`ESTABLISHING AN EFFICIENT AND COST-EFFECTIVE INFRASTRUCTURE FOR SMALL AND MEDIUM ENTERPRISES TO DRIVE THE DATA SCIENCE PROJECTS FROM PROTOTYPE TO PRODUCTION.

by

Hrishikesh Manohar Thakurdesai

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

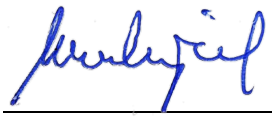
APRIL 2024.

ESTABLISHING AN EFFICIENT AND COST-EFFECTIVE INFRASTRUCTURE FOR SMALL AND MEDIUM ENTERPRISES TO DRIVE THE DATA SCIENCE PROJECTS FROM PROTOTYPE TO PRODUCTION.

by

Hrishikesh Manohar Thakurdesai

APPROVED BY:



Dissertation Chair

RECEIVED/APPROVED BY:

Admissions Director

Dedication

This thesis is dedicated to the small and medium-sized enterprises (SMEs) striving to integrate machine learning methodologies into their operations to address practical challenges, despite encountering obstacles related to infrastructure limitations, a shortage of skilled engineers, and budgetary constraints. It is our belief that this thesis will provide valuable insights to SMEs, assisting them in identifying and resolving challenges encountered during the execution of machine learning projects, thereby facilitating the attainment of effective efficient outcomes and cost.

Acknowledgements

In the course of my doctoral journey, I am indebted to numerous individuals who have played pivotal roles in my academic pursuit. Firstly, I express my profound gratitude to the divine for granting me the fortitude to complete this doctoral program.

I extend my sincere appreciation to my esteemed research advisor, Professor Mario Silic, whose unwavering support, invaluable guidance, insightful counsel, constructive critiques, and unwavering encouragement were instrumental throughout the development of this dissertation. Professor Mario, your guidance has been indispensable, and this research would not have reached fruition without your mentorship.

I would also like to acknowledge the leadership and staff of SSBM for affording me the privilege to study at a distinguished business school. Additionally, I would like to express my deep appreciation to my esteemed alma mater, the College of Engineering, Pune (COEP), whose exceptional educational environment and dedicated faculty not only inspired my academic journey but also equipped me with the knowledge and skills that have been invaluable in reaching this advanced stage of my doctoral studies.

Finally, I dedicate this doctoral degree to my beloved family, Manohar and Meghana Thakurdesai, for their unwavering support throughout my academic journey. Their steadfast commitment and moral support have been pivotal in the successful completion of this doctoral research. I thank them profoundly and invoke blessings upon each of them.

ABSTRACT

ESTABLISHING AN EFFICIENT AND COST-EFFECTIVE INFRASTRUCTURE FOR SMALL AND MEDIUM ENTERPRISES TO DRIVE THE DATA SCIENCE PROJECTS FROM PROTOTYPE TO PRODUCTION.

Hrishikesh Manohar Thakurdesai
2024

Eric Schmidt, Google's Executive Chairman, perceptively remarked that the amount of data generated from the dawn of civilization until 2003, estimated at 5 Exabyte, now materializes every two days, underscoring the data-driven paradigm of our times. While big data unquestionably occupies a pivotal role in data science, its significance extends beyond sheer volume. Big data solutions prioritize data organization and preprocessing over extensive analysis. In recent years, Data Science has assumed a pivotal role across diverse industries, spanning agriculture, risk management, fraud detection, marketing optimization, and public policy. Within these domains, Data Science leverages machine learning, statistical analysis, data preparation, and predictive modeling to tackle multifaceted challenges and provide actionable. Currently, numerous enterprises are fervently pursuing data-driven models, necessitating transitions to cloud and Spark clusters. However, the lack of comprehensive expertise often leads to increased costs and suboptimal infrastructure outcomes. Cultivating a holistic understanding of contemporary technologies and their seamless integration with existing infrastructure is imperative. This holistic approach enables the establishment of an adaptive platform tailored to

specific organizational needs, facilitating exploration and implementation of machine learning-based solutions. Attaining this goal hinges on meticulous examination of the advantages and limitations inherent in current big data technologies.

Through our research, we aspire to offer recommendations and construct a comprehensive infrastructure framework tailored to small and medium-sized enterprises (SMEs). This endeavor aims to bolster their capabilities in proficiently developing machine learning solutions, aligning them with the evolving landscape of data-driven innovation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Introduction

Renowned business author and data expert, Thomas Redman, aptly asserted that "Where there is data smoke, there is a business fire (Redman, 2008)". In today's era of Big Data, an unfathomable amount of data is being generated, surpassing even the value of oil. According to the International Data Corporation, global data is projected to reach a staggering 175 zettabytes by 2025 (Coughlin, 2018). Consequently, the significance of data science has surged exponentially, with companies of all sizes seeking insights from data and adopting data-driven models for root cause analysis to enhance their operations (Castelluccio, 2020; Waller, 2013).

Data Science has revolutionized industries across the spectrum by efficiently handling vast data from diverse sources and extracting invaluable insights, facilitating data-driven decision-making. It has left an indelible mark on sectors including marketing, healthcare, finance, banking, insurance, and more (Lee and Shin, 2020). With the aid of Data Science, industries not only decipher their challenges but also devise effective solutions. Automation of repetitive tasks, utilizing historical data for training machines, has notably reduced labor-intensive efforts. Machine Learning has empowered industries to tailor products and services to individual customers. For instance, e-commerce websites employ Recommendation Systems to provide personalized suggestions based on user history. Healthcare has witnessed significant improvements through early-stage tumor detection made feasible by machine learning (Sharma and Kaur, 2014).

Moreover, small-scale industries leveraged machine learning in marketing and customer acquisition (Neneh, 2018). Machine learning algorithms processed customer data to pinpoint target audiences, tailor marketing campaigns, and enhance advertising expenditure efficiency. By targeting the right audience with tailored messages, small businesses maximized their marketing ROI and attracted more customers. Following table shows examples of business applications of AI in SME-dominated sectors (OECD SME and Entrepreneurship Outlook, 2019).

Table 1. *Examples of business applications of AI in SME-dominated sectors (OECD Library, 2019)*

Sector	AI Applications Examples	Potential Benefits due to AI
Retail	Inventory management, personalized marketing, chatbots	Reduced stockouts, targeted promotions, improved customer service
Manufacturing	Predictive maintenance, quality control, supply chain optimization	Reduced downtime, improved product quality, enhanced efficiency
Healthcare	Patient management systems, diagnostic tools, telemedicine	Improved patient care, faster diagnosis, increased accessibility
Finance	Fraud detection, customer support via chatbots, loan processing automation	Reduced fraud, better customer service, faster loan approvals
Real Estate	Property value estimation, virtual tours, customer lead management	Accurate pricing, enhanced property showcasing, better lead conversion
Hospitality	Personalized guest experiences, demand forecasting, dynamic pricing	Increased guest satisfaction, optimized occupancy, improved revenue

Agriculture	Crop monitoring, automated irrigation, yield prediction	Enhanced crop yields, efficient water use, better resource management
Education	Personalized learning, automated grading, administrative task automation	Customized education, reduced workload for educators, streamlined operations
Logistics	Route optimization, warehouse automation, demand forecasting	Faster deliveries, reduced operational costs, improved inventory management
Professional Services	Client management, document automation, predictive analytics	Better client relationships, reduced administrative workload, informed decision-making

However, the adoption of machine learning algorithms presents challenges. An Algorithmia survey revealed that while organizations increased their AI/ML budgets and hired more data scientists, the deployment time for models extended, with 64% taking a month or longer (Paleyes, 2022). Setting up the necessary infrastructure, especially for modern data requirements, poses a significant hurdle.

This thesis embarks on a comprehensive exploration of the imperative task of "Establishing an efficient and cost-effective infrastructure for Small and Medium Enterprises (SMEs) to drive Data Science projects from prototype to production." SMEs, often operating within resource constraints, stand to benefit immensely from streamlined approaches that optimize their utilization of data-driven technologies.

The journey of a Data Science project encompasses numerous phases, each laden with its unique challenges and requirements. The initial stages involve data collection, cleaning, and exploration, often leading to the development of prototype models. While this phase is crucial for understanding the potential of data-driven insights, the ultimate

value lies in transitioning from prototypes to production-ready solutions that can drive real-world impact.

In this pursuit, SMEs grapple with multifaceted challenges, encompassing infrastructure limitations, budget constraints, the shortage of in-house expertise, and the need for scalable and efficient systems. The aim of this thesis is to dissect these challenges comprehensively and provide strategic insights and recommendations. Our focus extends beyond the theoretical to the practical, offering actionable guidance rooted in real-world applicability.

We will delve into the key components of an effective Data Science infrastructure, exploring aspects such as data storage, processing frameworks, cloud adoption, scalability, security, and integration with existing systems. Through meticulous examination and case studies, we aim to illuminate the pathway toward establishing a robust infrastructure capable of propelling SMEs from Data Science experimentation to the full realization of its potential.

1.2Research Problem

Small and medium-sized enterprises (SMEs) are crucial drivers of the global economy, making substantial contributions to job creation, innovation, and overall economic development. (Lukacs, 2015). In an era defined by data, the ability of SMEs to harness data science is crucial for their competitiveness and sustainability. **However, a pressing problem arises as SMEs often grapple with numerous challenges in establishing an efficient and cost-effective infrastructure for driving data science projects from the prototype stage to full-scale production.**

One of the central issues is infrastructure limitations (Parker, 2012). Many SMEs lack the necessary computing power, storage capacity, and network resources required for data-intensive data science tasks. The absence of scalable infrastructure inhibits their ability to process large datasets and implement machine learning algorithms effectively. Budget constraints represent another substantial challenge. SMEs typically operate with limited financial resources, making it challenging to invest in cutting-edge data science infrastructure. The cost of acquiring and maintaining the hardware, software, and skilled personnel necessary for data science initiatives can be prohibitive. Expertise Shortages pose a significant hurdle. Data science demands a specialized skill set encompassing data analysis, machine learning, programming, and domain knowledge. SMEs often struggle to attract and retain data science professionals, hindering their capacity to leverage data effectively.

The lack of a Clear Roadmap is another problem (Schelter, 2015). SMEs may have a desire to implement data science projects but often lack a comprehensive strategy. They require guidance on where to begin, what technologies to adopt, and how to progress from initial prototypes to production-ready solutions. Data Security and Compliance represent critical concerns (Kumar, 2017). Mishandling of sensitive data can lead to regulatory violations, legal issues, and reputational damage. Many SMEs lack the expertise to establish robust data security and compliance measures. Moreover, the Complexity of Data Ecosystems presents a challenge. SMEs often deal with heterogeneous data sources, ranging from structured databases to unstructured text and sensor data. Integrating and managing this diverse data landscape can be daunting without the right infrastructure and expertise. Following figure summarizes the data availability challenges occurred while adapting the machine learning techniques (GAO analysis of expert discussions, GAO-20-215SP).

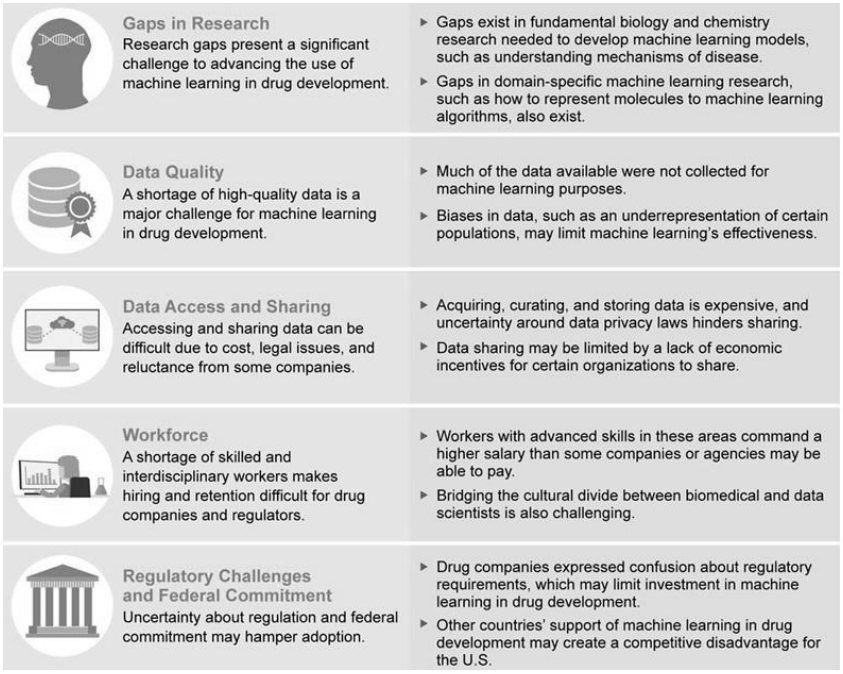


Figure 1 : *Challenges in Data availability (GAO analysis of expert discussions, GAO-20-215SP).*

The overarching research problem, therefore, revolves around finding solutions to these multifaceted challenges. It is imperative to address these issues comprehensively to empower SMEs to unlock the transformative potential of data science. This study seeks to explore innovative strategies, technology solutions, and best practices that can guide SMEs in establishing the efficient and cost-effective data science infrastructure necessary to drive their projects from the prototype stage to full-scale production.

1.3 Purpose of Research

The primary aim of this research is to explore and establish efficient infrastructure strategies for small and medium-scale enterprises (SMEs) looking to integrate machine

learning and data science into their operations. As a part of this research deliverable, we have developed a proof of concept chatbot application which can be deployed and used by data scientists to take guidance on machine learning queries. The details for the same are mentioned in the Results - section 4.7. Along with this POC, we also aim to design and study the research questions (exact questions are mentioned in the upcoming sections) to identify the essential infrastructure requirements, common challenges, critical cost factors, performance metrics, skill gaps, and strategic approaches necessary for SMEs to successfully adopt and manage these advanced technologies. By understanding the fundamental infrastructure requirements, SMEs can make informed decisions about the necessary investments in hardware, software, and data management systems. This ensures that resources are allocated efficiently and effectively. This research highlights common challenges faced by SMEs in data science infrastructure. By recognizing these challenges early on, SMEs can develop strategies to mitigate them, thus avoiding common pitfalls and ensuring smoother project implementation. Insights into the key components and factors in cost analysis enable SMEs to budget more accurately and manage their resources prudently. This helps in optimizing costs without compromising the quality and scope of their machine learning projects.

By identifying relevant metrics and KPIs, SMEs can effectively track and evaluate the performance of their machine learning infrastructure. This continuous monitoring allows for timely adjustments and improvements, leading to more successful project outcomes. Addressing the knowledge and skill gaps within the workforce is crucial for the effective utilization and management of machine learning infrastructure. This research provides guidance on training and development initiatives that can help SMEs build a competent and capable team.

In summary, this research aims to equip SMEs with the knowledge and tools needed to successfully integrate machine learning and data science into their operations. By following the findings and recommendations provided, SMEs can achieve greater efficiency, innovation, and competitiveness in the rapidly evolving technological landscape.

1.4 Significance of the Study

The study on "Establishing an efficient and cost-effective infrastructure for Small and Medium Enterprises (SMEs) to drive data science projects from prototype to production" is significant for multiple stakeholders and the broader landscape of data-driven business operations (Naradda, 2020). In today's data-intensive world, SMEs constitute a substantial and dynamic segment of the global economy. Their agility and capacity to innovate are key drivers of economic growth and employment generation. However, SMEs often face significant challenges in harnessing the full potential of data science due to limitations in infrastructure, budget constraints, and expertise shortages. This study is of critical importance as it aims to bridge these gaps and empower SMEs to leverage data science effectively.

The significance of this study extends to the very core of economic growth. SMEs, collectively, contribute significantly to a nation's GDP and job creation. By enabling SMEs to implement data science projects efficiently, this study not only enhances their operational capabilities but also fuels economic expansion. This, in turn, leads to increased employment opportunities, higher living standards, and overall prosperity within communities. Furthermore, the competitive edge in today's business landscape is defined by the ability to make data-driven decisions. SMEs that can harness the power of data science gain a substantial advantage. This study is significant because it equips SMEs with the knowledge and strategies needed to not only compete but excel in

their respective industries. It empowers them to make informed decisions, optimize processes, and enhance customer experiences. Innovation is another cornerstone of this study's significance. Efficient data science infrastructure serves as a catalyst for innovation. SMEs can develop novel products, services, and business models that are driven by data insights. By fostering a culture of innovation, this study contributes to the continuous evolution of industries. Resource optimization is a pivotal concern for SMEs, often operating within tight budgets. This study's significance lies in its ability to assist SMEs in maximizing resource utilization. It guides them in making informed investments in data science infrastructure, ensuring that every resource is leveraged effectively to yield maximum returns. Data security and compliance are significant in today's digital landscape. This study addresses these concerns comprehensively, reducing the risks associated with data breaches and legal non-compliance. By providing strategies and best practices for data security and compliance, it safeguards the reputation and integrity of SMEs.

The significance of this study also extends to **knowledge transfer**. As data science expertise is shared and disseminated, it contributes to the overall knowledge pool. SMEs can learn from best practices and lessons learned by others, fostering a collaborative and knowledge-sharing environment.

In summary, this study has implications that extend beyond the scope of individual small and medium-sized enterprise. It has the potential to impact economies, drive innovation, improve competitiveness, and shape the broader landscape of data science implementation. By assisting SMEs in their efforts to develop efficient and affordable data science infrastructure, this study plays a crucial role in paving the way for a data-driven future for SMEs and the global business community.

1.5 Research Questions

This research intends to provide guidance and insights to Small and Medium Enterprises (SMEs) on **establishing an efficient and cost effective infrastructure for driving data science projects from prototype to production**. To achieve this overarching purpose, several key research questions will be addressed:

1 - What are the fundamental infrastructure requirements for SMEs to effectively integrate machine learning and data science into their operations?

In the contemporary business landscape, the integration of machine learning (ML) and data science has emerged as a pivotal factor for SMEs for maintaining competitive advantage (Domínguez, 2020). SMEs which are often characterized by limited resources compared to their larger counterparts, stand to gain significantly from leveraging these advanced technologies. However, the successful implementation of ML and data science is contingent upon having robust and efficient infrastructure. This research seeks to elucidate the fundamental infrastructure requirements that SMEs must establish to harness the full potential of ML and data science in their operations. By identifying and addressing these needs, SMEs can better navigate the complexities of technological adoption and achieve sustainable growth.

The choice of this research question is driven by the critical role that infrastructure plays in the successful adoption of ML and data science. Unlike large enterprises, SMEs frequently encounter challenges such as budget constraints, limited technical expertise, and scalability issues. Understanding the specific infrastructure requirements is essential for these companies to make informed decisions about their technological investments. By honing in on SMEs, this study aims to provide actionable insights tailored to the unique constraints and opportunities faced by these businesses.

The findings from this research can offer numerous benefits to SMEs. Firstly, by identifying the essential infrastructure components, such as data storage solutions, computational power, and networking capabilities, SMEs can prioritize their investments more effectively. This prioritization helps in minimizing costs while maximizing the impact of their technological adoption. Secondly, a clear understanding of infrastructure needs facilitates smoother and faster integration of ML and data science, reducing the time to market for innovative solutions and services. Thirdly, with the right infrastructure in place, SMEs can enhance their data analytics capabilities, leading to better decision-making, improved customer insights, and more efficient operations. Ultimately, this research aims to empower SMEs to leverage ML and data science not just as tools for survival, but as catalysts for growth and innovation in an increasingly data-driven world.

2- What are the commonly faced challenges and issues by SMEs in data science infrastructure?

Establishing the necessary infrastructure to support ML and data science initiatives presents numerous challenges for SMEs (Walcott, 2021). This research focuses on identifying the commonly faced challenges and issues that SMEs encounter in building and maintaining data science infrastructure. By understanding these obstacles, SMEs can develop strategies to overcome them, ensuring successful implementation and sustainable use of data science capabilities. Selecting this research question is crucial due to the unique position SMEs hold in the business ecosystem. Unlike large corporations, SMEs typically operate with constrained budgets, limited technical expertise, and fewer resources, which can make the establishment of sophisticated data science infrastructure particularly daunting. By pinpointing the specific challenges faced by SMEs, this research aims to fill a significant gap in the literature, providing insights that are directly applicable to smaller enterprises. This focus is essential because existing research often

overlooks the nuanced difficulties SMEs encounter, concentrating instead on solutions tailored for larger organizations. Addressing this gap can lead to more targeted and effective support for SMEs embarking on data science initiatives. The insights gained from this research can deliver several key benefits to SMEs. Firstly, by clearly identifying the challenges in data science infrastructure, SMEs can better prepare for the hurdles they are likely to encounter, enabling proactive planning and resource allocation. Secondly, understanding these challenges allows SMEs to seek out and implement specific solutions that have been proven effective in similar contexts, thereby avoiding common pitfalls and reducing trial-and-error efforts. Thirdly, this research can highlight areas where external support, such as partnerships, consulting services, or training programs, might be necessary, helping SMEs to make informed decisions about where to invest their resources. Lastly, overcoming these challenges will enable SMEs to fully leverage ML and data science, resulting in improved operational efficiencies, more accurate decision-making, and enhanced competitive positioning. Ultimately, this research aims to empower SMEs by providing them with the knowledge and tools needed to successfully integrate data science into their operations, fostering innovation and growth.

3- What components / factors play a key role while doing the cost analysis of the project in SMEs?

One of the critical aspects of this integration is the cost analysis of establishing the necessary infrastructure (Abdullahi, 2015). Conducting a thorough cost analysis is essential for SMEs, as it enables them to allocate resources efficiently and ensure the sustainability of their technological investments. This research focuses on identifying the key components and factors that play a pivotal role in the cost analysis of ML and data

science projects within SMEs. By understanding these factors, SMEs can make informed financial decisions that support their growth and innovation goals.

Choosing this research question is crucial due to the financial constraints that typically characterize SMEs. Unlike larger enterprises, SMEs often operate with limited budgets, making it imperative to carefully analyze and manage costs associated with new technology implementations. Identifying the key components that influence cost analysis helps SMEs to forecast expenses accurately, avoid unforeseen financial burdens, and optimize their investment strategies. This research question addresses a critical gap in the existing literature, which often focuses on the technical aspects of ML and data science without delving into the financial implications for smaller enterprises. By focusing on cost analysis, this study aims to provide practical insights that are directly applicable to the financial planning processes of SMEs. The findings from this research can offer several benefits to SMEs. Firstly, a detailed understanding of the cost components involved in ML and data science projects can help SMEs to budget more effectively. This includes identifying direct costs such as hardware, software, and personnel, as well as indirect costs like training, maintenance, and data management. Secondly, recognizing these cost factors enables SMEs to evaluate the potential return on investment (ROI) more accurately, thereby supporting better financial decision-making and prioritization of projects. Thirdly, by highlighting cost-saving opportunities, such as leveraging open-source tools or cloud-based solutions, SMEs can reduce their financial outlay while still achieving their technological objectives. Finally, a comprehensive cost analysis can aid SMEs in securing funding or investment by providing clear and detailed financial projections to stakeholders. Overall, this research aims to empower SMEs to implement ML and data science infrastructure in a financially sustainable manner, enhancing their competitiveness and innovation capabilities in the long term.

4- What metrics and KPIs should SMEs track to evaluate the performance of their machine learning infrastructure?

As SMEs adopt advanced technological solutions, establishing a robust and efficient infrastructure becomes important (Melin, 2011). A critical component of this infrastructure is the ability to monitor and evaluate its performance effectively. This raises the important research question regarding the study of KPIs and performance metrics. Understanding and implementing the right performance indicators is essential for ensuring that machine learning initiatives deliver the desired outcomes and contribute positively to the business goals of SMEs (Okudan, 2022). Addressing this question is important because it enables SMEs to systematically assess the effectiveness and efficiency of their machine learning infrastructure. Without proper metrics and KPIs, it is challenging to determine whether the deployed models and algorithms are functioning optimally, utilizing resources efficiently, and providing valuable insights. This lack of clarity can lead to wasted resources, suboptimal performance, and missed opportunities for improvement. By identifying and tracking relevant metrics, SMEs can gain actionable insights into their infrastructure's performance, identify bottlenecks, and make informed decisions to enhance the effectiveness of their machine learning projects.

The benefits of these findings for SMEs are substantial. By establishing clear metrics and KPIs, SMEs can achieve greater transparency and control over their machine learning initiatives. This allows for continuous performance monitoring, which is essential for iterative improvement and long-term success. Moreover, tracking the right performance indicators helps in optimizing resource allocation, reducing costs, and maximizing the return on investment. Additionally, these metrics can provide valuable feedback that informs strategic planning and decision-making, ensuring that machine learning efforts are aligned with broader business objectives. Ultimately, these insights

empower SMEs to leverage machine learning and data science more effectively, driving innovation and sustaining competitive advantage in a dynamic market environment.

