about data lakes: An improved definition and possible architectures" has compared the data lake definitions suggested in various researches (Couto et al., 2019) and Giebler et al. (2019) investigated about various data lake architectures and has discussed about common challenges in building data lake storages and suggested governance and meta data management are a key factor for the success of Enterprise data lake architectures.

Rihan et.al. (2023) in their research "Beyond the hype: Big data concepts, methods, and analytics" has discussed about data lakes architectures and metadata management. Sawadogo et.al. (2019) in their research "Metassdata Systems for Data Lakes: Models and Features" have discussed about the open-source technologies involved in the development of data lake systems. Pivotal et.al (2013) in research "The Technology of the Business Data Lake" proposed an architecture for business implementation of data lakes that consisted of Ingestion tier, Insight tier and Action tier. Muratov et.al (2023) in their research

"Framework architecture of a secure big data lake" have delved into the development of a Secured Data Lake Architecture Framework (SDLAF) aimed at tackling security concerns and enhancing data management quality. With the projection that data lakes will store over 175 zettabytes of data by 2025, securing these vast repositories becomes paramount. The authors argue for a comprehensive framework to secure data lakes, pointing out that traditional data lake architecture frameworks (DLAF) are inadequate and require modifications to effectively address emerging threats. In this design it has been suggested that the ingestion layer more of a data storage system to ingest the raw data from various sources while the insights layer to be used for interactive analysis of the data and Action tier for the applications to connect and consume the insights. Many criticized the data lakes approach and mentioned if the metadata management and data governance principles are not strictly followed then there is a high probability of data lake turning into data swamps.

Hartzell (2023) in his research "Comparison of Big Data SQL Engines in the Cloud" has discussed how serverless computing is revolutionizing data processing architectures. The benefits of using cloud-based services such as AWS Glue and Azure Databricks for big data analytics have highlighted scalability, flexibility, and cost-effectiveness. The research also touched on the critical role of metadata management in big data environments, discussing the need for robust metadata standards and efficient querying mechanisms to effectively discover the data and do analysis. Proper metadata management will ensure that data is easily accessible and can be exploited to its full potential.

Hassan Alrehamy et.al. (2015) in their research “Personal Data Lake With Data Gravity Pull” has stated that the data privacy and security are critical factors in the development of Data Lakes and success of the data lake is dependent on the metadata management of the system. Muratov (2023) in his research “Framework architecture of a secure big data lake” presents a modified DLAF that integrates machine learning algorithms and checksum calculations to detect anomalies and prevent malicious operations. This framework, termed SDLAF, includes a Global Monitoring Task (GMT) that consolidates log records from various sources to identify patterns and predict potential security breaches. This approach aims to create a more resilient and secure data management environment by leveraging advanced technologies. Michael Armbrust et.al. in their research “Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics” discussed about the lake house architecture which helps in addressing the major challenges with the data warehouses like data staleness, reliability, cost of ownership and use case limitations (Armbrust et al., 2021). In the survey conducted by Fivetran have mentioned that 86% of the analysts use out-of-date data due to limitations in data warehouses while 82% of companies are making decisions based on stale information (Fivetran, 2022).

Hellman (2023) in his research “Study and Comparison of Data Lakehouse Systems” has discussed in their research about several important aspects of data lakes, iceberg tables, hudi tables, SQL queries, and Delta Lake dummy rows, all within the context of modern data processing architectures. The significance of data lakehouses and their ability to integrate with various data sources seamlessly while offering scalable

solutions for big data processing has been a important factor for organizations to move towards the lakehouse systems. In the research, the primary finding highlighted the data lake houses can greatly improve the data analysis at scale offering decoupled storage and compute systems. The efficiency in the analysis is achieved by minimizing the need for manual data integration and real-time scaled insights. The research also discussed about the advantages of using Iceberg tables and their superior performance and scalability compared to the traditional hive tables. This is a critical factor for organizations dealing with large volumes of data where performance and scalability are critical. Hudi tables also provide am effective way to store and process substantial amounts of data and offer significant performance leading to a reduced storage costs and making them an important tool for big data analytics. However, the research has not focused on the critical factors involved in data engineering and the importance of data governance and security to ensure overall success of the data lake houses and their applications in the big data processing. Hartzell ” (2023) in his research “Comparison of Big Data SQL Engines in the Cloud discussed the importance of data governance and security within Big Data SQL engines. The authors highlighted the risks associated with uncontrolled data growth, such as data breaches and unauthorized access, and the necessity of effective access control mechanisms and comprehensive data governance policies. These measures are crucial for maintaining the integrity and security of big data environments.

The purpose of the big data systems is being to make the data available and accessible to all so that meaningful insights can be generated. A good supporting strategy and architecture is needed to achieve this purpose. Medallion architecture is one such data

design pattern used logically to organize data in a lake house system. As per the architecture, the data from the source will be landed to bronze layer as is with a focus of quick-change data capture, while the silver layer is for cleansing and conformed data and the bronze layer is for curated business level tables. Hellman through his research "Study and Comparison of Data Lakehouse Systems" Hellman, 2023) has stressed the necessity of using scalable and efficient query execution mechanisms to manage large datasets effectively. The research has discussed the use of Apache Spark's SQL module as a means to execute complex queries efficiently and that has been crucial for processing and analyzing big data.

There are multiple challenges that could arise in a big data processing systems. Benvenuti et.al (2023) in their research “A Reference Data Model to Specify Event Logs for Big Data Pipeline Discovery” have discussed about the importance of process oriented solutions to smooth the big data operationial issues emphasizing the importance of data awareness in handling complex data pipelines. The authors have shared an universally applicable data model for big data pipeline executions and its practical applicability to demonstrate it. The term ‘dark data’ has been coined by the authors to reffer hidden or lost data within the organization. Process mining based techniques could be used to understand the performance information and data manipulation details capturing data pipeline execution details. Big data pipelines become too complex to manage if no proper job level lineage management is not done and data pipeline executions are not recorded, logged and audited. A process oriented solution is essential for managing big data solutions and pipelines. Using the process mining techniques one should be able to discover any hidden

or lost data within organization. However, the potential challenges and limitations involved in the process mining is not discussed in the research. The research also does not discuss about the key metrics involved in the process mining or dataops, mlops or devops practices. The study primarily emphasized the importance of the data awareneess and scheduling of jobs through process orientation but has not discussed how such models can be scaled, improvided and measured. Though the proposed model for analyzing big data pipelines is promising in identifying the bottlenecks, inefficiencies and risks but it requires further research to develop robust frameworks for evaluating data pipeline frameworks and enhanced resource management.

Even after successful implementations there are potential chances of data being in silos because of the organization structure, and certain business units would be unable to make the data accessible to all due to various reasons like sensitivity, development and ownership of the data and others. So, it is important to ensure to have an organization level strategy which would ensure the data would not result in silos. Data Mesh is one such structure proposed for avoiding these data silos by connecting all the organization level data into one single whole data unit which is governed and maintained by individual business units but ensuring the data is available and accessible to the required users (Machado et al., 2022).

In the research conducted by multiple researchers about data spaces and sharing, the researchers discussed about the challenges and opportunities involved in creating trustworthy data sharing for Common European Data Spaces (CEDS) (Scerri, S., Tuikka, T., de Vallejo, I.L., Curry, 2022). The authors have discussed the importance and need for

standardization, coordination and experimentation for trust worthy data sharing. Standardization is the crucial element for building trust worthy systems as it requires organizations to have systems that are interoperable and compatible between data storing and data accessing across platforms and business units. Exploration allows organizations to be creative, explore different areas of scope, be innovative and simplify the approaches with regular and iterative feedback loops. The teams should also be coordinating to align the efforts to achieve the integration of various data sources and services. The authors in their resarch for common european data spaces have discussed about the importance of having a common vsision to achieve the required state. In their research they have also proposed clear action items to achieve this goal and stressed the need for building a trustworthy data sharing platforms.

From the research, here are some key takeaways for building a trust worthy data sharing platform.

1. Standardization is critical for data sharing, it needs to be ensured for standard data formats are consistent and interoperable between the systems so the data can be shared and utilized between teams and platforms
2. Trustworthy data sharing requires a technical competence and strategic alignment and cooperation between the stakeholders
3. For building scalable and adaptable solution teams should be willing to experiment and be innovative

In the research “An Organizational Maturity Model forData Spaces: A Data Sharing Wheel Approach” done by the researchers Curry et.al, (2022) the researchers discussed the

challenges and opportunities of creating common European Data Spaces (EDS). The authors emphasized the need for building common blocks, need for coordinated actions, and map exisiting initiatives to this.The research also discussed about the challenges involved like EU's position in data sharing, investing in strategic cooperation and developing technical competence. The research also discusses about the importance of experimenation for piloting the identified data sharing usecases in safe and dynamic environment to share across regions. Concerns about coordination, standardization and not willingness to share the data due to privacy and security concerns can become potential blockers for implementing data sharing across various units or regions.

### 2.7 Navigating Complexity in Machine Learning Applications

Machine Learning applications are slightly complex compared to Data Engineering applications. Muratov et.al in his research "Framework architecture of a secure big data lake" (Muratov S. Y, 2023) has highlighted challenges and complexity involved in integrating machine learning algorithms with existing data management systems. This integration may necessitate substantial changes to data architectures, posing a considerable hurdle for implementation. Despite these challenges, the potential benefits of enhanced security and data quality make the effort worthwhile. They have many interlocking analytical components beyond training of the ML model. D.Sculley et al. (2015) in their research "Hidden Technical Debt in Machine Learning Systems" argued that the ML applications have hidden technical debt due to their additional requirements on the ML requirements. They also mentioned that they are hard to detect as they are at a system level than at code level and have focused on the system level interactions where the technical

debts can quickly accumulate. Dominik Kruezburger et al. (2022) in their research "Machine Learning Operations (MLOps): Overview, Definition, and Architecture" conducted a mixed method research and provided an aggregated view of the various components present in the ML systems such as architecture, workflows and components which helped in understanding the problem of MLOps and productization of ML applications. Doris Xin et al. (2021) in his research "Production Machine Learning Pipelines: Empirical Analysis and Optimization Opportunities; Production Machine Learning Pipelines: Empirical Analysis and Optimization Opportunities" analyzed provenance graphs of 3000 production ML models at Google to understand the complexity of the ML models which helped in understanding the complexities involved, topology of the various ML pipelines and their granularities. Their analysis revealed that data management techniques can be used to optimize the ML pipelines and they also identified that the models were trained but are wasted without being deployed into the systems causing wastage in computing and storage. Nikil Muralidhar et al. (2021) in their research "Using AntiPatterns to avoid MLOps Mistakes; Using AntiPatterns to avoid MLOps Mistakes" described and discussed about the anti-patterns in ML applications similar to Design Patterns in the software applications and suggested the cataloging of these anti patterns for the future of MLOps maturity. Cedric Renggli et al. (2021) in his research "A Data Quality-Driven View of MLOps" has discussed about the processes involved in the ML model and also the traditional software and have highlighted that the different ML components can be efficiently designed from a technical and theoretical perspective. Eric Breck et al. (2021) in their research "What's your ML Test Score? A rubric for ML

production systems" has opinionated that the testing and monitoring are the key considerations for checking on the production readiness of the ML systems. The researchers have presented 28 specific tests and monitoring needs to reduce the technical debt and improve the production readiness of the ML systems. Fig. 2.9 represents the encapsulation of the software engineering-based components into ML architecture that represents the key components required in any ML architecture development.

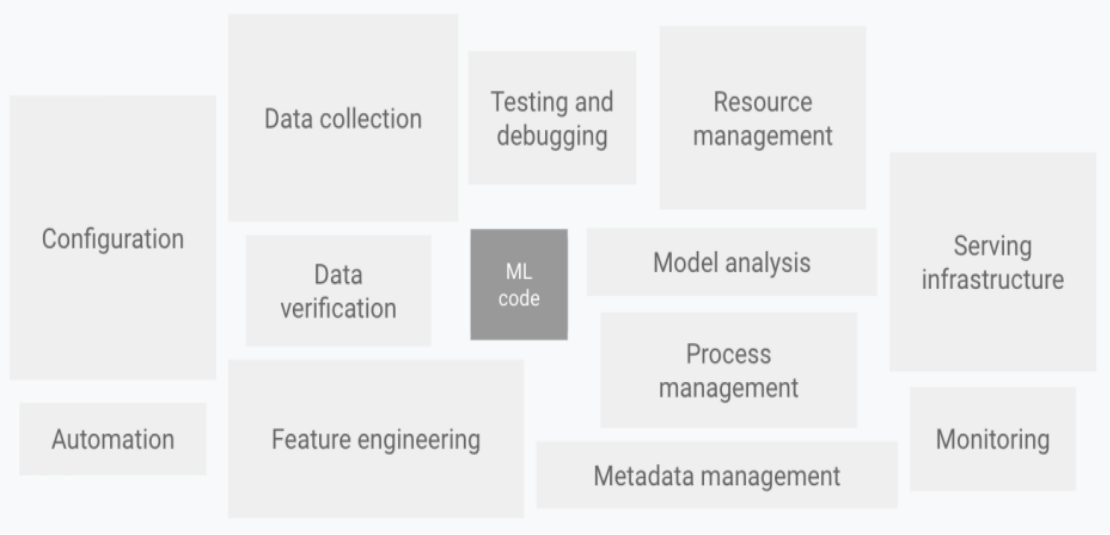


Figure 2.9 ML Components Diagram (Source: Felipe and Maya, 2016, p. 3)

A high level of coordination is required between various systems and efficient tools, technologies with strong guidelines are required for processing and accessing the data from the big data storage systems. The real challenge is with ensuring the data is fed into these systems on daily basis without failures otherwise it would lead into data inconsistencies and would soon turn the high value data to a dead or garbage data. Along with that strong guidelines and principles are required in maintaining and accessing these data. Data

Security and Governance play a critical role and if there is not a thoughtful access management strategy for the organization needs then it would soon turn a difficult challenge to maintain the access of the data and if there is no strong metadata management it would also lead to a state where it would be difficult for users to understand what data is present in the system. Assigning DataOps teams in organization with clear roles and responsibilities will bridge the missing gaps and ensure quality data productionized at a faster rate and accessible with relevant information.

Widad Elouataoui et.al. (2022) in their research "An Advanced Big Data Quality Framework Based on Weighted Metrics" proposed Bigdata Quality Framework Assessment based on 12 metrics to ensure accurate data quality. Timmerman et.al. proposed rule-based measurement for measuring the data quality which allows the handling of uncertainty. Goutam et al. (2019) in their research "An Automated Big Data Accuracy Assessment Tool" suggested a model to choose the optimal one using word embeddings and linkage of records as part of their big data quality assessment. Taleb et al. (2018) in their research "Big Data Quality: A Survey" suggested a framework that suggests to store the valuable project information, data quality rules and profiling. There has been various research conducted in the regard of the data quality and it is one of the important areas for building a successful data organization. However, from the literature review done we have noticed that the various areas of the data products development are measured in the individual areas like data quality, code development and others, however the study for an integrated metric system that measures the overall end to end success of a big data product from business perspective are not dwelled deeply and I strongly feel there is a need for

such study that can help organizations to measure their data success. Irene et.al (2024) in their research "KPIs for Quality and Availability of Data in an Industrial Setting" provides valuable insights into the significance of considering feature utility coefficients when determining the success of data completion exercises. Its findings have important implications for organizations aiming to improve their data quality and availability. By adopting the proposed metrics and developing context-specific KPIs, organizations can optimize their data collection processes and achieve more reliable and complete datasets.

Lucy et.al (2019) in their research "A Taxonomy of Software Engineering Challenges for Machine Learning Systems: An Empirical Investigation" conducted on mastering ML models, the authors delved into the engineering demanding situations of growing and operating and mastering machine learning systems in real-time organization settings. The authors have proposed a taxonomy depicting the stages of integrating ML additives into software-engineering practices, aiming to provide a complete framework for understanding the evolution and operation of those systems. The authors discussed one of the important issues is the importance of thinking about software engineering (SE) components that are beyond the algorithmic troubles. The authors stress the need of a structured techniques for development process, robust infrastructure to experiment, and strategies for evaluation to evaluate model performance comprehensively. These elements are essential for ensuring that ML structures may be reliably and efficiently integrated into real time software environments.

In the research, the authors discussed several demanding situations encountered at each level of an ML systems lifecycle. These challenges range from dataset creation and

problem statement creation and making sure there is high-quality data for training and deploying models in production environments of the organization. The authors also discuss some specific set of obstacles at each stage that need to be addressed to achieve a successful and scalable ML applications. Hellman (2023) in the research “Study and Comparison of Data Lakehouse Systems” have discussed the usage of Delta Lake dummy rows, and their importance of generating scaled datasets by creating dummy data to simulate real-world scenarios. This approach can enhance performance and decrease storage costs, providing a practical solution for testing and optimizing data processing workflows. However, the authors have failed to discuss about the critical factors such as data quality, freshness and data usability which are essential for the success of the data operations and machine learning operations. The metrics related to the data quality are vital for ensuring the data is accurate, up-to-date and useful for decision making processes.

One of the important problems discussed is building highly scalable ML pipelines that can process data in real time and generate insights. The complexity of scaling those systems regularly includes overcoming issues related to data processing and management, computation infrastructure, and the seamless integration of various components. Also there is a uncompromisable requirement for reproducing models to make the models compliance adherent, and is also critical for preserving consistency and reliability in ML packages. Debugging of the machine learning and deep learning models also pose a significant challenge due to closed nature of these models not allowing the users how the insights or predictions made be the systems.

In the research, the authors discussed several demanding situations encountered at each level of an ML systems lifecycle. These challenges range from dataset creation and problem statement creation and making sure there is high-quality data for training and deploying models in production environments of the organization. The authors also discuss some specific set of obstacles at each stage that need to be addressed to achieve a successful and scalable ML applications.

One of the important problems discussed is building highly scalable ML pipelines that can process data in real time and generate insights. The complexity of scaling those systems regularly includes overcoming issues related to data processing and management, computation infrastructure, and the seamless integration of various components. Also, there is a uncompromisable requirement for reproducing models to make the models compliance adherent, and is also critical for preserving consistency and reliability in ML packages. Debugging of the machine learning and deep learning models also pose a significant challenge due to closed nature of these models not allowing the users how the insights or predictions made be the systems.

Multiple researchers (Kai Hartzell, 2023; Muratov S. Y, 2023) have discussed the below key important metrics:

1. Data Processing Efficiency: The ability to process large-scale datasets quickly and efficiently is paramount for modern businesses. Efficient data processing frameworks enable organizations to derive insights faster and make data-driven decisions more effectively.

2. Query Optimization: Implementing effective query optimization techniques is crucial for managing complex queries and delivering real-time insights. Optimized queries can significantly reduce processing time and improve the overall performance of data operations.

3. Serverless Computing: The use of cloud-based services like AWS Glue and Azure Databricks offers substantial benefits for big data analytics, including scalability, flexibility, and cost savings. Serverless computing allows organizations to handle varying data loads without the need for extensive infrastructure management.

4. Metadata Management: Establishing robust metadata standards and efficient querying mechanisms is necessary for effective data discovery and analysis. Proper metadata management ensures that data is well-organized and easily accessible, facilitating better data utilization.

5. Data Governance and Security: Implementing effective access control mechanisms and robust data governance policies is essential for ensuring the integrity and security of big data environments. These measures protect against data breaches and unauthorized access, maintaining the trust and reliability of the data.

6. Data Quality: The SDLAF framework emphasizes maintaining high-quality data management practices. Ensuring data accuracy, consistency, and reliability is fundamental to the framework's effectiveness.

7. Adaptability: Developing a framework that can adapt to changing data sources and logging formats is vital. This adaptability ensures that the framework remains relevant and effective as data environments change.

8. Automation: Automated feature extraction methods and machine learning algorithms can help reduce dependencies on specific log sources and expand the framework's applicability. This automation is key to managing large-scale data environments efficiently.

## 2.8 DataOps Principles and Maturity Model

DataOps aims to improve collaboration among data scientists, engineers, and technologists, ensuring that every team works harmoniously to leverage data more efficiently and expediently. Fig 2.10 represents the foundational architecture in a typical organization setup which is architected for flexibility, quality, rapid development and real time monitoring.

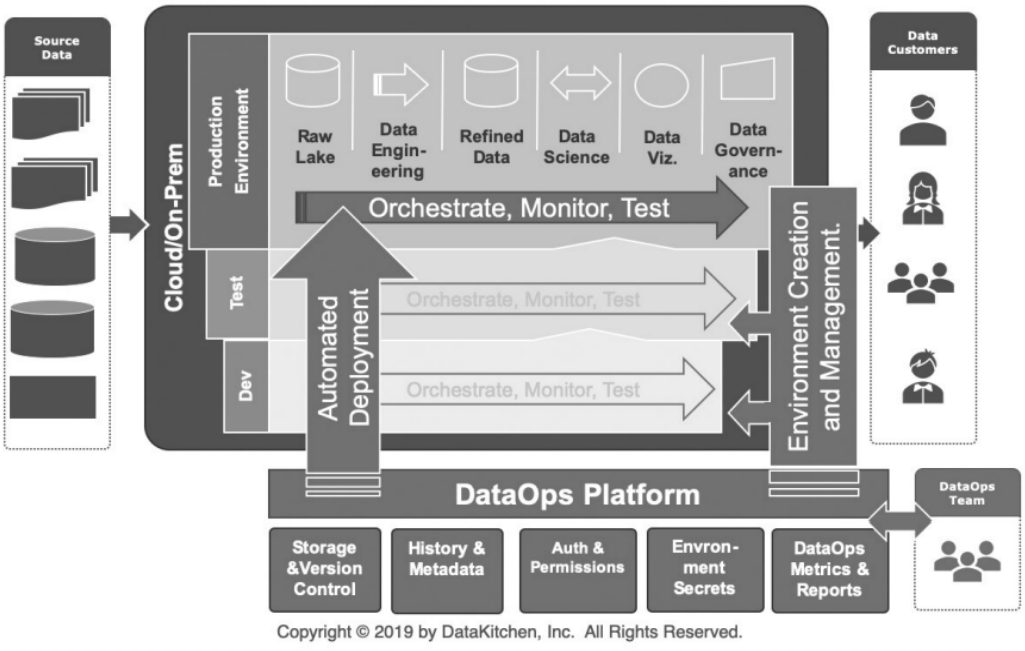


Figure 2.10 DataOps Foundational Data Architecture

(Source: DataOps Manifesto, 2021, p. 171)

**DataOps Principles:**

The DataOps Manifesto defines the following DataOps principles

**a)** Continually satisfy your customer: Early and fast delivery of valuable analytic insights by collaboratively iterating with the customer.

**b)** Value working analytics: Provide fast analytical insights to customers by integrating precise data with robust framework.

**c)** Embrace change by adapting to the evolving customer requirements which could be internal and external.

**d)** A diverse talent with experience in skills and tools will help in innovating and productivity.

**e)** On a day-to-day basis have a constant communication between customer, analytics teams and operations teams.

**f)** Self-organize to quickly meet the changes in the requirements to achieve a common goal.

**g)** Reduce single dependencies by creating sustainable teams and process.

**h)** Do retrospectives at regular intervals to take feedbacks and operational statistics.

**i)** Deliver insight by using appropriate tools to access, integrate model or visualize data.

**j)** Orchestrate the end-to-end pipeline which is a critical key driver for analytic success.

**k)** Develop applications with reproducibility, versioning, configuration and reusable efforts.

**l)** Disposable technical environments to be strategized to minimize efforts in development and testing.

**m)** Importance to be given technical excellence and smart design to focus on the essentials to complete.

**n)** Follow lean manufacturing principles. Reduce waste and continuously improve the analytical insights.

**o)** Quality of data is utmost importance so managing quality and performance is critical.

**p)** Improve monitoring and performance to detect anomalies in monitoring quality, security and performance.

**q)** Reduce repeated work.

**r)** Reduce the time and effort to reduce the cycle time in developing the customer idea to an analytics solution.

Jakobsen (2023) in his research "Study of DataOps as a concept for Aker BP to enable data-driven assets" proposed DataOps maturity model based on the core principles discussed in DataOps Manifesto. It has been developed to analyze an organization readiness for matured data engineering operations and proposed 5 level maturity guidelines detailed in Fig 2.11. The maturity level 1 is the state where there is no stable environment and processes are usually ad-hoc and un-organized. At maturity level 2, there are processes

and requirements in place to perform and develop analytical insights and ensure data quality. At maturity level 3, best practices and standards are established across organization with clean data accessible to users. At maturity level 4, there is universal confidence in the data and resulting insights in the organization. At maturity level 5, architectures are defined to process large volumes of data with efficiency and can rapidly respond to changes.

**DataOps Maturity Model**

| Stage | 1. Initial | 2. Develop | 3. Define | 4. Measure | 5. Optimize |
|---|---|---|---|---|---|
| Data Management & Analytics | • Ad-hoc ways of working.<br>• Few datasets or products have been onboarded to the organization.<br>• Data is used mainly for reporting purposes. | • Data insights are used to inform business decisions.<br>• The organization can combine data from different sources.<br>• Some established ways of working.<br>• Data to be defined as an important asset. | • Dedicated teams focused on analytical value delivery and improvement.<br>• DataOps practices are characterized for the organization and proactive.<br>• Defined data catalogs used to manage data inventory and products. | • Data delivery and insight is measured and controlled.<br>• Analytical insight and performance is measured and continuously improved.<br>• Certain data management practices are automated across the organization. | • DataOps processes are fully automated, regularly improved and optimized.<br>• New technology and analytics is developed upon an architecture designed for efficiency and large volumes of data. |
| Culture & Collaboration | • The use of data is based on individual efforts.<br>• The business lacks a coherent data architecture and strategy.<br>• No business unit is working on improving how the organization works with data. | • Business goals are known and communicated.<br>• Activities are deliberate and documented.<br>• Data is part of measuring results. | • DataOps practices are aligned with an enterprise-wide strategy.<br>• Business users have cleaned data available to enable data-driven insight.<br>• Deliberate ways of collaborating across business units. | • Decision makers are enabled with the results of data analysis to maximize business outcomes.<br>• Service catalog of agreed technologies and high-level framework for teams to adopt new technologies. | • The business is constantly looking to leverage and combine data from new sources.<br>• Automated AI/ML algorithms and data pipelines continuously improve business objectives. |
| Data Governance | • Governance is largely manual and lacks consistency. | • Processes are in place to protect data quality across the organization. | • Best practices are defined and evangelized, and teams enabled to evolve practices. | • There is universal confidence in the data and resulting insights. | • Data Governance is integrated into all business processes. |

Figure 2.11 DataOps Maturity Model (Source: Jakobsen, 2023, p. 35)

## 2.9 MLOps and its Maturity

Similar to DataOps, MLOps accelerates the development and deployment of machine learning models with improved model tracking and reliability. As per the definition of Microsoft (2022), MLOps at level 0 maturity is at a state where there is no MLOps practices and difficult to manage machine learning model life cycle. At level 1, there is DevOps but no MLOps, releases of applications are less painful then no MLOps

state but still difficult to trace how the models perform in productions. At level 2, automated training of the machine learning models are in place and models are fully managed and traceable. At level 3, releases of ml model with low friction with full traceability from deployment back to original data. At level 4, systems are fully automated and easily monitored with zero down time system.

Adoption of DevOps principles in machine learning applications helps in continuous development, deployment and delivery of machine learning models. In the research "Towards MLOps: A Framework and Maturity Model" the authors have proposed a framework that identifies the activities involved in adopting MLOps and stages of evolution (John et al., 2021). The authors also have validated three software intensive embedded systems to explain the integration of MLOps into large scale developments. The authors have explained the five stages of MLOps maturity consisting ad-hoc, centralized, standardized, automated and industrialization which companies can use to evolve in their MLOps practices by following the maturity levels. The research also discusses on the challenges associated with MLOps, including data quality issues, model interpretability and scalability. The research suggest that to address these challenges associated with MLOps requires a combination of technical experience and adoption of MLOps pratices within the organization. The study has provided valuable insights about adopting the DevOps principles in ML systems. The proposed framework by the researchers can serve organizations as a guideline to integrate MLOps into their development processes. However, further research is needed to address the challenges and develop effective practices for scaling up ML Models. Overall the researchers have provided comprehensive

overview of the current state of MLOps research and its applications in software intensive applications. It also discussed the potential benefits of DevOps principles in ML systems including increased efficiency, improved collaboration and enhanced innovation for faster development of data products.

In today's data-driven organizations, the success of DataOps (Data Operations) and MLOps (Machine Learning Operations) heavily relies on efficient ETL processes. The proposed framework by Liu (2014) can play a crucial role in optimizing these processes, enabling organizations to improve their data processing capabilities, reduce costs, and enhance decision-making.

The study “Representations of epistemic uncertainty and awareness in data-driven strategies” done by Mario et.al., (2023) proposed a classification system for maturity models, which is critical in understanding the development process of data science capability maturity models. The study also discussed the need for a survey-based approach to develop the data science capability maturity model. The studies suggest that organizations should focus on developing their MLOps capabilities to realize value from big data and asks to focus on the development of MLOps capabilities to improve decision making for organizations.

### 2.10 Importance of Key Performance Indicators

Key performance indicators are quantifiable measurements used by organizations to reflect the critical success factors. An effective organization knows that if they do not have measurements to track their process or products, they can’t control it. DataOps and MLOps practices also need to be measured on regular basis to assess and take corrective

actions as and when necessary to meet the organization goals. Wang et.al (1996) in their study “Beyond Accuracy: What Data Quality Means to Data Consumers” has deep dived the importance of the data quality metrics has extensively survyed to understand the importance of the data quality metrics from consumers point of view. Accuracy of the data, Relevancy of the data, Representation of the data and Accessibility of the data are the key areas that has been focused as part of the Wang reaseach. It has been proposed in this study the importance of having a data quality framework to measure, analyze and improved the validity of the data. The hierarchial data quality conceptual framework was proposed with four categories Intrinsic data quality, contextual data quality, representational data quality and accessibility data quality groupig the metrics related to data quality. Generally, the key performance indicators are defined by subject matter experts who has a deep understanding on the business process. Table 2.1 has the details of the various KPIs that are widely used in the industry which has acted as the basis for the survey questions in this research.

| **Key Performance Indicator (KPI)** | **Definition** |
|---|---|
| Data Accuracy | Data accuracy is a measure of the extent to which data represents the true value of the attribute it is intended to measure (Peralta, 2006). It is crucial for reliable decision making and operational efficiency. Accurate data minimizes errors, reduces the costs and also increases the credibality of the reports. For data driven insights it is very crucial to maintain data accuracy. |

| | |
|---|---|
| Data Completeness | It is a measure that checks the extent to which the data elements are provided with no missing values (Peralta, 2006). This parameter checks whether all the data is available and the data collection process, integration process and validation processess are in place to ensure data is collected correctly without any misses. Incomplete data can lead to missed opportunities and flawed conclusions. Regular audit and data management proactices are essential for maintaining data completeness. |
| Data Freshness | Checks the freshness of the data whether the data that is there in the system is upto date as per the source (Peralta, 2006). Maintaining fresh data ensures the systems and reports are up to date where as stale data can result in out dated decisions and will result in being out of the market. For business reporting, fresh data is required for timely decision making and informed strategy decisions. Implementing real time pipelines and regular updates helps maintains data freshness and ensuring the reliability of the reports. Maintatining real time pipelines is challenging for organizations and support teams. |

| | |
|---|---|
| Data Consistency | Checks the level of consistency of the data between various data sources. This KPI helps in avoiding confusion and discrepancies in decision making (Peralta, 2006) (Ruf et al., 2021). Ensuring data accuracy is across various databases, applications and process is important for avoiding discrepancies, reducing data conflicts, and reliable integration. Consistent data supports similar metrics reported across various business reporting applications avoiding confusions, errors and misinformed decisions. Having robust data governance, standardized formats and synchronization mechanism is essential for data cosnistency. |
| Data Availability | Checks the data up time, ensuring data is accessible and available to the users (Peralta, 2006). It is crucial for evaluating the reliability and performance of IT supporting systems. Monitoring this KPI helps to identify and mitigate issues that could lead to downtime or data inaccessibility. Ensuring data availabiliy is crucial for business continuity and timely decision making process. This KPI is helpful indicator to ensure the proportion of the time the data system is accessible to end user. |

| | |
|---|---|
| Data Quality Score | Measures the overall quality of the data based on data accuracy, completeness,consistency and freshness (Peralta, 2006) (Ruf et al., 2021). Generally each dimension of the data quality fields are assessed and assigned a score which is then aggregated to calculate the overall score. This metric helps in providing a quantifiable measure on the data health, helping organizations to identify the areas for improvement. Maintaining high data quality score is important for any organization for reliable analytics and informed decision making. Regular monitoring of this metric will ensure the data health is maintained and remains trust worthy and valuable. |
| Data Timeliness | This measures the availability of the data at the time needed. Delay in availbility of the data results in impactung decision making (Peralta, 2006). This KPI is crucial for tracking and ensuring the data decisions that are being made are on the latest data. Achieving timeliness involving optimization of data pipelines and process that can reduce latency and enable access to data in real time. Regular updates and real time data |

| | |
|---|---|
| | integration are key practices for maintaining data timeliness. |
| Data Integrity | Data integrity involves maintaining the accuracy, reliability and consistency of data over its entire lifecycle (Peralta, 2006). To ensure data integrity, robust data error detection, data validation and secure handling mechnanisms to be implemented. Any compromises in this can lead to faulty data decision and could potential lead to financial losses. Having strong data governance practices and regular audits are essential for uploading data integrity. |
| Mean Time To Detect Data Errors | Average time taken to detect data errors. This is one of the crucial metric as the longer mean time to detect generally means the identification of errors are taking longer durations and could potential impact decision making due to data errors. Organizations should target low mean time to detect errors through effective proactive monitoring process, automated alerts, regular audits and robust data validation process. |
| Data Drift | Data drift refers to the unexpected and undocumented changes in the data characteristics (Ruf et al., 2021). It can be caused due to various reasons such as change in |

| | |
|---|---|
| | customer behaviours, market changes, modification in data collections methods or government regulations. Data drifts can have adversary effects on the model performance and accuracy of machine learning models which would have been initially trained on a dataset that did not had data drift changes. Continuous monitoring of data drift changes is critical for detecting these changes and reducing the impact on to the business decisions. Regular model training with the latest data implementing drift supportive algorithms can mitigate the impact due to data drifts. Addressing data drift is critical for maintaing the relevancy of the reports and ensure business decisions are reliable and trustable to the end business users. |
| Data Deduplication ratio | This metric measures the data duplication in the system. This is critical to ensure one source of truth and avoid unnecessary storage, infra costs and avoid confusions in product developments. By ensuring low / no dat depluication will enhance and the backup and recovery speeds and in overall improves data management efficiency. Data Engineers should regularly analyze the data and categorize data to identify redundant data patterns. Tracking and tracking this KPI will help in |

| | |
|---|---|
| | taking necessary actions into storage cost redutions and allocation improvements and ensure the organizations are cost effective. |
| Schema Drift Detection | This metric checks the number of times schema drift occurred due to changes in the source. This is an important metric as it would result in data pipeline failures if there are frequent schema drifts. Early detection help in proactive corrective actions, preventing data quality failures and pipeline failures. This is crucial for business reporting to ensure business decision are made on the right and accurate data.<br>To Data engineers should implement automated schema validation and monitoring systems to notify when such changes occur at the source side. Regular schema reviews, incorporating data governance and effective communications between upstream and downstream will help in informed decision making and operational efficiency. |
| Time to Market | This metric measures the different aspects of development and delivery and how much organizations are matured for accelerated poducts deployment into markets.<br>Demonstrating faster time to market is critical for gaining |