5- What are the knowledge and skill gaps SMEs need to address within their workforce to effectively utilize and manage infrastructure for machine learning and data science?

As the integration of machine learning (ML) and data science continues to reshape industries, small and medium-sized enterprises (SMEs) are increasingly seeking to leverage these technologies to enhance their operations, drive innovation, and remain competitive. However, the effective utilization and management of ML and data science infrastructure demand a workforce with specific knowledge and skills. This research focuses on identifying the knowledge and skill gaps that SMEs need to address within their workforce to successfully harness these technologies. By pinpointing these deficiencies, SMEs can implement targeted training and development programs to build the necessary competencies, ensuring they can fully capitalize on the benefits of ML and data science.

Choosing this research question is crucial due to the significant impact that workforce capabilities have on the successful adoption of ML and data science. Unlike larger corporations, SMEs often face constraints in recruiting and retaining highly specialized talent due to limited resources. As a result, understanding the specific knowledge and skill gaps within their existing workforce is essential for SMEs. This research question addresses a critical aspect of technological adoption that is frequently overlooked: the human factor. By focusing on the workforce, this study aims to provide insights that are directly actionable for SMEs, enabling them to develop their human capital in alignment with their technological goals. The insights derived from this research can offer multiple benefits to SMEs. Firstly, by identifying the precise

knowledge and skill gaps, SMEs can tailor their training and development programs to address these areas effectively. This targeted approach ensures that employees gain the specific competencies needed to manage and utilize ML and data science infrastructure, thereby enhancing overall efficiency and productivity. Secondly, investing in workforce development can lead to higher employee satisfaction and retention, as employees feel more valued and equipped to handle new challenges. Thirdly, a skilled workforce can drive more innovative and data-driven decision-making processes, leading to improved business outcomes. Lastly, by closing these knowledge and skill gaps, SMEs can reduce reliance on external consultants and services, which can be cost-prohibitive. This self-sufficiency not only reduces costs but also fosters a culture of continuous learning and adaptation, positioning SMEs for long-term success in an increasingly data-centric business environment.

In summary, the research purpose is to provide practical answers to these fundamental questions, empowering SMEs with the knowledge and strategies they need to navigate the data science landscape efficiently and cost-effectively.

6- What general infrastructure strategies should be employed by the SMEs while managing a machine learning project?

This research question seeks to uncover the best practices and approaches that SMEs can utilize to effectively support their machine learning initiatives. Understanding these strategies is essential for SMEs to successfully leverage the advanced technologies within their existing operational frameworks. Without a well-planned and robust infrastructure, SMEs may face significant challenges such as inadequate computational resources, data management issues, and integration difficulties. These challenges can lead to inefficiencies, increased costs, and project delays. By identifying and employing effective infrastructure strategies, SMEs can overcome these obstacles and create a strong

foundation that supports the scalability and efficiency of their machine learning projects. This is crucial for ensuring that these projects deliver meaningful business value and are sustainable in the long run.

Effective infrastructure strategies enable SMEs to optimize their resource utilization, ensuring that computational power, data storage, and network capabilities are aligned with project demands. This optimization not only reduces costs but also enhances the performance and reliability of machine learning models. Additionally, a well-structured infrastructure facilitates seamless integration of machine learning tools with existing systems, improving overall operational efficiency. Furthermore, these strategies provide SMEs with the agility to adapt to evolving technological trends and business needs, thereby maintaining their competitive edge in the market. Ultimately, by adopting the right infrastructure strategies, SMEs can maximize the return on their machine learning investments, driving innovation and achieving sustained growth.

CHAPTER II:

REVIEW OF LITERATURE

In the initial phase of the literature review, we analyzed the importance of infrastructure to run the data driven projects. The importance of infrastructure in running a machine learning based project cannot be overstated. Infrastructure provides the foundation upon which data can be collected, stored, processed, and analyzed effectively. It ensures the availability and reliability of computing resources, databases, and data storage, allowing for the seamless execution of data-driven tasks. A robust infrastructure enables scalability to handle large datasets and high computational demands, which is crucial as data volumes continue to grow. Furthermore, it facilitates real-time processing, which is essential for making timely decisions based on up-to-date information. Security measures within the infrastructure safeguard sensitive data, protecting it from breaches and ensuring compliance with data protection regulations. In essence, infrastructure is the backbone of any data-driven project, and investing in the right infrastructure is fundamental to the success and performance of such initiatives.

2.1 Key Features of the data science infrastructure.

In today's data-driven world, having a robust data science infrastructure is essential for organizations aiming to leverage their data effectively (Pine, 2015). Drawing insights from extensive research papers and articles, we have identified several key features that define an optimal data science infrastructure. These features ensure that the infrastructure is not only capable of handling complex data processes but also supports the seamless execution of data science tasks.

One of the most critical aspects of data science infrastructure is its capacity to facilitate quick and efficient debugging (Demchenko, 2012). An ideal system should

generate intuitive and comprehensive logs for all errors, which can then be seamlessly transported to a governance dashboard for easy access and review. This capability allows data scientists and engineers to rapidly identify and address issues, minimizing downtime and enhancing overall productivity. However, many current providers fall short in offering such user-friendly logging mechanisms. In these cases, users often have to resort to cloning repositories and manually obtaining log tails to diagnose problems, a process that is both time-consuming and cumbersome. Moreover, an effective data science infrastructure should not only log errors but also provide detailed, granular insights into them. The system should be able to classify and group errors based on their nature and origin, thereby simplifying the process of error analysis and resolution. By categorizing errors, the infrastructure enables more efficient troubleshooting and helps in pinpointing root causes more accurately. Unfortunately, some systems fail to deliver this level of detail in their error logs, making it challenging for users to debug and fix issues promptly. Comprehensive and well-organized error logging is thus a fundamental feature that significantly enhances the robustness and usability of data science infrastructure.

When considering the scalability of data science infrastructure, the emphasis often falls on speeding up data processing, but true scalability involves much more. It requires the ability to scale across multiple dimensions, such as data transfer speed, processing speed, file transfer speed, number of ports, processing power, and the capability to parallelize workflows. These aspects are essential for handling increasing data loads and complexities effectively. Scalability also means being able to manage unexpected growth in data volumes dynamically, ensuring resources can be scaled up without delays. Cloud providers like AWS offer flexibility with a range of instance sizes, from economical small instances to powerful large ones, allowing organizations to adjust resources based on their needs. However, while cloud services provide scalability and flexibility, they

may become cost-prohibitive over the long term compared to maintaining an in-house infrastructure. For sustained projects, investing in a proprietary data science infrastructure can be more cost-effective. Ultimately, scalability in data science infrastructure is about more than just processing speed; it involves adapting to data growth, managing resources efficiently, and making informed choices between cloud and in-house solutions to build a robust, scalable system that meets evolving demands.

Ensuring robust security in data science infrastructure is crucial. Unfortunately, some providers fall short in offering adequate transparency regarding their data handling processes, leading to potential security vulnerabilities. Frequent authentication errors can often indicate that the security layer is effectively preventing unauthorized access by quickly rejecting unexpected requests. It is essential to prioritize the strength and robustness of security protocols over the convenience of user access to maintain a secure environment. Moreover, a key aspect of secure data science infrastructure is minimizing human interference. Once a model is deployed, it should be protected from unauthorized access and modifications. No individual should have the ability to tamper with or extract data from the system. However, it is important to allow for the extraction of specific metrics to monitor and govern the model's performance. Despite this necessity, the infrastructure must rigorously safeguard the model's engineering frameworks and proprietary information, ensuring that sensitive data and intellectual property are protected at all costs. To further enhance security, the infrastructure should incorporate advanced encryption methods, secure access controls, and continuous monitoring for potential threats. Implementing multi-factor authentication and conducting regular security audits can help identify and mitigate vulnerabilities. By prioritizing these security measures, organizations can build a resilient data science infrastructure that not

only protects data but also fosters trust and compliance with industry standards and regulations.

In the realm of data science infrastructure, prioritizing automation is important. The infrastructure's primary goal should be to facilitate automated connections with crucial service providers, databases, and machines integral to the business operations. Insufficient support for services like Mongo-DB or Postgre-SQL can impede the scalability of models and hinder their connection to vital data sources. Additionally, the infrastructure must be capable of accommodating various hosting services like Docker or Kubernetes, streamlining the deployment and management processes for models. Limited support for these services could transform the infrastructure from an asset into a liability. Therefore, a robust infrastructure should prioritize automation to optimize efficiency and effectiveness in data science endeavors.

An essential aspect of data science infrastructure lies in its ability to facilitate the smooth transfer of information from various sources. However, if the infrastructure lacks standardized practices, such as utilizing out-dated versions of operating systems like Red Hat, transferring models and data can become cumbersome. Instances built on out-dated systems may have expired packages and may not support the new packages or services necessary for the model's functionality. It is imperative to conduct thorough research into the deployment process to grasp the intricacies involved. Resources like articles detailing deployment procedures, such as "How to deploy model xyz on an EC2 instance," provide valuable insights into the deployment journey. Moreover, documentation plays a pivotal role in ensuring efficient maintenance and deployment processes. Without comprehensive documentation, tasks related to maintenance and deployment can become significantly more challenging. Access to thorough documentation beforehand streamlines these processes, providing clear guidelines and instructions for seamless deployment and on-

going maintenance. Therefore, it is essential to prioritize the creation and maintenance of detailed documentation to alleviate potential obstacles and ensure the smooth operation of the data science infrastructure.

2.2 Analysis of existing approaches for deployment of ML Projects:

In the second phase of Literature review, we analyzed the existing ways to productionize the Data Science projects along with their advantages and Limitations.

A - Hardware Infrastructure Based Analysis:

1 - Local Environment:

When initiating a data science project, it's common practice to construct a proof of concept model using sample data on a local machine equipped with tools like Python, Jupyter, and Tableau Desktop. This approach offers advantages such as cost savings in the initial stages before determining project feasibility (Vecchio, 2018). Additionally, utilizing an internal secured network ensures data access security, while the absence of server-wide impacts simplifies adjustments. However, challenges arise when relying solely on local machines. Data scientists often face limitations in computing power and memory, necessitating the use of smaller sample datasets for model training. While these samples aid project initiation, they introduce complexities later in the data science lifecycle. Issues also arise concerning data updating and quality maintenance. Relying on local data copies risks of building predictions on out-dated or inaccurate representations of real-world scenarios. Accessing larger, more comprehensive datasets from centralized sources mitigates this risk and enhances model accuracy. The recent surge in artificial intelligence and machine learning interest is fuelled by the ability to process vast volumes of structured, unstructured, and semi-structured data efficiently. Machine learning thrives on larger, more diverse datasets, particularly those combining semi-structured interaction data, such as website logs and event data, with unstructured data like email or online

reviews, and structured transactional data from ERP, CRM, or order management systems.

To unlock significant business value from machine learning, enterprises should harness extensive datasets that merge transactional and interaction data. However, integrating laptops into this data processing mix introduces bottlenecks and delays due to their limited processing capabilities. As data scales continue to grow, cloud computing or large on-premises clusters emerge as preferred environments for processing. Minimizing reliance on laptops in data science workflows is essential to maintain efficiency and streamline data processing workflows effectively. In the current landscape, data scientists have a plethora of open-source machine learning frameworks at their disposal, including R, Sci-kit Learn, Spark ML, Tensor Flow, MXNet, and CNTK. However, managing the infrastructure, configurations, and environments for these frameworks can be arduous and time-consuming, particularly when conducted on laptops or on-premises servers. This additional overhead in infrastructure management detracts from core data science activities, highlighting the pivotal role of data logistics in AI/ML operations. Data logistics encompasses tasks such as collecting, labelling, categorizing, and managing datasets that mirror the real-world scenarios being modelled with machine learning. In enterprises with multiple data users, the challenge is exacerbated by the proliferation of local copies of datasets among various users; further complicating data management and governance.

The growing concerns surrounding security and privacy underscore the necessity for enterprise data processes to adhere to stringent data privacy and security regulations. Centralizing data repositories not only streamlines data management and governance but also ensures data consistency and facilitates model's ability to audit. However, these

factors collectively contribute to a delayed time-to-value in laptop-based data science workflows.

In a typical workflow for a data scientist operating from their laptop, several manual steps are involved. Initially, they must sample and manually download datasets onto their laptops or establish connections to databases using ODBC drivers. Subsequently, they need to install and maintain a plethora of required software tools and packages, including RStudio, Jupyter, Conda distributions, and various machine learning libraries across different programming languages like R, Python, and Java. Upon completion, deploying the model to production involves handing it over to an ML engineer, who faces the task of either converting the code to a production language like Java, Scala, or C++, or optimizing the code and integrating it with the existing application infrastructure. This optimization process encompasses rewriting data queries into ETL jobs, profiling the code to identify bottlenecks, and incorporating essential production-level capabilities such as logging and fault tolerance. Each of these steps presents potential bottlenecks that can lead to delays, especially if inconsistencies arise in software or package versions between development and production environments. Additionally, code developed in Windows or Mac environments may encounter compatibility issues when deployed onto Linux platforms, further complicating the deployment process.

In contrast, the software-as-a-service (SaaS) model in the cloud significantly reduces the overhead associated with infrastructure management. The usage-based pricing model offered by cloud services is particularly advantageous for machine learning workloads characterized by bursts of activity. Additionally, the cloud facilitates seamless experimentation with various machine learning frameworks, as cloud vendors provide comprehensive model hosting and deployment options. Furthermore, leading cloud

service providers such as Amazon Web Services, Microsoft Azure, and Google Cloud offer a range of intelligent capabilities as services. This approach effectively lowers barriers to integrating advanced functionalities into new products or applications, enhancing flexibility and innovation within the data science ecosystem.

2 - Cloud Based Environment:

Cloud computing has rapidly emerged as a transformative technology, prompting a widespread shift among companies from traditional on-premise systems. Whether opting for public, private, or hybrid solutions, cloud computing has become indispensable for organizations striving to remain competitive in today's dynamic business landscape. (Mendez, 2019; Hwang, 2017). As a part of the literature review, we have studied the advantages and disadvantage of adapting cloud infrastructure which are discussed in the sections below.

In the realm of cloud infrastructure, several advantages have been identified, each contributing significantly to the modernization and efficiency of businesses. Firstly, cost efficiency stands out as a primary driver behind the migration to cloud computing, offering substantial savings compared to on-premise technologies. The cloud's vast storage capabilities eliminate the need for expensive disk storage, optimizing resource utilization and reducing operational costs. Furthermore, cloud computing boasts high speed and agility, enabling rapid deployment of services with minimal effort. This agility translates into swift access to resources, ensuring that businesses can scale their systems efficiently to meet evolving demands. The unparalleled accessibility of cloud-stored information empowers users to access data anytime, anywhere, fostering flexibility and productivity in the digital era.

Moreover, the cloud streamlines data backup and recovery processes, alleviating the complexities and time constraints associated with on-premise technologies. With

automatic backups and seamless recovery mechanisms, businesses can ensure the integrity and continuity of their operations. Additionally, cloud providers handle infrastructure maintenance and updates, freeing organizations from the burdens of IT management and enabling them to focus on core business objectives. Another notable advantage of cloud computing lies in its ability to accommodate sporadic batch processing, allowing businesses to scale resources according to fluctuating workloads (Avram, 2014). This scalability minimizes resource wastage and optimizes cost efficiency. Furthermore, cloud infrastructure grants businesses a strategic edge by providing access to cutting-edge applications and services without the need for extensive investments in installation and maintenance. Cloud hosting also simplifies implementation, offering businesses the flexibility to retain existing applications and processes without grappling with backend technicalities. With no physical storage requirements, cloud infrastructure eliminates the need for on-site hardware, although backups are recommended for disaster recovery purposes.

Moreover, cloud computing facilitates automatic software integration, streamlining the customization and integration of applications to align with business preferences. The reliability of cloud hosting ensures consistent performance and seamless updates, bolstering operational efficiency and stability. Additionally, cloud services support workforce mobility, enabling employees to access resources from any location with an internet connection. Furthermore, the virtually unlimited storage capacity of the cloud ensures scalability and cost-effective storage solutions for businesses of all sizes. Collaborative capabilities inherent in cloud platforms facilitate secure and efficient collaboration among geographically dispersed teams, fostering innovation and productivity. Overall, cloud infrastructure represents a transformative technology that empowers businesses with agility, scalability, and enhanced collaboration capabilities,

driving growth and competitiveness in the digital age.In addition to the benefits already mentioned, cloud computing offers numerous other advantages such as on-demand self-service, multi-tenancy, resilient computing, rapid and efficient virtualization, cost-effective software solutions, advanced online security, high availability, automatic scaling to meet demand, pay-per-use models, web-based control interfaces, and API access. While these advantages highlight the significant benefits of cloud technology, it is crucial to also consider the potential drawbacks before implementation. In the next section, we will explore the disadvantages associated with cloud computing to provide a balanced perspective. By evaluating both the positive and negative aspects, businesses can make well-informed decisions about adopting cloud technologies that best align with their unique needs and goals.

Firstly, vulnerability to cyber-attacks is a significant concern, as storing sensitive data in the cloud increases the risk of unauthorized access and information theft. Despite robust security measures, the inherent online nature of cloud storage leaves organizations susceptible to security breaches. Additionally, cloud computing relies heavily on network connectivity, making reliable and high-speed internet access imperative for optimal performance. The dependency on internet connectivity exposes businesses to disruptions in service, leading to downtime that can impact operations and productivity. Moreover, vendor lock-in poses a considerable challenge when migrating between cloud platforms, as differences in vendor technologies may result in compatibility issues, configuration complexities, and increased expenses.

Furthermore, cloud customers often face limited control over their deployments, as cloud services are hosted on remote servers managed by service providers. This lack of control can hinder customization and optimization efforts, limiting the flexibility of the infrastructure. Additionally, some cloud providers offer only limited features, potentially

restricting the capabilities available to users. Moreover, concerns regarding redundancy and backup arise, as cloud servers may not inherently provide redundancy or backup capabilities. Investing in a redundancy plan is essential to mitigate the risk of data loss and ensure business continuity. Bandwidth issues may also arise, particularly when packing numerous storage devices and servers into a limited number of data center, potentially leading to performance bottlenecks and increased costs.

Furthermore, cloud computing companies may lack adequate customer support, relying on FAQs and online resources for assistance. Varied performance levels may be experienced in a shared cloud environment, as resources are allocated to multiple businesses simultaneously; making the infrastructure susceptible to performance fluctuations. Lastly, technical issues and outages are inherent risks in cloud technology, despite rigorous maintenance standards. It's essential for organizations to weigh these challenges against the benefits of cloud computing and implement mitigation strategies to address potential risks effectively.

B - Technology / Software Based Analysis:

1 - Pandas:

Pandas stands as a prominent software library crafted specifically for the Python programming language, dedicated to facilitating data manipulation and analysis tasks (Yudin, 2021). Notably, it provides a rich array of data structures and functionalities tailored for manipulating numerical tables and time series data with ease and efficiency. Being open-source, Pandas is freely available to users, released under the permissive three-clause BSD license. Widely recognized for its robustness and versatility, this library offers a plethora of advantages to users across various domains and industries. (McKinney, 2015; Bernard, 2016). Following are some of the advantages that are at the very core of the Pandas Library:

- **Excellent data representation**: Pandas emerges as an indispensable tool for aspiring data scientists and analysts due to its versatile capabilities in data representation and organization. Effective data organization is important for meaningful analysis and interpretation, making Pandas' functionality indispensable in this regard. With Pandas, users can effortlessly streamline data manipulation tasks, achieving more in fewer lines of code compared to traditional languages like C++ or Java. This efficiency not only simplifies data processing but also saves valuable time for both beginners and seasoned professionals in the field of data science. The ability to reduce the coding burden is particularly advantageous in a field like data science, where continuous practice and exploration are key to success. By leveraging Pandas, individuals can focus on honing their analytical skills and deriving insights from data, rather than getting bogged down by cumbersome coding tasks.
- **Efficient big-data handling:** Pandas is engineered to excel in handling vast datasets with unparalleled speed and precision. Its ability to swiftly and effectively analyze copious amounts of data sets it apart as a cornerstone tool for data scientists and analysts worldwide.
- **Extensive features availability:** The Pandas library stands out for its robustness, offering users a comprehensive suite of commands and features for seamless data analysis. With Pandas, users can effortlessly filter, segregate, and segment data according to their specific criteria and preferences. This versatility empowers data scientists and analysts to manipulate and explore datasets with unparalleled ease and flexibility, making Pandas an invaluable asset in the field of data science.
- **Seamless compatibility with Python:** Harnessing the power of Pandas within the Python ecosystem allows users to leverage a plethora of libraries and

functionalities, unlocking the full potential of Python's rich ecosystem. Among the most widely utilized libraries are NumPy, Matplotlib, and SciPy, each offering unique tools and capabilities to enhance data analysis and visualization tasks within the Pandas framework. This integration of Pandas with Python empowers data scientists and analysts to efficiently manipulate and explore data, driving insights and innovation in the field of data science.

While the benefits of Pandas are evident, it's essential to acknowledge that there are also drawbacks to consider (Bantilan, 2020). These will be explored in the forthcoming section, providing a comprehensive understanding of the library's nuances and limitations. By examining both the strengths and weaknesses of Pandas, users can make informed decisions about its suitability for their specific data science tasks. In the context of Pandas limitations, one notable drawback is its lack of support for scaling, particularly as datasets expand in size. With larger datasets, users may encounter challenges related to memory usage and CPU performance, necessitating additional hardware resources for efficient processing. While this simplicity may appeal to some users, it can also pose limitations on the library's scalability and performance as datasets grow. Moreover, Pandas exhibits a steep learning curve, which can be daunting for newcomers to the framework. While initial usage may seem straightforward, delving deeper into Pandas' intricacies can prove challenging for those unfamiliar with its underlying principles. Overcoming this learning curve requires determination and access to comprehensive learning resources, enabling users to fully harness the capabilities of the library.

Another significant issue with Pandas is its poor documentation, particularly for novice users. Well-documented libraries provide invaluable guidance on features and

functionalities, empowering users to explore and utilize the library effectively. However, the lack of comprehensive documentation in Pandas may hinder users' understanding of its full potential, limiting its accessibility to a subset of users and hindering broader adoption.Furthermore, Pandas' compatibility with three-dimensional matrices is lacking, posing limitations for users working with such data structures. While Pandas excels in handling two-dimensional matrices, transitioning to three-dimensional matrices necessitates the use of alternative libraries like NumPy. This restriction may introduce complexities and dependencies, detracting from the seamless workflow provided by Pandas for two-dimensional data analysis.

Despite these limitations, Pandas remains a versatile and flexible library within the Python ecosystem, particularly for smaller datasets that fit within memory constraints. While it may not leverage multiple CPUs efficiently and is limited by available RAM, Pandas continues to be a valuable tool for data manipulation and analysis in the Python machine learning ecosystem.

2 – Apache Spark:

Apache Spark stands as a prominent Big Data processing framework, initially crafted in Scala but offering versatile language bindings for Java, Python, and R, thus catering to a diverse user base. Unlike Pandas, Spark's inception was primarily geared towards addressing the challenges posed by massive datasets that surpass the memory capacity of a single machine or even an entire computing cluster. Initially, Spark introduced the Resilient Distributed Datasets (RDDs) API, requiring developers to model their data as classes. However, as the framework evolved, Spark integrated the data frame API, inspired by successful paradigms from Pandas and R, emerging as the preferred choice for data manipulation tasks. Similar to Pandas, Spark data frames operate on columnar and row-based structures, with columns collectively defining a schema shared

across all rows, thus facilitating seamless data processing and analysis within the Spark ecosystem. (Assefi, 2017). Following figure (figure 2) shows the overall architecture of spark.

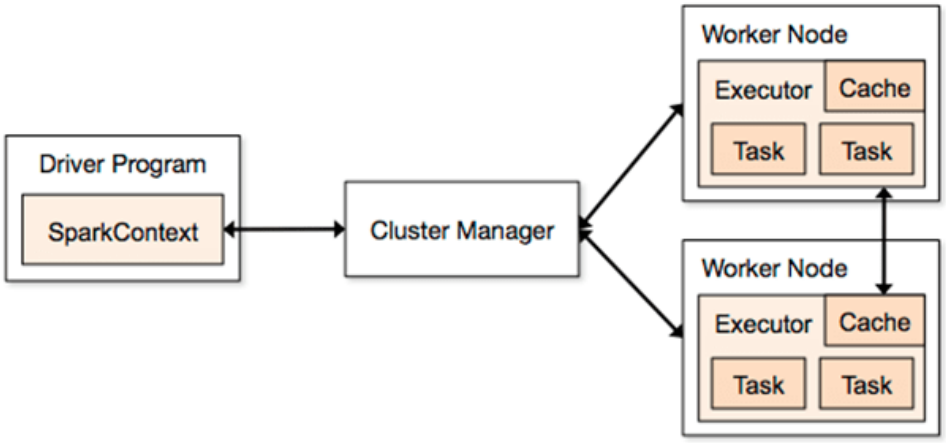


Figure 2: *Basic architecture of Spark (Raviyanshu, 2022)*

As we can see in the figure above, the core components of Spark include the driver, executors and cluster manager. The Driver orchestrates the execution of tasks by converting them into a directed acyclic graph (DAG) of stages. Executors are distributed across the cluster to execute these tasks and store data for future operations. The Cluster Manager, which can be Spark's own standalone manager, YARN, or Mesos, oversees resource allocation and task scheduling. Spark's use of RDDs provides fault tolerance and efficient data processing by allowing transformations and actions to be performed in memory. This architecture offers significant benefits such as high processing speed, scalability, and ease of use, making it ideal for SMEs looking to leverage big data technologies for their machine learning and data science projects.