| | |
|---|---|
| | customer confidence (Forsgren and Kersten, 2018). Shorter time to market allows business to capitalize on market opportunities and reach customer demands on time. Faster response to market trends and innovation are crucial benefits for Time to Market KPI. To target this metric, the focus should be on adopting agile methodologies, streamlined workflows and work with cross functional teams to accelerate the development and deployment processes. Regularly reviewing the project timelines using project management tools can keep a healthy check on the timelines. Closely monitoring this metric will help in identifying the areas of improvments, identify process efficiencies and address any impediments or bottlenecks in achieving this targeted goal of an organization. |
| Number of releases | This metric measures the releases completed over a given span of time. This helps in identifying if there is an imporvement or decline in the release frequency. Implementing robust CI / CD (continuous improvement / continuous deployment) pipelines, automated testing of the features that are in scope of the releases and defect free smooth releases are crucial for this metric. Having a |

| | |
|---|---|
| | clear guidelines on the communications between the development team and the operations team would ensure the release management is completed effectively. Tracking this metric will help in understanding the development team velocity, helps in identifying the areas of optimization and helps to develop the culture of continuous improvement and agility. |
| Development Cycle Time | This metric measures the development speed from the start of the work to its delivery time. Shorter development cycle time provides a quicker time for delivery of value for customers, faster feedback loops ensuring early corrections incase of deviations and enhanced opportunity to adopt to maket changes. Adopting agile practices, enhanced collaboaration across teams and leveraging the automated methods for testing and deployment will help in improving this metric. Tracking development cycle times can help in bringing process efficiencies and remove any impediments in the delivery of data products on time. |
| Defect Rate | This metric measures the number of defects raised in the system post deployment. This helps in understanding the efficiencies in testing and development. Lower defect rate |

| | |
|---|---|
| | means higher data product quality, clear understanding on the requirements, reduced costs and delays in identifying and fixing the issues and also improved business satsiaction. Implementing rigourous testing and adopting practices such as Test Driven Developments (TDD) and Continuous Integration can help in improving the defect rate metric. Regular code reviews and maintaining robust code quality assurance frameworks can help in reducing the defects. Monitoring the defect rate can help in providing insights into product quality, areas of improvement in the development process and support decision making that are aimed at improving the product reliability and trust. |
| Customer Feedback Response | This measures the trust the customer is showing on the data product and processes established. Higher the metric higher is the trust in the system. Ensuring feedback management system which is continuously monitored and responded would improve the satisfaction of the end users. Performing topic modeling on the customer feedbacks can help in identifying the various topics that the end users are satisfied or dissatisfied and that are required for improvement can be identified. Addressing the critical and |

| | |
|---|---|
| | negative feedbacks should be of a top priority for organizations to ensure the brand reputation is not impacted because of the negative feedbacks. Positive customer experience and strong customer relationships should be the primary target for the customer experience teams in the organizations. |
| Data Ingestion Time | This metric measures the time it took to ingest any new tables into the system. Lower the metric, higher is the impact in accelrating the data product developments. Data is sourced from multiple data sources, having the connectors developed to connect to these sources and addressing any challenges related to performance and connectivity issues would be critical to have a reduced data ingestion timelines. Having an automated system that can capture the source details and build a pipeline can be tagrted by organization to reduce any overhead involved in the data ingestion pipeline developments. Having efficient data ingestion practices are required for timely, reliable and readily data availablity for the product teams. |
| Data Processing Time | The average time it took to process certain volume of records (Liu, 2014). Its represents the infrastructure capability and code efficiency practices followed in the |

| | |
|---|---|
| | system. Lower the value, higher is the efficiency. Data voume is also another important factor that could impact the data processing times. Ensuring the right business rules are applied is one of the key factor that can improve the processing times. Higher the processing times, higher is the infrastructure usage cost which could lead to platform budget constraints. In my opinion based on my experience with various organizations, the data processing should be within 5 mins for atleast 80% of the data pipelines and the rest 20% should be within the 30 mins time frame. All data pipelines should be developed to support incremental load of data rather than the full load of data every day to avoid higher data processing, economically feasible and technical the ideal way to load as there would be any changes to most of the data and it would have been already loaded in the previous days executions. |
| Pipeline Uptime | This metric measures the time the pipeline is operational and available for data processing. It is an indicator on how reliable are the data pipelines in the day to day operations. Organizations will have hundreds of pipelines for data extraction and transformation and these pipelines should |

| | |
|---|---|
| | be smoothly operating without any failures to avoid any disruptions to the business reporting. Robust monitoring and alerting systems needs to be enabled to detect and resolve any failures wih the pipelines. Non operational pipelines should be disabled and decommissioned to improve efficincies. This metric helps in understanding the reliability of the infrastructure and identify any recurring issues and support continuous improvements. Also, organizations should practice to use only specificed or identified set of tools and technologies and keep the technology spread in control to avoid any challenges with the maintenance of the systems. |
| Data Throughput | This metric measures the data movement over a period of time. Higher througputs are good for organizations as they indicate efficient data handling & faster data processing which are the key features of a high reliable data systems. Scalable infrastructure along with the high performance data processing frameworks and efficient parallel processing techniques to achieve high data throughputs. This metric gives insight about the performance of the data systems, identifies bottlenecks and ensure data operations can meet growing needs of the organizations. |

| | |
|---|---|
| Error Rate | This metric measures the amount of errors noticed during the data transfer due to data and infrastructure issues. Lower error rate indicate higher reliability and quality systems. Monitoring the error rate is essentential for maintaining the system integrity and accuracy of the business process. Adopting good testing practices, automated testing and continous integration can help in identifying and addressing the issues early. System performance and quality of the system can be efficiently trracked through this KPI. |
| Deployment Frequency | This metric measures the number of times the teams are able to deploy the application into production over a period of time. Higher the metric higher is the teams agility and delivery of data product |
| Deployment Success Rate | This metric measures the success rate of a deployment process compared to number of deployments initiated over a period of time. Higher the metric higher is the trust and reliability in the deployment process. This metric is crucial for maintaining service continuity, reducing roll backs and delivering features with improved user experience. Adopting to continuous integration and continuous deployment practices, automated testing and |

| | |
|---|---|
| | doing pre-deployment validations would help in improving the deployment success rate. This metric will help in identifying the areas of improvement in the deployment process, enhancing the deployment pipelines for increased success rate and also improve the alignment between development and operations team to deliver high quality and reliable data products o the end users. |
| Testing Coverage | This metric measures the percentage of automated test coverage of a data product. Higher the percentage, higher is the efficiencies in the system. Regular monitoring of the metric helps in identifying the issues and defects early. Having strcuture testing approaches for both manual and automated will improve and optimize the testing coverage. Prioritizing the testing activities based on the critical functionalities and potential risks would help in ensuring the key functionalities are delivered withot any defects. Testing coverage can be improved using autoated testing frameworks, code coverage tools and continuous testing practices. Higher metric will ensure the systems are thoroughly tested and the quality of the system is high before it gets delivered to the end users. |

| | |
|---|---|
| Deployment Down Time | This metric measures the down time it has caused for a data product due to deployment planning. This has to be low to have positive customer experience. Minimizing the deployment downtime is crucial for maintaining service quality and ensuring any service disruptions to end users. Practices like blue green deployment which will alow new version to be deployment along with the older versions and allowing a minimal down time during the switch will ensure minimal disruptions during the releases. The deployment teams needs to have pre-written deployment and roll back steps, incase if the deployment fails, application teams needs to be ready to avoid any major disruptions because of the issues in the deployment process. Human errors in the deployment process can be reduced by followng automated deployment practices. Doing a thorough deployment steps in the lower environments like test can help in identifying the risks or challenges linked with the deployment steps which could be due to infrastructure or network or application code itself and have the mitigation steps identified before the production deployment. Objective of this metric is to |

| | |
|---|---|
| | ensure to minimize the downtime and maximize the uptime for the end users. |
| Median Response Time | Median time it took to respond to an incident by the application support team. Mean response time could be skewed and so the median response time would be a metric to measure the support team alertness to react to end user concerns. The response time to acknowledge an incident should be lower to ensure the reported incidents are swiftly acted by the support team. This metric needs to be validated againsts the service level agreements (SLA) to measure how the support team is performing. It helps in identifying the gaps in the support process and ensure the improve the areas to bring in trust to the end users. |
| Median Resolution Time | Median time it took to resolve an issue or icnident by the application support team. Providing resolution within the agreed service level agreements is critical for any organization to ensure trust and confidence on the data product to the end users. Incidents should be resolved with proper root cause and the details of the failures has to be notified to end users to have higher confidence on the systems. Also organizations should compare these metrics against the previous time periods to understand if the |

| | |
|---|---|
| | improvements made to the system are on the right track to improve.Regular review of these metrics needs to be conducted and a review has to be done on all the issues that have breached the service level agreements so that any corrections can be done on to the processess or service level agreements. |
| Incident Reopen Rate | Number of incidents that are reopened after resolution. This metric helps in understanding the quality of the resolution. High incident open rate is an indication of the issues with the initial resolution or incomplete resolution for the underlying problem. Conducting thorough root cause analysis, improving the knowledge articles about the issues that occurred for the data prodcut will help the application support and development teams to provide quality resolutions. Identifying common issues and implementing preventive measures through active monitoring will reduce the incident re-open rate. This metric is crucial in tracking the effectiveness of the support operations team and also the customer satisfaction levels. |
| SLA Compliance Rate | This metric measures the percentage of incidents resolved within SLA time. High SLA indicates the operations |

| | |
|---|---|
| | teams is able to meet the agreed service levels of agreements which is a critical factor to customer satisfaction. This metric helps in tracking the potential areas of improvement which are causing potential breaches to the service level agreements and take necessary actions to improvise the processes and tool to improve the compliance rate. Corrective and preventive actions sshould be taken whenever service level agreement breaches happen and clear accountability should be maintained to ensure there service operations are effective. Regular reviews are to be conducted with all the associated parties to ensure the service quality and reliability. |
| RCA Coverage | This metric measures for how many number of incidents or problems do we have Root Cause Analysis (RCA) completed. This is helpful to measure whether the teams are able to identify the underlying root cause of an issue and work on a permanent resolution. High RCA indicates proactive approach to identify and address the underlying issues. This will lead to improved system reliability, reduced reccurence of the problems and enhanced overall quality. Regularly review RCA process and identify the |

| | |
|---|---|
| | patterns for the common failures and having problem tickets raised for them will help in improving the systems. Tracking RCA coverage provides insights into an organization commitment to increased quality and continuous improvement. RCA for all incidents may not be feasible , the support operations team may need to identify the priority and recurring incidents and ensure the RCA is completed for them to reduce the volume of incidents and resolve these reoccuring issues permanently. |
| Model Serving Latency | This metric measures the model serving response time. Low latency is one of the critical aspects of good user experience (Ruf et al., 2021). To optimize model latency teams should use efficient architectures, optimized code and deploy on scalable infrastructure. Longer the latency period there is a potential chance of the requests getting dropped and if not handled properly it would lead to negative user experience and in some cases it might result in unusable data products. Monitoring and analyzing the latency trends help to identify performance bottle necks and address them accordingly. Machine learning models are complex and due to its high capacity they cannot be lazy loaded and so the web servers are memory are |

| | |
|---|---|
| | consumed by the model itself. Quantization optimization methods can be used to improve the model performance with little compromise on the model accuracy. |
| Model Performance | This metric checks if the current model is good enough for the given task. Metrics like accuracy, precision, recall, false positive rate, true positive rate and confusion matrix are some metrics to assess model performance (Ruf et al., 2021). To improve the model performance, the operations and development teams needs to regularly train the models with the updated data, fine tune the hyperparamaeters and validate against the validation datasets. Continuous monitoring and evaluation help in maintaining high performance and detect any potential drifts. Tracking the model performance ensures the reliability and effectiveness of AI driven solutions which can support strategic goals and enhance operational efficiency of the data products there by improving the decision making process. |
| Bias/Fairness | The trained model could be biased towards one category resulting in higher predictions on that category creating a bias in the system. It is important to reduce the bias in the models (Ruf et al., 2021). To optimize the fairness, the |

| | |
|---|---|
| | operations teams should perform regular bias audits and use fairness metrics like disparate impact and other to minimize the biasness with the machine learning models. Tracking model biasness not only improves the model reliability but also improves the compliance with the regulations thereby improving the trust with the AI applications. This metric helps in maintaining ethical standards and preventing any unknown biases with the models. Regularly updating the models with the diverse and representative data helps in minimizing the biases and promoting fairness across segments in the data. |
| Sprint Velocity | This metric measures how much an agile team produces during their normal sprint cycle. High sprint velocity indicates the high performance of the teams and regular and consistent progress made by the team to achieve the project timelines. Teams generally should focus on the good sprint planning considering the various tasks and planned releases in the sprint and should prioritize to remove any impediments quickly. Tracking this metric provides valuable insights about team capacity, usage, timelines and also better resource allocation. Regular |

| | |
|---|---|
| | retrospectives can help in identifying the areas of improvement and increase productivity. |
| Sprint Burndown | This chart provides a visual representation of the remaining work versus time within a sprint. This helps to visualize the progress made by the team, identify potential bottle necks and ensure timely completion of the tasks. To optimize the sprint burn down, a clear definitions of the tasks, regular update of the charts and addressing the impediments will help. This metric helps in maintaining transparency, improving team accountability, and facilitate effective sprint planning and adjustments to the sprint deliverables. |
| Sprint Retrospective Completion | This activity helps in assessing what went well and what did not and assess the performance of the team members which help in planning and ensuring the team members are working to the best of their abilities. This activity will help to reflect on the teams performance, identify areas of improvement and have any actionable steps for the upcoming sprints. Having regular sprint retrospectives fosters a culture of continuous improvement, ensures to have open communications and also improve teams collaboration. This will help in teams to have a self |

| | |
|---|---|
| | assessment which will help in leading higher efficiencies and better project outcomes. |
| CI/CD Pipeline Efficiency | This metric shows the reliability of the CI/CD pipelines. These pipelines are critical for automated deployment of data products and are to be reliable as they would disturb the entire data product planning thereby affecting the organization goals. High efficient pipelines indicates streamlined and better workflows, optimized feedback loops and rapid deploment of code changes to production with the faster delivery. Teams should target for automated testing and deployment stages to reduce the time to market. Regular monitoring of the pipelines and promptly resolving any bottle necks related to the pipelines will improve the integration and deployment processes. |
| Backlog Health | This metric measures whether there are enough user stories in ready state with a cummulative number of story points greater than average velocity. This metric measures the balance between the incoming work and completed work, focuses on the prioritization of the tasks, track the aged tasks and also checks the overall deliverables are in line with the project goals and targets. Monitoring the |

| | |
|---|---|
| | backlog health will help in giving visibility to the upcoming tasks for the teams and also set expectations with the stakeholders. This activity involves regular grooming of the backlog items, updating priorities based on the business demands and addressing dependencies and blockers to maintain the speed in the delivery of the systems. |
| Resource Utilization Rate | This metric measures the effective utilization of the team members and is required to ensure the team members bandwidth is considered in both directions for planning the activities. This metric helps in understanding the bandwidth utilization and check on the engagement of the team members in the project deliverables. Re-prioritization, re-assigning of the tasks, capacity planning, scheduling the tasks effectively and increasing the speed of deliverables can be targeted with the help of this metric. This helps in maximizing the efficiency of the resources, reduce cost and overall improve project performance and deliverables. |
| Data Product Adoption | This metric shows the growth in the user base for the application. It is critical for managements to make the decision whether the data product is reaping the intended |

| | |
|---|---|
| | benfits or not. Low adoption rate data products have fundamental issues, they would require a revisit to fix the problem or they would need to be removed from the system. Adoption of the data product can be increased by effectively marketing the data product to the end users, offering training in the local languages, and explaining the benefits of its usage to the end users. It is also important to have a leadership level alignment to ensure their respective teams can actively use the data product and have improved opertational efficiency to have a competitive advantage. Any challenges related to the adoption can be identified for the products with low adoption rate and the respective teams should take necessary measure to improve the adoption. |
| Compliance and Governance Adherance | This metric measures the adherence to the compliance and regulatory requirements of the data product. It is critical in organizations like pharma where adherence to these principals is critical to ensure all the required practices are followed without failure as it is tightly related to the safety of the systems and users. Tracking this metric helps in mitigating the legal and reputational risks associated with the business process. Having clear policies and |

| | |
|---|---|
| | gudielines, training the team members on these policies, regular monitroing and auditing of these policies, implementing risk management practices and continuous improvement will ensure adherence to the compliance requirement. This metric will ensure if the organization aligns with the legal and compliance requirements, industry standards, internal policies and ethical standards that needs to be followed as part of the business process. |
| Data Product Down Time | This metric measures the average down time of a data product over a period of time. It is critical to ensure the reliability of the data product for customers. Minimizing the down time is crucial for ensuring less business disruptions, higher data accessibility and meet the service level agreements. Having fail over mechanisms or ssytems for the applications which are crtical for the organizations would lessen the impact. Setting up monitoring and alerting systems and automated recovery processes and disaster recovery planning systems can ensure the impact due to the application down time on the end users. If a data product has higher down times then effective measures should be taken for maintenance of the systems to reduce these downtimes. Identifying opportunities in improving the |

| | |
|---|---|
| | infrastructure, processes and resposne proceedures help in reducing the disruptions in the business operations due to these down times. This metric is essential to improve customer satisfaction and also have a competitive advantage in data driven environments. |

Table 2.1 KPI Metrics definitions

Organizations can achieve significant improvements in their data product development lifecycle by measuring and tracking the critical KPIs required for the project. Enhancements based on the insights can lead to better decision-making and improved customer outcomes, ultimately driving business success.