In contrast to Pandas, Spark's data frame schema entails specifying the data type for each column, akin to traditional databases, where columns enforce fixed data types on all records. While newer No-SQL databases may offer more flexibility, the concept of

type enforcement remains robust and pertinent. Furthermore, Spark facilitates easy schema inspection, allowing users to effortlessly visualize the columns and their respective data types. Unlike Pandas, Spark lacks support for indexing to enable efficient access to individual rows within a DataFrame. Instead, Spark approaches tasks requiring indexing with a brute-force methodology, as transformations are applied to all records, with data reorganization performed dynamically as necessary. Unlike Pandas, Spark doesn't treat columns and rows interchangeably, owing to its scalability focus on the number of rows rather than columns. While Spark seamlessly handles datasets with billions of rows, it's advisable to limit the number of columns to maintain optimal performance, typically within the range of hundreds to a few thousand.

An important distinction between Spark and Pandas lies in their handling of columns and rows, with Spark lacking the interchange ability of these elements. Unlike Pandas, transposing operations aren't straightforward in Apache Spark. Initially, Pandas may seem to offer greater flexibility and integration with the Python data science ecosystem. However, delving deeper into Apache Spark reveals its intrinsic capabilities.

In contrast to Pandas' immediate execution approach, Spark employs a lazy execution model. Transformations applied to a Data-frame aren't processed instantly; instead, they are recorded in a logical execution plan, forming a graph where nodes represent operations. Spark optimizes the entire plan before execution, enhancing efficiency. Additionally, Spark's inherent multithreading and design for large clusters enable it to utilize machine cores and cluster resources effectively.

Spark's scalability extends beyond multiple machines in a cluster, leveraging disk storage to handle vast amounts of data. Unlike Pandas, which may encounter memory limitations, Spark's design allows for data processing without requiring the complete materialization of the dataset in RAM. Instead, data is divided into smaller chunks,

processed independently, and stored in smaller units, eliminating the need to fit all data into RAM simultaneously. This approach ensures efficient resource utilization, making Spark a robust solution for big data processing tasks. Following figure (figure 3) represent the compute time analysis for the dataframe in Spark and Pandas environments.

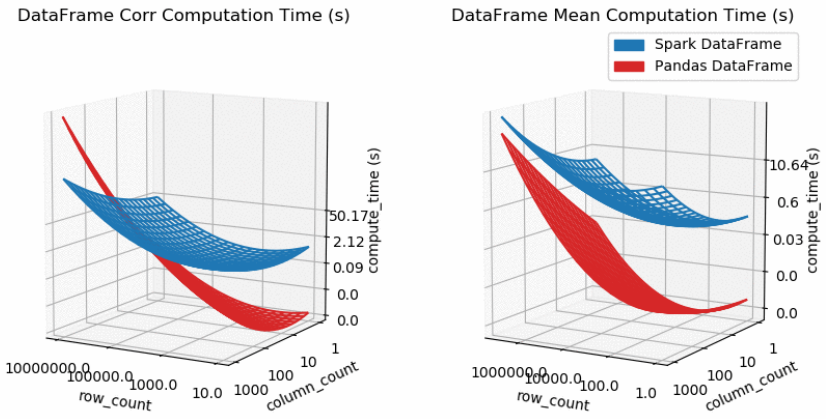


Figure 3: Dataframes correlation and compute time analysis in spark and pandas. (*Lee K.C. , 2021****)***

As we can see from this figure which compares the computation time for processing a dataframe in Spark and Pandas, it is observed that Spark generally outperforms Pandas for large datasets due to its distributed computing capabilities. For example, when performing tasks such as filtering, grouping, and aggregating a dataframe with a million rows, Spark efficiently distributes the workload across multiple nodes, significantly reducing the processing time. In contrast, Pandas operates in-memory on a single machine, which can lead to slower performance and higher memory usage as data size increases. However, for smaller datasets, Pandas may be faster and more convenient due to its simpler setup and lower overhead.

Conclusion of the discussion on Spark: In examining Spark and Pandas, it's evident that while Spark may have slightly less flexibility compared to Pandas, its

unparalleled scalability compensates for this limitation. Designed for addressing different problem domains, Spark stands out for its ability to handle massive datasets and utilize resources efficiently. While Pandas excels in data reshaping tasks, Spark's strength lies in its capacity to work with enormous datasets by leveraging disk space, RAM, and scaling across multiple CPU cores and machines in a cluster. Spark's versatility makes it a powerful tool for manipulating large datasets, particularly those with a fixed schema, even if they contain billions of rows. Its scalability, coupled with connectors to various storage systems, positions Spark as an exceptional choice for Big Data engineering and data integration endeavours. Despite the trade-offs in flexibility, Spark's ability to tackle extensive datasets with ease makes it a compelling solution for handling complex data processing tasks at scale.

The following figure (figure 4) illustrates the time analysis results obtained when processing data of approximately 0.5 GB size, along with the code for a fundamental task. Additionally, it depicts the practical time analysis conducted using both Spark and Pandas for two distinct file sizes. Notably, the x-axis of the graph represents the duration of time in seconds.

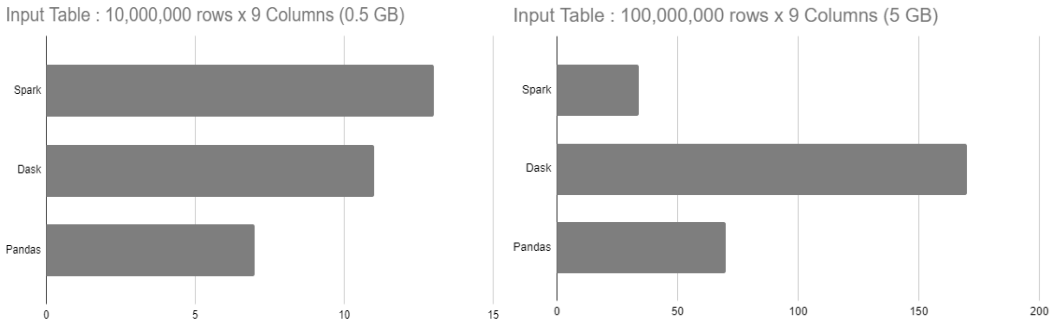


Figure 4: *Time comparison of frameworks based on input size. (Censius AI , 2021)*

Upon repeating the basic task twice, it was observed that Dask and Pandas yielded nearly identical results, while Apache Spark exhibited a slight difference of approximately 2 seconds. In the graph provided below, Apache Spark averaged around 13 seconds, whereas Dask and Pandas averaged 11 and 7 seconds, respectively. Consequently, Pandas demonstrated superior performance when handling straightforward tasks with smaller data sets. However, if the data size is increased, say to around 5GB, the outcome may vary. As depicted in the figure below, when confronted with substantial data volumes, Apache Spark notably outperforms, completing the task in less time, whereas Dask requires approximately 170 seconds.

C - Programming Languages based Analysis:

When contrasting Spark and Pandas, it's essential to consider the programming languages supported by each framework. While Pandas is exclusively limited to Python, Spark offers compatibility with Scala, Java, Python, and R, with additional bindings under development by their respective communities (Shanahan, 2015). Selecting the appropriate programming language for machine learning projects in small and medium-sized enterprises (SMEs) carries significant implications. Python stands out due to its extensive library support for data science and its code readability, making it well-suited for data scientists with strong mathematical backgrounds but limited programming expertise. Leveraging PySpark, data scientists can handle large datasets exceeding local machine memory limits while retaining access to crucial Python libraries, provided data can be downscaled or aggregated to render these tools feasible again. Conversely, many data engineers advocate for using Spark with Scala for data processing and engineering tasks. Proficiency in Scala is essential for data engineers, as it offers better performance and extensive extensibility over Python. Scala's native integration with Spark allows

developers to implement custom transformations not available in Spark, accessing even its internal developer APIs for more comprehensive functionality.

Following figure shows the memory usage statistics on loading parquet files of different sizes. It can be seen from the figure that as the file size goes on increasing, the code becomes more prone to out of memory errors if we don't go with distribute computing using PySpark. On the other side, if the data is smaller, we can consider Pandas due to its ease to set-up and run on any local single core machine.

Figure 5: *Pandas and Spark run time comparison (Tri Juhari, Medium, 2021)*

From all the above discussions, it is clear to selection of a programming language is a multifaceted decision, necessitating consideration of both functional and non-functional requirements. While Python excels in data science, Scala emerges as the preferred choice for data engineering tasks within the Spark ecosystem, enabling SMEs to optimize their machine learning projects effectively.

D – Other factors affecting the project cost for SMEs:

In the realm of small and medium-sized enterprises (SMEs), the execution of data science projects is often influenced by a multitude of factors beyond the traditional considerations of hardware and software expenses. Understanding these diverse factors is crucial for SMEs to effectively manage project costs and ensure successful outcomes.

This literature review aims to explore the various dimensions that impact the financial aspects of data science projects in SMEs, shedding light on factors that extend beyond hardware and software investments. By examining these factors comprehensively, this review seeks to provide valuable insights into the nuanced financial implications associated with undertaking data science initiatives in SME environments.

One of the key factors affecting project costs for SMEs is the acquisition and retention of skilled talent in the field of data science. According to surveys conducted by industry Small Business Administration (SBA), SMEs often face challenges in recruiting and retaining professionals with expertise in data analysis, machine learning, and statistical modeling (Smaldone, 2022). The demand for skilled data scientists exceeds the available talent pool, leading to competitive salaries and recruitment expenses. Additionally, investing in ongoing training and development programs to up-skill existing staff further adds to the overall project costs. A survey conducted by McKinsey & Company found that 60% of SMEs reported talent shortages as a significant barrier to implementing data science projects. Ensuring the quality and reliability of data is another critical factor influencing project costs for SMEs engaged in data science initiatives. Surveys conducted by organizations like Gartner have highlighted the challenges faced by SMEs in managing and maintaining data quality. SMEs may incur expenses related to data acquisition from external sources, licensing fees for proprietary datasets, and data cleansing processes to eliminate inconsistencies and errors. Poor data quality can compromise the accuracy and reliability of analytical insights, leading to erroneous decision-making and increased project costs over time. Therefore, SMEs must allocate resources towards implementing robust data quality management practices to enhance the reliability and usefulness of their data assets.

Compliance with regulatory standards and data security considerations pose significant cost implications for SMEs engaged in data science projects. Surveys conducted by PricewaterhouseCoopers (PwC) have highlighted the increasing regulatory burden faced by SMEs in various industries. SMEs must allocate resources towards ensuring compliance with data protection laws, industry regulations, and privacy mandates, which may involve conducting compliance audits, implementing data governance frameworks, and investing in legal counsel. Moreover, safeguarding sensitive data against cyber threats, unauthorized access, and breaches requires investments in cyber security technologies, encryption protocols, and security infrastructure. Non-compliance with regulatory requirements can result in severe financial penalties, legal liabilities, and reputational damage, underscoring the importance of prioritizing investments in compliance and security measures.

2.3 Original Contribution to Knowledge:

When delving into the specifics of the Pandas library, its simplicity, flexibility, and accessibility make it an ideal tool for data exploration and preliminary experimentation, provided the data fits into memory. This is particularly true for machine learning projects, where Pandas serves as the starting point due to its seamless integration with powerful ML libraries. Even when dealing with large datasets, Pandas offers a user-friendly and adaptable solution for initial experiments with a data subset. However, it's important to note that due to Python's dynamically typed nature and the associated weaker correctness guarantees, Pandas is not recommended for production workloads. Nonetheless, the extensive array of essential ML libraries available makes both Python and Pandas valuable for certain production applications today.

In scenarios where Pandas struggles, Spark excels due to its impressive scalability across CPUs, machines, and especially data volumes. Unlike Pandas, Spark can handle virtually unlimited data, with time being the primary constraint rather than resources. Additionally, Spark offers an extensive range of connectors for various data sources and destinations, making it highly proficient at integrating diverse data types. Its use of Scala, a statically typed and compiled language, enhances the robustness of Spark code compared to Python. This robustness, coupled with its scalability and integration capabilities, makes Spark—especially when used with Scala—a highly advantageous choice for production environments.

Spark is engineered for handling massive datasets, excelling in performance over its predecessor Hadoop. However, for smaller datasets where Pandas operates efficiently within seconds, Spark can be slower. Although Spark includes some machine learning (ML) algorithms, it cannot rival the extensive library available in Python. It’s essential to remember that Spark is built around relational algebra for cluster computing, which differs fundamentally from the matrix algebra operations (like matrix multiplication and decomposition) often required in ML algorithms. Consequently, implementing ML algorithms in Spark involves converting numerical problems into map/reduce tasks suitable for distributed processing. This conversion is more complex and partly explains the slower pace of new ML feature development in Spark. While Spark is ideal for ETL/ELT tasks, it is often a secondary choice for ML projects due to its limited algorithm offerings. Nevertheless, for extremely large datasets where sub-sampling isn’t viable, Spark remains a viable solution.

Up to this point, we've focused on the advantages of using either Pandas or Spark for various tasks but haven't explored the potential of combining them within a single application. This integration is feasible through PySpark, the Python interface for Spark.

PySpark not only provides a Python API for Spark but also increasingly supports the seamless inclusion of Pandas code. The motivation behind this integration is the understanding among Spark developers that Spark cannot, and perhaps should not, completely replace Pandas, especially in certain machine learning contexts. Instead, Spark's development is geared towards facilitating integration with tools like Pandas and TensorFlow.

2.4 Conclusion:

In this literature review, we studied libraries, languages and frameworks along with their limitations. This study will help the companies to make wise decisions while building the data science platform. The Final aim is to recommend and establish a complete infrastructure in the SMEs so that they can develop the machine learning requirements efficiently.

CHAPTER III:

METHODOLOGY

3.1 Overview of the Research Problem:

In today's data-driven landscape, Small and Medium Enterprises (SMEs) are increasingly recognizing the importance of harnessing the power of data science to gain a competitive edge. Data science projects hold the potential to unlock valuable insights, optimize operations, and enhance decision-making processes. However, SMEs often face significant challenges when attempting to transition these projects from the prototype stage to full-scale production. This transition involves overcoming various technical, operational, and financial hurdles that can be formidable barriers for resource-constrained organizations.

One of the primary obstacles SMEs encounter is the lack of an efficient and cost-effective infrastructure tailored to their specific needs. Unlike larger enterprises with substantial budgets and dedicated data science teams, SMEs often operate on tight resource constraints. They may lack the technical expertise and financial resources required to establish and maintain a robust data science infrastructure. As a result, promising data science prototypes often remain stuck in a development phase, preventing SMEs from realizing the full potential of their investments in data-driven technologies.Moreover, the absence of a well-structured infrastructure can lead to inefficient processes, data silos, and suboptimal collaboration among teams. This not only hampers the scalability of data science initiatives but also diminishes the overall return on investment for SMEs. To address these challenges, it is imperative to research and develop a framework that allows SMEs to efficiently and cost-effectively transition their data science projects from the experimental phase to production.

This research problem, therefore, centers on the urgent need to establish a practical and tailored infrastructure solution that empowers SMEs to navigate the complexities of data science project development. This infrastructure should encompass not only the technical aspects of data management, analytics, and deployment but also the organizational and financial considerations specific to SMEs. By addressing these challenges comprehensively, SMEs can unlock the true potential of data science, fostering innovation, growth, and competitiveness in an increasingly data-centric business environment. This thesis aims to delve into the methodologies and strategies required to tackle this problem and provide valuable insights for SMEs looking to harness the transformative power of data science in their operations.

3.2Operationalization of Theoretical Constructs:

In order to deliver cost-effective recommendations, it's essential to delve into existing projects encompassing the most prevalent **machine learning tasks**. Additionally, a thorough examination of the official documentation of widely used **libraries** and **frameworks** is crucial. By scrutinizing both, gained insights into best practices, efficient implementations, and innovative approaches, thereby enhancing the ability to devise optimized solutions tailored to specific needs and constraints. Apart from these aspects, it is also important to study the **cloud services** and its benefits in terms of cost and maintenance. Following is the list of sources which allowed us to operationalize the theoretical constructs for this research.

- **Journal of Machine Learning Research (JMLR**): Comprehensive survey articles on various machine learning topics.
- **ACM Computing Surveys**: Surveys covering a wide range of computer science topics, including machine learning.

- **ICML (International Conference on Machine Learning)**: Surveys on machine learning theory, algorithms, and applications.
- **CVPR (Conference on Computer Vision and Pattern Recognition)**: Surveys on computer vision and related machine learning topics.
- **Google Scholar:** Scholarly articles, including surveys and reviews.
- Websites including Kaggle, Towards Data Science, and Medium feature articles, tutorials, and surveys related to machine learning and data science.
- **Pandas Documentation**: Pandas, a Python library, offers easy-to-use data structures and data analysis tools, making it ideal for data manipulation and analysis tasks.
- **Apache Spark Documentation**: Apache Spark, an open-source distributed computing system, facilitates large-scale data processing and analytics across clusters with speed and fault tolerance.
- **TensorFlow Documentation**:TensorFlow, developed by Google, is a comprehensive machine learning platform that enables building, training, and deployment of various ML models, including neural networks.
- **PyTorch Documentation**:PyTorch, an open-source deep learning framework, offers a dynamic computational graph and seamless GPU acceleration, making it popular among researchers and practitioners.
- **Scikit-learn Documentation** : Scikit-learn, a Python library, provides simple and efficient tools for data mining, machine learning, and data analysis, making it suitable for both novice and expert users.
- **Amazon Web services Documentation**: AWS is a comprehensive and widely-used cloud computing platform offered by Amazon. It provides a vast array of services, including computing power, storage solutions, database management,

machine learning tools, and more. With global infrastructure and scalability, AWS is favored by businesses of all sizes for building and deploying various applications and services.

- **Microsoft Azure Documentation**: Azure is Microsoft's cloud computing platform that offers a broad set of services and solutions for building, deploying, and managing applications and services through Microsoft's global network of data centers. Azure provides a wide range of services, including virtual machines, databases, AI and machine learning tools, IoT services, and more. It is popular among enterprises due to its integration with other Microsoft products and services

To obtain comprehensive and up-to-date information about cloud services offered by AWS (Amazon Web Services), Azure (Microsoft Azure), we visited their official documentation and websites. On these official websites, we found detailed information about the cloud services they offer, including but not limited to Compute services (e.g., virtual machines, containers, server less computing), Storage and database services, Networking services (e.g., virtual networks, load balancers), Machine learning and AI services, Analytics and big data services, DevOps and developer tools, Identity and security services, IoT (Internet of Things) services, Migration and hybrid cloud solutions, Industry-specific solutions.

We should note that cloud services are regularly updated and expanded, so it's important to refer to the official documentation for the most accurate and current information.

3.3Research Purpose and Questions

The purpose of this thesis is to investigate and propose practical solutions for establishing an efficient and cost-effective infrastructure tailored specifically to the needs of Small and Medium Enterprises (SMEs) aiming to transition their data science projects from the prototype stage to full-scale production. As a part of this thesis, we also aim to deliver a ML assistance chat bot application, "**Smart ML assistance**", which can answer queries as ML expert. In today's highly competitive business landscape, data science has emerged as a key driver of innovation and competitiveness, and SMEs, in particular, stand to benefit significantly from its application. However, the journey from a promising data science prototype to a fully operational and productive system is fraught with challenges, and this paper seeks to address this critical issue.

The primary research objective is to comprehensively **analyze the existing barriers** that hinder SMEs from effectively progressing their data science initiatives. These barriers encompass technical, operational, and financial aspects, which, when left unaddressed, can impede the growth and sustainability of data-driven projects within SMEs. By understanding these obstacles in depth, this paper aims to provide actionable insights and strategies to overcome them, ultimately enabling SMEs to harness the full potential of data science.

Furthermore, this research seeks to outline a clear and adaptable framework for SMEs to establish an infrastructure that aligns with their unique requirements and constraints. The study will encompass not only the technological aspects of data science but also the organizational and financial considerations that are critical for SMEs. It aims to serve as a practical guide for SMEs looking to make data science an integral part of their business strategy, enabling them to navigate the complexities of data project development efficiently.In addition to addressing the challenges and providing solutions,

this thesis aspires to highlight the tangible benefits that SMEs can reap by implementing an efficient data science infrastructure. These benefits include improved decision-making, enhanced operational efficiency, and increased competitiveness in the market. By showcasing real-world examples and case studies, this paper aims to demonstrate the return on investment that SMEs can achieve when they successfully transition data science projects from prototype to production. Ultimately, the research purpose of this thesis is to contribute to the body of knowledge surrounding data science adoption in SMEs and offer practical guidance to decision-makers, IT professionals, and entrepreneurs within these organizations. By doing so, it aspires to empower SMEs to leverage data science as a transformative tool, fostering innovation and growth in a data-centric business environment.

In pursuit of a comprehensive understanding of how to establish an efficient and cost-effective infrastructure for SMEs to drive data science, this thesis addresses several key research questions:

1. What are the fundamental infrastructure requirements for SMEs to effectively integrate machine learning and data science into their operations?
2. What are the commonly faced challenges and issues by SMEs in data science infrastructure?
3. What components / factors plays a key role while doing the cost analysis of the project in SMEs?
4. What metrics and KPIs should SMEs track to evaluate the performance of their machine learning infrastructure?
5. What are the knowledge and skill gaps SMEs need to address within their workforce to effectively utilize and manage infrastructure for machine learning and data science?

6. What general infrastructure strategies should be employed by the SMEs while managing a machine learning project?

By addressing these research questions, this thesis aims to provide a comprehensive and actionable guide for SMEs seeking to establish a data science infrastructure that optimizes their capabilities, resources, and potential for innovation in a data-driven world.

3.4Research Design

The research design for addressing the topic of "Establishing an efficient and cost-effective infrastructure for Small and Medium Enterprises (SMEs) to drive data science projects from prototype to production" is carefully crafted to provide a systematic and comprehensive exploration of the subject matter. The chosen research design combines qualitative and quantitative research methods to ensure a well-rounded analysis and practical recommendations (Pfeuffer, 2023). This research adopts a mixed-methods approach that integrates both qualitative and quantitative research techniques. This approach allows for a deeper understanding of the challenges and opportunities associated with data science infrastructure in SMEs. Following are the key steps which we followed to conduct this research.

1. Literature Review:

We have conducted an extensive literature review to explore existing research on data science infrastructure, specifically focusing on studies related to small and medium enterprises (SMEs), infrastructure establishment, cost-effectiveness, and project management. This step allowed us to identify gaps, trends, and best practices to inform our research design.

2. Define Research Objectives:

We have defined clear research objectives aligned with the overarching goal of establishing an efficient and cost-effective infrastructure for SMEs to drive data science projects from prototype to production. These objectives serve as guiding principles for our study, directing our efforts towards addressing pertinent research questions.

3. Develop Research Questions:

Based on the defined objectives, we have formulated specific research questions to guide our investigation. These questions are designed to be clear, concise, and directly relevant to our research objectives, shaping the focus of our data collection and analysis efforts.

4. Select Research Methodologies:

We have adopted a mixed-methods approach, combining qualitative and quantitative methods to achieve a comprehensive understanding of the research topic. This approach includes conducting interviews, surveys, case studies, and employing data analysis techniques to gather and analyze data from multiple perspectives.

5. Design Data Collection Instruments:

We have developed tailored data collection instruments to align with our chosen methodologies. For qualitative research, we have designed interview guides and case study protocols, while for quantitative research, we have created survey questionnaires and data collection protocols. These instruments are designed to effectively capture relevant data in line with our research questions.

6. Full scale Data Collection:

We have carried out data collection activities according to our designed methodologies. This involved conducting interviews, administering surveys, and

gathering information through case studies. Throughout the data collection process, we ensured adherence to ethical principles, including informed consent, confidentiality, and data protection.

7. Data Analysis and Interpretation:

We analyzed the collected data using appropriate techniques and tools. For qualitative data, we employed thematic analysis to identify patterns, themes, and insights. For quantitative data, we used statistical analysis to examine relationships, trends, and correlations. We integrated findings from both qualitative and quantitative analyses for a comprehensive understanding. Upon completion of data analysis, we interpreted the results in relation to our research questions and objectives. We synthesized the findings to draw conclusions and insights regarding the establishment of efficient and cost-effective infrastructure for SMEs to drive data science projects from prototype to production.