# CHAPTER 3 :

# METHODOLOGY

## 3.1 Overview of the Research Problem

Informed decision making based on data has been critical for all organizations in this modern era. Data products development and delivery has been the utmost importance. Challenges with the development and delivery of the data products has need to be resolved on priority to squeeze the benefits of a data product being in production. The current practices of DataOps and MLOps are helping organizations in following the best practices that can accelerate the data product developments. However, there is a requirement for identifying the KPI metrics that can measure the progress and give insights to organizations to take course corrections if required. Organizations would need to prioritize and focus only on the relevant KPIs that can deliver high impact and business value. This research is aimed to provide insights on to the importance of the KPIs and the insights they provide to organizations in ensuring the timely delivery and effective utilization of data products.

## 3.2 Research Design

The research design for this study adopted mixed methods approach, incorporating both quantitative survey method and qualitative literature review. This approach enabled to do a comprehensive exploration on the key performance indicators for business alignment during the data product development.

A systematic review of the literature was conducted to identify the key concepts, challenges, benefits related to KPI governance. The literature review helped in understanding the research topic, identifying gaps and areas in this topic which require exploration and also in development of survey questionnaire.

A structured survey is developed based on the insights and understanding from the literature review. The survey is designed to collect both quantitative and qualitative data on the importance of the KPIs in the data product development from diverse set of industry professionals working in the areas of data engineering and machine learning products. The survey included closed ended questions to gather the quantitative data on the types of KPIs used in the industry and the open-ended questions are included to collect qualitative insights, experiences and recommendations from the industry experts. Industry experts where directly reached to collect the data for the survey questions.

### 3.3 Population and Sample

This research has focused on two approaches, qualitiative and quantitative approaches. The qualitative approach has focused on analyzing 62 research papers and based on the quality of the content and the information that supports this research reduced to 40 research papers on the topics of DataOps and MLOps. The quantitative research targeted 30 industry experts, however the responses received was from 15 experts with good experience in these domain. Out of the 15 participants, 7 were product managers of different senior levels who are managing the data product delivery and development, 5 were Data Engineers, 1 Delivery Managers and 2 ML Engineers. The selection of the

participants has been chosen to ensure to have diversity in technologies and product management skills. Each participant has been explained about the survey process, intention of the survey and also taken their consent for the survey.

### 3.4 Participant Selection

The participants for the quantitative study where primarily selected from the personal network comprising people with industry knowledge and their acquaintances who are engaged in this domain. The participants where primarily with data science and data engineering domain in Healthcare and Finance domains. On average, the participants had over 13 years of industry experience, ensuring that the data collected was from experienced professionals who could offer deep insights into the challenges and opportunities within these fields. Their extensive experience enabled them to provide nuanced perspectives on data operations, governance, and analytics, which are crucial for the study.

The focus was to gather quality data from people who not only possessed substantial industry experience but also had a proven track record of working on data-related projects. This selection criterion was essential to ensure that the feedback and insights were both relevant and actionable. By targeting individuals who are actively engaged in the field, the study aimed to capture the latest trends, practices, and pain points that these professionals encounter in their day-to-day work. The diverse perspectives brought by participants from Healthcare and Finance sectors has added a layer of depth to the research. The sectors have different data privacy concerns, regulatory environments and operational challenges, providing a diverse view of the data landscape. The inclusion of these different

perspectives helped in identifying common themes and unique challenges which has been useful for the research.

## 3.5 Instrumentation

The quantitative survey consisted of 9 survey questions to gather information on the importance of 46 KPIs which have been asked to rate them on a numeric scale where numeric 1 represents high priority and the priority decreases as the number value increases. There are 2 questions asking about the maturity of the DataOps and MLOps practices in their current projects. There are another 2 questions about the user experience and their current role.

The acceptance scale for the KPI survey questions were closed ended questions on a numeric scale designed on the principles followed from the research "Questionnaire Designing for a Survey" (Roopa and Rani, 2012).

The KPI related questions are grouped into the following categories

a) Data Quality Metrics
    a. Data Accuracy
    b. Data Completeness
    c. Data Consistency
    d. Data Timeliness
    e. Data Integrity
    f. Data Drift Detection
    g. Schema Shift Rate

b) Data Operations Efficiency Metrics
    a. Data Ingestion Time
    b. Data Processing Time
    c. Pipeline Uptime
    d. Data Throughput
    e. Error Rate
c) Deployment Performance Metrics
    a. Deployment Frequency
    b. Deployment Success Rate
    c. Testing Coverage
    d. Deployment Downtime
d) Product Development Efficiency Metrics
    a. Time to Market
    b. Number of Releases
    c. Development Cycle Time
    d. Defect Rate
    e. Customer Feedback Response
e) Support Operations Performance Metrics
    a. Median Response Time to Incidents
    b. Median Response Time for Resolution
    c. Incident Reopen rate
    d. SLA Compliance Rate

    e. RCA Coverage

f) Data Observability Metrics

    a. Data Availability

    b. Data Freshness

    c. Data Quality Score

    d. Mean Time To Detect Data Errors

    e. Data Deduplication ratio

g) Machine Learning Model Performance Metrics

    a. Model Accuracy

    b. Model Latency

    c. Model Drift Detection

    d. Model Versioning

h) Agile Development Metrics

    a. Sprint Velocity

    b. Sprint Burndown Chart

    c. Sprint Retrospective Completion

    d. CI/CD Pipeline Efficiency

    e. Backlog Health

    f. Resource Utilization Rate

i) Data Product Performance Metrics

    a. Time to Market

    b. Data Product Adoption

c. Customer Satisfaction

d. Compliance and Governance Adherence

e. Data Product Down Time

All the survey participants were asked to arrange them in the order of their importance using numbers 1 to 5 based on their experience and understanding in the projects. Number 1 signifies highest importance while 5 signifies lower importance.

### 3.6 Data Collection Procedures

For the quantitative analysis, 12 participants have been interviewed directly by meeting in person and 3 of them have been consulted via phone. Each meeting lasted for 30 to 45 mins and their responses where recorded against the survey questions in an excel file. It has given a chance to understand the level of technical expertise of the respondeds and also understand their view point of their answer for the key performance indicators (KPIs). The interview was structured to delve into the practical experience of the experts to get deeper insights into the view points of the technical experts.

The qualitative analysis on the research papers have helped in identifying the important technical factors involved in data product development and helped in identifying the key factors such as data quality management, algorithm section, scalability considerations and integration challenges. These were instrumental in formulating the key performance indicators related to the data product performance and usability.

## 3.7 Data Analysis

The responses from the experts has been analysed manually to understand the highest rated KPIs in each category of KPI segment using an excel file. The respondents questions were also analysed based on the roles they hold to check if the preference of choosing a KPI is influenced based on their roles. However, the results have been analysed and published using the overall ratings as the role level ratings may not be significant as the survey was conducted on a small set of audience who are expert in these domains.

## 3.8 Research Design Limitations

In the exploration of the Key Performance Indicators, its crucial to list the following potential limitations which were there in the research design:

a) Sample size constraints: It is a challenge to get enough representations from different organizations to participate in the study. So, the representations that are presented in the research are primarily based on the view presented by the members participated in the survey.

b) Resource limitations: Constraints on the time, budget and resources may limit the depth and breadth of the data collection and analysis. Certains aspects of the KPI governance may have not received comprehensive coverage in the research.

c) Self-reporting bias: There is a likely chance to introduce an unknown bias into the respondeds regarding their KPIs as they may overstate or understate their adherences to the best practices. Since the responses are based on their

exposure into the system, not having a 360 degree view of the system could be a reason for having biases with in the responses of the participants.

d) Contextual factors: Variations in organizational culture, industry norms, regulatory environment and operational model could influence the interprestation and generalizability of the findings. Some organizations would have preference towards certain tools and technologies due to their vendor level agreements and also due to their internal decisions made at the organization level. Addressing these issues would require additional resources, time and methodological considerations which would be beyond the scope of this research.

e) Organization maturity: The views and the details shared by the industry experts is limited to the industry they operate and the maturity of their current organizations. Each experts opinion is driven by the unique challenges and opportunities faced by the organization, as well as exisiting processes and technological sophistication of the organization. Experts from matured organization might stress the importance of advanced data integration techniques and highly sphisticated analytical capabilities, while experts from less matured organization might focus on the foundational capabilities of the data integration and their initial steps towards digitization of their data products. The variability in the opinions of the experts needs to be contextualized in the operating organization context and accurately understand and interpret the insights shared by them.

f) External validity: The applicability of this research is limited to the constraints this research has been conducted. Applicability of this research outside the research setup would need to be used with caution.

Recognizing these limitations is required to understand the context in which the research has been conducted. The validity and reliability of the study is primarily within the above constraints and it needs to be applied with caution if it needs to be applied with caution. However, this study focused to generalize the concepts to ensure it can be applied across organizational context and industries.

## 3.9 Conclusion

This chapter has discussed about the research design that had been followed in this study which discussed on the popultation, sampling proceedure and data collection proceedures. Research design limitations have also been discussed in this discussion.

# CHAPTER 4 :

# RESULTS

This section describes the outcome of the work that was done to identify and address the research questions that was aimed for this research. The process involved a detailed and extensive literature review, quantitative surveys, and qualitative interviews with industry experts. The goal was to gather a wide range of data and insights to thoroughly understand the challenges and opportunities within the field of key performance indicators for measuring data products maturity.

## 5.1 Research Question One

The research question one is targeted to understand the key performance indicators that can be help in effectively governing the DataOps and MLOps principles for a successful and seamless delivery of quality data products.

The effective governance of DataOps and MLOps principles is of utmost importance for organizations and it has been agreed with the industry experts from the survey responses. In this research we have understood the importance of these KPIs and how these are to be tracked to ensure successful outcomes.

Through a survey conducted with industry experts supported through literature review, the research gained insights into the pirority KPIs that drive the governance of DataOps and MLOps principles. This study focused on identifying the important KPIs across the dimensions in data product development which includes data quality, agile

practices, data pipeline monitoring and efficiency, machine learning model development, customer satisfcation and compliance.

High quality data is the fundamental to the success of DataOps initiatives. Based on survey responses represented in Fig 4.1, KPIs that belong data accuracy and data completeness are the top two important KPIs from a data quality standpoint. Data Engineers prioritized the metrics such as data accuracy, completeness, consistency and timeliness in sequence as critical to ensure the reliability of the data products. This reflects the essential aspects of data management that directly impact the day-to-day operations and decision-making processes.

ML Engineers focus was on metrics like drift detection and model accuracy which is obvious as the machine learning development is highly impacted with drift in data and data quality can be to an extent is sufficient for the model development. These metrics are crucial for maintaining the performance and reliability of machine learning models over time. It should be understood that model retraining schedules are often dictated by the detection of data drift, making it a vital metric for ML engineers.

The product managers focus was primarily on data accuracy, completeness and consistency. These KPIs are essential for ensuring that the data products meet user expectations and support business objectives. Product managers often act as a bridge between technical teams and business stakeholders, and therefore their emphasis on data quality aligns with overall business goals.Schema drift KPI which tracks the number of times the schema has impacted is of low importance which is meaningful as this occurs once in a while and is not of top concern ompared to other metrics. However, it's important

to monitor schema drift periodically to avoid any potential disruptions in data workflows. By focusing on these critical KPIs, organizations can better support their DataOps and MLOps initiatives, leading to more effective data-driven decision-making and improved business outcomes.

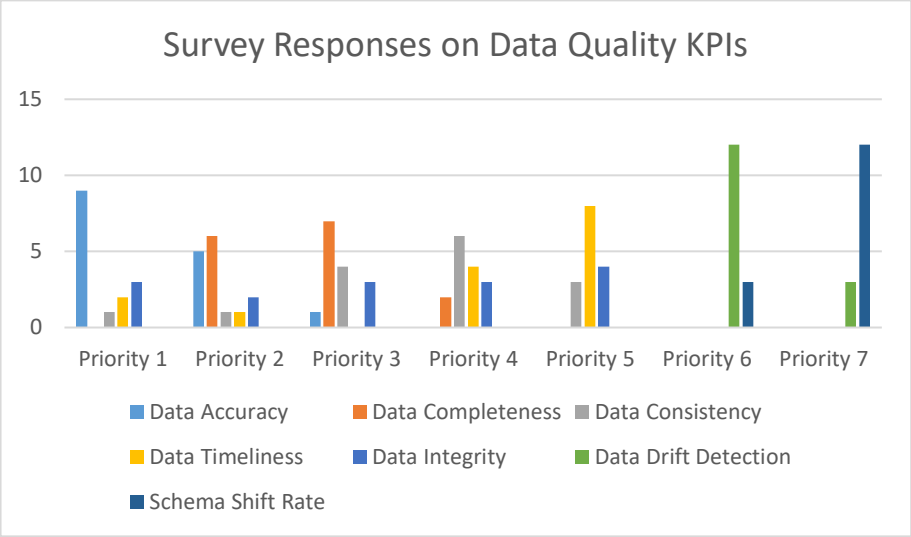


Figure 4.1 Survey Responses on Data Quality KPIs

In the product development metrics it can be noticed from Fig 4.2 that the experts opinioned that the Time to Market, Number of Releases and Development Cycle Time are the key metrics which can be understood because of their direct impact on the efficiency of the product development process and help in managing the development timelines effectively.

Time to Market is essential as it measures the duration from the initial concept to the final product launch, reflecting the organization's agility and ability to respond to market demands. A shorter Time to Market can provide a competitive advantage by allowing companies to capitalize on emerging trends and customer needs more quickly.

The Number of Releases metric indicates how frequently updates or new versions of the product are made available. This is crucial for maintaining customer engagement and satisfaction, as regular updates can introduce new features, fix bugs, and improve performance. A higher frequency of releases typically suggests a more responsive and adaptive development process.

Development Cycle Time, on the other hand, measures the total time taken to complete a single development cycle, from planning to deployment. This metric is vital for identifying bottlenecks and inefficiencies within the development process. By analyzing Development Cycle Time, organizations can streamline workflows, allocate resources more effectively, and enhance overall productivity.

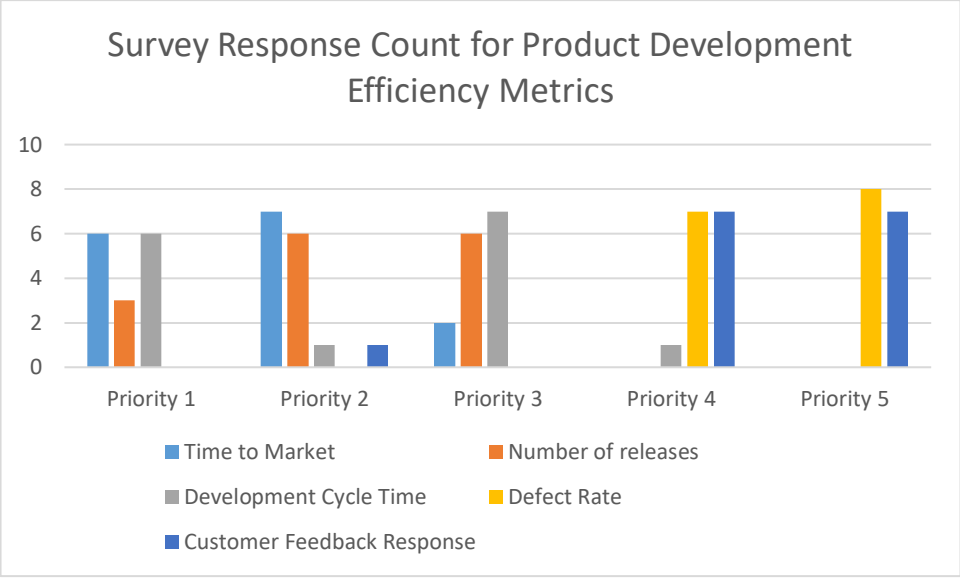


Figure 4.2 Survey Responses for Product Development Efficiency Metrics

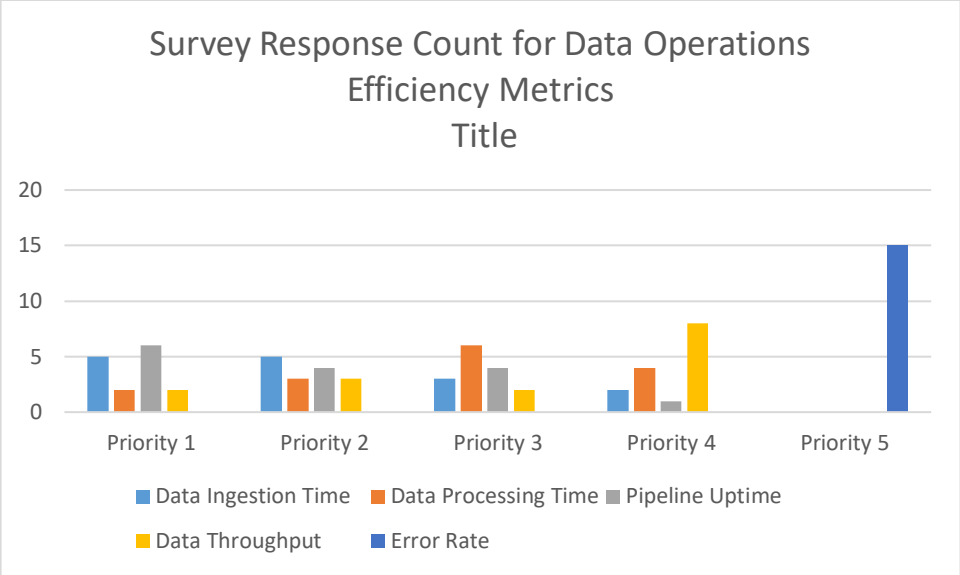


Figure 4.3 Survey Responses for Data Operations Efficiency Metrics

In data operations metrics, data ingestion time and pipeline uptime are critical as per the survey. This can be understood because of their pivotal role in ensuring the timely flow of data. Efficient data ingestion reduces delays in data processing and analysis directly impacting the overall responsiveness and agility of data-driven operations. By minimizing the time taken to ingest data, organizations can accelerate the availability of fresh data for analysis, enabling quicker insights and more timely decision-making. High pipeline uptime ensures uninterrupted data availability which are essential for maintaining operational efficiency and supporting timely decision-making processes. It can also be understood that error rate and data throughput are comparitively not of that significant. Continuous data flow without interruptions is essential for maintaining operational efficiency, as any downtime can disrupt data processing, delay analysis, and impair the decision-making process. Reliable pipeline uptime supports the consistent availability of data, which is

critical for applications that require real-time or near-real-time data, such as monitoring systems, predictive analytics, and automated decision-making processes.