8. Validation and Verification:

We validated our research findings through peer review, expert consultation, and comparison with existing literature. We verified the accuracy and reliability of the results by ensuring transparency in our data collection, analysis, and interpretation processes.

9. Conclusion and Recommendations:

Finally, we will summarize the key through these key steps in our research design, we aim to conduct a systematic and rigorous investigation into the establishment of an efficient and cost-effective infrastructure for SMEs to drive data science projects from prototype to production.

3.5 Data Collection Procedures

For the research on establishing an efficient and cost-effective infrastructure for Small and Medium Enterprises (SMEs) to drive data science projects from prototype to

production, a comprehensive approach was adopted, incorporating interviews, surveys, literature review, and discussions (Whang, 2023). Additionally, popular machine learning (ML) datasets were analyzed to provide practical insights into the challenges and requirements faced by SMEs in this context.

- **Interviews**: S To gain the insights for the first research question, where we explore the fundamental infrastructure requirements for SMEs, We have employed interviews with 12 experienced data scientists who are currently working in SMEs and have exposure in data science related projects. The interview guide and interview questions can be found in the appendix section C and D. These interviews aim to capture rich, in-depth insights into the challenges, opportunities, and strategies related to establishing infrastructure for data science initiatives. The qualitative data gathered through interviews provided valuable context and nuanced understanding of the factors influencing infrastructure implementation in SMEs.

- **Case Study**: To explore the metrics and KPIs to be monitored by SMEs, we explored the case studies from two SMEs, Influx Data and Stitch Fix. Stitch Fix company functions as an online personal styling service, utilizing machine learning algorithms to tailor clothing recommendations for its customers. Influx data provides time series solutions to different industries. Additionally, we also studied the challenges encountered by "Molecule 53," a startup in the cosmetic products industry during their journey in implementing data science projects, and how they navigated through these challenges to achieve success. These case studies gave the insights on infrastructure related metric and KPIs to be tracked to maximize the return on investment.

Surveys: To gain the deeper understanding of the infrastructure challenges, we conducted an online survey named "Challenges and Issues in Data Science Infrastructure for SMEs". This survey was mainly targeted to the data scientists working within SMEs. The details of this survey along with the questionnaire can be found in the Appendix E section. It aimed to identify the most commonly faced issues and obstacles related to data handling and infrastructure, offering valuable insights directly from professionals at the forefront of these endeavors. We received the responses from 24 candidates which are summarized in "Results" section. The quantitative data obtained from surveys are analyzed to identify trends, patterns, and statistical relationships relevant to our research objectives. **Discussions with Machine Learning Experts**: To explore the gaps in knowledge and skill sets for data science resources, this thesis gathered the insights through discussions with 4 data science leads from SMEs, each providing a unique perspective on the hurdles their organizations face and the competencies they deem essential for success. These leads were involved in resource allocation procedure in the project hence were well aware of the knowledge and skills gaps in data science. They also had an exposure in cost-analysis of the projects, providing the practical insights for key components which analyzing the cost.

- **Exploration of Success Stories, Research Papers and Online Forums**: Success stories, and use cases of SMEs that had successfully implemented data science projects were examined. These real-world examples provided practical insights into infrastructure setup, challenges encountered, and strategies employed to drive data science projects from prototype to production. Technical articles , online forums, and community platforms where data scientists, industry experts, and

SMEs shared insights, experiences, and best practices related to data science infrastructure and project management were reviewed. Participation in discussions, questions, and engagement with the community facilitated gaining practical insights and advice. Guidance was sought from academic advisors, research mentors, and subject matter experts in the field of data science, machine learning, and business management. They provided recommendations for relevant literature, insights into emerging trends and research directions, and feedback on the research approach.

By leveraging a combination of interviews, surveys, literature review, discussions, valuable insights were obtained into the challenges and opportunities for SMEs in establishing efficient and cost-effective infrastructure.

3.6 Dataset Validation:

For this research, various data collection methods, including interviews, discussions, and surveys with data scientists from different companies, were conducted to validate the collected data, several methods were employed.

One of the methods was Triangulation. It involved cross-verifying information gathered from multiple sources, such as interviews, discussions, and surveys. By comparing data from different methods, inconsistencies could be identified and addressed, enhancing the validity of the findings. This approach ensured that the data obtained from different sources converged on similar conclusions, strengthening the overall reliability of the research outcomes. We also did the **member checking**. It was utilized to validate the accuracy of the data collected through interviews, discussions, and surveys. Participants were given the opportunity to review transcripts, summaries, or survey responses to confirm their accuracy and provide feedback. This process ensured

that the data accurately reflected the perspectives and experiences of the data scientists interviewed, enhancing the credibility of the research findings.

Another method of validation was **Peer debriefing**. This involved seeking feedback from colleagues or peers familiar with the research topic and methodology. Colleagues reviewed interview transcripts, survey instruments, and data analysis procedures to identify any biases, errors, or misinterpretations. Their input helped validate the data interpretation and improve the rigor of the research process, ensuring that the findings were robust and trustworthy. An **audit trail** was maintained throughout the research process to document decisions made during data collection and analysis. This detailed record of interview protocols, survey instruments, and analytical decisions ensured transparency and accountability, enabling external reviewers to replicate the study and validate the data. It provided a clear documentation of the research process, enhancing the trustworthiness of the research outcome. **Saturation** was monitored to determine when data collection reached theoretical saturation, indicating a comprehensive understanding of the research topic. This occurred when no new themes or insights emerged from subsequent interviews, discussions, or surveys. Achieving saturation ensured that the collected data adequately represented the diversity of perspectives among data scientists, validating its reliability and validity for informing the research objectives.

Apart from the data collected from interviews and discussions, we also referred the proven sources for accessing the publications and research papers. These sources include Google Scholar, CVPR (Conference on Computer Vision and Pattern Recognition) and JMLR (Journal of Machine Learning Research) etc. Research papers published in conferences such as ICML and CVPR undergo a rigorous peer-review process where experts in the field evaluate the quality, significance, and validity of the

research presented. This peer review ensures that the data and findings presented in these papers are of high quality and credibility. We also performed Cross-referencing to verify the consistency and accuracy of the data presented in multiple research papers. Data from different papers discussing similar topics or methodologies were compared to identify any discrepancies or inconsistencies. Consistent findings across multiple papers increased the reliability and validity of the data.

3.7 Data Analysis:

To analyze the data collected from interviews, surveys, and discussions for this research the diverse steps were taken. The collected data was cleaned to remove any errors, inconsistencies, or missing values. It was organized into a structured format suitable for analysis, ensuring consistency in data formats across different sources (interviews, surveys, discussions).The distribution of key variables, such as project size, budget, timeline, and infrastructure requirements, was explored. Patterns, trends, and outliers in the data were identified to inform further analysis. The qualitative data obtained from interviews and discussions was analyzed using thematic analysis techniques. Recurring themes, patterns, and insights related to infrastructure challenges, project workflows, resource allocation, and stakeholders' perspectives were identified.

Similar responses were grouped into thematic categories to understand common challenges and opportunities. The quantitative data obtained from surveys was analyzed to quantify responses and trends. Descriptive statistics, such as means, medians, and standard deviations, were calculated for numerical variables. Frequency tables and charts were created to visualize responses for categorical variables. Statistical tests (chi-square test, t-test) were performed to assess relationships between variables and identify significant findings. Findings from interviews, surveys, and discussions were integrated to provide a comprehensive understanding of infrastructure challenges and requirements.

Insights from different data sources were compared and contrasted to identify converging or diverging perspectives. The analysis results were synthesized to identify key insights, challenges, and opportunities related to establishing an efficient and cost-effective infrastructure for SMEs' data science projects. Actionable recommendations and strategies were formulated for addressing identified challenges and optimizing infrastructure setup and management. Recommendations were prioritized based on their potential impact and feasibility of implementation within SMEs' constraints. The analysis findings were validated through peer review, expert consultation, or comparison with existing literature and industry best practices. Iterations were made on the analysis process as needed to refine insights, address feedback, and ensure the validity and reliability of the research findings. Through these steps, the data collected from interviews, surveys, and discussions was effectively analyzed to inform the research on establishing an efficient and cost-effective infrastructure for SMEs to drive data science projects from prototype to production.

3.9 Research Design Limitations:

While conducting this research for SMEs, it's important to acknowledge and address potential limitations that may affect the scope and validity of the study. These limitations include:

The research findings may not be fully generalizable to all SMEs due to variations in industry, size, location, and technological maturity. The study's scope may focus on specific segments of SMEs, potentially limiting the broader applicability of the proposed infrastructure solutions. Data collected from surveys, interviews, and case studies may be subject to response bias, as participants may provide information that reflects their experiences or opinions. Efforts will be made to minimize this bias, but it cannot be

entirely eliminated. The research may be constrained by limitations in terms of time, funding, and access to SMEs. This could impact the depth and breadth of data collection and analysis, potentially leading to a partial representation of the challenges faced by SMEs.

The field of data science and technology infrastructure is rapidly evolving. Infrastructure solutions and best practices available at the time of the study may become outdated in the future. Therefore, the proposed solutions may need to be adapted to keep pace with technological advancements. The financial aspects of SMEs, including their budget allocations for data science projects, can vary significantly. While the research will analyze budget considerations, SMEs' financial situations may change over time, affecting their infrastructure decisions. The research may encounter limitations related to the collection and analysis of sensitive data, including compliance with data privacy regulations. Ethical and privacy considerations will guide the handling of such data.

The case studies selected for this research may not represent the full spectrum of SME experiences. The selection of cases may be influenced by factors such as availability, willingness to participate, and relevance to the research objectives.The success of infrastructure implementation can be influenced by factors such as leadership, organizational culture, and the skills of the workforce. While the research aims to address technical aspects, the human element may not be comprehensively covered.

Addressing these limitations is essential to ensure the research's transparency and reliability. Mitigation strategies, such as robust data collection and analysis techniques, clear research objectives, and acknowledging the scope of the study, will be implemented to minimize the impact of these limitations on the validity and usefulness of the research findings.

3.10 Conclusion:

In conclusion, this research has yielded invaluable insights into the establishment of a cost-effective infrastructure to the needs of SMEs. Through the Operationalization of theory using interviews and surveys, dataset validation, and the specification of research objectives, we have comprehensively addressed critical questions surrounding the value of data science, frequently used machine learning tasks, pain areas where data science may be ineffective, challenges faced by smaller companies in data and infrastructure, key components influencing cost analysis in projects with limited budgets, and general infrastructure recommendations. These findings are instrumental in guiding SMEs as they navigate the complexities of implementing data science initiatives, empowering them to make informed decisions, optimize resource allocation, and maximize the impact of their data science endeavors.

The methodology section, consisting of interviews, surveys, and dataset validation, serves as the foundation upon which this research is built. By employing rigorous research methodologies, we ensured the reliability, validity, and credibility of our findings, thereby enhancing the robustness of our conclusions and recommendations. This methodology section is pivotal in providing transparency and accountability in our research process, allowing stakeholders to understand the rigor and integrity with which the study was conducted. Additionally, the use of interviews and surveys enabled us to capture diverse perspectives and insights from industry practitioners, enriching the depth and breadth of our analysis.

CHAPTER IV:

RESULTS

The research focused on the establishment of an efficient and cost-effective infrastructure tailored for Small and Medium Enterprises (SMEs) to facilitate the transition of data science projects from the prototype stage to production. Through a comprehensive research design and data analysis, several key findings and results have emerged, shedding light on the challenges, best practices, and actionable recommendations for SMEs in this domain.

4.1 Research Question One: What are the fundamental infrastructure requirements for SMEs to effectively integrate machine learning and data science into their operations?

After aggregating the outcomes of interview discussions and study of research papers, the fundamental infrastructure requirements are divided into these main sections: Hardware, Software, Data Management, Security, Training and Scalability. We have provided the summary findings for each these sections and also the aggregated result in the form of table.

When it comes to hardware requirements, the interview participant 1 mainly emphasized on computing resources, data storage solutions, and development tools. He said, “These resources are essential for tasks such as training ML models, processing large datasets, coding and testing algorithms”. Several options are available to meet these hardware requirements. Companies can choose to invest in physical servers and storage devices, which provide dedicated resources but often require substantial upfront investments and ongoing maintenance costs. Alternatively, cloud-based solutions offered by providers like Amazon Web Services (AWS), Google Cloud Platform (GCP), or

Microsoft Azure offer virtual machines and storage services on a pay-as-you-go basis. This allows companies to access scalable resources without the need for physical infrastructure. Cost-effective options for hardware requirements include leveraging cloud-based solutions. Cloud platforms offer virtual machines and storage services with flexible pricing models, allowing companies to pay only for the resources they use. This eliminates the need for upfront investments in physical hardware and reduces maintenance costs. For example, AWS EC2 instances and S3 storage offer scalable computing and storage capabilities at affordable rates. By opting for cloud-based solutions, smaller companies can access the necessary hardware resources without breaking the bank, enabling them to focus on their ML and data science initiatives without financial constraints.

In the discussion of software requirements, most of the participants suggested the fundamental software tools for data processing, model development, and deployment. These tools include ML frameworks, data management systems, and integrated development environments (IDEs). Such software enables companies to handle data efficiently, build and train models, and deploy solutions seamlessly. Available options for these software requirements vary widely. Companies can choose proprietary software solutions that often come with robust features and dedicated support but can be expensive. Examples include MATLAB, SAS, and Microsoft Azure's ML Studio. Alternatively, there are open-source options like TensorFlow, PyTorch, Apache Hadoop, and various IDEs such as Jupyter Notebooks and VS Code, which provide comprehensive functionalities without the hefty price tag. Cost-effective software options are generally open-source tools. These tools are free to use and offer extensive community support and regular updates. For instance, TensorFlow and PyTorch are powerful ML frameworks used for developing and training models, while Jupyter

Notebooks provides an interactive environment for coding and visualizing data. Using these open-source tools, smaller companies can minimize software costs while still accessing high-quality, reliable software for their ML and data science projects. This approach allows them to allocate resources more efficiently and focus on project development and innovation.

Participant 3 and participant 4, who were mainly working into data pipelines, emphasized on the need for robust data management systems to handle data collection, storage, preprocessing, and analysis Efficient data management is crucial for ensuring data quality and accessibility, which directly impacts the performance and accuracy of ML models. The requirements include databases for storing structured and unstructured data, tools for data cleaning and transformation, and systems for data integration and retrieval. Several options are available for data management. Companies can choose commercial solutions like Oracle, Microsoft SQL Server, or IBM Db2, which offer advanced features and support but often come with high licensing fees. Alternatively, there are open-source databases such as MySQL, PostgreSQL, and Mongo-DB, which provide powerful capabilities without the associated costs. For data cleaning and transformation, tools like Talend, Informatica, and Apache NiFi are available, with varying costs and complexities.Cost-effective data management options include open-source tools, which are both powerful and free to use. MySQL and PostgreSQL are excellent choices for relational databases, offering robust performance and reliability for structured data. For handling unstructured data, MongoDB is a popular open-source option. For data cleaning and transformation, Apache NiFi is a versatile open-source tool that supports data flow automation and integration. Using these open-source tools, smaller companies can manage their data efficiently without incurring significant costs. For example, PostgreSQL offers advanced features like ACID compliance, full-text

search, and custom extensions, making it a strong, cost-effective choice for managing structured data in ML and data science projects.

Apart from these basic requirements, we have also discussed on the aspect of scalability with almost all the participants. The significance of making scalable solutions lies in the ability to handle growth efficiently and cost-effectively. As the proof of concepts evolves with time, they often require processing larger datasets and running more complex models, which demand increased computational resources. Scalable solutions, such as cloud computing and containerization, allow companies to adjust their resources dynamically based on current needs without investing in expensive infrastructure. This flexibility ensures that companies can maintain performance and efficiency while keeping costs under control, enabling them to adapt quickly to changing project demands and business growth. Smaller companies need scalability solutions that allow them to handle increasing amounts of data and growing computational demands without significant downtime or performance loss. Scalability requirements include flexible computing resources, storage solutions that can expand with data needs, and tools for managing and deploying ML models efficiently across different environments. Several options are available to meet these scalability requirements. Companies can invest in high-end physical infrastructure with powerful servers and large-scale storage solutions, but this approach can be costly and inflexible. Cloud computing platforms such as Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure offer scalable virtual machines and storage services on a pay-as-you-go basis. Additionally, containerization tools like Docker and orchestration platforms like Kubernetes can help manage and deploy applications in a scalable and efficient manner.

While discussing with one of the interviewee who was cloud expert, he said “Smaller companies must consider leveraging the cloud for cost-effective scalability

solutions. This approach provides the advantage of enhanced flexibility and faster deployment times." According to his opinion, cloud platforms provide the flexibility to scale resources up or down based on demand, which helps control costs by only paying for what is used. For example, using AWS Auto Scaling, a company can automatically adjust its compute capacity to maintain consistent performance at the lowest possible cost. Additionally, containerization with Docker allows applications to run consistently across various environments, and Kubernetes can automate the deployment, scaling, and management of these containerized applications. These tools reduce the need for extensive physical infrastructure, thereby lowering capital expenditure while ensuring that scalability needs are met efficiently. By adopting these solutions, smaller companies can handle their ML and data science workloads effectively, even as their data and computational needs grow.

One participant, who was mainly involved in the expenses management of the project, said , "To manage the budget and track the expenses effectively, it is important to adapt the cost management software. These requirements include budget tracking systems, expense management software, and tools for financial analysis and reporting". As per his opinion, these tools enable companies to monitor spending, forecast costs, and make data-driven decisions to control and reduce expenses. There are several options available for cost-management tools. Companies can opt for comprehensive enterprise resource planning (ERP) systems like SAP or Oracle, which offer advanced features for financial management but often come with high licensing fees and implementation costs. Alternatively, there are specialized expense management tools like Expensify and Concur, which provide robust functionalities for tracking and managing expenses. For financial analysis and reporting, tools like QuickBooks and Xero are popular choices that offer a range of features tailored to small and medium-sized businesses. Cost-effective

options for cost-management tools often involve using open-source software or affordable cloud-based solutions. For example, Wave Accounting is a free, cloud-based accounting software that offers comprehensive features for expense tracking, invoicing, and financial reporting, making it ideal for smaller companies with limited budgets. Additionally, tools like Google Sheets can be used for budget tracking and financial analysis with customizable templates and easy sharing capabilities. These cost-effective tools provide the necessary functionalities without significant expenses, allowing smaller companies to manage their finances effectively and focus their resources on developing their ML and data science projects.

In summary, adapting a data science in SMEs requires addressing several fundamental needs across hardware, software, data management, and scalability. All the above discussed aspects of infrastructure requirements are summarized into the table shown below (Table 2). This table provides the clear overview of the infrastructure requirements as guided by the data scientists working in SMEs.

Table 2: *Summary of infrastructure requirements for SMEs (Collected from interviews).*

Requirement Category	Key Components	Cost-Effective Options	Purpose
Hardware	High-performance servers/computers	Cloud-based virtual machines (e.g., AWS, Google Cloud)	To avoid high upfront costs of physical servers and ensure scalability.
	GPUs (Graphics Processing Units)	Cloud GPU instances (e.g., AWS EC2 P3, Google Cloud)	To provide necessary computational power for ML tasks without purchasing expensive hardware.
	Data storage solutions	Cloud storage (e.g., AWS S3, Google Cloud Storage)	For scalable, on-demand storage solutions, avoiding high costs of

			physical storage infrastructure.
Software	ML and data science frameworks	Open-source frameworks (e.g., TensorFlow, PyTorch)	For developing and deploying ML models without licensing costs.
	Data management and processing tools	Open-source tools (e.g., Apache Hadoop, Spark)	To handle and process large datasets efficiently at minimal cost.
	Development environments and tools	Free tools (e.g., Jupyter Notebooks, VS Code)	For coding, testing, and visualizing ML models without additional costs.
Data Management	Data storage systems	Open-source databases (e.g., MySQL, PostgreSQL)	For efficient data organization and retrieval with no licensing fees.
	Data pre-processing tools	Open-source ETL tools (e.g., Talend, Apache NiFi)	For preparing data for analysis and modeling without incurring costs.
Scalability	Cloud computing services	Pay-as-you-go cloud services	To scale computational resources as needed, avoiding large upfront investments.
	Containerization and orchestration tools	Open-source tools (e.g., Docker, Kubernetes)	For managing and deploying ML applications efficiently at minimal cost.
Cost Management	Budget planning tools and financial management software	Free budgeting tools (e.g., Mint, Wave)	For tracking and optimizing expenditures related to ML and data science projects without additional costs.
	Cost-effective resource allocation strategies	Serverless computing (e.g., AWS Lambda)	To maximize the value obtained from infrastructure

			investments by only paying for what is used.

During the discussion with one of the data scientist, another important aspect came in which explores the strategies that SMEs could adopt to ensure **data security and** **"**SMEs should have utilized role-based access control, also called as RABC, to restrict data access based on employees' roles within the organization. This measure would have minimized the risk of data breaches by ensuring that only authorized personnel could access sensitive information". Furthermore, the data scientist recommended data Anonymization techniques, where personal identifiers were removed or obfuscated to protect individual privacy. This process would have made it difficult to trace data back to specific individuals, thereby safeguarding personal information.

The interviewee emphasized on using the secure data storage solutions that complied with industry standards and regulatory requirements. The interviewee said, "Whether using on-premise servers or cloud services, it is crucial to ensure these solutions are secure and capable of protecting sensitive data". Lastly, the data scientist stressed the importance of regular security audits and continuous monitoring of systems for any suspicious activities. By identifying vulnerabilities and addressing them promptly, SMEs could have maintained a high level of data security and privacy in their machine learning and data science operations.

4.2 Research Question 2: What are the commonly faced challenges / issues by smaller companies in Data and Infrastructure?

The findings gained from the survey, which addresses this question, are divided into two sections; firstly we talk about types of data generally used in machine learning task along with the data challenges. Secondly we talk about the infrastructure challenges. The survey questioner can be referred from the appendix section. Kindly note that the

results obtained are collected from the survey responses plus referring the research papers for additional insights.

As a part of survey, we also collected the type of datasets in which the candidates have worked. This exploration is useful to identify the data type specific problems in the project. As we understood from the survey findings and research papers, various types of data play a critical role, each bringing its unique attributes and applications (Kumar, 2017; Agrawal, 2020). Firstly, the popular data type, Numerical data, which consists of quantifiable numbers, can be either continuous or discrete. Continuous numerical data includes examples such as temperature and income, while discrete numerical data encompasses values like age. This type of data is foundational for predictive modelling and statistical analysis. Moreover, categorical data represents distinct categories or labels and is often depicted using strings or integers. This data type is essential for classification tasks, such as identifying colours, gender, or types of vehicles. The ability to categorize data into defined groups enables machine learning models to make sense of and predict categorical outcomes efficiently.In addition to numerical and categorical data, text data is a crucial form of unstructured data comprising sentences, paragraphs, or documents. Examples of text data include product reviews, news articles, and social media posts. This type of data is central to natural language processing (NLP) tasks, where understanding and generating human language is key. Furthermore, image data, represented as arrays of pixel values, is fundamental in the realm of computer vision. This data type involves visual information in the form of images or pictures and is utilized in applications such as object detection, image classification, and facial recognition. The ability to interpret visual data allows for advanced machine learning models to perform complex image-related tasks. Similarly, audio data involves auditory information captured in recordings or signals, often represented by waveforms or

spectrograms. This type of data is vital for tasks such as speech recognition, music analysis, and audio classification. By analysing sound patterns, machine learning models can extract meaningful insights from audio data. Time series data is another critical data type, comprising sequential data points collected over time intervals and indexed by timestamps. This data is indispensable for forecasting and trend analysis, with common applications including stock price prediction, weather monitoring, and analysing sensor readings. Understanding the temporal patterns in time series data enables accurate predictions over time. Lastly, spatial data is associated with geographic locations or spatial coordinates, such as GPS data, maps, and satellite images. This type of data is pivotal for geospatial analysis, providing insights based on geographic contexts and enabling applications in navigation, mapping, and environmental monitoring. By leveraging spatial data, machine learning models can uncover patterns and relationships tied to specific locations.

These diverse types of data in ML experiments bring unique challenges that must be addressed to build effective models. From ensuring data quality and scalability to managing computational resources and achieving model interpretability, each data type demands specialized techniques and considerations. Understanding and overcoming these challenges is crucial for the successful application of machine learning across various domains. One of the primary hurdles faced by practitioners is **missing data**, which can skew results and compromise model integrity if not handled properly (Gupta, 2019). **Outliers**, those pesky data points lying far outside the norm, can similarly disrupt the learning process, demanding meticulous identification and management. Moreover, **imbalanced datasets**, where certain classes are significantly overrepresented or underrepresented, pose another common challenge, necessitating strategic techniques like oversampling or undersampling to ensure fair representation and robust model

performance across all classes. **Data quality** emerges as a critical concern, as poor-quality data laden with inaccuracies and inconsistencies can lead to flawed models. Consequently, diligent preprocessing steps such as cleaning and normalization become imperative to enhance data reliability. Equally crucial is the art of feature selection and extraction—choosing the most relevant features to empower the model with meaningful insights while sidestepping irrelevant noise. Additionally, vigilance against data leakage, whereby extraneous information contaminates model training, is essential to uphold the model's capacity for generalization and real-world applicability. Furthermore, issues like overfitting and underfitting loom large, demanding a delicate balance between model complexity and generalization prowess (Mohammed, 2020). Consistent data formats and standardized scaling techniques are foundational for seamless integration and model convergence. Yet, perhaps one of the most pressing concerns is that of biased data, which can perpetuate societal inequities if left unaddressed. Thus, a conscientious approach to assessing and mitigating bias is indispensable for fostering fair, ethical, and inclusive machine learning systems. By confronting these common data challenges head-on and implementing sound strategies for mitigation, practitioners can chart a course towards more reliable, robust, and equitable machine learning outcomes.

Apart from survey findings, we have also referred the relevant research papers to summarize the data related issues with some of the proven techniques to tackle them (Chowdhury, 2019; Samala, 2020, Sze, 2017; Garg, 2022; Kettimuthu, 2018). Following table (table 3) provides us the clear view of data related challenges and ways to resolve them.

***Table 3**: Summary of general problems in datasets and mechanisms to handle them (Chowdhury, 2019).*

Data Related Issue	Description	Resolution Methods

Missing Data	Incomplete datasets with missing values can lead to biased results and reduced model performance.	1. Imputation: Replace missing values with estimated substitutes like mean, mode, or median. 2. Eliminate rows or columns with missing data. - Use of specialized algorithms to handle missing values.
Outliers	The data points which are significantly different from other observations and can skew model results.	1. Identification: Statistical methods including z-score, IQR to detect outliers. 2. Handling: Consider removing outliers, transforming data, or using robust algorithms. - Domain knowledge to differentiate between genuine outliers and errors.
Imbalanced Data	Imbalance occurs when one class is overrepresented or underrepresented, leading to biased models.	1. Oversampling: Increase the instances for minority class samples. 2. Undersampling: Decrease instances of majority class samples. - Using specialized algorithms like SMOTE designed for imbalanced data.
Data Quality	Poor data quality, including inaccuracies and inconsistencies, can compromise model reliability.	1. Data Cleaning: Remove duplicates, correct errors, and standardize formats. 2. Data Normalization: Scale data to a similar range to improve consistency. 3. Domain expert consultation to verify data accuracy.
Feature Selection/Extraction	Choosing relevant features and extracting meaningful information from raw data is crucial for model accuracy.	1. Feature Selection: Identify and choose the most relevant features for model training. 2. Feature Extraction: Transform raw data into a more compact representation with meaningful information. 3. Dimensionality reduction techniques like PCA or t-SNE.
Data Leakage	Model training is heavily influenced by the information outside of the training data.	1. Strict separation of training and testing data. 2. Feature engineering to remove features prone to leakage. 3. Careful validation techniques to detect leakage during model evaluation.

Overfitting/Underfitting	Overfitting (model memorizes training data) and underfitting (model too simple) hinder model generalization.	1. L1/ L2 regularization techniques 2. Cross-validation to tune model complexity. 3. Ensemble methods like bagging, boosting to improve generalization.
Inconsistent Data Format	Varied data formats across sources/features complicate pre-processing and analysis.	1. Standardization: Convert all data into a consistent format. 2. Data Transformation: Modify data into a unified structure. 3. Automated data integration tools to streamline format conversion.
Feature Scaling	Features havingdiverse scales units impacts the overall model performance.	1. Min-Max Scaling: Scale features to range like 0 to 1 2. Robust scaling methods to handle outliers
Biased Data	Biased datasets perpetuate bias in models, leading to unfair outcomes.	1. Bias Assessment: Identify and quantify biases in the dataset. 2. Bias Mitigation: Modify dataset or algorithm to reduce bias.

Apart from data related issues, smaller companies often encounter unique infrastructure-related challenges when venturing into machine learning projects due to resource constraints and limited expertise (Hopkins, 2021). One of the primary challenges SMEs face is hardware limitations. Unlike large corporations, SMEs often lack access to powerful computing resources necessary for training and deploying complex ML models. The cost of purchasing and maintaining high-performance hardware, such as GPUs and large-scale servers, can be prohibitive. As a result, SMEs may struggle with slower training times and reduced model performance. A cost-effective solution to this problem is utilizing cloud computing services like AWS, Google Cloud, or Microsoft Azure, which offer scalable resources without the need for significant upfront investment. Renting GPU instances or using managed ML services can provide

the computational power required for advanced data science tasks. Another significant issue for SMEs is the shortage of skilled professionals in data science and machine learning. Attracting and retaining top talent can be challenging due to high competition and limited financial resources. SMEs may find it difficult to hire experienced data scientists, ML engineers, and data engineers. To address this, SMEs can invest in upskilling their existing workforce through training programs and online courses. Collaborating with freelancers or consultants on a project basis can also provide access to expertise without the need for full-time hires. Additionally, leveraging open-source tools and participating in knowledge-sharing communities can help bridge the talent gap.

From the survey findings, we found another challenge in maintain effective data management system. SMEs often face challenges related to data governance, storage, and compliance. Managing large volumes of data, ensuring data quality, and adhering to regulatory requirements can be daunting. Implementing cost-effective data management solutions such as open-source databases (e.g. PostgreSQL, MySQL) and data lakes (e.g., Apache Hadoop, AWS S3) can help. Utilizing encryption, access controls, and monitoring tools ensures data security and compliance. Collaborating with third-party vendors for specialized data management services can further enhance data handling capabilities.As data science initiatives grow, SMEs may struggle to scale their infrastructure to accommodate increasing data volumes and computational demands. Limited scalability can lead to performance bottlenecks and hinder the deployment of large-scale ML models. Cloud-based solutions that offer scalability and flexibility are essential for addressing this issue. Using managed services and auto-scaling features allows SMEs to dynamically adjust resources based on their needs, ensuring that their infrastructure can grow with their data science projects. Smaller companies may not have established DevOps practices, which are essential for automating the deployment,

monitoring, and management of ML models and infrastructure. The absence of these practices can result in inefficient workflows and increased operational overhead. Introducing DevOps methodologies gradually, with a focus on automation, continuous integration/continuous deployment (CI/CD), and infrastructure as code (IaC), can streamline processes. Open-source tools and cloud services can facilitate the adoption of DevOps practices, making it easier to manage ML infrastructure efficiently. Integrating machine learning models into existing business processes and systems can be complex, especially for smaller companies with limited technical resources. Utilizing APIs and Microservices architecture can facilitate seamless integration with existing systems. Collaborating with IT teams to align machine learning initiatives with business goals and technical requirements is essential. Starting with small, manageable integrations and scaling gradually can ensure successful implementation and minimize disruption.

Survey candidates also described the challenges in ensuring data security throughout the machine learning process. This aspect is crucial, but smaller companies often lack the resources and expertise to implement robust security measures. Basic security practices such as encryption, access controls, and regular audits are essential to protect sensitive data. Utilizing cloud-based security services and compliance frameworks can simplify security management. SMEs can also benefit from collaborating with cyber security experts to enhance their security posture and ensure compliance with regulations. Budget constraints are a common issue for SMEs, limiting their ability to invest in the necessary infrastructure, tools, and talent for successful data science projects. Prioritizing investments based on project requirements and potential return on investment (ROI) is crucial. Utilizing open-source tools and cost-effective cloud services can help stretch limited budgets. Exploring alternative funding options such as grants,

partnerships, or venture capital can provide additional financial support for data science initiatives.

Following figure (figure 4) shows the summary of the bottlenecks while adapting the AI technology across various sectors. These bottlenecks encompass factors such as access to data, expertise, and computational resources. Limited availability of high-quality data, especially labeled data required for training machine learning models, poses a significant challenge for organizations looking to leverage AI effectively. Additionally, the shortage of skilled professionals with expertise in data science and machine learning further exacerbates the bottleneck, as organizations struggle to recruit and retain top talent in these fields. Furthermore, the computational resources required for training and deploying AI models, including hardware infrastructure and cloud computing resources, can be prohibitively expensive for many organizations. Overcoming these bottlenecks requires concerted efforts to address data challenges, invest in talent development, and explore innovative solutions for optimizing computational resources.

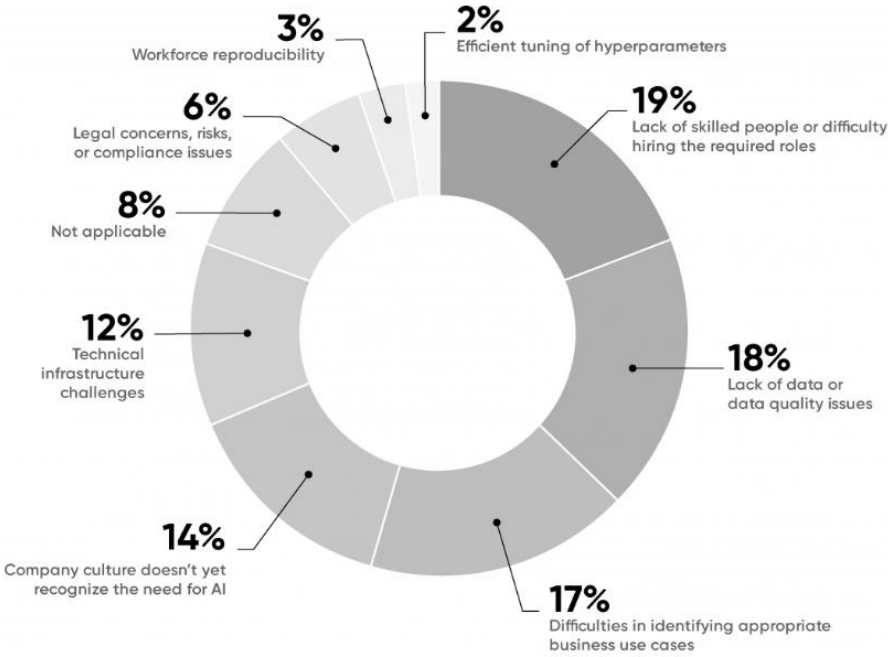


***Figure 6**: Bottlenecks of AI adaption (Loukides M, 2021).*

In conclusion**,** SMEs face a range of challenges when establishing and maintaining efficient data science infrastructure. From hardware limitations and talent shortages to data management and security concerns, these issues can hinder the effective deployment and utilization of machine learning and data science technologies. By leveraging cloud-based solutions, investing in employee training, adopting DevOps practices, and utilizing open-source tools, SMEs can overcome these obstacles and build robust data science infrastructure. Addressing these challenges effectively enables SMEs to harness the full potential of machine learning and data science, driving innovation and growth in a competitive market.

4.3 Research Question 3: What components / factors play a key role while doing the cost analysis of the projects in SMEs?

Review of research papers along with the discussions with data science team have provided practical insights and helped validate the theoretical findings from the literature. All data science leads invited for the discussion were of opinion that cost analysis conduction for ML projects involves many critical components especially when the budget is limited. Participant 1 and participant 2 mainly talked about understanding and defining the **project scope**. One of them said, “Clearly defining the objectives, deliverables, and expected outcomes helps identify the specific tasks and resources required.” Breaking down the project into manageable components helps in granular cost assessment, which enables organizations to prioritize essential activities within budget constraints. Additionally, establishing realistic expectations and goals helps align cost estimates with the project's strategic objectives, ensuring that resources are allocated efficiently to maximize value. Secondly, **assessing resource requirements** is crucial.

Recruiting data scientists, ML engineers, and data analysts can be expensive. SMEs must balance between hiring full-time employees, consultants or outsourcing. Existing staff may need training to keep up with the latest ML and data science techniques, adding to the overall cost. Personnel costs, including salaries, benefits, and training expenses for data scientists and engineers constitute a significant portion of the budget. Another important aspect which came out from the discussion with remaining 2 participants was regarding the initial investment in technology and hardware. Considering the computational resources, software licenses, and data storage infrastructure required for model development and deployment is essential for budget planning. Implementing machine learning algorithms requires a significant initial investment in both hardware (high-performance servers, GPUs) and software (ML frameworks, data processing tools). The cost associated with data storage solutions, including databases and data warehouses can vary based on the volume of data and the chosen technology. After the infrastructure provision, the next stage is data acquisition and preparation. Acquiring relevant datasets can involve purchasing third-party data or investing in data collection infrastructure including sensors, web scraping tools. Preparing data for analysis often requires significant time and resources to clean, preprocess, and annotate data, which can be labor-intensive and costly. Ensuring data privacy and compliance with regulations like GDPR (General Data Protection Regulation) can add to costs, particularly if the data contains sensitive information.

One of the participants, who was working as a data science manager, said that“SMEs should initially focus on leveraging cost-effective solutions and technologies rather than directly opting for the paid cloud plans. Open-source tools and libraries, such as TensorFlow, scikit-learn, and PyTorch, offer powerful capabilities for machine learning development without the huge price tags associated with proprietary software”.

Cloud computing platforms, such as AWS, Google Cloud Platform, and Microsoft Azure, provide scalable infrastructure and pay-as-you-go pricing models, enabling organizations to access advanced computing resources without upfront capital investment. By strategically utilizing cost-effective solutions and optimizing resource utilization, organizations can minimize expenses while delivering impactful machine learning projects. Lastly, prioritizing model simplicity and efficiency can mitigate costs associated with complex algorithms and data processing pipelines. While sophisticated deep learning models may offer superior performance in certain scenarios, they often come with higher computational requirements and longer development cycles. Simplifying model architectures, reducing feature dimensions, and implementing streamlined data preprocessing techniques can help minimize resource consumption and accelerate time-to-value. Additionally, focusing on incremental improvements and iterative development allows organizations to deliver tangible results while conserving resources and mitigating project risks.

To get the clear idea of the factors which should be considered, we assumed we have a budget of Ten thousand dollars ($10000) for a particular machine learning project. we requested the cost-analysis expert (participant 3) to provide tentative distribution of the funds and collected the response .Following table (table 4) summarizes the findings for this exercise which contains key expenses and tentative budget allocation):

***Table 4**: Summary of overall cost-estimation key parameters in machine learning.*

Area of Expenses	Budget Allocation (Out of $10000)	Reason of Allocation	Examples of Expenses

Computational Resources	$2,000	Computational resources are essential for training and deploying machine learning models. This amount ensures access to sufficient CPU or GPU resources for model development, training, and deployment tasks.	Cloud computing services, GPU rental fees, virtual machine instances
Data Storage and Processing	$1,500	Effective data storage and processing are crucial for managing project datasets efficiently. This allocation covers the costs of acquiring storage solutions, such as cloud-based databases or data lakes, ensuring data can be stored, processed, and analysed effectively.	Cloud storage fees, database subscription plans, data pre-processing tools
Software and Tools	$2,000	Software tools and libraries are indispensable for model development, training, and evaluation. Allocating a substantial amount ensures access to essential tools and platforms, including proprietary machine learning software or specialized development environments.	Machine learning frameworks (TensorFlow, PyTorch), IDEs, data visualization tools
Infrastructure Setup	$1,500	Infrastructure setup is critical for supporting the project's technical requirements, whether on-premises or through cloud computing services. This allocation covers the costs associated with setting up hardware, networking, and software configurations.	Hardware components (servers, GPUs), networking equipment, cloud platform setup fees

Model Development and Iteration	$1,500	Model development requires resources for data pre-processing, feature engineering, and hyper parameter tuning. This allocation covers the costs associated with comprehensive model development and optimization, including personnel salaries, data acquisition expenses, and software licensing fees.	Data labelling services, feature extraction tools, model training platforms
Scalability and Performance	$1,000	Planning for scalability and performance is essential for handling large-scale datasets and improving model performance over time. This allocation enables flexibility to scale up resources and optimize algorithms, ensuring efficient handling of computational tasks and improved model performance.	Cloud auto-scaling features, distributed computing frameworks, algorithm optimization tools
Monitoring and Maintenance	$500	On-going monitoring and maintenance are critical for ensuring project reliability. This allocation covers the costs of performance troubleshooting efforts, software updates, and security patches, facilitating proactive maintenance and optimization throughout the project lifecycle.	Monitoring software subscriptions, IT support services, software update licenses

Data Privacy and Security	$500	Data privacy and security measures are essential to protect sensitive information and ensure regulatory compliance. This allocation allows for the implementation of data encryption, access controls, and compliance measures, safeguarding project data against breaches and legal liabilities.	Data encryption software, access management tools, compliance consulting services
Training and Skill Development	$500	Investing in training enhances team capabilities in machine learning. This allocation supports participation in training programs, workshops, and certifications to strengthen team expertise and skills, ensuring that team members are equipped with the knowledge and skills needed to effectively contribute to the project.	Machine learning courses, certification programs, workshops on specific tools and techniques

Another aspect which discussed was regarding ongoing maintenance. It is a crucial aspect of cost analysis for ML within SMEs. Ensuring that models and infrastructure remain functional and efficient over time requires a systematic approach and investment in several key areas. Here's a breakdown of the some more areas which might not be thought of commonly but are important where costs are involved:

- Investment in monitoring tools (Datadog, Prometheus) to track model performance metrics. Costs typically include licensing fees for these tools.
- Salaries of data scientists or ML engineers dedicated to monitoring models, averaging 10-20 hours per week.

- Implementation of drift detection algorithms and integration with existing monitoring systems. This may incur a one-time development cost.
- Cost involved in buying the subscription to backup services (AWS Backup, Google Cloud Storage).
- Cost involved in project management tools like Asana, Jira, or Trello. Subscriptions to other tools for time management like Toggl or Harvest.
-

By understanding and addressing the above cost factors points, SMEs can better leverage machine learning and data science to achieve their business goals while maintaining financial health.

4.4 Research Question 4: What metrics and KPIs should SMEs track to evaluate the performance of their machine learning infrastructure?

Generally, when we evaluate the performance of the machine learning tasks, we always talk about accuracy, precision, recall or F1-score (Erickson, 2021). Here, our goal is to evaluate the infrastructure performance which allows driving these tasks. This evaluation is important to ensure that it meets the desired business goals and operates effectively. Tracking relevant metrics and key performance indicators (KPIs) allows SMEs to ensure that their machine learning models are both effective and efficient, thereby maximizing the return on investment. This systematic approach helps in identifying areas for improvement, optimizing resources, and ultimately driving better business outcomes.

Through the discussions with an SME entrepreneur, who founded the SME “Molecule 53” and exploration of 2 case studies, we've garnered essential insights into the metrics and KPIs discussed in the sections below:

- **CPU and GPU Usage:**

CPU and GPU usage metrics measure the percentage of time the central processing units (CPUs) and graphics processing units (GPUs) are actively processing data. Monitoring CPU and GPU usage helps in identifying whether the computational resources are being effectively utilized. High utilization indicates efficient use, but consistently high levels may suggest a need for additional or upgraded hardware to avoid bottlenecks and ensure smooth operations. Sometimes, multiple GPUs encounter the issue during run time where some GPUs may suddenly stop handling any computational load. Initially, both GPUs may start performing computations, but after a few minutes, all the load might be redirected to a single GPU. This can be tracked using the wandb python package. This package is one of the examples which allows to track GPU, CPU, memory usage and other metrics over time by adding two lines of code. Following figure shown the example of the scenario where GPU usage tracking gave clear indication that single GPU is handling all the load.

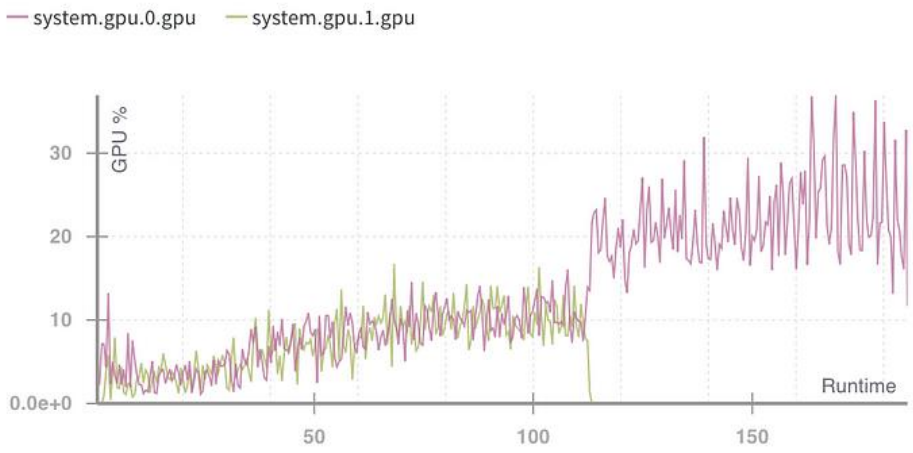


Figure 7: *Monitoring the multiple GPUs usage (Lukas Biewald, 2019)*

Detecting and resolving such bugs are crucial to ensure that all GPUs are effectively used, optimizing performance and cost efficiency.

- **Memory Utilization:**

Memory utilization tracks the amount of RAM being used by the machine learning applications. Ensuring adequate memory utilization is crucial for handling large datasets and complex models without performance degradation. Monitoring this metric helps prevent slowdowns or crashes, enabling seamless data processing and model training.

- **Storage Utilization:**

Storage utilization measures the amount of disk space being used by data and models. Efficient storage management ensures there is enough space for data processing and quick data retrieval. This metric helps in planning for storage expansion and optimizing data management practices to avoid storage-related bottlenecks.

- **Training Time:**

Training time measures the duration required to train a machine learning model. Reducing training time without sacrificing accuracy leads to more agile development cycles and quicker deployment of models. This metric helps in optimizing computational resources and improving overall workflow efficiency. This metric allows to visualize which model is sufficient to use so that SMEs can save their costs in choosing complex and expensive models. Following figure gives an example of the scenario where we plot the cost associated with training time along with training cost and validation cost.

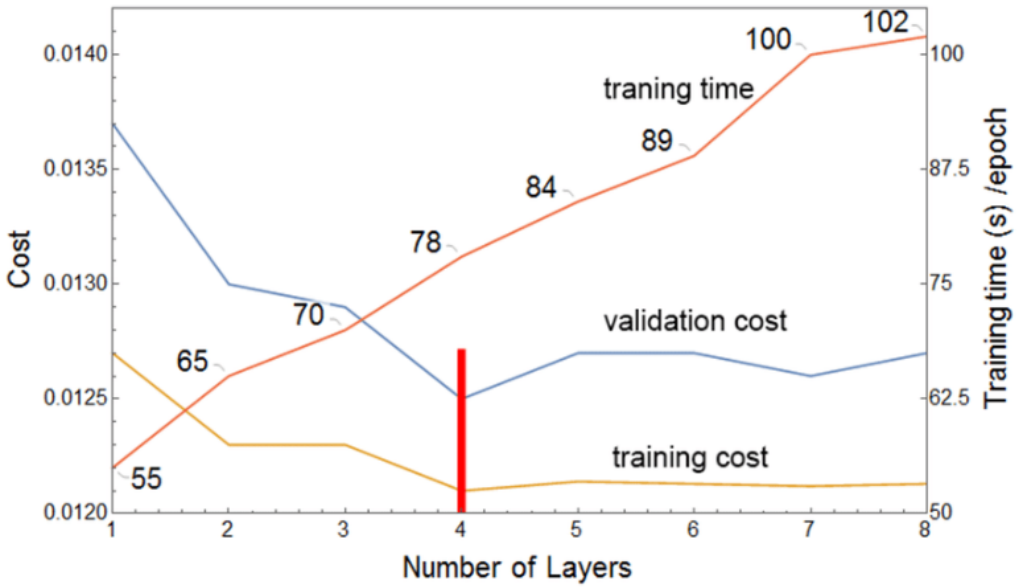


Figure 8: *Number of layers of DNN v/s training cost (Sun Hui, 2019)*

It is clearly seen from the figure that 4 layer DNN would be enough to extract the required features efficiently.

- **Inference Time:**

Inference time is the duration taken by the model to make predictions once trained. Lower inference times are crucial for applications requiring real-time or near-real-time predictions. This metric ensures that the deployed models can meet the performance requirements of time-sensitive applications.

- **Scalability:**

Scalability metrics evaluate the infrastructure's ability to handle increasing amounts of data and more complex models. Assessing scalability helps in planning for future growth and ensuring that the infrastructure can adapt to changing demands. It includes both horizontal (adding more machines) and vertical (enhancing machine capabilities) scalability.

- **Cost per Prediction:**

This metric calculates the cost associated with each prediction made by the model. Understanding the cost per prediction helps in managing budgets and ensuring the economic efficiency of machine learning applications. It allows SMEs to optimize spending and allocate resources effectively.