Error rate and data throughput were comparatively less significant by survey respondents. While these metrics are still important, they do not have as immediate or critical an impact as data ingestion time and pipeline uptime. Error rate, which measures the frequency of errors during data processing, is essential for maintaining data quality but can often be managed through robust error-handling mechanisms and data validation steps. Data throughput, which measures the volume of data processed over time, is important for understanding the capacity and efficiency of data pipelines but may be less critical in environments where timely data availability and uninterrupted processing are the primary concerns.

Data ingestion time and pipeline uptime directly influence other aspects of data operations, such as data accuracy, completeness, and consistency. By ensuring that data is ingested quickly and pipelines remain operational, organizations can maintain high standards of data quality and reliability. This, in turn, supports the broader goals of DataOps and MLOps initiatives, fostering an environment where data-driven insights can be generated efficiently and effectively.

Overall, the focus on data ingestion time and pipeline uptime reflects a prioritization of metrics that ensure the smooth and timely flow of data, which is fundamental for maintaining operational efficiency and supporting timely decision-making processes. This approach aligns with the need to keep data operations agile and responsive in the face of evolving business demands and data-driven opportunities.

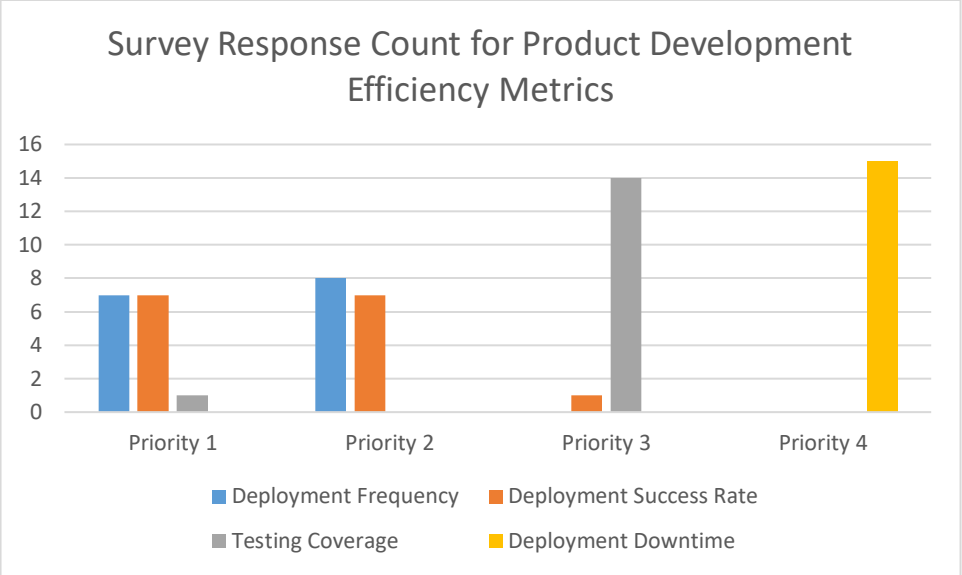


Figure 4.4 Survey Responses for Product Development Efficiency Metrics

From Fig 4.4, it can be understood from the survey results that deployment frequency and deployment success rate are of high importance, while test coverage and deployment downtime are comparatively less significant. It is evident that a high deployment rate facilitates rapid iteration and releases, allowing organizations to respond quickly to market demands and customer feedback. This agility is crucial in today's fast-paced business environment, where being able to deploy updates and new features swiftly can provide a competitive edge.

The deployment success rate is another critical metric, as it determines the quality of the deployments. A high success ratio is directly correlated with deployment frequency, as frequent deployments with a high success rate indicate a well-oiled, reliable deployment process. This metric ensures that deployments are not only frequent but also stable and

reliable, minimizing the risk of introducing errors or downtime into the production environment.

The comparatively lower significance placed on test coverage suggests that organizations may be prioritizing successful deployments by adopting a risk-based testing approach. This approach focuses on testing the most critical aspects of the application to ensure stability and functionality while keeping testing efforts minimal and efficient. By concentrating on high-risk areas and essential functionalities, organizations can streamline their testing processes, reduce time-to-market, and still maintain a high level of deployment success.

Deployment downtime, while still important, is seen as less significant compared to deployment frequency and success rate. This indicates that organizations might be more concerned with maintaining a steady flow of updates and ensuring that each deployment is successful, rather than minimizing the time taken for each deployment. Minimizing deployment downtime is still a valuable goal, but it appears to take a back seat to ensuring that deployments are frequent and successful.

Overall, the emphasis on deployment frequency and success rate reflects a strategic focus on agility and reliability. By iterating quickly and ensuring high-quality deployments, organizations can adapt more swiftly to changing market conditions and customer needs. The risk-based testing approach further supports this by optimizing testing efforts and focusing resources on areas that are most likely to impact deployment success. This balanced approach allows organizations to achieve a high deployment frequency and success rate while managing testing and deployment downtime efficiently.

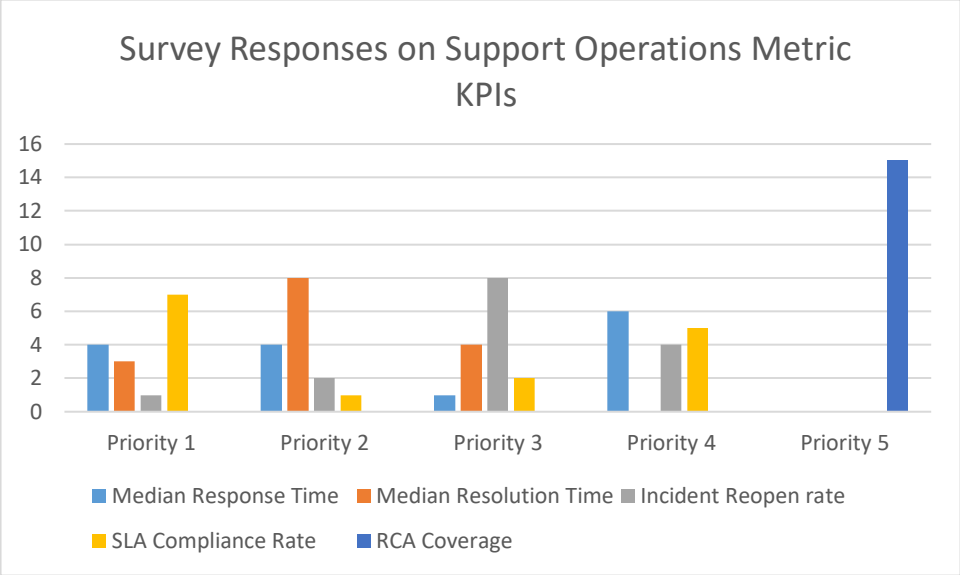


Figure 4.5 Survey Responses on Support Operations Metric KPIs

Production Support process is another critical area in the context of DataOps. It encompasses several key aspects essential for maintaining the stability and performance of data products. Proactive monitoring is vital, as it helps in the early detection of potential issues, enabling teams to address problems before they impact users. Automated response and resolution methods are equally important, as they streamline the process of fixing issues, reducing downtime, and minimizing the need for manual intervention.

Cross-team collaboration is a cornerstone of effective production support. It ensures that various teams can work together seamlessly to resolve issues, share knowledge, and implement best practices. This collaboration is crucial for addressing complex problems that require input from different areas of expertise. Additionally, resilience and reliability are fundamental to maintaining infrastructure stability. Ensuring that systems can

withstand and recover from failures quickly is essential for delivering continuous service to users.

Service Level Agreement (SLA) compliance rate and median resolution time are identified as the key metrics in this area. SLA compliance ensures that the agreed timelines for issue resolution are met, which helps build trust with customers and enhances the reliability of the services provided. It is a measure of how well the production support team meets its commitments, reflecting the overall effectiveness of the support process.

A low median resolution time indicates prompt handling and resolution of issues, showcasing the efficiency of the production support team. This metric is crucial as it demonstrates the team's ability to quickly restore normal operations, minimizing the impact of disruptions on users. Efficient issue handling not only improves user satisfaction but also reduces the risk of prolonged outages and the associated business impact.

Together, these metrics provide a comprehensive view of the production support process's effectiveness. They highlight the team's ability to deliver timely and reliable support for data products, ensuring that any issues that arise are swiftly addressed and resolved. By focusing on proactive monitoring, automated response, cross-team collaboration, resilience, and reliability, organizations can maintain high standards of service and ensure the continuous stability of their data infrastructure.

In summary, the Production Support process in DataOps is vital for ensuring the ongoing stability and reliability of data products. Key metrics such as SLA compliance rate and median resolution time provide insights into the effectiveness of the support process. By prioritizing proactive monitoring, automated response, cross-team collaboration, and

resilience, organizations can deliver timely and reliable support, maintaining high levels of customer trust and service reliability.

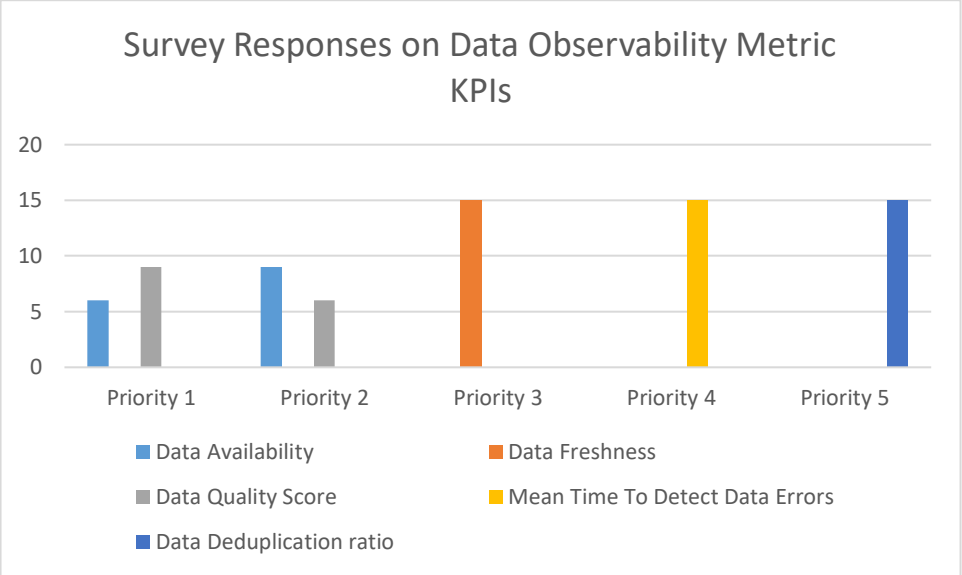


Figure 4.6 Survey Responses on Data Observability Metric KPIs

From Fig 4.6, the survey respondents prioritized data availability and data quality score over data freshness, data de-duplication ratio and mean time to detect errors. This can be understood that the top priority is to ensure the quality data availabile at all times to data products which is critical for any data product to operate.

In the agile metrics from fig 4.7, the respondents have chosen backlog health as the priority over the other KPIs. Without a healthy backlog items, it would not be possible to plan the workload, task prioritization and delivery. In my opinion, the sprint velocity and CI/CD pipeline efficiency are top factors as it helps in measuring the productivity and efficiency of the development teams and deployment processes. It provides a great insight

into t eam performance, capacity planning and project forecasting. Tracking resource utilization is of lower importance wich could be due to the efficiencies are already tracked in the sprint velocity and burndown charts.

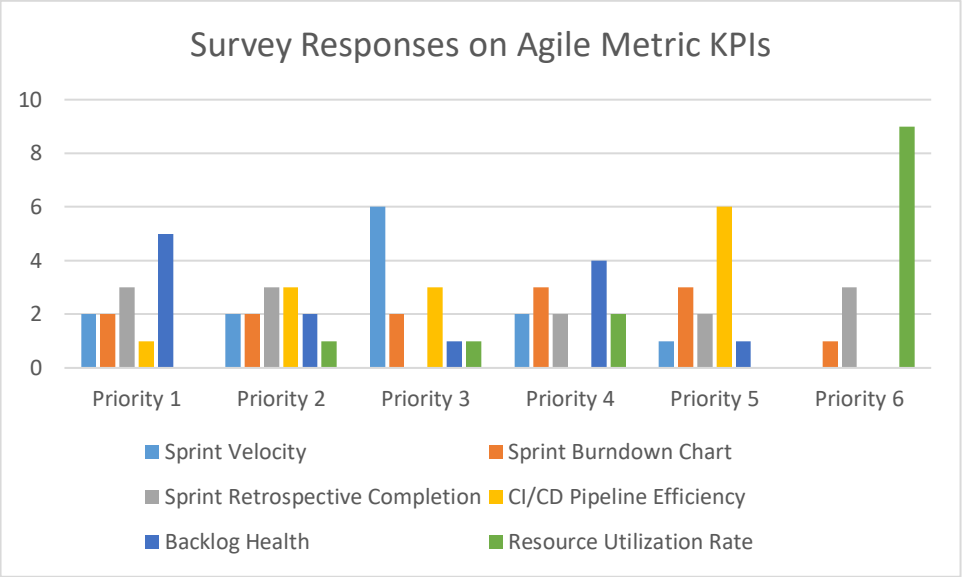


Figure 4.7 Survey Responses on Agile Metric KPIs

From fig 4.8, customer satisfaction is a top factor from survey respondents stand point which is obvious as the success of the data producy is dependent on the end user usability and adoption to his business processes. Data product adoption is another top factor as the organizations can measure the success of data prodcut through its user adoptions. Data product down time is another factor as the customer satisfaction and adoption are directly correlated to this metric so it needs to maintained at lower levels for the success of any data product.

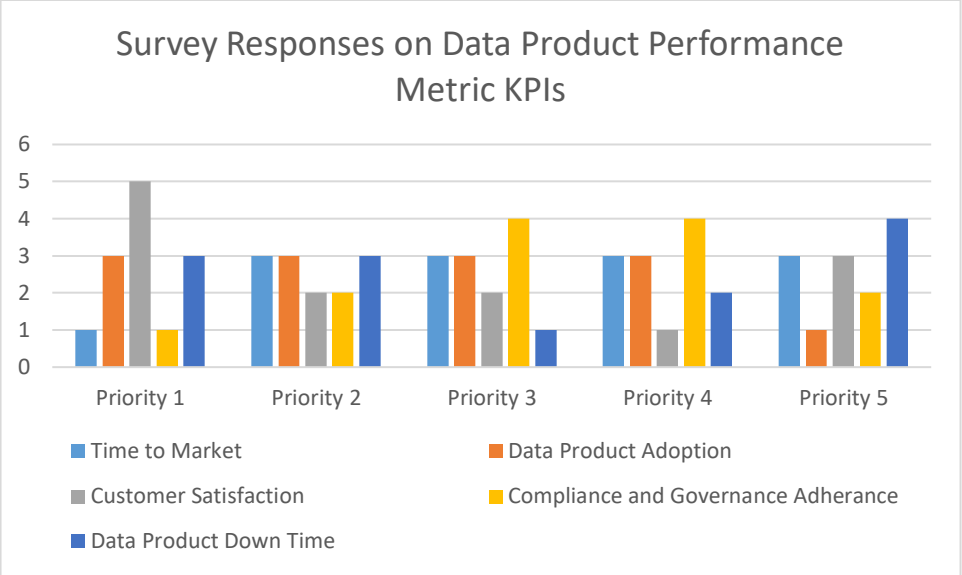


Figure 4.8 Survey Responses on Data Product Performance Metric KPIs

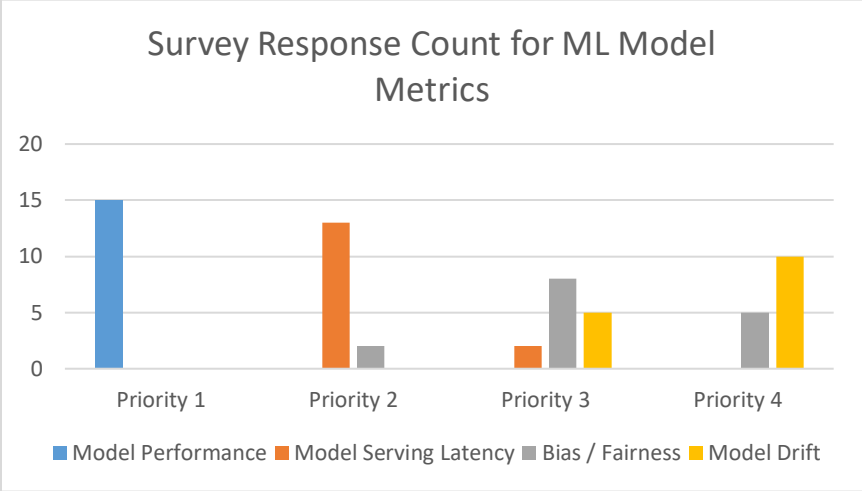


Figure 4.9 Survey Response Count for ML Model Metrics

Fig 4.9 represents the survey responses from the ML metrics. It can be noticed that all the participants acknowledged that model performance is the top factor among the other metrics. Model serving latency is the next critical factor which could be because of higher latency will have a negative impact on the user experience with the data product. Bias and

Fairness is critical in terms of complaince and ethical requirements and it needs to be ensured that all models are bias free. Model drift is a conceptual drift and is not generally noticed regularly unless there is a major changes in the business process or factors that are influencing the business process.

## 5.2 Research Question Two

Research question two is targeted in understanding the technical and non-technical challenges involved and the current practices followed for implementation of DataOps and MLOps principles in developing data products.

Based on the literature review and survey conducted with the experts in the organizations, there are diverse technical and non-technical challenges involved in implementing DataOps and MLOps principles in developing data products. The practices followed by organizations to address these challenges are varying depending the organization maturity, business requirements and data integration maturity between business functions.