- **Infrastructure cost:**

Infrastructure cost includes the total expenditure on cloud services, hardware, software licenses, and other resources. Tracking infrastructure costs provides insights into the financial sustainability of machine learning projects. It helps in identifying cost-saving opportunities and ensuring that investments align with business objectives.

- **Return on Investment (ROI):**

ROI measures the financial return generated by machine learning initiatives relative to their cost. A high ROI indicates that the machine learning infrastructure is delivering significant value to the business. This metric helps in justifying investments and demonstrating the impact of machine learning on the company's bottom line.

- **Data Completeness and Consistency:**

Data completeness measures the extent to which datasets are complete and free of missing values. High data completeness ensures that models are trained on robust and comprehensive datasets, leading to more accurate and reliable predictions. Data consistency checks whether the data follows the same format and standards across different sources. Consistent data is crucial for reliable model performance. Ensuring data consistency helps in maintaining data integrity and reducing errors during data processing and analysis.

- **User satisfaction and adaption rate:**

User satisfaction is measured through surveys and feedback forms to gauge how satisfied end-users are with the machine learning solutions. High user satisfaction often correlates with successful implementation and practical utility of machine learning applications. It indicates that the solutions are meeting user needs and adding value to their workflows. It is also advantageous to track how widely and frequently the machine learning tools are used within the organization. We call it as ML adaption rate. A high adoption rate indicates that the tools are valuable and effectively integrated into business processes. It reflects the practical impact and relevance of machine learning solutions in daily operations.

- **Infrastructure downtime and Incident Response Time:**

Downtime measures the amount of time the machine learning infrastructure is unavailable due to maintenance or failures. Minimizing downtime is crucial for maintaining continuous operations and ensuring the reliability of machine learning applications. This metric helps in identifying areas for improvement in infrastructure stability. Incident response time measures how quickly issues are resolved once they are identified. Faster incident response times indicate robust support processes and minimize disruptions to operations. This metric ensures that the infrastructure can quickly recover from failures and maintain high availability.

By systematically tracking these metrics and KPIs, SMEs can gain comprehensive insights into the performance and impact of their machine learning infrastructure. This approach not only aids in optimizing current operations but also in planning future enhancements and scaling efforts, ensuring sustainable growth and competitiveness in the market.

4.5 Research Question 5: What are the knowledge and skill gaps SMEs need to address within their workforce to effectively utilize and manage infrastructure

for machine learning and data science? What are the cost-effective suggestions to address these gaps for SMEs?

The findings shared from the data science team leaders who have experience of handling the teams in different domains are presented in this section. During the discussion with participant 1, we identified several critical areas for improvement within the organization in which that candidate was working. A primary concern is the foundational understanding of machine learning concepts among employees. Many resources lack the basic knowledge necessary to contribute meaningfully to data-driven projects, which hinders problem-solving and innovation. Another significant gap is found in data management skills, including data cleaning, preprocessing, and handling large datasets efficiently. Improving these skills would enhance data quality and streamline processing efforts. Additionally, he emphasized the need for technical proficiency in tools such as Python, R, TensorFlow, and cloud platforms like AWS and Azure. Mastery of these tools would facilitate faster development cycles and scalable solutions. Furthermore, integrating data science into business processes is crucial for ensuring projects align with organizational goals and add value. Lastly, we agreed that stronger collaboration skills are essential for effective multidisciplinary teamwork and smooth project execution.

Participant 2 highlighted key knowledge and skill gaps that need to be addressed. He said, “A robust understanding of statistics is often missing among employees, which is vital for interpreting data correctly and building reliable models”. Enhancing statistical knowledge would improve model accuracy and the reliability of insights. Another significant gap is the proficiency in deploying and monitoring models. Employees need to develop skills to ensure solutions remain effective over time. Programming proficiency, particularly in Python and SQL, is also lacking, which hampers efficient data

manipulation and model development. Furthermore, there is a shortage of knowledge in data engineering practices, such as building and maintaining data pipelines. This knowledge is essential for ensuring data quality and integrity. Lastly, Emily emphasized the importance of domain expertise, which helps in understanding the context of data and designing relevant models. This expertise leads to more accurate and applicable insights tailored to business needs.

During the discussion with participant 3 and 4, several areas were pointed out for filling the knowledge gaps while working on ML projects. Employees need a comprehensive understanding of the machine learning lifecycle, from data collection to model evaluation and deployment. This understanding is crucial for successful project execution and better resource allocation. Participant also highlighted the need for improved skills in cloud computing platforms such as AWS, Google Cloud, and Azure. He said, “Proficiency in these platforms is necessary for scalable and efficient machine learning operations, providing scalable resources and reducing infrastructure costs.” Another identified gap is the lack of knowledge in big data technologies like Hadoop and Spark, which are essential for handling large datasets efficiently. Interviewee emphasized the importance of project management skills to handle the complexities of data science projects, including time management and resource allocation. Effective project management ensures timely project completion and better coordination among team members. Lastly, he noted the need for strong communication skills to explain complex data science concepts to non-technical stakeholders, facilitating better decision-making and alignment with business objectives.

Employees often lack a deep understanding of various machine learning algorithms, which is crucial for selecting the right approach for different problems. This understanding leads to better model performance and more accurate predictions. One of

the lady interviewee also pointed out the need for skills in data visualization tools like Tableau and Power BI. These skills are essential for presenting data insights clearly and effectively, enhancing the interpretability of data and aiding in decision-making processes. That interviewee emphasized the importance of knowledge in ethics in AI and data science. Developing responsible and fair models is critical for compliance with regulations and building trust with stakeholders. Another identified gap is in problem-solving skills, which are essential for addressing unique challenges in data science projects. Strong problem-solving skills promote innovative solutions and improve the ability to overcome project hurdles. Lastly, she stressed the importance of continuous learning to keep employees updated with the latest advancements in the field, ensuring the team remains competitive and capable of leveraging new techniques and technologies.

During the discussion with last participant, we spoke on acquiring the skills such as deep learning and reinforcement learning. He mentioned, “acquiring skills such as deep learning can lead to the development of more sophisticated models capable of tackling complex problems”. Another critical area is scalability solutions, where knowledge in building scalable machine learning solutions is crucial for handling growing data and user demands. This ensures models can handle increased loads without performance degradation. We also highlighted the need for understanding security and privacy concerns related to data handling. Protecting sensitive information and ensuring compliance with data protection regulations are important, mitigating the risks of data breaches. Additionally, optimization techniques were noted as a gap, as they are necessary for improving model efficiency and performance, leading to faster and more resource-efficient models. Finally, the importance of business acumen, which is essential for aligning data science projects with strategic business goals. This alignment ensures

that data science initiatives drive tangible business value and support organizational objectives.

After summarizing the discussion, we received following top areas which require fulfilling the gaps in terms of knowledge and skills within their workforce. We have also mentioned the cost effective ways suggested by the team leaders in data science field.

- Employees need a basic comprehension of machine learning principles to contribute effectively to projects. This can be cost-effectively addressed by encouraging employees to take advantage of free or low-cost online courses from platforms such as Coursera, Udemy, and edX. Additionally, promoting self-study groups and peer learning sessions within the company can facilitate collaborative learning and deeper understanding.
- A Grasp of statistics is essential for accurate data interpretation and building reliable models. SMEs can address this gap by utilizing free online resources and textbooks on statistics. Internal workshops or webinars led by employees with strong statistical backgrounds can provide practical, hands-on learning. Encouraging participation in online forums and communities where statistical problems are discussed can also enhance employees' skills.
- Proficiency in data cleaning, preprocessing, and handling large datasets is crucial for maintaining data quality and efficiency. Free coding platforms like Codecademy, freeCodeCamp, and Khan Academy can be used to enhance programming skills. Organizing internal coding bootcamps or hackathons can also promote hands-on learning and practical application of these skills, fostering a culture of continuous improvement.
- Knowledge of programming languages such as Python and SQL, along with familiarity with tools like TensorFlow, is necessary for model development and data

manipulation. Utilizing open-source tools such as Docker and Kubernetes for hands-on practice can further enhance these skills. Participation in online communities and forums focused on model deployment best practices can also be beneficial.

- Expertise in cloud platforms such as AWS, Google Cloud, and Azure is a for scalable and cost-effective machine learning operations.Skills in deploying models into production and monitoring their performance are vital to ensure ongoing model effectiveness. SMEs can leverage free tier services offered by cloud platforms such as AWS, Google Cloud, and Azure for training purposes. Encouraging employees to complete free online certifications and tutorials provided by these cloud service providers, and promoting internal knowledge-sharing sessions, can be an effective strategy.
- Aligning data science projects with business goals and understanding ethical implications are crucial for driving value and building stakeholder trust. SMEs can integrate business strategy and ethics training into regular team meetings or lunch-and-learn sessions. Using case studies and real-world examples can illustrate the importance of these skills. Encouraging continuous learning through free online resources and webinars focused on business and ethics in data science can help build a well-rounded and ethically aware workforce.

By addressing these knowledge and skill gaps through cost-effective strategies, SMEs can build a competent workforce capable of effectively utilizing and managing infrastructure for machine learning and data science. This approach not only enhances technical capabilities but also aligns data science initiatives with broader business objectives, fostering innovation and growth.

4.6 Research Question 6: What general infrastructure strategies should be employed by the SMEs while managing a machine learning project?

As a part of this thesis, the essential infrastructure strategies that SMEs should employ while managing machine learning projects are derived by studying the research papers and articles. This study aimed to bridge the gap between advanced technological capabilities and practical implementation. By focusing on tailored strategies, SMEs can optimize resource allocation, enhance data handling capabilities, and ensure robust model deployment, ultimately leading to improved decision-making and business outcomes.

- **Recommendations for resource hiring and training:** When it comes to hiring and training resources effectively for machine learning projects in smaller companies, it's essential to prioritize both technical expertise and practical experience (Sarkar, D., 2018). When hiring, look for candidates with a strong foundation in **mathematics, statistics, and programming**, as these are fundamental skills for machine learning. Additionally, candidates should possess experience with relevant programming languages and frameworks commonly used in machine learning, such as Python, TensorFlow, or PyTorch. In terms of training, providing hands-on experience with real-world projects can be invaluable. Consider assigning new hires to work on practical machine learning tasks under the guidance of experienced team members. Encourage participation in online courses and workshops focused on machine learning and data science to enhance their skillsets and stay updated with the latest advancements in the field. Moreover, fostering a culture of continuous learning and knowledge sharing within the team can help create a dynamic and collaborative environment conducive to skill development and innovation.
- **Recommendations for creating the Cost-Effective proof of concepts:** When creating a Proof of Concept for a machine learning project with a focus on cost-effectiveness, several strategies can be employed. Utilize **open-source tools** and libraries like TensorFlow or scikit-learn for model development, reducing the need

for expensive proprietary software. Take advantage of **cloud computing** platforms such as AWS or GCP (Zhang, 2020), which offer scalable resources and pay-as-you-go pricing models, minimizing upfront infrastructure costs. Explore freely available datasets from sources like Kaggle or UCI Machine Learning Repository to avoid data acquisition expenses. Begin with a minimal viable product (MVP) approach, focusing on specific use cases to reduce complexity and iteration time. Invest time in data preparation internally, leveraging cross-functional collaboration to maximize existing expertise and resources. Consider using automated machine learning (**AutoML**) tools to streamline the development process and minimize manual intervention. Finally, opt for minimalistic visualization and reporting solutions to communicate key insights effectively without investing in elaborate visualization tools. By implementing these strategies, a cost-effective POC can be developed to demonstrate the value of the machine learning project.

- **Recommendations on Selecting the On-Premise infrastructure OR Cloud Computing:** When determining whether to utilize on-premises infrastructure or a cloud platform for cost-effectiveness, it's essential to evaluate factors such as scalability, initial investment, operational costs, and data privacy. On-premises infrastructure may be preferable when the project demands predictable resource requirements and has long-term stability, allowing companies to leverage existing investments in hardware and software licenses without incurring additional costs. Additionally, for organizations with stringent data privacy and compliance requirements, maintaining control over sensitive data within the confines of on-premises infrastructure ensures adherence to regulatory standards and mitigates the risk of data breaches. Conversely, cloud platforms offer unparalleled scalability and flexibility, enabling companies to rapidly scale resources up or down based on project

demands without the need for significant upfront investment. This agility is particularly beneficial for projects with fluctuating computational requirements or those operating within dynamic environments. Moreover, cloud platforms often include managed services for tasks such as automated backups, security patches, and maintenance, reducing the operational burden on internal IT teams and potentially lowering overall costs. Additionally, cloud platforms facilitate collaboration and accessibility, allowing teams to work remotely and access resources from anywhere. Following figure (Figure 9) shows the comparison of on-premise and cloud infrastructure.

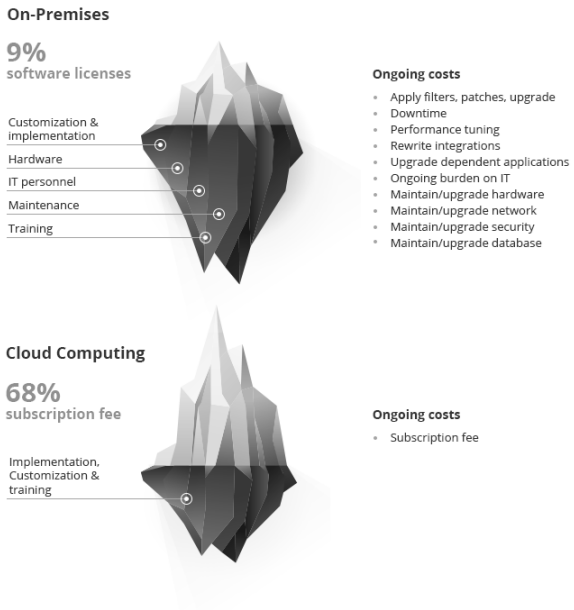


***Figure 9**: Comparison of On-premise and Cloud infrastructure (Intellias, 2021)*

Ultimately, the decision between on-premises infrastructure and cloud platforms hinges on the specific requirements, budget constraints, and strategic objectives of the machine learning project.

- **Recommendations on using popular frameworks**: Firstly, let us discuss on use of Spark v/s Pandas. PySpark, a distributed computing framework built on top of Apache Spark, is ideal for **handling large-scale datasets** that exceed the memory capacity of a single machine. It excels in distributed processing tasks and is well-suited for parallelizing computations across a cluster of machines. Therefore, PySpark is recommended when dealing with big data scenarios, where processing efficiency and scalability are important, or when the dataset cannot fit into memory. On the other hand, pandas, a Python library, are **suitable for smaller to medium-sized datasets** that can comfortably fit into memory. Pandas offer a user-friendly and intuitive interface for data manipulation and analysis, making it a preferred choice for exploratory data analysis, data preprocessing, and feature engineering tasks. It is efficient for single-machine processing and provides extensive functionality for data cleaning, transformation, and statistical analysis. Thus, pandas is preferred in scenarios where the dataset is relatively small, and there is no need for distributed processing or scalability. Apart from these, we have various Deep Learning libraries to support machine learning tasks. The selection of these deep learning library hinges on project specifics, with each offering distinct advantages tailored to different use cases. **TensorFlow**, renowned for its scalability and robust deployment capabilities, is ideal for large-scale projects requiring efficient distributed computing across multiple devices. Its extensive ecosystem, including TensorFlow Serving and TensorFlow Extended (TFX), facilitates seamless model deployment and production-level integration, making it well-suited for industry applications (Jain, 2019) **PyTorch**, characterized by its dynamic computation graph and intuitive interface, is favored for research and experimentation due to its flexibility and ease of use. It excels in rapid prototyping and exploration of novel architectures, enabling

researchers to iterate quickly and efficiently. Keras, a high-level API built on top of TensorFlow and capable of running on both TensorFlow and Microsoft Cognitive Toolkit (CNTK) backends, is preferred for beginners and those seeking simplicity and abstraction. Its user-friendly interface and modular design allow for easy model building and rapid development, making it an excellent choice for educational purposes and quick proof-of-concept implementations. Ultimately, the decision on which deep learning library to use should be based on project requirements, expertise, and objectives, with each offering unique advantages tailored to different scenarios and preferences.

4.7 Research Deliverable: Smart ML Assistance for providing the quick support to SMEs.

As a part of this thesis, we have also developed a Smart Chat assistance which acts as a Machine learning expert. Any user who is working on Data Science project can use this chat application to get a fair idea on the required resources, algorithms and what techniques would be useful. Also, he can take the support of this app in the intermediate stages to clarify the doubts OR take any advice.

Full development of this application is NOT in the scope of this research, but we have built a proof of concept for the same (Figure 10). This application is designed in such a way that it is able to answer machine learning related queries as a Data science expert. For queries other than ML, it detects and informs that it cannot answer those. We have designed two selectors, one is for framework selection and other is for machine learning tasks. Kindly refer to the screenshot of the chat application (Developed under this research thesis), "**SmartML Assistance**" below:

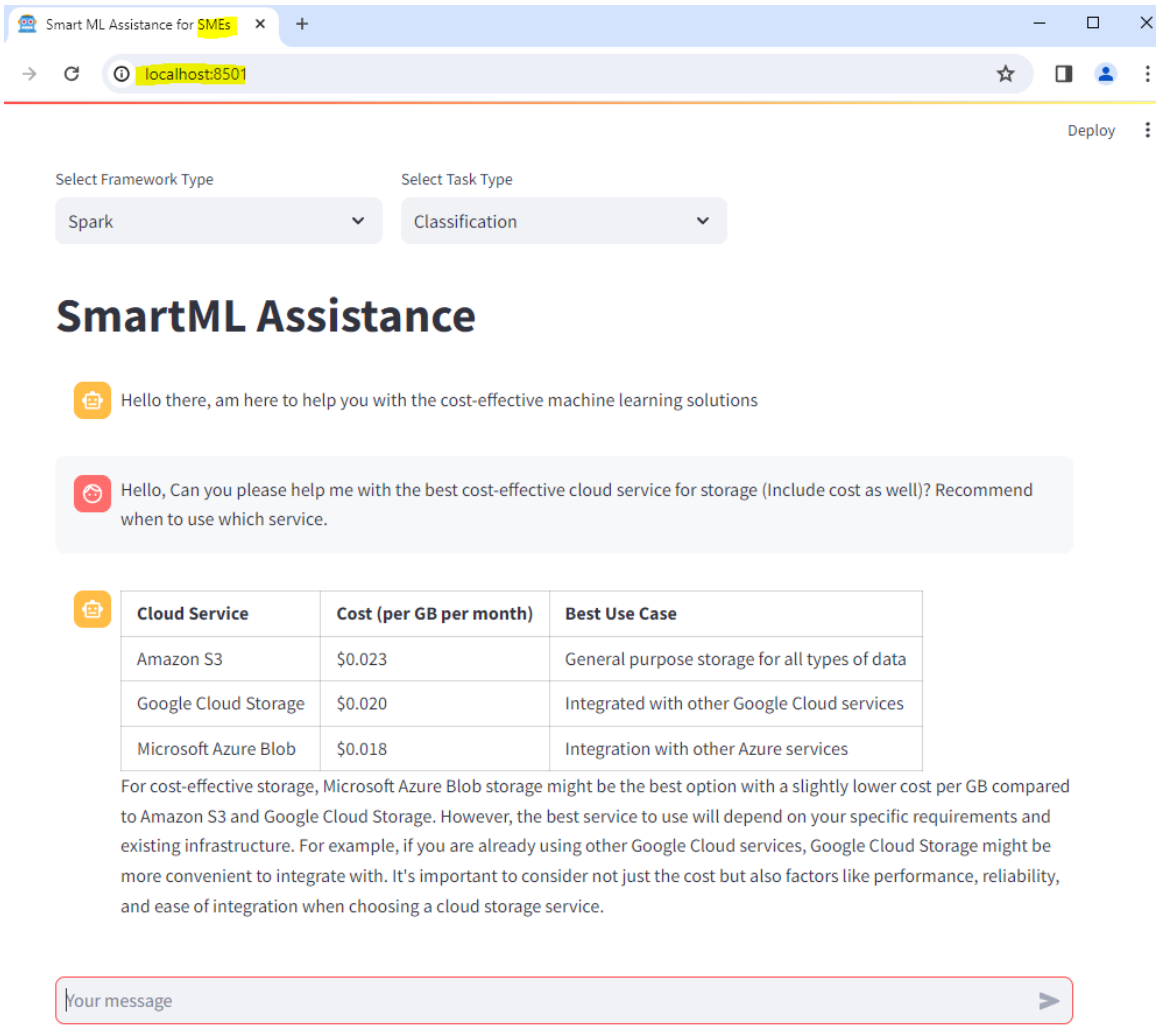


Figure 10: *Smart ML Assistance (POC Development as a part of this research)*

CHAPTER V:

DISCUSSION

5.1 Discussion of Results

The discussion section of the research delves into the findings and implications of the investigation into establishing an efficient and cost-effective infrastructure for small and medium enterprises (SMEs) to drive data science projects from prototype to production. The research revealed that data science has emerged as a valuable asset in the current era due to several factors. The abundance of data generated across various domains, coupled with advancements in technology and computing power, has created opportunities for organizations to harness data-driven insights for informed decision-making. In terms of machine learning tasks prevalent in ML-driven projects, the research identified classification, regression, clustering, and natural language processing as among the most frequently used tasks.

However, despite the growing significance of data science, the research also uncovered pain areas and scenarios where data science is still not fully effective. Issues such as data quality and availability, model interpretability, ethical considerations, and organizational barriers were identified as challenges hindering the effectiveness of data science initiatives. Addressing these challenges requires concerted efforts from organizations to enhance data governance practices, promote transparency and accountability in model development, and foster a data-driven culture across the organization. Furthermore, the research shed light on the commonly faced challenges and issues by smaller companies in data and infrastructure. Limited resources, lack of expertise, scalability concerns, and budget constraints emerged as prominent challenges inhibiting the successful implementation and management of data science projects in smaller companies. Overcoming these challenges requires strategic planning, investment

in talent development, and leveraging cost-effective solutions such as cloud computing and open-source tools. Finally, the discussion highlighted the key components and factors that play a crucial role in conducting cost analysis for data science projects with a limited budget. Infrastructure costs, software licensing fees, personnel expenses, data acquisition and storage costs, and potential return on investment were identified as critical factors to consider when allocating resources and ensuring project success within budgetary constraints.

5.2 Discussion on Research Question 1: What are the fundamental infrastructure requirements for SMEs to effectively integrate machine learning and data science into their operations?

To address this research question, the interview discussions were conducted with twelve data scientists working in various SMEs. This section discusses the insights gleaned from these interviews, the common themes that emerged, and the implications for SMEs aiming to adopt ML and DS technologies. The interview process spanned over two months, from December 2023 to February 2024. Each interview lasted approximately 30 minutes per candidate and was conducted via online calls. The data scientists were selected based on their experience in implementing ML and DS projects within their respective organizations. Following are the details of the interview candidates with their profiles.

Table 5*: Summary of interview candidates from data science teams.*

Participant Name	Interview Date	Interview Time	Interview Profile
Participant 1	05-Dec-23	10:00 AM	Associate Data Scientist
Participant 2	12-Dec-23	11:30 AM	Data Engineer
Participant 3	19-Dec-23	2:00 PM	Machine Learning Engineer
Participant 4	03-Jan-24	9:15 AM	Data Scientist

Participant 5	10-Jan-24	4:00 PM	Data Analyst
Participant 6	17-Jan-24	1:45 PM	Senior Data Scientist
Participant 7	24-Jan-24	3:30 PM	Business Intelligence Analyst
Participant 8	31-Jan-24	11:00 AM	Data Engineer
Participant 9	07-Feb-24	10:45 AM	Machine Learning Engineer
Participant 10	14-Feb-24	2:15 PM	Data Scientist
Participant 11	21-Feb-24	9:30 AM	Associate Data Scientist
Participant 12	28-Feb-24	12:00 PM	Data Analyst

A semi-structured interview format was utilized, ensuring a consistent set of core questions while allowing flexibility for participants to elaborate on their unique experiences and perspectives. One of the most prominent themes that emerged from the discussions was the necessity for robust data infrastructure. All participants highlighted the importance of a reliable data pipeline, which includes data collection, storage, and processing capabilities. Effective data management systems were deemed critical for handling the large volumes of data required for training ML models. Additionally, the data scientists emphasized the need for scalable storage solutions and efficient data pre-processing tools to ensure that the data fed into ML algorithms was clean and well-organized. Another key point raised was the requirement for adequate computational resources. The majority of interviewees pointed out that high-performance computing resources, such as GPUs and cloud-based services, were essential for training complex ML models within a reasonable timeframe. Several data scientists noted that their organizations had transitioned to cloud platforms to leverage scalable computing power and reduce the costs associated with maintaining on-premises hardware. This shift not only facilitated faster model training and deployment but also allowed for more flexible resource management. Lastly, the importance of a skilled workforce and supportive tools was a recurring theme. The data scientists discussed the need for continuous professional development and training to keep up with the rapidly evolving ML and DS fields. They

also stressed the significance of user-friendly tools and frameworks that enable efficient model development and deployment. Tools that facilitate collaboration and version control, such as Jupyter notebooks and Git, were mentioned as particularly beneficial for maintaining productivity and ensuring the reproducibility of results.