One of the major challenges for organizations is ensuring data quality and governance throughout the data life cycle. From the survey it is understood that internal tools like data quality framework have been built to monitor and provide data correction rules for automated data correction to ensure the quality of the data is up to date. Data quality monitoring dashboards are also in place to monitor and alert to proactively identify and address issues.

Ensuring data pipelines built are stable and do not create failures on a daily basis is another important factor for a successful data product. Integration of dashboards to monitor the pipeline performance on a day-to-day basis and also with automated recovery and alerting mechanisms are being developed for reducing the pipeline failures. From the survey, experts shared their opinion that with the use of cloud-based infrastructure, applications can scale up or down the infrastructure based on the needs. Deployment of the infrastructure using infrastructure-as-code practices with right processes would eliminate the dependency with the teams and also will accelerate the development process. Experts also mentioned use of CI/CD practices which are capable of automated testing and deployment is also helping to reduce the time to market for a data product.

Establishing processes and practices to ensure smooth collaboration between the teams and developing knowledge sharing platforms for easy access of information help the teams to improve and reduce the delays in cross communications. Upskilling the employees to address skill gaps with the employees will also help the teams to mature in understanding the importance of the practices.

In summary, organizations are addressing challenges related to data quality, pipeline stability, infrastructure management, and deployment efficiency through the adoption of various tools, practices, and technologies. These efforts aim to improve the reliability, scalability, and agility of data product development processes, ultimately enhancing the organization's ability to deliver high-quality data products efficiently.

## 5.3 Research Question Three

What best practices that are being currently used for successful implementation of DevOps can be adapted to DataOps and MLOps?

From the survey results and industry insights, the following best practices from DevOps that can be leveraged for the successful implementation of DataOps and MLOps.

**a)** Adopting CI/CD practices that automates the testingm integration and depoyment of code is a critical factor to reduce the time involved in the development and deployment of applications

**b)** Automated infrastructure as code and code versioning practices from DevOps will accelerate the data product development. These practices help in improved collaboration, reduce cross team dependencies between the team

**c)** Automated monitoring and alerting system enables proactive identification of failures and alerting users during failures. All data pipelines should be integrated with automated email alerting mechansim and also automated incident raising mechanism will help in getting immediate attention and reduce down time for the applications. Data quality monitoring tools will help in automated monitoring of the data quality checks on the data being loaded on daily basis, this wwill esnure the confidence on the data to the end users. Automated monitoring of the model drifts will help users to reduce the risk in model degraded performances and the data quality

monitoring will help in reduction of the data drift scenarios thereby improving the reliability of the applications.

**d)** Challenges related to the cross functional teams working can be improved with the setting process and practices. Each team should have a clear boundary on the roles and responsibilities they own and there should be accounability within the teams and the organizatons should promote shared responsibility. Also, organizations promoting data democratization would need to ensure there are processes on data ownership and accountability are in place for effective data sharing and responsible utilization of data. All data products should have clear support ownership and should align with the organization level data modeling teams to avoid duplication in data and also ensure the standards are maintained across teams.

**e)** Automation which is a priority practice from DevOps should be the focus for all teams. Automated pipeline orchestration, model training, model deployment, data product deployment and monitoring process should be adopted to reduce manual efforts and reduce errors.

To summarize, the adoption of best practices from DevOps into DataOps and MLOps will improve the reliability and efficiency of the data products. These approaches provide a structure framework to bring in efficiencies and reducing the risks to deliver tangible benefits to organizations.

## 5.4 Research Question Four

Research question four is targeted to measure the KPIs and help the organizations using these KPI to mature.

KPIs are the important indicators to understand if the DataOps practices and MLOps practices are meeting the targets set by their organization. All organization would need these KPIs to be at their respective max level which is an ideally state and would be difficult to achieve. Trageting the highest level of maturity can help organizations to have automated systems that can scale and have process efficiencies minimizing the inder dependencies between the cross functional teams. Organizations should try to ensure the practices and processes that are established work in the direction to reach the max level of each KPI to achieve success and have a competitive advantage. There is no single threshold that a KPI can have, depending on the organization and business process the thresholds for each of the KPIs would differ and so the industry standards have not been clearly defined in the literatures. The following base lines for each of the KPIs are proposed which is based on the extensive literature study done and from my personal experiences in the industry. These metrics are just a baseline and organizations are flexible to adjust them based on their current organization maturity levels. From personal experience, any huge variation on the negative side from the KPI base line mentioned in the Table 4.1 requires rethink on the organization strategy for those specific areas to achieve an efficient state.

| Metric | Threshold | Justification |
|---|---|---|
| Data Accuracy | ≥ 95% | High accuracy ensures data reliability and maintain credibility with the applications. Maintaining data accuracy at or above 95% is critical for ensuring the reliability of the data used in the business applications. Additionally, achieving these targets supports any compliance and regulatory standards. It also improves organizations reputation for data integrity. |
| Data Completeness | ≥ 98% | Incomplete data could result in biased conclusion, less accurate predictions which would impact business decisions. Ensuring the data completeness at above 98% is crucial for generating unbiased insights. Incomplete data can lead to significant gaps in analysis and biased and inaccurate predictions. High data completeness supports accurate reporting which helps in effective business operations and strategic planning. |
| Data Consistency | ≥ 90% | Data should be consistent across sources, else results in conflicting insights and unreliable insights. Ensuring the data consistency at or above >90% is essential for maintaining integrity and reliability of the applications. This helps in minimizing the discrepancies and errors, ensuring seamless data integration and improving the overall quality of the business reports. |

| | | |
|---|---|---|
| Data Timeliness | ≤ 24 hours | Organizations should target for real time data refreshes at-least daily refreshes are required for timely decision making. |
| Data Integrity | ≥ 99% | Data needs to be reliable, if not maintained it will result in flawed reports. |
| Data Drift Detection | ≤ 5% drift rate | Data drift should be kept at low else the model performance would be degraded. |
| Schema Shift Rate | ≤ 2% shift rate | Schema shift should be avoided to ensure stability of the data pipelines. |
| Time to Market | <60 days | Rapid development and release to the market of a minimal viable product within 60 days is critical to gain a market advantage. |
| Number of releases | >=1 per month | Smaller releases would show continuous improvements and user features released in the market for faster validation by users. |
| Development Cycle Time | < 30 days | Delayed market releases and lost opportunities can be avoided. |
| Defect Rate | < 5% | Defects to be at low for quality data products. |
| Customer Feedback Response | >90% | Better satisfaction means greater product. |
| Data Ingestion Time | < 3 days | Organizations should target to ingest within less than 3 days from the time of request to PROD. It should be automated as much as possible. |
| Data Processing Time | <5 mins/ job | Pipelines should be optimized to efficiently run. |
| Pipeline Uptime | >95% | Unavailability of a data pipeline results in loss of data processing causing delays in report refreshes which should be avoided. |

| | | |
|---|---|---|
| Data Throughput | >1000 records / second | Infrastructure and pipelines should be able to scale to these levels of data processing. |
| Error Rate | <5% | Error rate should be kept at minimal to avoid data discrepancies. |
| Deployment Frequency | 1 per sprint | Frequent deployments in Dev and Test environments will help organizations to rapidly test and provide user feedbacks and acceptance. |
| Deployment Success Rate | >90% | Provides confidence on the team's ability and code quality. |
| Testing Coverage | >90% | Testing should be automated and the coverage should be achieved to ensure systems are fully tested and defects are kept at minimal. |
| Deployment Down Time | < 2hours / month | Deployment should be carried during non-business hours and down time should be kept at minimal to avoid business disruptions. |
| Median Response Time | < 5mins | Response time should be targeted to be less than 5 mins to ensure good user experience and also provide immediate attention to users. |
| Median Resolution Time | SLA to be met | Lesser the resolution time greater the experience. |
| Incident Reopen Rate | <10% | Quality of the incident solution should be high so that the solutions are effective and the issues re-occurrences are avoided. |
| SLA Compliance Rate | >90% | Organizations should target is met to ensure good experience to users which also improves trust and credibility with users. |

| | | |
|---|---|---|
| RCA Coverage | >80% | RCA analysis should be completed for issues to ensure the root cause is found and mitigated, so that the issues do not repeat. |
| Model Serving Latency | <50 milliseconds | Model responses should be delivered at a faster rate to provide good experience to user. |
| Model Performance | >85% accuracy | Models depending on the use cases should at-least target 85% accuracy. >95% would be an ideal state. |
| Bias/Fairness | <5 % Disparity | Disparity should be avoided to ensure the ethical policies are met. |
| Sprint Velocity | 30 story points / person /sprint | Team's efficiency to be maintained to ensure delivery is on track. |
| Sprint Burndown | Linear trend | High risk items should be targeted first and by the time sprint closes all the activities should be closed. |
| Sprint Retrospective Completion | 100% | Every sprint should be followed by a retrospective meeting to ensure all hurdles and challenges are faced are mitigated in the next releases. |
| CI/CD Pipeline Efficiency | >90% success | Automated deployment pipelines bring efficiency and agility. So, it should always target for >90%. |
| Backlog Health | <5% aging issues | Less aging issues ensures the backlogs are actively maintained and managed. |
| Resource Utilization Rate | >80% | Sprint planning should be targeted to ensure there is sufficient activities planned for each team member and can deliver value. |

| | | |
|---|---|---|
| Data Product Adoption | >4x CAPEX and OPEX costs | Data product adoption should be maximized, it should provide a potential 4X times of the actual development and maintenance costs. |
| Compliance and Governance Adherance | 100% | It is critical measurement, as failing to meet these would result in heavy penalties and also would affect company reputation and sales. Organizations should have controls and regular audits to ensure the product teams are compliant and teams can identify their gaps in the processes and mitigate before they are identified during the audits. |
| Data Product Down Time | < 2hours / month | Target should be zero hours as business disruptions would be caused if there is a downtime. Having observability systems that can track the up time and downtime of the applications and alert users when something goes down would help in achieving these targets. |

Table 4.1Proposed KPI baseline metrics

## 5.4 Conclusion

In conclusion, the results section has helped in understanding the various technical and non-technical challenges and the importance of KPI metrics in an organization. This research has also provided valuable insights about the complexities involved in data decision making. By addressing challenges with data pipelines, quality, stability, deployment practices and team collaboration, organizations can streamline their data product development process to achieve organizational goals. It is critical for organizations

to have key performance indicators identified and baseline established to have actionable insights and monitor their efficiency levels. It is essential for organizations to adopt industry best practices and evolving technologies with a focus on continuous improvement to innovate and accelerate decision making.

# CHAPTER 5 :

# DISCUSSION

## 5.1 Discussion of Results

This section discusses on the results obtained from the research which focused on the research questions on the implementation of DataOps and MLOps principles in data product development. The study has discussed about the various technical and non technical challenges and the importance of KPIs in measuring the standards of the organization. It followed a mixed-method approach incorporating literature review and expert inteviews and this section discusses on the results obtained from the research. Overall the discussion presented in this section discusses the observations and its importance.

## 5.2 Discussion of Research Question One

Research problem discussed on the survey results on the various KPI metrics and their order of preferences for an organization based on participants experiences in their projects and organizations. It is noticed that the KPI priorities are subjective to change based on the role and level in which the person works. But it can be clearly understood from the results that the data quality metrics are of importance which is acknowledged by all the experts during the survey. Data being at the core for any data product, maintaining the data quality becomes crucial for a data product to be successful. Metrics that track the team collaboration and support have highlighted the importance of standard operating processes to ensure smooth and clear collaboration between the teams. Delivery metrics

have highlighted the importance of accelerated development and deployment processes which are required for rapid delivery of data products into markets for getting a competitive advantage and reap the benefits of the data products. Organizations should also focus on the data product adoption, as the success of the data product is tightly tied with the data product adoption of the users. Automation, collaboration, continuous integration and deployment and infrastructure as code are other key areas where the focus needs to be given by organizations. In overall, it is understood that the KPIs provide a clear measurement on how the organizations are marching towards success using their DataOps and MLOps practices.

## 5.3 Discussion of Research Question Two

Research problem discussed the technical and nontechnical challenges that are impacting the adoption of DataOps and MLOps principles in the development of data products. Technical challenges are multifacted and could result in a barrier in development and deployment of data product. Challenges like infrastructure scalability, disparate data sources, orchestraton of complex data pipelines, pipeline stability, data quality, integration complexities and deployment practices are critical concerns especially in the development of large scale data products. Democraticization of data and access management are other critical areas which would restrict the data product development acceleration. A robust framework that ensures the infrastructure stability, practices that ensure data quality and monitoring framworks to uphold data integrity and policy standards would need to be strategically defined for a successful delivery of the data products.

Non-Technical challenges significantly influence the development of the data products using DataOps and MLOps principles. Organization business units setup, cross team communications, prioritizations, cultural barries and skill shortages are critical factors that influence the progress of the development and delivery of data products. Resistance to change within organizations, frequent changes in the organization structure, silo coomunications often obstruct the smooth adoption of dataops and mlops practices.

To mitigate the technical challenges, organizations have adopted cloud based infrastructure that can accommodate the dynamic workloads and infrastructure scalability. Implementation of data quality frameworks and development of observability frameworks are also developed to mitigate data quality issues. CI/CD practices and automated testing strategies from DevOps are being considered along with code version tools are helping to accelerate the deployment processes. However, it should be considered that these strategies would require a continuous evolution to keep adjusted with the evolving technologies and practices. Future research should focus on the new AI based technologies and promote innovation and optimizations in data product development and deployment process.

## 5.4 Discussion of Research Question Three

Research question establishing the common areas between DevOps, DataOps and MLOps methodologies and cross use of best practices between the methodologies. The key DevOps principles like collaboration, automation and continuous improvement are at the core of DevOps can be easily incorporated into DataOps and MLOps. Defining processes and practices with different teams is required to establish a standard operating procedure

and smooth functioning of the teams. Organizations are lacking these processes and having no RACI (Responsibility, Accountability, Consulted, Informed) would result in to and fro between teams resulting in delays and not being aligned to organizational goals. Best practices like continuous integration and continuous deployment (CI/CD) and automated testing which are at the core of DevOps have showed significant improvement in the deployment life cycles are already adopted into DataOps and MLOps life cycle improving the operational efficiency. Organizations should have a CI/CD strategy defined starting with the release branching strategy, tool strategy and defined deployment patterns to reduce confusions and establish a standard pattern for accelerated deployments. Deploying practices around infrastructure as code which is one of the DevOps principles reduces the dependency between the teams in the infra providing an automated process driven deployments then having dependency between the teams. Continuous integration and deployment practices from DevOps are adopted with the automated testing practices making organizations to reach to the markets at a faster rate.

Automated testing, continuous monitoring and feedback, cross team collaboration practices from the Devops practices would straightly imply to DataOps and MLOps which helps in removing the hurdles and bottle necks with iterative improvements can accelerate product developments. It is clear to mention that the DevOps principles can be adopted to unique challenges that any data product development.

## 5.5 Discussion of Research Question Four

With increasing complexity and multifaceted problems faced by organizations, working towards a single task is a daunting task. The research has discussed about the various KPI metrics that are important for any organization to track and improve their data delivery practices. Data quality is the core fundamental which is obvious and also from the expert's opinion tracking of these metrics is crucial for successful of a data product in production. Data being the fuel of the data product, any challenges with the data would be a daily daunting task for any organization. Measuring the quality of the data on a daily basis would help the organizations in establishing trust and reliability of the data product. The infrastructure monitoring KPIs helps organizations to track and monitor the infrastructure stability and reliability. For reliable and trustable data products, these KPIs will help in a realistic view of the current state of the data product from a 360-degree view thereby supporting organizations to take necessary precautions to improve and accelerate the data product development. The KPIs generally need to be targeted based on the organization standards and business process. Here in the research a base level KPI thresholds have been defined, however, they would need to be updated based on the needs of the organizations.

# CHAPTER 6 :

# SUMMARY, IMPLICATIONS, AND RECOMMENDATIONS

## 6.1 Summary

This study discussed on the effective governance of DataOps and MLOps principles using Key Performance Indicators to ensure smooth delivery of data products. From the literature reviews and expert surveys, the research discussed into the current practices, challenges and best practices involved with implementing of DataOps and MLOps in organizations. The findings have highlighted the importance of the KPIs aligning to business objectives and the standard governance to track and monitor these KPIs. The study also discussed about the importance of continuous improvement, cross functional collaboration and risk management in the process of maturing data products that are being developed to align with the organizational goals. The research has also discussed on the approaches and significance of automation and monitoring frameworks in achieving scalable data operations. In overall the study has provided details about the implementing the important Key Performance Indicators for effective DataOps and MLOps governance for data driven product developments.

## 6.2 Implications

The findings provide insights for organizations to enhance their with the key performance indiscators that can be used for tracking the data product development process. Implementing and tracking these KPIs will help in measuring the organization

progress and would help the organizations to take effective measures accordingly to bring efficiencies thereby providing a competitive edge.

Researchers can leverage the study's findings to do a next level research in this field. Cross disciplinary research can further deepen the understanding the importance of DataOps and MLOps principles and drive innovation. For professionals working in these fields this study would help to understand the concepts and the importance of the principles in multi direction perspective thereby advancing and improving themselves in these areas. Organizations can use these study to refine their data governing KPIs that can improve their data strategies, optimize the processes for an effective data product development. This study acts as a foundational resource for both academic and practical advancements in the areas of data product development and operations.

### 6.3 Recommendations for Future Research

DataOps and MLOps terminologies are existing in the industry for near to a decade, however, organizations are still finding difficulties in implementing them and make them as a standard practice. This study has discussed on the key performance indicators (KPIs) which the organizations can track for a successful delivery of a data product. From the literature review and discussion with the industry experts, it was understood that the democratization of data to all consumers in a cloud environment setup is still a challenge following the organizational constraints. A deeper study is required in this area to establish guidelines in different scenario setups.