In conclusion, the findings from the interviews underscore the critical infrastructure components necessary for SMEs to successfully integrate ML and DS into their operations. Robust data management systems, adequate computational resources, and a skilled workforce equipped with supportive tools are fundamental to achieving this goal. These insights provide valuable guidance for SMEs looking to harness the power of machine learning and data science to drive innovation and growth.

5.3 Discussion on Research Question 2: What are the commonly faced challenges and issues by SMEs in data science infrastructure?

To address this research question, a survey was conducted among data scientists working in various small and medium-sized enterprises (SMEs). The survey gathered 25 responses, providing valuable insights into the specific obstacles these organizations encounter when implementing data science projects. This discussion section will analyse the key themes that emerged from the survey responses, highlighting the prevalent challenges and their implications for SMEs.

A major challenge identified by the survey respondents was the lack of adequate computational resources. Many data scientists reported that their organizations struggled with limited access to high-performance computing (HPC) resources, such as GPUs and cloud computing services. This limitation often resulted in prolonged training times for machine learning models, hindering the timely deployment of data-driven solutions. Additionally, the high cost associated with scaling computational infrastructure was cited as a significant barrier, particularly for SMEs with constrained budgets.

Another prevalent issue was the difficulty in managing and processing large volumes of data. Respondents indicated that inefficient data pipelines and storage solutions often led to bottlenecks in data processing workflows. Poor data quality and lack of proper data governance practices further exacerbated these problems, making it challenging to maintain the integrity and reliability of the datasets used for machine learning. Several data scientists emphasized the need for robust data management systems that could handle the complexities of collecting, cleaning, and organizing diverse data sources. The survey also revealed that many SMEs faced challenges related to the integration of data science tools and technologies. Data scientists reported that their organizations often struggled with adopting and integrating new tools due to compatibility issues with existing systems. This lack of interoperability not only slowed down the implementation of data science projects but also increased the complexity of maintaining and updating the infrastructure. Respondents highlighted the necessity of adopting flexible and scalable solutions that could seamlessly integrate with their current technological stack. A significant theme that emerged was the shortage of skilled personnel in data science. Many data scientists indicated that their organizations faced difficulties in recruiting and retaining talent with the requisite skills in machine learning, data engineering, and data analysis. This skills gap often led to an over-reliance on a small number of individuals, creating bottlenecks and reducing the overall productivity of data science teams. The respondents stressed the importance of investing in training and development programs to up-skill existing employees and attract new talent.

In conclusion, the survey responses highlighted several key challenges faced by SMEs in establishing and maintaining effective data science infrastructure. Limited computational resources, inefficient data management, integration issues, and a shortage of skilled personnel were identified as major obstacles. Addressing these challenges

requires a strategic approach that includes investing in scalable computational solutions, improving data governance practices, adopting flexible tools, and enhancing workforce capabilities through targeted training and development initiatives. These insights provide a foundation for SMEs to develop robust data science infrastructure that can support their growth and innovation goals.

5.4 Discussion on Research Question 3: What components / factors play a key role while doing the cost analysis of the project in SMEs?

To address this research question, an review of relevant articles and research papers was conducted. Additionally, discussions with data science managers involved in cost analysis were held to gain practical insights. This discussion section will synthesize the key factors identified from these sources, highlighting their importance and implications for SMEs undertaking data science and machine learning projects.

One of the primary factors identified was the cost of computational resources. Both literature and discussions emphasized that the expenses associated with acquiring and maintaining high-performance computing infrastructure, such as GPUs and cloud services, constitute a significant portion of the overall project cost. Data science managers pointed out that the decision between on-premises and cloud-based solutions greatly influences cost structures. Cloud services offer scalability and flexibility, but the costs can accumulate quickly with extensive usage, making cost management and forecasting crucial. Another critical factor was data acquisition and storage costs. Research articles and managers alike noted that collecting and storing large volumes of data necessary for training machine learning models can be expensive. Costs can arise from data purchase, storage hardware, and cloud storage services. Effective data management practices, such as data cleaning and pre-processing, also incur costs but are essential for ensuring data quality and reliability. Managers stressed the importance of

balancing the investment in data infrastructure with the anticipated value derived from the data.

The cost of human resources emerged as a significant component in the cost analysis. Skilled personnel in data science, including data engineers, data scientists, and machine learning experts, command high salaries due to the demand for their expertise. Additionally, on-going training and professional development are necessary to keep the team updated with the latest advancements in the field. Managers highlighted that while human resource costs are substantial, investing in a competent team is critical for the successful implementation and maintenance of data science projects. Software and tool expenses were also highlighted as important factors in cost analysis. The choice of software, whether open-source or commercial, can impact the overall cost. While open-source tools like Python and R can reduce expenses, commercial software often comes with support and additional features that can enhance productivity and efficiency. Managers indicated that licensing fees, subscription costs, and the potential need for custom software development should be carefully considered when budgeting for data science projects.Lastly, opportunity costs and return on investment (ROI) were frequently mentioned as essential components of cost analysis. Literature and managers emphasized the importance of evaluating the potential ROI of data science projects to justify the expenditures. Opportunity costs, such as the potential benefits of alternative projects or the cost of not implementing a data science solution, must be taken into account. Managers highlighted the need for thorough cost-benefit analysis to ensure that the projects undertaken provide significant value to the organization.

In conclusion, the study of articles, research papers, and discussions with data science managers revealed several key factors in the cost analysis of data science projects in SMEs. Computational resources, data acquisition and storage, human resources,

software and tools, and opportunity costs are all crucial components that must be carefully considered. Understanding and managing these factors can help SMEs make informed decisions, optimize their investments, and successfully integrate machine learning and data science into their operations.

5.5 Discussion on Research Question 4: What metrics and KPIs should SMEs track to evaluate the performance of their machine learning infrastructure?

To address this research question, case studies published by Stitch Fix and Influx Data were analysed, alongside relevant supporting articles. This discussion section synthesizes the insights from these sources, highlighting the key metrics and KPIs that SMEs should monitor to ensure their machine learning infrastructure is effective and efficient.

One of the primary metrics identified from the Stitch Fix case study was model accuracy. Stitch Fix emphasized the importance of measuring how well their machine learning models predicted customer preferences and recommended clothing items. High model accuracy indicates that the models are effectively learning from the data and making reliable predictions. This metric is crucial for SMEs to ensure that their models are performing as expected and providing value to the business by making accurate predictions or classifications. Another critical KPI discussed in both case studies was system latency. Influx Data highlighted the significance of tracking the response time of their machine learning systems, particularly for real-time applications. Low latency is essential for providing timely insights and maintaining a positive user experience. SMEs need to monitor latency to ensure that their infrastructure can handle the required processing speed, especially for applications that demand real-time analysis and decision-making.

Resource utilization was also a key focus in the Influx Data case study. Monitoring how effectively computational resources, such as CPUs, GPUs, and memory, are being used can help SMEs optimize their infrastructure. Efficient resource utilization ensures that the systems are not overburdened and that the available resources are used cost-effectively. This metric can aid in identifying bottlenecks and scaling resources appropriately to meet the demands of machine learning workloads. In addition to these technical metrics, both case studies underscored the importance of business impact metrics. Stitch Fix, for instance, tracked customer engagement and satisfaction to assess the success of their recommendations. Similarly, Influx Data considered the impact of their machine learning models on operational efficiency and cost savings. For SMEs, evaluating the business impact of their machine learning initiatives is crucial to justify investments and demonstrate the tangible benefits of their infrastructure. Lastly, the importance of monitoring the maintainability and scalability of the machine learning infrastructure was evident. Both Stitch Fix and Influx Data highlighted the need for tracking metrics related to system maintenance, such as the frequency of updates and the ease of integrating new models. Scalability metrics, such as the ability to handle increased data volumes and user requests, are essential for ensuring that the infrastructure can grow with the business needs. SMEs should focus on these metrics to ensure long-term sustainability and adaptability of their machine learning systems.

In conclusion, the analysis of case studies from Stitch Fix and Influx Data, along with supporting articles, identified several critical metrics and KPIs for evaluating the performance of machine learning infrastructure in SMEs. Model accuracy, system latency, resource utilization, business impact, and maintainability and scalability were highlighted as key areas to monitor. By tracking these metrics, SMEs can ensure their

machine learning infrastructure is effective, efficient, and capable of supporting their growth and innovation goals.

5.6 Discussion on Research Question 5:What are the knowledge and skill gaps SMEs need to address within their workforce to effectively utilize and manage infrastructure for machine learning and data science?

To address the research question, interviews were conducted with 4 data science tech leads responsible for overseeing data science resources in their organizations. This section discusses the insights derived from these interviews, focusing on the prevalent skill gaps and their implications for SMEs.A significant knowledge gap identified was the lack of expertise in cloud computing platforms. Most tech leads indicated that while their teams were proficient in traditional on-premises computing environments, they struggled with the complexities of cloud-based infrastructure. This included difficulties in managing cloud resources, optimizing costs, and ensuring security. As cloud platforms offer scalability and flexibility critical for machine learning operations, bridging this knowledge gap was seen as essential for SMEs to leverage these technologies effectively.

Another critical skill gap was in the area of data engineering. The tech leads emphasized that their teams often lacked the necessary skills to build and maintain robust data pipelines. This included tasks such as data extraction, transformation, and loading (ETL), as well as ensuring data quality and integrity. Without strong data engineering capabilities, the efficiency and reliability of machine learning models could be compromised. Addressing this gap was deemed vital to ensure that data science projects were built on solid and dependable data foundations. Proficiency in advanced machine learning techniques was also highlighted as a gap. While basic machine learning concepts were generally well understood, the tech leads noted a shortage of expertise in more sophisticated areas such as deep learning, natural language processing (NLP), and

reinforcement learning. This limitation restricted the types of problems that their teams could tackle and the sophistication of the solutions they could develop. Investing in training and up-skilling in these advanced techniques was seen as a priority for expanding the scope and impact of data science initiatives.

The interviews also revealed a gap in project management skills tailored to data science projects. Traditional project management approaches were often inadequate for the iterative and experimental nature of machine learning projects. The tech leads reported challenges in managing project timelines, resources, and stakeholder expectations. They emphasized the need for their teams to adopt agile methodologies and develop a better understanding of how to manage the lifecycle of data science projects effectively, from conception to deployment and monitoring. Lastly, communication and collaboration skills were identified as crucial areas needing improvement. The tech leads pointed out that data scientists often worked in silos and lacked the ability to effectively communicate their findings and needs to non-technical stakeholders. This gap hindered the integration of data science insights into broader business strategies and decision-making processes. Enhancing these soft skills was seen as essential for fostering a collaborative environment where data science could drive significant business value.

In conclusion, the interviews with data science tech leads highlighted several key knowledge and skill gaps that SMEs need to address within their workforce to effectively utilize and manage machine learning and data science infrastructure. Expertise in cloud computing, data engineering, advanced machine learning techniques, project management, and communication were identified as critical areas for development. Addressing these gaps through targeted training and upskilling initiatives will enable SMEs to harness the full potential of their data science capabilities and drive innovation within their organizations.

5.7 Discussion on Research Question 6: What general infrastructure strategies should be employed by the SMEs while managing a machine learning project?

To address this research question, a review of research papers and notable project documents from various SMEs was conducted. This discussion synthesizes the key strategies identified from these sources, emphasizing their practical implications and relevance for SMEs undertaking machine learning projects.

From the study, it became clear that starting with a cost-effective proof of concept is crucial for SMEs venturing into machine learning. Research emphasized the importance of leveraging existing open-source tools and pre-trained models to minimize initial costs. SMEs were advised to focus on small-scale projects that can demonstrate the potential value of machine learning with limited investment. Additionally, iterative development and validation were recommended to ensure that the PoC aligns with business objectives before committing significant resources. By adopting a lean approach, SMEs could validate their hypotheses without substantial financial risks.

The study highlighted the significance of hiring the right talent and providing continuous training. Given the competitive nature of the data science field, SMEs were encouraged to seek versatile candidates who possess a combination of technical expertise and business acumen. It was found that investing in training programs and fostering a culture of continuous learning were vital for keeping the team updated with the latest advancements in machine learning. Collaboration with academic institutions and participation in industry workshops were also recommended as cost-effective strategies to enhance the skill set of the workforce. SMEs that prioritized professional development were better positioned to innovate and maintain a competitive edge. Deciding between on-premise infrastructure and cloud computing emerged as a critical consideration. The

reviewed documents suggested that cloud computing offers several advantages, including scalability, flexibility, and reduced upfront costs. For SMEs with limited IT resources, cloud services such as AWS, Google Cloud, and Azure were recommended due to their ability to scale resources according to project demands. Conversely, for SMEs with stringent data security requirements or those handling highly sensitive information, on-premise solutions could be more appropriate. The decision should be guided by a thorough cost-benefit analysis, considering factors such as data security, compliance requirements, and long-term scalability needs.

In conclusion, the findings from the research papers and project documents provided valuable insights into effective infrastructure strategies for SMEs. By focusing on cost-effective proof of concepts, strategic hiring and training, careful selection between on-premise and cloud computing, and adopting hybrid approaches, the smaller enterprises can establish robust and scalable machine learning infrastructures.

CHAPTER VI:

SUMMARY, IMPLICATIONS, AND RECOMMENDATIONS

6.1 Summary:

The thesis, "Establishing an Efficient and Cost-effective Infrastructure for Small and Medium Enterprises to Drive Data Science Projects from Prototype to Production," delves into the multifaceted landscape of data science infrastructure in the context of Small and Medium Enterprises (SMEs). This summary encapsulates the key findings and recommendations arising from the paper, addressing critical issues such as machine learning challenges, cost considerations, skill sets, budget constraints, cloud complexities, and other pertinent concerns faced by SMEs. The research underscores the pervasive challenges that SMEs encounter in the realm of machine learning. These challenges range from limited access to quality data and lack of in-house expertise to scalability constraints and issues related to model deployment. The thesis acknowledges that these hurdles are substantial and often deter SMEs from harnessing the transformative potential of machine learning.

A central theme of the thesis revolves around the financial aspects of data science infrastructure. It emphasizes the critical importance of cost-effectiveness for SMEs, given their typically constrained budgets. The research recognizes that mismanagement of costs, especially in cloud environments, can be a significant barrier to successful project implementation. The paper elucidates the significant role of skill sets and expertise in the success of data science endeavors. It underscores the shortage of data science professionals in the job market and the need for SMEs to invest in training and up skilling their workforce. The acquisition of data science talent is highlighted as a strategic imperative.

In addition to financial considerations, the thesis addresses the allocation of resources. It acknowledges the complexities associated with resource allocation and the need for judicious planning to ensure that limited resources are optimally utilized for data science projects.The research delves into the challenges posed by cloud computing, which is increasingly adopted by SMEs for its scalability and flexibility. It recognizes the need for SMEs to navigate the complexities of cloud environments efficiently, manage costs, and ensure data security and compliance. Throughout the thesis, a multitude of concerns faced by SMEs in their data science journey are explored, from data privacy and compliance to organizational culture and resistance to change. The paper concludes with a comprehensive set of recommendations that guide SMEs in establishing a practical, effective, and budget-conscious data science infrastructure.

In conclusion, this thesis serves as a roadmap for SMEs seeking to overcome the challenges associated with data science infrastructure. It underscores the imperative of striking a delicate balance between technological advancement, budget considerations, skillet development, and strategic alignment with organizational goals. By addressing these concerns and embracing the recommended best practices, SMEs can embark on a journey to harness the full potential of data science and drive innovation and competitiveness in their respective domains.

6.2 Implications

The thesis, "Establishing an Efficient and Cost-effective Infrastructure for Small and Medium Enterprises to Drive Data Science Projects from Prototype to Production," delves into critical implications for Small and Medium Enterprises (SMEs) arise from the challenges and considerations surrounding data science infrastructure in the current machine learning era. Understanding cost-effective infrastructure allows SMEs to

optimize their budget allocation for machine learning initiatives. By identifying affordable solutions and minimizing unnecessary expenses, SMEs can make the most of their limited financial resources. This efficiency ensures that SMEs can invest strategically in crucial aspects of their machine learning projects, such as talent acquisition, data acquisition, and model development, without exceeding their budgetary constraints.Cost-effective infrastructure empowers SMEs to compete more effectively in the increasingly data-driven marketplace. By leveraging advanced analytics and machine learning capabilities, SMEs can drive innovation, improve operational efficiency, and differentiate themselves from competitors. Access to cost-effective infrastructure levels the playing field, enabling SMEs to harness the power of data-driven insights to enhance decision-making and gain a competitive edge in their respective industries.

Cost-effective infrastructure solutions, such as cloud computing platforms, offer scalability and flexibility tailored to the needs of SMEs. These platforms provide on-demand access to scalable resources, allowing SMEs to expand or contract their infrastructure based on project requirements and budget constraints. This scalability ensures that SMEs can adapt to changing business needs and technological advancements without incurring significant upfront costs, enabling them to grow their machine learning initiatives organically over time. Studying cost-effective infrastructure encourages SMEs to optimize resource utilization and minimize wastage. By implementing efficient resource management practices, SMEs can maximize the efficiency of their infrastructure, ensuring that computing resources are utilized effectively without unnecessary overhead. This optimization not only reduces operational costs but also enhances overall productivity and performance, enabling SMEs to achieve more with less.

This study also fosters innovation and growth for SMEs in the machine learning space. By lowering the barrier to entry for machine learning projects, SMEs can experiment with new ideas, develop innovative solutions, and drive business growth. Access to affordable infrastructure empowers SMEs to explore new opportunities, develop predictive models, and extract actionable insights from data, positioning them for sustained success and growth in the dynamic machine learning era.

6.3 Recommendations for Future Research:

The thesis, "Establishing an Efficient and Cost-effective Infrastructure for Small and Medium Enterprises to Drive Data Science Projects from Prototype to Production," provides valuable insights into addressing the challenges and considerations surrounding data science infrastructure for SMEs. As the field of data science and machine learning continues to evolve, there are several promising areas for future exploration and development. The following recommendations outline potential avenues for future research and expansion of the concepts presented in this paper:

- Advanced Machine Learning Solutions: Future research can delve deeper into advanced machine learning techniques and their applicability to SMEs. Exploring cutting-edge algorithms, such as deep learning and reinforcement learning, can help SMEs stay at the forefront of data science innovation.
- AI Ethics and Governance: With the growing emphasis on ethical AI, future studies can investigate the ethical implications of data science projects in SMEs. This includes bias mitigation, fairness, transparency, and ethical data collection and usage.
- Predictive Analytics and Prescriptive Modeling: Expanding research into predictive analytics and prescriptive modeling can empower SMEs to make data-

driven decisions with greater precision. These techniques can enhance forecasting, risk management, and optimization strategies.

- Interdisciplinary Collaboration: Future research can explore the benefits of interdisciplinary collaboration between data scientists, domain experts, and business professionals within SMEs. This approach can lead to more effective problem-solving and better alignment with business goals.
- Cost Optimization Strategies: Investigating advanced cost optimization strategies, including auto-scaling mechanisms and dynamic resource allocation, can further help SMEs manage their budgets effectively in cloud-based environments.
- Skills Development Programs: Expanding on the skills development aspect, future research can explore innovative approaches to training and up skilling the SME workforce in data science, ensuring a sustainable pipeline of talent.

6.4 Conclusion:

This thesis explored the critical aspects of establishing efficient infrastructure for the adoption of machine learning and data science within small and medium-scale enterprises (SMEs). The research was structured around six key questions, each addressing a different facet of this complex integration process. The findings from a combination of literature review, survey responses, interviews, and case studies provided a comprehensive understanding of the requirements, challenges, and strategies pertinent to SMEs.

Firstly, the investigation into the fundamental infrastructure requirements revealed that robust data management systems, scalable computational resources, and advanced data processing tools were essential for effectively integrating machine learning and data science. SMEs needed reliable data pipelines, storage solutions, and high-performance

computing capabilities to support the data-intensive nature of these technologies.Secondly, the research identified common challenges faced by SMEs in data science infrastructure. Limited access to high-performance computing resources, inefficiencies in data management, difficulties in integrating new tools with existing systems, and a shortage of skilled personnel were highlighted as significant obstacles. These challenges underscored the need for strategic investment and planning in infrastructure development.The third research question focused on the critical components for cost analysis in data science projects. It was found that computational resources, data acquisition and storage costs, human resource expenses, software and tool investments, and opportunity costs were key factors that needed careful consideration. Effective cost management required a balanced approach to ensure that investments were aligned with the anticipated benefits.

For the next research question, the metrics and KPIs essential for evaluating machine learning infrastructure performance were identified. Metrics such as model accuracy, training time, computational efficiency, data throughput, and cost-effectiveness were crucial for monitoring and optimizing the infrastructure. These metrics helped SMEs assess the effectiveness of their infrastructure and make informed decisions for improvements.The fifth research question addressed the knowledge and skill gaps within the SME workforce. The study revealed that SMEs needed to invest in training programs to bridge gaps in machine learning, data engineering, and data analysis skills. Continuous professional development and collaborations with academic institutions were recommended to ensure the workforce remained adept at managing and utilizing the infrastructure effectively.Lastly, the exploration of general infrastructure strategies for managing machine learning projects emphasized the importance of starting with cost-effective proofs of concept, strategic hiring, and training, and making informed decisions

between on-premise and cloud computing solutions. Hybrid approaches that combined the strengths of both on-premise and cloud solutions were also recommended for optimal flexibility and cost management.

In conclusion, this thesis provided an examination of the infrastructure requirements, challenges, cost factors, performance metrics, skill gaps, and strategic approaches necessary for SMEs to successfully integrate and manage machine learning and data science projects. By addressing these areas comprehensively, SMEs can develop robust, scalable, and efficient infrastructures that support their data-driven initiatives and drive innovation.

APPENDIX A

SURVEY COVER LETTER

Dear All,

I extend my sincere gratitude to everyone for participating in the interviews conducted as part of my thesis research on "Establishing an Efficient and Cost-Effective Infrastructure for Small and Medium Enterprises to Drive Data Science Projects from Prototype to Production."

Your insights and experiences shared during the survey / interviews have been invaluable in gaining a deeper understanding of the challenges and opportunities surrounding the transition of data science projects from the prototype stage to production within small and medium enterprises (SMEs). Your expertise has contributed significantly to the richness and depth of this research endeavor.I want to assure you that all information provided during the interviews will be treated with the utmost confidentiality and will only be used for research purposes. Your anonymity will be maintained, and any identifying information will be kept strictly confidential.

The data gathered from these interviews will be instrumental in identifying common pain points, successful strategies, and areas for improvement in establishing an efficient and cost-effective infrastructure for data science initiatives within SMEs. Your contribution will directly influence the recommendations and insights presented in this thesis.

Once again, I express my heartfelt appreciation for your time, expertise, and willingness to share your perspectives. Your participation has been instrumental in advancing knowledge in this field and driving progress in the realm of data science infrastructure for SMEs.

Thank you for your invaluable contribution to this research endeavor.

Sincerely,

Hrishikesh Thakurdesai.

APPENDIX B:

INFORMED CONSENT

Title: Informed Consent for Participation in Interview Research.

ResearchTitle:

Establishing an Efficient and Cost-Effective Infrastructure for Small and Medium Enterprises to Drive Data Science Projects from Prototype to Production.

Principal Investigator: Dr. Mario Silic

Affiliation: Swiss school of Business Management, Geneva.

Contact Information: Email - hrishikesh.thakurdesai.92@gmail.com

Introduction:

You are being invited to participate in an interview as part of a research study conducted by Hrishikesh Thakurdesai on the topic of establishing an efficient and cost-effective infrastructure for small and medium enterprises (SMEs) to drive data science projects from prototype to production. This informed consent form provides information about the purpose of the study, the procedures involved, and your rights as a participant. Please read this document carefully before deciding to participate.

Purpose of the Study:

The purpose of this research study is to explore the challenges and opportunities associated with transitioning data science projects from the prototype stage to production within small and medium enterprises. The study aims to gather insights from individuals with experience in data science projects within SMEs to identify strategies for improving infrastructure efficiency and cost-effectiveness.

Procedures:

If you agree to participate in this study, you will be invited to take part in an interview conducted by the researcher. The interview will involve discussing your experiences, perspectives, and insights related to data science projects within SMEs. The interview will be audio-recorded for accuracy and analysis purposes. The duration of the interview is expected to be approximately [insert estimated time].

Risks and Benefits:

Participation in this study involves minimal risks, such as potential discomfort in discussing sensitive topics related to your experiences with data science projects. However, the benefits of participating include contributing to the advancement of knowledge in the field of data science infrastructure for SMEs and potentially informing future initiatives and strategies in this area.