There has been an observation particularly from data engineers where it has been mentioned that the evolving data requirements is a challenge and the data requirements should be standard as the business processes related to the data are already established and there should be a framework for defining the requirements and success criteria for data modeling requirements. Organizations should scale to automate their data ingestion and transformation frameworks and also have real time data ingestion capabilities can accelerate the data product development and insights generation.

Also, it is understood that there isn't enough analysis done on the non-functional requirements for data engineering products and also the initial assessment on the infrastructure accessibility is not completely done before the start of the product development. Research can be conducted on the variety of the infrastructure challenges and propose a framework of guidelines for data engineering and ml product development.

It is also observed that the data engineering and ml development teams are working in silos and there have been architecture level challenges noticed during the integration of components between data engineering and ml components. A comprehensive study to be done in establishing the architectural guidelines on the compatibility, adaptability, scalability and reliability of the data product with these components to improve the collaboration between machine learning and data engineering teams.

## 6.4 Conclusion

To conclude, effective governance of DataOps and MLOps principles is critical for organizations and tracking them on a regular basis is required for effective planning,

delivery and success of a data product throughout its life cycle. By adopting best practices, addressing  challenges and would help in decision making process providing a competitive advantage in the current data drivem landscape.

# APPENDIX A

## SURVEY COVER LETTER

Dear Sir,

I am a research student in the Swiss School of Business Administration in Geneva, conducting a research under the supervsion of Dr. Anna Provodnikova. I am researching on the topic “Govern Relevant Key Performance Indicators for Business Alignment While Developing Data Products”.

Previous researches have studied the principles of DataOps and MLOps and have proposed best practices and improvements along with maturity levels. However, the key performance indicators that govern them needed a research in-order to provide a method to measure and provide insights on the improvements that can be planned to improve the data product development.

So, I request your response to complete the questionnaire and be assured that the data collected will be kept confidential, and no firm, organization and individual will be identified in the thesis or in any report based on this research.

Thanks in advance for your co-operation.

Yours Sincerely,
Leela Ravi Shankar Dhulipalla

APPENDIX B

INFORMED CONSENT

**GOVERN RELEVANT KEY PERFORMANCE INDICATORS FOR BUSINESS ALIGNMENT WHILE DEVELOPING DATA PRODUCTS**

I <participant name>, agree to participate in the research project titled "Govern Relevant Key Performance indicators for business alignment while developing data products", conducted by Leela Ravi Shankar Dhulipalla who has discussed the research project with me.

I have received, read and kept a copy of the information letter/plain language statement. I have had the opportunity to ask questions about this research and I have received satisfactory answers. I understand the general purposes, risks and methods of this research.

I consent to participate in the research project and the following has been explained to me:

- the research may not be of direct benefit to me
- my participation is completely voluntary
- my right to withdraw from the study at any time without any implications to me
- the risks including any possible inconvenience, discomfort or harm as a consequence of my participation in the research project
- the steps that have been taken to minimise any possible risks
- public liability insurance arrangements
- what I am expected and required to do

- whom I should contact for any complaints with the research or the conduct of the research
- I am able to request a copy of the research findings and reports
- security and confidentiality of my personal information.

In addition, I consent to:

- Publication of results from this study on the condition that my identify will not be revealed.

**Name:** _____ (please print)

Signature:

_____

Date: _____

APPENDIX C

INTERVIEW GUIDE

Here is the interview guide related to the project:

1. Title: Govern Relevant Key Performance Indicators for Business Alignment while Developing Data Products
2. Candidate Introduction
    a. What best describes your current position?
        i. ML Engineer
        ii. Data Engineer
        iii. Product Manager
        iv. Architect
        v. Others
    b. How many years of experience do you have in the industry?
    c. Do you have established DataOps practices in your project?
    d. Do you have established MLOps practices in your project?
3. Data Quality KPIs: Can you please order the following Data Quality KPIs based on your experience and understanding (Rate them with numbers, 1 being the highest important and 7 being the lowest importance)
    a. Data Accuracy
    b. Data Completeness
    c. Data Consistency
    d. Data Timeliness
    e. Data Integrity

f. Data Drift Detection
g. Schema Shift Rate

4. Product Development Efficiency Metrics: Can you please order the following KPIs based on your experience and understanding? (Rate them with numbers, 1 being the highest important and 5 being the lowest importance)
    a. Time to Market
    b. Number of releases
    c. Development Cycle Time
    d. Defect Rate
    e. Customer Feedback Response
5. Data Operation Efficiency Metrics: Can you please order the following KPIs based on your experience and understanding? (Rate them with numbers, 1 being the highest important and 5 being the lowest importance)
    a. Data Ingestion Time
    b. Data Processing Time
    c. Pipeline Uptime
    d. Data Throughput
    e. Error Rate

6. Deployment Performance Metrics: Can you please order the following KPIs based on your experience and understanding? (Rate them with numbers, 1 being the highest important and 4 being the lowest importance)
    a. Deployment Frequency
    b. Deployment Success Rate
    c. Testing Coverage

    d. Deployment Downtime

7. Support Operations Performance Metrics: Can you please order the following KPIs based on your experience and understanding? (Rate them with numbers, 1 being the highest important and 5 being the lowest importance)
    a. Response Time to Incidents
    b. Median Response Time for Resolution
    c. Incident Reopen rate
    d. SLA Compliance Rate
    e. RCA Coverage
8. Data Observability Metrics: Can you please order the following KPIs based on your experience and understanding? (Rate them with numbers, 1 being the highest important and 5 being the lowest importance)
    a. Data Availability
    b. Data Freshness
    c. Data Quality Score
    d. Mean Time To Detect Data Errors
    e. Data Deduplication ratio
9. Machine Learning Model Performance Metrics: Can you please order the following KPIs based on your experience and understanding? (Rate them with numbers, 1 being the highest important and 4 being the lowest importance)
    a. Model Performance
    b. Model Serving Latency
    c. Bias / Fairness
    d. Model Drift

10. Agile Development Metrics: Can you please order the following KPIs based on your experience and understanding? (Rate them with numbers, 1 being the highest important and 6 being the lowest importance)
    a. Sprint Velocity
    b. Sprint Burndown Chart
    c. Sprint Retrospective Completion
    d. CI/CD Pipeline Efficiency
    e. Backlog Health
    f. Resource Utilization Rate
11. Data Product Performance Metrics: Can you please order the following KPIs based on your experience and understanding? (Rate them with numbers, 1 being the highest important and 5 being the lowest importance)
    a. Time to Market
    b. Data Product Adoption
    c. Customer Satisfaction
    d. Compliance and Governance Adherance
    e. Data Product Down Time

Below table has the responses received by the participants for the survey for the questions 2 to 11.

| KPI Metric / Person Name (Initial) | S | B | N | P | D | R | O | T | O | G | S | S | H | S | A |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Person Role | Product Manager | Associate Director, Product Management | Product Manager | Product Manager | Senior Product Manager | Senior Product Manager | Associate Director, Product Management | Data Engineer | Principal Data Engineer | Principal Data Engineer | Principal Data Engineer | Data Engineer | Senior ML Engineer | ML Engineer | Delivery Manager |
| | | | | | | | | | | | | | | | |
| **Can you please order the following Data Quality KPIs based on your experience and understanding?** | | | | | | | | | | | | | | | |
| Data Accuracy | 1 | 3 | 2 | 1 | 2 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 |
| Data Completeness | 2 | 4 | 3 | 2 | 3 | 2 | 3 | 3 | 4 | 3 | 3 | 2 | 2 | 3 | 2 |
| Data Consistency | 4 | 1 | 4 | 3 | 5 | 3 | 2 | 5 | 3 | 4 | 4 | 3 | 4 | 4 | 5 |
| Data Timeliness | 5 | 2 | 5 | 4 | 1 | 5 | 4 | 1 | 5 | 5 | 5 | 4 | 5 | 5 | 4 |
| Data Integrity | 3 | 5 | 1 | 5 | 4 | 4 | 5 | 4 | 1 | 1 | 2 | 5 | 3 | 2 | 3 |
| Data Drift Detection | 6 | 6 | 7 | 6 | 6 | 7 | 6 | 6 | 6 | 6 | 7 | 6 | 6 | 6 | 6 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Schema Shift Rate** | 7 | 7 | 6 | 7 | 7 | 6 | 7 | 7 | 7 | 7 | 6 | 7 | 7 | 7 | 7 |
| | | | | | | | | | | | | | | | |
| **Product Development Efficiency Metrics** | | | | | | | | | | | | | | | |
| **Time to Market** | 1 | 3 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 2 | 2 | 1 | 2 |
| **Number of releases** | 3 | 1 | 1 | 2 | 2 | 2 | 2 | 3 | 3 | 2 | 3 | 3 | 3 | 2 | 1 |
| **Development Cycle Time** | 2 | 4 | 3 | 3 | 3 | 3 | 3 | 1 | 1 | 1 | 1 | 1 | 1 | 3 | 3 |
| **Defect Rate** | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 4 | 5 | 4 | 4 |
| **Customer Feedback Response** | 4 | 2 | 4 | 4 | 4 | 4 | 4 | 5 | 5 | 5 | 5 | 5 | 4 | 5 | 5 |
| | | | | | | | | | | | | | | | |
| **Data Operation Efficiency Metrics** | | | | | | | | | | | | | | | |
| **Data Ingesti** | 3 | 4 | 2 | 1 | 3 | 2 | 4 | 1 | 3 | 1 | 2 | 1 | 2 | 2 | 1 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **on Time** | | | | | | | | | | | | | | | |
| **Data Processing Time** | 4 | 3 | 3 | 4 | 4 | 4 | 3 | 2 | 2 | 3 | 3 | 2 | 1 | 1 | 3 |
| **Pipeline Uptime** | 1 | 2 | 1 | 2 | 1 | 1 | 2 | 3 | 1 | 2 | 1 | 3 | 3 | 3 | 4 |
| **Data Throughput** | 2 | 1 | 4 | 3 | 2 | 3 | 1 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 2 |
| **Error Rate** | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| | | | | | | | | | | | | | | | |
| **Deployment Performance Metrics** | | | | | | | | | | | | | | | |
| **Deployment Frequency** | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 2 | 1 |
| **Deployment Success Rate** | 1 | 2 | 2 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 1 | 2 | 3 | 1 | 2 |
| **Testing Coverage** | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 1 | 3 | 3 |
| **Deployment Downtime** | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| | | | | | | | | | | | | | | | |
| **Support Operations Perfor** | | | | | | | | | | | | | | | |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **mance Metrics** | | | | | | | | | | | | | | | |
| **Response Time to Incidents** | 2 | 4 | 1 | 4 | 2 | 4 | 1 | 2 | 4 | 1 | 2 | 1 | 4 | 4 | 3 |
| **Median Response Time for Resolution** | 3 | 2 | 2 | 3 | 1 | 2 | 2 | 3 | 1 | 2 | 3 | 2 | 2 | 1 | 2 |
| **Incident Reopen rate** | 4 | 3 | 3 | 1 | 3 | 3 | 3 | 4 | 2 | 3 | 4 | 3 | 3 | 2 | 4 |
| **SLA Compliance Rate** | 1 | 1 | 4 | 2 | 4 | 1 | 4 | 1 | 3 | 4 | 1 | 4 | 1 | 3 | 1 |
| **RCA Coverage** | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| | | | | | | | | | | | | | | | |
| **Data Observability Metrics** | | | | | | | | | | | | | | | |
| **Data Availability** | 2 | 1 | 2 | 2 | 2 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 |
| **Data Freshness** | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| **Data Quality Score** | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 1 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Mean Time To Detect Data Errors** | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| **Data Deduplication ratio** | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| | | | | | | | | | | | | | | | |
| **Machine Learning Model Performance Metrics** | | | | | | | | | | | | | | | |
| **Model Performance** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **Model Serving Latency** | 2 | 2 | 2 | 2 | 2 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 3 |
| **Bias / Fairness** | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 4 | 3 | 3 | 4 | 4 | 4 | 4 | 2 |
| **Model Drift** | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 4 | 4 | 3 | 3 | 3 | 3 | 4 |
| | | | | | | | | | | | | | | | |
| **Agile Development Metrics** | | | | | | | | | | | | | | | |
| **Sprint Velocity** | 3 | 2 | 1 | 4 | 2 | 3 | 1 | 5 | 4 | 3 | 3 | 3 | No | No | 3 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Sprint Burndown Chart** | 2 | 4 | 3 | 1 | 4 | 2 | 3 | 6 | 5 | 4 | 5 | 5 | No | No | 1 |
| **Sprint Retrospective Completion** | 4 | 5 | 6 | 2 | 5 | 4 | 6 | 2 | 1 | 6 | 1 | 1 | No | No | 2 |
| **CI/CD Pipeline Efficiency** | 5 | 3 | 5 | 3 | 3 | 5 | 5 | 1 | 2 | 5 | 2 | 2 | No | No | 5 |
| **Backlog Health** | 1 | 1 | 2 | 5 | 1 | 1 | 2 | 4 | 3 | 1 | 4 | 4 | No | No | 2 |
| **Resource Utilization Rate** | 6 | 6 | 4 | 6 | 6 | 6 | 4 | 3 | 6 | 2 | 6 | 6 | No | No | 6 |
| | | | | | | | | | | | | | | | |
| **Data Product Performance Metrics** | | | | | | | | | | | | | | | |
| **Time to Market** | 1 | 2 | 3 | 4 | 5 | 2 | 4 | 4 | 5 | 3 | 2 | 5 | | | 3 |
| **Data Product Adoption** | 2 | 3 | 4 | 1 | 3 | 3 | 1 | 5 | 4 | 2 | 1 | 4 | | | 2 |
| **Customer Satisfaction** | 3 | 1 | 1 | 5 | 4 | 1 | 5 | 1 | 2 | 5 | 3 | 2 | | | 1 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Compliance and Governance Adherance** | 4 | 5 | 2 | 3 | 1 | 5 | 3 | 2 | 3 | 4 | 4 | 3 | | | 4 |
| **Data Product Down Time** | 5 | 4 | 5 | 2 | 2 | 4 | 2 | 3 | 1 | 1 | 5 | 1 | | | 5 |
| | | | | | | | | | | | | | | | |

REFERENCES

Alrehamy, H. and Walker, C. (2015) 'Personal Data Lake With Data Gravity Pull', In: *IEEE Fifth International Conference on Big Data and Cloud Computing 2015*, pp.26–28.

Alvord, M.M., Lu, F., Du, B. and Chen, C.-A. (2020) 'Big Data Fabric Architecture: How Big Data and Data Management Frameworks Converge to Bring a New Generation of Competitive Advantage for Enterprises', *Google Scholar*. [online] Available at: https://eapj.org/wp-content/uploads/2020/11/Big-Data-Fabric-Architecture.pdf [Accessed 11 Jul. 2023].

Anon (2019) *2020 State of Enterprise Machine Learning*. [online] Available at: https://cdn2.hubspot.net/hubfs/2631050/0284 CDAO FS/Algorithmia_2020_State_of_Enterprise_ML.pdf [Accessed 26 Dec. 2021].

Ansyori, R., Qodarsih, N. and Soewito, B., (2018) 'A systematic literature review: Critical Success Factors to Implement Enterprise Architecture'. *Procedia Computer Science*, 135, pp.43–51.

Armbrust, M., Ghodsi, A., Xin, R., Zaharia, M. and Berkeley, U., (2021) 'Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics'. [online] Available at: https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf.

Ballard, C., Herreman, D., Schau, D.F., Bell, R., Kim, E. and Valencic, A., (1999) 'Data Modeling Techniques for Data Warehousing'. *Google Scholar*. [online] Available at: https://eddyswork.synthasite.com/resources/Data Modeling Tech For Data Warehouseing.pdf [Accessed 12 Jul. 2023].

Barr Moses, (2024) *What is Data Downtime?* [online] Monte Carlo. Available at: https://www.montecarlodata.com/blog-the-rise-of-data-downtime/ [Accessed 28 Mar. 2024].

Benvenuti, D., Marrella, A., Rossi, J., Nikolov, N., Roman, D., Soylu, A. and Perales, F.,

(2023) 'A Reference Data Model to Specify Event Logs for Big Data Pipeline Discovery'. In: *Business Process Management Forum*. [online] pp.38–54. Available at: https://link.springer.com/chapter/10.1007/978-3-031-41623-1_3.

Bou Ghantous, G., Gill, A. and Bou, G., (2017) 'Association for Information Systems AIS Electronic Library (AISeL) DevOps: Concepts, Practices, Tools, Benefits and Challenges Recommended Citation'. *PACIFIC-ASIA CONFERENCE ON INFORMATION SYSTEMS (PACIS 2017)*, [online] p.1. Available at: http://aisel.aisnet.org/pacis2017/96 [Accessed 11 Jun. 2023].

Breck, E., Cai, S., Nielsen, E., Salib, M. and Sculley, D., (2021) 'What's your ML Test Score? A rubric for ML production systems'. *Reliable Machine Learning in the Wild - NIPS 2016 Workshop (2016)*. [online] Available at: https://storage.googleapis.com/pub-tools-public-publication-data/pdf/45742.pdf [Accessed 9 Jul. 2022].

Bucena, I. and Kirikova, M., (2017) 'Simplifying the DevOps Adoption Process'. *BIR Workshops 2017*. [online] Available at: http://ceur-ws.org/Vol-1898/paper14.pdf [Accessed 19 Dec. 2021].

Couto, J., Borges, O., Ruiz, D., Marczak, S. and Prikladnicki, R., (2019) 'A mapping study about data lakes: An improved definition and possible architectures'. *Proceedings of the International Conference on Software Engineering and Knowledge Engineering, SEKE*, [online] 2019-July, pp.451–458. Available at: https://ksiresearchorg.ipage.com/seke/seke19paper/seke19paper_129.pdf [Accessed 12 Jul. 2023].

Curry, E., Scerri, S. and Eds, T.T., (2022) 'An Organizational Maturity Model forData Spaces: A Data Sharing Wheel Approach'. *Data Spaces*. [online] Available at: https://doi.org/10.1007/978-3-030-98636-0_2.

DataOps Manifesto, (2021) *The DataOps Manifesto - Read The 18 DataOps Principles*. [online] Available at: https://datakitchen.io/wp-content/uploads/2023/08/DataKitchen-DataOps-Cookbook-Version-3-2023.pdf [Accessed 19 Dec. 2021].

Davenport, T.H. and Dyché, J., (2013) 'Big Data in Big Companies'. *Baylor Business Review*, [online] 321, pp.20–21. Available at: https://www.mendeley.com/catalogue/08631913-4a06-34b9-a772-6c40a981bd0d/ [Accessed 2 Dec. 2022].

Dixon, J., (2010) *Hadoop and Data Lakes*. [online] Available at: https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/ [Accessed 12 Mar. 2023].

Elouataoui, W., El Alaoui, I., El Mendili, S. and Gahi, Y., (2022) 'An Advanced Big Data Quality Framework Based on Weighted Metrics'. *Big Data and Cognitive Computing 2022*.

Ereth, J., (2018) 'DataOps-Towards a Definition'. *Lernen, Wissen, Daten, Analysen*. [online] Available at: http://ceur-ws.org/Vol-2191/paper13.pdf [Accessed 11 Dec. 2021].

Feijter, R., Overbeek, S., van Vliet, R., Jagroep, E. and Brinkkemper, S., (2018) 'DevOps competences and maturity for software producing organizations'. *Lecture Notes in Business Information Processing*, [online] 318, pp.244–259. Available at: https://link.springer.com/chapter/10.1007/978-3-319-91704-7_16 [Accessed 11 Jun. 2023].

Felipe, A. and Maya, V., (2016) 'The State of MLOps'. *Universidad de los Andes*. [online] Available at: http://hdl.handle.net/1992/51495 [Accessed 27 Apr. 2023].

Fivetran, (2022) *Over 80 Percent of Companies Rely on Stale Data for Decision-Making*. [online] Available at: https://www.fivetran.com/press/over-80-percent-of-companies-rely-on-stale-data-for-decision-making [Accessed 14 Dec. 2022].

Forsgren, N. and Kersten, M., (2018) 'DevOps metrics'. *Communications of the ACM*, [online] 614, pp.44–48. Available at: https://dl.acm.org/doi/10.1145/3159169 [Accessed 11 Jun. 2023].

Gandomi, A. and Haider, M., (2015) 'Beyond the hype: Big data concepts, methods, and

analytics'. *International Journal of Information Management*, 35, pp.137–144.

Giebler, C., Gröger, C., Hoos, E., Schwarz, H. and Mitschang, B., (2019) 'Leveraging the Data Lake: Current State and Challenges'. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, [online] 11708 LNCS, pp.179–188. Available at: https://link.springer.com/chapter/10.1007/978-3-030-27520-4_13 [Accessed 13 Jul. 2023].

Granlund, T., Kopponen, A., Stirbu, V., Myllyaho, L. and Mikkonen, T., (2021) 'MLOps Challenges in Multi-Organization Setup: Experiences from Two Real-World Cases'. [online] Available at: https://oraviz.io/ [Accessed 11 Dec. 2021].

Hai, R., Koutras, C., Quix, C. and Jarke, M., (2023) 'Data Lakes: A Survey of Functions and Systems'. *IEEE Transactions on Knowledge and Data Engineering*, 35.

Hellman, F., (2023) *Study and Comparison of Data Lakehouse Systems*. [online] *Doria*. Vaasa Abo Akademi University. Available at: https://www.doria.fi/bitstream/handle/10024/187408/hellman_fredrik.pdf?sequence=2&isAllowed=y.

Inmon, W.H., (2005) *Building the data warehouse*. 4th ed ed. [online] New York: Wiley Pub. Available at: https://www.google.co.in/books/edition/Building_the_Data_Warehouse/QFKTmh5IFS4C?hl=en&gbpv=1&dq=inauthor:%22W.+H.+Inmon%22&printsec=frontcover.

Irene O'Callaghan, Andriy Hryshchenko, K.B.& D.O., (2024) 'KPIs for Quality and Availability of Data in an Industrial Setting'. *SSRN*.

Jakobsen, A.S., (2023) *Study of DataOps as a concept for Aker BP to enable data-driven assets*. [online] University of Stavanger. Available at: https://hdl.handle.net/11250/3019262 [Accessed 21 Jul. 2023].

John, M.M., Olsson, H.H. and Bosch, J., (2021) 'Towards MLOps: A Framework and

Maturity Model'. In: *2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. pp.1–8.

Kai Hartzell, (2023) *Comparison of Big Data SQL Engines in the Cloud*. [online] University of Helsinki. Available at: https://helda.helsinki.fi/server/api/core/bitstreams/73b60661-8528-47e5-b14e-febd06e17cfb/content.

Kaisler, S., Armour, F., Espinosa, J. and Money, W., (2013) 'Big Data: Issues and Challenges Moving Forward'. *Proceedings of the Annual Hawaii International Conference on System Sciences*, pp.995–1004.

Kim, S.-Y. and Kim, T.-H., (2005) 'Implementing Data warehouse Methodology Architecture: From Metadata Perspective'. *International Commerce and Information Review*, [online] 71, pp.55–74. Available at: https://koreascience.kr/article/JAKO200530159714270.pdf.

Kreuzberger, D., Kühl, N. and Hirschl, S., (2022) 'Machine Learning Operations (MLOps): Overview, Definition, and Architecture'. *IEEE access*, 11, pp.31866–31879.

Leite, L., Rocha, C. and Kon, F., (2019) 'A Survey of DevOps Concepts and Challenges'. *ACM Computing Surveys Volume 52 Issue 6*, [online] 526, pp.1–35. Available at: https://doi.org/10.1145/3359981 [Accessed 11 Dec. 2021].

Liu, X., (2014) 'Optimizing ETL Dataflow Using Shared Caching and Parallelization Methods'. *CoRR*, [online] abs/1409.1. Available at: http://arxiv.org/abs/1409.1639.

Lucy Ellen Lwakatare Aiswarya Raj, Bosch, J., Olsson, H.H. and Crnkovic, I., (2019) 'A Taxonomy of Software Engineering Challenges for Machine Learning Systems: An Empirical Investigation'. *Springer*. [online] Available at: doihttps://doi.org/10.1007/978-3-030-19034-7_14.

Machado, I.A., Costa, C. and Santos, M.Y., (2022) 'Data Mesh: Concepts and Principles of a Paradigm Shift in Data Architectures'. *Procedia Computer Science*, 196, pp.263–

271.

Mainali, K., Ehrlinger, L., Himmelbauer, J. and Matskin, M., (2021) 'Discovering DataOps: A Comprehensive Review of Definitions, Use Cases, and Tools'. *The Tenth International Conference on Data Analytics*. [online] Available at: https://www.researchgate.net/publication/355107036 [Accessed 11 Dec. 2021].

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A., (2011) *Big data: The next frontier for innovation, competition, and productivity*. [online] Available at: https://personal.utdallas.edu/~muratk/courses/cloud11f_files/MGI-full-report.pdf.

Mario Angelelli and Massimiliano Gervasi, (2023) 'Representations of epistemic uncertainty and awareness in data-driven strategies. *CoRR'*, [online] abs/2110.1. Available at: https://arxiv.org/pdf/2110.11482.

Martín, N., Biddle, H., Ribeiro, V., Sainudiin, R. and Magnani, M., (2023) *Lakehouse architecture for simplifying data science pipelines: data engineering and graph data mining explorations in Trase.earth for the traceability of supply chains driving deforestation*. [online] Uppsala Universitet. Available at: https://github.com/nmartinbekier/ds_de_thesis.

Marz, N. and Warren, J., (2015) *Big Data: Principles and Best Practices of Scalable Realtime Data Systems*. 1st ed. [online] USA: Manning Publications Co. Available at: https://mitpressbookstore.mit.edu/book/9781617290343.

Merelda Wu, (2021) *What the Ops are you talking about? | by Merelda Wu | Towards Data Science*. [online] Medium. Available at: https://towardsdatascience.com/what-the-ops-are-you-talking-about-518b1b1a2694 [Accessed 28 Mar. 2023].

Michael Huttermann, (2012) *Michael Hüttermann. DevOps for Developers.* 1st ed ed. [online] USA: Apress Berkeley, CA. Available at: http://huettermann.net/devops/ [Accessed 19 Dec. 2021].

Microsoft, (2022) *Machine Learning operations maturity model - Azure Architecture Center | Microsoft Learn*. [online] Available at: https://learn.microsoft.com/en-us/azure/architecture/ai-ml/guide/mlops-maturity-model [Accessed 21 Apr. 2023].

Munappy, A.R., Mattos, D.I., Bosch, J., Olsson, H.H. and Dakkak, A., (2020) 'From Ad-Hoc data analytics to DataOps'. *Proceedings - 2020 IEEE/ACM International Conference on Software and System Processes, ICSSP 2020*, [online] 20, pp.165–174. Available at: http://dx.doi.org/10.1145/3379177.3388909 [Accessed 15 Dec. 2021].

Murali, A., (2021) *MLOps – 5 Steps you Need to Know to Implement a Live Project*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/08/mlops-5-steps-you-need-to-know-to-implement-a-live-project/ [Accessed 13 Jul. 2023].

Muralidhar, N., Muthiah, S., Butler, P., Jain, M., Yu, Y., Burne, K., Li, W., Jones, D., Arunachalam, P. and Ramakrishnan, N., (2021) 'Using AntiPatterns to avoid MLOps Mistakes; Using AntiPatterns to avoid MLOps Mistakes'. *CoRR*. [online] Available at: https://arxiv.org/pdf/2107.00079.pdf [Accessed 19 Dec. 2021].

Muratov S. Y, M.S.B., (2023) 'Framework architecture of a secure big data lake'. In: *Procedia Computer Science*. [online] Russia: Elsevier BV, pp.39–46. Available at: https://shorturl.at/s4t7Y.

Mylavarapu, G., Thomas, J.P. and Viswanathan, K.A., (2019) 'An Automated Big Data Accuracy Assessment Tool'. *2019 4th IEEE International Conference on Big Data Analytics, ICBDA 2019*, pp.193–197.

Narayanan, S., S, M. and Zephan, P., (2024) 'Real-Time Monitoring of Data Pipelines: Exploring and Experimentally Proving that the Continuous Monitoring in Data Pipelines Reduces Cost and Elevates Qualit'. *EAI.EU*.

Nargesian, F., Pu, K.Q., Zhu, E., Ghadiri Bashardoost, B. and Miller, R.J., (2020) 'Organizing Data Lakes for Navigation'. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, [online] pp.1939–1950. Available at:

https://dl.acm.org/doi/10.1145/3318464.3380605 [Accessed 12 Jul. 2023].

Nybom, K., Smeds, J. and Porres, I., (2016) 'On the Impact of Mixing Responsibilities Between Devs and Ops'. *Lecture Notes in Business Information Processing*, 251, pp.131–143.

Östberg, P.-O., Vyhmeister, E., Castañé, G.G., Meyers, B. and Van Noten, J., (2022) 'Domain Models and Data Modeling as Drivers for Data Management: The ASSISTANT Data Fabric Approach'. *IFAC-PapersOnLine*, 5510, pp.19–24.

Peralta, V., (2006) *Data Freshness and Data Accuracy: A State of the Art*. [online] Uruguay. Available at: https://www.fing.edu.uy/inco/grupos/csi/esp/Publicaciones/2006/tr0613-vp.pdf [Accessed 28 Apr. 2023].

Pivotal and Capegemini, (2013) *The Technology of the Business Data Lake*. [online] Available at: https://www.capgemini.com/wp-content/uploads/2017/07/pivotal-business-data-lake-technical_brochure_web.pdf.

Poe, V., Klauer, P. and Brobst, S., (1998) *Building a data warehouse for decision support*. 2nd ed. [online] *ACM Digital Library*. USA: Prentice Hall PTR. Available at: https://dl.acm.org/doi/10.5555/550486 [Accessed 12 Jul. 2023].

Power, K. and Conboy, K., (2014) 'Impediments to Flow: Rethinking the Lean Concept of 'Waste' in Modern Software Development'. *Lecture Notes in Business Information Processing*, [online] 179 LNBIP, pp.203–217. Available at: https://link.springer.com/chapter/10.1007/978-3-319-06862-6_14 [Accessed 19 Dec. 2021].

Provost, F. and Fawcett, T., (2013) 'Data Science and Its Relationship to Big Data and Data-Driven Decision Making'. *Big Data*, [online] 1, pp.51–9. Available at: https://pubmed.ncbi.nlm.nih.gov/27447038/.

Ralph Kimball, M.R., (2011) *The Data Warehouse Toolkit: The Complete Guide to*

*Dimensional Modeling*. 2nd edn ed. [online] United States of America: John Wiley & Sons. Available at: https://books.google.co.in/books?hl=en&lr=&id=XoS2oy1IcB4C&oi=fnd&pg=PA1&dq=kimball+data+warehouse&ots=1DMilGeLjD&sig=qo_U5q5y0Gu0V7Ya9vcWW0Bf_rQ&redir_esc=y#v=onepage&q=kimball data warehouse&f=false.

Renggli, C., Rimanic, L., Merve Gürel, N., Karlaš, B., Wu, W., Zhang, C. and Zurich, E., (2021) 'A Data Quality-Driven View of MLOps'. *CoRR*. Available at: arXiv:2102.07750v1 [Accessed 10 Jul. 2023]

Rodriguez, M., Jonatã, L., De Araújo, P. and Mazzara, M., (2020) 'Good practices for the adoption of DataOps in the software industry'. *Journal of Physics: Conference Series*, [online] 1694, p.12032. Available at: https://iopscience.iop.org/article/10.1088/1742-6596/1694/1/012032/pdf [Accessed 19 Dec. 2021].

Roeleven, S., (2010) 'Why Two Thirds of Enterprise Architecture Projects Fail: An Explanation for The Limited Success of Architecture Projects'. *IDS-Scheer White paper*, [online] December, p.12. Available at: https://www.cio.com/whitepaper/370709/why-two-thirds-of-enterprise-architecture-projects-fail/?type=other&arg=0&location=featured_li.

Roopa, S. and Rani, M.S., (2012) 'Questionnaire Designing for a Survey'. *The Journal of Indian Orthodontic Society*, 464, pp.273–277.

Ruf, P., Madan, M., Reich, C. and Ould-Abdeslam, D., (2021) 'Demystifying MLOps and Presenting a Recipe for the Selection of Open-Source Tools'. *Applied Sciences 2021, Vol. 11, Page 8861*, [online] 1119, p.8861. Available at: https://www.mdpi.com/2076-3417/11/19/8861/htm [Accessed 19 Dec. 2021].

Sawadogo, P.N., Scholly, É., Favre, C., Ferey, É., Loudcher, S. and Darmont, J., (2019) 'Metadata Systems for Data Lakes: Models and Features'. *Communications in Computer and Information Science*, [online] 1064, pp.440–451. Available at: https://link.springer.com/chapter/10.1007/978-3-030-30278-8_43 [Accessed 12 Jul.

2023].

Scerri, S., Tuikka, T., de Vallejo, I.L., Curry, E., (2022) 'Common European Data Spaces: Challenges and Opportunities'. *Data Spaces*. [online] Available at: doihttps://doi.org/10.1007/978-3-030-98636-0_16.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F. and Dennison, D., (2015) 'Hidden Technical Debt in Machine Learning Systems'. *NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, [online] pp.2503–2511. Available at: https://proceedings.neurips.cc/paper_files/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf [Accessed 10 Jul. 2023].

Senapathi, M., Buchan, J. and Hady, O., (2019) 'DevOps Capabilities, Practices, and Challenges: Insights from a Case Study'. *CoRR*, [online] 1907.10201. Available at: https://www.researchgate.net/publication/326029173_DevOps_Capabilities_Practices_and_Challenges_Insights_from_a_Case_Study [Accessed 25 Dec. 2021].

Shahin, M., Babar, M.A. and Zhu, L., (2016) 'The Intersection of Continuous Deployment and Architecting Process: Practitioners' Perspectives'. *International Symposium on Empirical Software Engineering and Measurement*, [online] 08-09-September-2016. Available at: https://dl.acm.org/doi/10.1145/2961111.2962587 [Accessed 11 Jun. 2023].

Sivarajah, U., Kamal, M., Irani, Z. and Weerakkody, V., (2017) 'Critical analysis of Big Data challenges and analytical methods'. *Journal of Business Research*, 70, pp.263–286.

Statista.com, (2022) *Total data volume worldwide 2010-2025 | Statista*. [online] Available at: https://www.statista.com/statistics/871513/worldwide-data-created/ [Accessed 21 Jan. 2023].

Susan, M., (2018) *How To Create A Business Case For Data Quality Improvement*. [online] Available at: https://www.gartner.com/smarterwithgartner/how-to-create-a-business-case-for-data-quality-improvement [Accessed 31 Dec. 2021].

Taleb, I., Serhani, M.A. and Dssouli, R., (2018) 'Big Data Quality: A Survey'. *Proceedings - 2018 IEEE International Congress on Big Data, BigData Congress 2018 - Part of the 2018 IEEE World Congress on Services*, pp.166–173.

VB Staff, (2019) *Why do 87% of data science projects never make it into production? | VentureBeat*. [online] Available at: https://venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production/ [Accessed 25 Dec. 2021].

Wang, R.Y. and Strong, D.M. (1996) 'Beyond Accuracy: What Data Quality Means to Data Consumers'. *Journal of Management Information Systems*, [online] 124, pp.5–33. Available at: https://api.semanticscholar.org/CorpusID:205581875 [Accessed 20 May 2023].

Xin, D., Miao, H. and Parameswaran, A., (2021) 'Production Machine Learning Pipelines: Empirical Analysis and Optimization Opportunities; Production Machine Learning Pipelines: Empirical Analysis and Optimization Opportunities'. *SIGMOD '21, June 20–25, 2021*. [online] Available at: https://github.com/tensorflow/serving [Accessed 10 Jul. 2023].