Confidentiality:

All information provided during the interview will be treated with the strictest confidence. Your identity will be kept confidential, and any identifying information will be removed from the data collected. Only the researcher and authorized personnel involved in the study will have access to the interview recordings and data.

Voluntary Participation:

Participation in this study is entirely voluntary. You have the right to refuse to participate or withdraw from the study at any time without penalty.

Contact Information:

If you have any questions, concerns, or would like further information about the study, please contact Mr. Hrishikesh Thakurdesai.

Email: hrishikesh.thakurdesai.92@gmail.com

Phone: +91 8850559118

Consent:

By agreeing to participate in this interview, you acknowledge that you have read and understood the information provided in this consent form. You voluntarily agree to participate in the study and consent to the recording of the interview for research purposes.

(Researcher – Hrishikesh Thakurdesai)

Date: November 2023

APPENDIX C:

INTERVIEW GUIDE

Interview Guide: Establishing an Efficient and Cost-Effective Infrastructure for SMEs to Drive Data Science Projects from prototype to production.

Introduction:

Thank you for participating in this interview. The purpose of this interview is to gather insights and perspectives from data scientists working in startups and multinational companies regarding the establishment of an efficient and cost-effective infrastructure for small and medium enterprises (SMEs) to drive data science projects from prototype to production. Your expertise and experiences will provide valuable insights for our research on this topic. The interview will last approximately 45 minutes, and your responses will be kept confidential.

1. Background and Experience:

a. Can you provide a brief overview of your role and responsibilities as a data scientist in your current organization?

b. How long have you been involved in data science projects, and what types of projects have you worked on?

c. Have you had experience working with SMEs or startups on data science initiatives? If so, could you describe your involvement and the challenges you encountered?

2. Infrastructure Challenges in Data Science Projects:

a. In your experience, what are the main challenges SMEs face when establishing infrastructure for data science projects?

b. How do infrastructure constraints impact the scalability and efficiency of data science projects in SMEs?

c. Can you discuss any specific technical or resource limitations that SMEs encounter when transitioning from prototype to production in data science projects?

3. Cost-Effective Solutions:

a. What cost-effective strategies or solutions do you recommend for SMEs to establish infrastructure for data science projects?

b. How can SMEs leverage cloud services or open-source technologies to minimize infrastructure costs while maintaining scalability?

c. Are there any best practices or lessons learned from your own experience in optimizing infrastructure costs for data science projects in SMEs?

4. Balancing Efficiency and Effectiveness:

a. How do you balance the need for efficiency with the effectiveness of infrastructure solutions in driving data science projects for SMEs?

b. What criteria do you consider when evaluating the suitability of infrastructure solutions for SMEs?

c. Can you share examples of successful implementations where efficient and cost-effective infrastructure solutions contributed to the success of data science projects in SMEs?

5. Recommendations and Advice:

a. Based on your experiences, what recommendations would you offer to SMEs looking to establish infrastructure for their data science projects?
b. How important is it for SMEs to invest in infrastructure early on in the development of data science projects?
c. Are there any additional insights or advice you would like to share with SMEs embarking on data science initiatives?

Conclusion:

Thank you for sharing your insights and expertise on this topic. Your contributions will greatly contribute to our research on establishing an efficient and cost-effective infrastructure for SMEs to drive data science projects from prototype to production. If you have any further thoughts or insights you'd like to share, please feel free to do so.

APPENDIX D:

INTERVIEW QUESTIONS

Following are some of the questions which are discussed with Data Scientists from well-known multinatioal companies to get quality insights:

- Can you explain your understanding of why data science has become so valuable in the current era?
- What specific factors or trends do you believe have contributed to the rise in importance of data science?
- How does data science differ from other approaches to solving problems or extracting insights from data?
- In your opinion, what are the key advantages of using data science techniques compared to traditional methods?
- Can you provide examples of industries or sectors where data science has had a significant impact, and why do you think that is?
- How do you see the role of data scientists evolving in the future, especially in light of advancements in technology and data analytics?
- Are there any limitations or challenges associated with data science that you think are important to consider?
- What skills or qualifications do you believe are essential for someone pursuing a career in data science today?
- How do you approach communicating the value of data science to stakeholders who may not be familiar with its concepts or methodologies?

- Can you share a specific project or case study where data science played a crucial role in achieving a successful outcome, and explain why data science was the preferred approach in that scenario
- Questions to Data Science leads for knowledge and skill gaps:
    a. Can you discuss any difficulties you've observed in SMEs regarding collaboration between data science teams and other departments?
    b. Can you describe the most common challenges you face when deploying machine learning models in small and medium enterprises (SMEs)?
    c. What specific technical skills do you believe are most lacking among SME employees when it comes to supporting machine learning and data science projects?What training or resources do you think would be most beneficial for SME employees to improve their data science and machine learning capabilities?

- How do you identify which tasks or projects are suitable candidates for data science and machine learning approaches?
- In your experience, what are the key characteristics of tasks or projects that benefit the most from data science and machine learning?
- Are there specific industries or domains where data science and machine learning are especially valuable, and if so, why?
- Can you discuss any challenges or limitations you've encountered when applying data science and machine learning to certain types of tasks or projects?
- What role does data quality play in determining the effectiveness of data science and machine learning solutions?

- What are the challenges / issues in Data, Infrastructure and Skills competency
- What infrastructure-related challenges have you experienced when setting up or maintaining data science environments and tools?
- Have you encountered any scalability issues with your infrastructure when working on data science projects, and if so, how did you overcome them?
- What are the key challenges in ensuring security and privacy compliance when dealing with sensitive data in your data science projects?
- How do you assess and address the skills gap within your data science team or organization?
- Are there any specific technical skills or competencies that you believe are lacking or in high demand within the field of data science?
- How do you approach training and up skilling initiatives to improve the competency of your data science team members?
- Can you share any experiences or lessons learned from past projects where data, infrastructure, or skills competency issues posed significant challenges, and how you addressed them?
- Can you discuss any common barriers or obstacles that SMEs face in terms of data availability and quality for machine learning initiatives?
- How do resource constraints, such as limited budget or manpower, impact SMEs' ability to effectively manage machine learning projects?
- What role do technical expertise and skills play in SMEs' capacity to drive successful machine learning projects, and what challenges arise in this area?
- Are there specific cultural or organizational factors within SMEs that hinder the adoption and execution of machine learning initiatives?

- How do SMEs navigate regulatory compliance and data privacy concerns when implementing machine learning projects, and what challenges do they encounter in this regard?
- Can you share your experience transitioning data science prototypes to production environments, and what were the key challenges you encountered?
- What are some common pitfalls or mistakes to avoid when moving from prototype to productionisation in data science projects?
- In your opinion, what are the most critical considerations when designing data pipelines for production, and how do you ensure scalability and reliability?
- Can you discuss the importance of collaboration between data scientists, engineers, and other stakeholders during the productionization process, and how do you foster effective communication?
- What strategies do you employ to optimize model performance and resource utilization for production environments?
- How do you approach version control and reproducibility in data science projects, especially when transitioning from prototype to production?
- Can you provide insights into selecting the right infrastructure and deployment strategies for scaling data science projects in production?
- What are your recommendations for implementing monitoring and logging solutions to ensure the health and performance of data science applications in production?
- How do you address security and compliance considerations when deploying data science solutions in production environments?
- Can you share any lessons learned or best practices you've developed for managing the lifecycle of data science projects from prototype to production.

APPENDIX E:

SCREENSHOT OF THE CODE FOR SMART ML ASSISTANCE

```python
import streamlit as st
from langchain_community.chat_models import ChatOpenAI
from langchain.chains import LLMChain
from langchain.prompts import PromptTemplate
from langchain.memory import ConversationBufferWindowMemory

prompt = PromptTemplate(
    input_variables=["chat_history", "question"],
    template="""You are a machine Learning and Data Science Expert.
    You have an expertise in infrastructure recommendations and cost estimations for machine learning projects projects.
    You will received the queries regarding the Data science projects optimizations.
    Always provide the response in the tabular format
    Provide the detailed answer for the same.If the input query is not related to Machine Learning or Data Science, you strictly don't provide the reponsne "


    chat_history: {chat_history},
    Human: {question}
    AI:"""
)


llm = ChatOpenAI(openai_api_key="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")
memory = ConversationBufferWindowMemory(memory_key="chat_history", k=4)
llm_chain = LLMChain(
    llm=llm,
    memory=memory,
    prompt=prompt
)


st.set_page_config(
    page_title="Smart ML Assistance for SMEs",
    page_icon="🤖",
    layout="wide"
)

def get_ai_output(input_text):
    ai_response = llm_chain.predict(question="Provide the information regarding {} in machine learning".format(input_text))
    st.write(ai_response)


col1, col2, col3 = st.columns(3)
with col1:
    tasks_list = ["Spark", "Pandas"]
    result = st.selectbox("Select Framework Type ", tasks_list)
    #get_ai_output(result)

with col2:
    tasks_list = ["Classification", "Regression"]
    result = st.selectbox("Select Task Type ", tasks_list)
    #get_ai_output(result)

st.title("SmartML Assistance")


# check for messages in session and create if not exists
if "messages" not in st.session_state.keys():
    st.session_state.messages = [
        {"role": "assistant", "content": "Hello there, am here to help you with the cost-effective machine learning solutions"}
    ]


# Display all messages
for message in st.session_state.messages:
    with st.chat_message(message["role"]):
        st.write(message["content"])


user_prompt = st.chat_input()

if user_prompt is not None:
    st.session_state.messages.append({"role": "user", "content": user_prompt})
    with st.chat_message("user"):
        st.write(user_prompt)


if st.session_state.messages[-1]["role"] != "assistant":
    with st.chat_message("assistant"):
        with st.spinner("Loading..."):
            ai_response = llm_chain.predict(question=user_prompt)
            st.write(ai_response)
    new_ai_message = {"role": "assistant", "content": ai_response}
    st.session_state.messages.append(new_ai_message)
```

```
C:\Windows\System32\cmd.exe - streamlit  run app3.py
Microsoft Windows [Version 10.0.19045.4170]
(c) Microsoft Corporation. All rights reserved.

C:\Users\admin\Desktop\Doctorate Implementation>streamlit run app3.py

  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8501
  Network URL: http://192.168.1.9:8501

C:\Users\admin\AppData\Local\Programs\Python\Python311\Lib\site-packages\langchain_core\_api\deprecation.py:117: LangChainDeprecationWarning: The class `langchain
_community.chat_models.openai.ChatOpenAI` was deprecated in langchain-community 0.0.10 and will be removed in 0.2.0. An updated version of the class exists in the
 langchain-openai package and should be used instead. To use it run `pip install -U langchain-openai` and import as `from langchain_openai import ChatOpenAI`.
  warn_deprecated(
```

APPENDIX F:

QUESTIONARE FOR THE SURVEY

Survey Title: Challenges and Issues in Data Science Infrastructure for SMEs

Introduction: This survey aims to understand the commonly faced challenges and issues by small and medium-sized enterprises (SMEs) in establishing and maintaining data science infrastructure. Your responses will help us identify key pain points and develop solutions tailored to SMEs. The survey should take approximately 10 minutes to complete. All responses are anonymous and will be used solely for research purposes.

Section 1: General Information

Section 2: Current Data Science Infrastructure

- Does your company currently utilize data science or machine learning?
- If yes, what type of data science infrastructure does your company use?
- What are the primary data science tools and technologies used in your company?

Section 3: Challenges Faced

- What are the major challenges your company faces in setting up data science infrastructure?
- How significant are the following issues in your data science operations?
    - Data quality and cleaning
    - Data integration from multiple sources
    - Lack of clear business objectives
    - Limited budget
    - Insufficient computational resources
    - Keeping up with technology trends

- Regulatory compliance
- Have you encountered any specific technical difficulties with your data science infrastructure? If so, please describe.

Section 4: Future Plans and Needs:

- Is your company planning to invest more in data science infrastructure in the next 1-2 years?

  What additional support or resources would be most helpful for your company to enhance its data science capabilities?

REFERENCES

Abdullahi, M.S., Ghazali, P.L., Awang, Z., Tahir, I.M. and Ali, N., 2015. The effect of finance, infrastructure and training on the performance of small and medium scale enterprises (SMEs) in Nigeria. International Journal of Business and Technopreneurship, 5(3), pp.421-452.

Agarwal, B., Mittal, N., Agarwal, B. and Mittal, N., 2016. Machine learning approach for sentiment analysis. Prominent feature extraction for sentiment analysis, pp.21-45.

Agrawal, R. and Prabakaran, S., 2020. Big data in digital healthcare: lessons learnt and recommendations for general practice. Heredity, 124(4), pp.525-534.

Agrawal, S. and Jain, S.K., 2020. Medical text and image processing: applications, issues and challenges. Machine Learning with Health Care Perspective: Machine Learning and Healthcare, pp.237-262.

Al-Jarrah, O.Y., Yoo, P.D., Muhaidat, S., Karagiannidis, G.K. and Taha, K., 2015. Efficient machine learning for big data: A review. Big Data Research, 2(3), pp.87-93.

Assefi, M., Behravesh, E., Liu, G. and Tafti, A.P., 2017, December. Big data machine learning using Apache spark MLlib. In 2017 ieee international conference on big data (big data) (pp. 3492-3498). IEEE.

Avram, M.G., 2014. Advantages and challenges of adopting cloud computing from an enterprise perspective. Procedia Technology, 12, pp.529-534.

Azodi, C.B., Tang, J. and Shiu, S.H., 2020. Opening the black box: interpretable machine learning for geneticists. Trends in genetics, 36(6), pp.442-455.

Bantilan, N., 2020. pandera: Statistical Data Validation of Pandas Dataframes. In Proceedings of the Python in Science Conference (SciPy) (pp. 116-124).

Bernard, J., 2016. Python data analysis with pandas. In Python Recipes Handbook (pp. 37-48). Apress, Berkeley, CA.

Bertolini, M., Mezzogori, D., Neroni, M. and Zammori, F., 2021. Machine Learning for industrial applications: A comprehensive literature review. Expert Systems with Applications, 175, p.114820.

Castillo-Botón, C., Casillas-Pérez, D., Casanova-Mateo, C., Ghimire, S., Cerro-Prada, E., Gutierrez, P.A., Deo, R.C. and Salcedo-Sanz, S., 2022. Machine learning regression and classification methods for fog events prediction. Atmospheric Research, 272, p.106157.

Cearns, M., Hahn, T. and Baune, B.T., 2019. Recommendations and future directions for supervised machine learning in psychiatry. Translational psychiatry, 9(1), p.271.

Chowdhury, K.P., 2019. Supervised machine learning and heuristic algorithms for outlier detection in irregular spatiotemporal datasets. Journal of Environmental Informatics, 33(1), pp.1-16.

Dahiya, N., Gupta, S. and Singh, S., 2022. A review paper on machine learning applications, advantages, and techniques. ECS Transactions, 107(1), p.6137.

De la Hoz Domínguez, E.J., Herrera, T.J.F. and Mendoza, A.A.M., 2020. Machine Learning and SMEs: Opportunities for an improved deci-sion-making process. Investigación e InnovaciónenIngenierías, 8(1), pp.21-36.

Del Vecchio, P., Di Minin, A., Petruzzelli, A.M., Panniello, U. and Pirri, S., 2018. Big data for open innovation in SMEs and large corporations: Trends, opportunities, and challenges. Creativity and Innovation Management, 27(1), pp.6-22.

Demchenko, Y., Zhao, Z., Grosso, P., Wibisono, A. and De Laat, C., 2012, December. Addressing big data challenges for scientific data infrastructure. In 4th IEEE International

Erickson, B.J. and Kitamura, F., 2021. Magician's corner: 9. Performance metrics for machine learning models. Radiology: Artificial Intelligence, 3(3), p.e200126.

Ezugwu, A.E., Ikotun, A.M., Oyelade, O.O., Abualigah, L., Agushaka, J.O., Eke, C.I. and Akinyelu, A.A., 2022. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Engineering Applications of Artificial Intelligence, 110, p.104743.

Fu, J., Sun, J. and Wang, K., 2016, December. Spark–a big data processing platform for machine learning. In 2016 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII) (pp. 48-51). IEEE.

Furano, G., Meoni, G., Dunne, A., Moloney, D., Ferlet-Cavrois, V., Tavoularis, A., Byrne, J., Buckley, L., Psarakis, M., Voss, K.O. and Fanucci, L., 2020. Towards the use of artificial intelligence on the edge in space systems: Challenges and opportunities. IEEE Aerospace and Electronic Systems Magazine, 35(12), pp.44-56.

Garg, S., Sinha, S., Kar, A.K. and Mani, M., 2022. A review of machine learning applications in human resource management. International Journal of Productivity and Performance Management, 71(5), pp.1590-1610.

Gupta, S. and Gupta, A., 2019. Dealing with noise problem in machine learning data-sets: A systematic review. Procedia Computer Science, 161, pp.466-474.

Hair Jr, J.F. and Sarstedt, M., 2021. Data, measurement, and causal inferences in machine learning: opportunities and challenges for marketing. Journal of Marketing Theory and Practice, 29(1), pp.65-77.

Hoffmann, F., Bertram, T., Mikut, R., Reischl, M. and Nelles, O., 2019. Benchmarking in classification and regression. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(5), p.e1318.

Hopkins, A. and Booth, S., 2021, July. Machine learning practices outside big tech: How resource constraints challenge responsible development. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 134-145).

Hwang, K. and Chen, M., 2017. Big-data analytics for cloud, IoT and cognitive computing. John Wiley & Sons.

Islam, T. and Manivannan, D., 2017, June. Predicting application failure in cloud: A machine learning approach. In 2017 IEEE International Conference on Cognitive Computing (ICCC) (pp. 24-31). IEEE.

Jain, A., Awan, A.A., Anthony, Q., Subramoni, H. and Panda, D.K.D., 2019, September. Performance characterization of dnn training using tensorflow and pytorch on modern clusters. In 2019 IEEE International Conference on Cluster Computing (CLUSTER) (pp. 1-11). IEEE.

Jang, H., 2019. A decision support framework for robust R&D budget allocation using machine learning and optimization. Decision Support Systems, 121, pp.1-12.

Jayachandran, S., Biradavolu, M. and Cooper, J., 2023. Using machine learning and qualitative interviews to design a five-question survey module for women's agency. World Development, 161, p.106076.

Jordan, M.I. and Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science, 349(6245), pp.255-260.

Kandula, S. and Shaman, J., 2019. Reappraising the utility of Google flu trends. PLoS computational biology, 15(8), p.e1007258.

Kang, Z., Catal, C. and Tekinerdogan, B., 2020. Machine learning applications in production lines: A systematic literature review. Computers & Industrial Engineering, 149, p.106773.

Kechiche, L., 2021, March. Hardware acceleration for deep learning of image classification. In 2021 International Conference of Women in Data Science at Taif University (WiDSTaif) (pp. 1-5). IEEE.

Kettimuthu, R., Liu, Z., Foster, I., Beckman, P.H., Sim, A., Wu, K., Liao, W.K., Kang, Q., Agrawal, A. and Choudhary, A., 2018, June. Towards autonomic science infrastructure: Architecture, limitations, and open issues. In Proceedings of the 1st International Workshop on Autonomous Infrastructure for Science (pp. 1-9).

Kumar, A., Boehm, M. and Yang, J., 2017, May. Data management in machine learning: Challenges, techniques, and systems. In Proceedings of the 2017 ACM International Conference on Management of Data (pp. 1717-1722).

Kumar, A., Boehm, M. and Yang, J., 2017, May. Data management in machine learning: Challenges, techniques, and systems. In Proceedings of the 2017 ACM International Conference on Management of Data (pp. 1717-1722).

Lalwani, P., Mishra, M.K., Chadha, J.S. and Sethi, P., 2022. Customer churn prediction system: a machine learning approach. Computing, 104(2), pp.271-294.

Langley, P., 2011. The changing science of machine learning. Machine learning, 82(3), pp.275-279.

Le Quy, T., Roy, A., Iosifidis, V., Zhang, W. and Ntoutsi, E., 2022. A survey on datasets for fairness-aware machine learning. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12(3), p.e1452.

Lee, I. and Shin, Y.J., 2020. Machine learning for enterprises: Applications, algorithm selection, and challenges. Business Horizons, 63(2), pp.157-170.

Lee, I. and Shin, Y.J., 2020. Machine learning for enterprises: Applications, algorithm selection, and challenges. Business Horizons, 63(2), pp.157-170.

Liao, T., Taori, R., Raji, I.D. and Schmidt, L., 2021, August. Are we learning yet? a meta review of evaluation failures across machine learning. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

Lutfi, A., Alsyouf, A., Almaiah, M.A., Alrawad, M., Abdo, A.A.K., Al-Khasawneh, A.L., Ibrahim, N. and Saad, M., 2022. Factors influencing the adoption of big data analytics in the digital transformation era: Case study of Jordanian SMEs. Sustainability, 14(3), p.1802.

Makam VK. Continuous Integration on Cloud Versus on Premise: A Review of Integration Tools. Advances in Computing. 2020;10(1):10-4.

Malik, M.M., 2020. A hierarchy of limitations in machine learning. arXiv preprint arXiv:2002.05193.

Marin, I., Shukla, A. and Sarang, V.K., 2019. Big Data Analysis with Python: Combine Spark and Python to unlock the powers of parallel computing and machine learning. Packt Publishing Ltd.

McKinney, W. and Team, P.D., 2015. Pandas-Powerful python data analysis toolkit.Pandas—Powerful Python Data Analysis Toolkit, 1625.

McKinney, W., 2011. pandas: a foundational Python library for data analysis and statistics. Python for high performance and scientific computing, 14(9), pp.1-9.

Mendez, K.M., Pritchard, L., Reinke, S.N. and Broadhurst, D.I., 2019. Toward collaborative open data science in metabolomics using Jupyter Notebooks and cloud computing. Metabolomics, 15(10), pp.1-16.

Mohammed, R., Rawashdeh, J. and Abdullah, M., 2020, April. Machine learning with oversampling and undersampling techniques: overview study and experimental results. In 2020 11th international conference on information and communication systems (ICICS) (pp. 243-248). IEEE.

Naradda Gamage, S.K., Ekanayake, E.M.S., Abeyrathne, G.A.K.N.J., Prasanna, R.P.I.R., Jayasundara, J.M.S.B. and Rajapakshe, P.S.K., 2020. A review of global challenges and survival strategies of small and medium enterprises (SMEs). Economies, 8(4), p.79.

Nayak, A., Satpathy, I., Patnaik, B.C.M., Baral, S.K. and Khang, A., 2023. Impact of Artificial Intelligence (AI) on Talent Management (TM): A Futuristic Overview. In Designing Workforce Management Systems for Industry 4.0 (pp. 139-158). CRC Press.

Neneh, B.N., 2018. Customer orientation and SME performance: the role of networking ties. African Journal of Economic and Management Studies, 9(2), pp.178-196.

Olson, R.S., La Cava, W., Orzechowski, P., Urbanowicz, R.J. and Moore, J.H., 2017. PMLB: a large benchmark suite for machine learning evaluation and comparison. BioData mining, 10, pp.1-13.

Paleyes, A., Urma, R.G. and Lawrence, N.D., 2022. Challenges in deploying machine learning: a survey of case studies. ACM Computing Surveys, 55(6), pp.1-29.

Parker, C., 2012, August. Unexpected challenges in large scale machine learning. In Proceedings of the 1st international workshop on big data, streams and heterogeneous source mining: Algorithms, systems, programming models and applications (pp. 1-6).

Paullada, A., Raji, I.D., Bender, E.M., Denton, E. and Hanna, A., 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. Patterns, 2(11).

Pine, K. and Mazmanian, M., 2015. Emerging insights on building infrastructure for data-driven transparency and accountability of organizations. iConference 2015 Proceedings.

Rathore, M.M., Shah, S.A., Shukla, D., Bentafat, E. and Bakiras, S., 2021. The role of ai, machine learning, and big data in digital twinning: A systematic literature review, challenges, and opportunities. IEEE Access, 9, pp.32030-32052.

Rawindaran, N., Jayal, A. and Prakash, E., 2021. Machine learning cybersecurity adoption in small and medium enterprises in developed countries. Computers, 10(11), p.150.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L. and Zhong, C., 2022. Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistic Surveys, 16, pp.1-85.

Sadeeq, M.M., Abdulkareem, N.M., Zeebaree, S.R., Ahmed, D.M., Sami, A.S. and Zebari, R.R., 2021. IoT and Cloud computing issues, challenges and opportunities: A review. QubahanAcademic Journal, 1(2), pp.1-7.

Samala, R.K., Chan, H.P., Hadjiiski, L. and Koneru, S., 2020, March. Hazards of data leakage in machine learning: a study on classification of breast cancer using deep neural networks. In Medical Imaging 2020: Computer-Aided Diagnosis (Vol. 11314, pp. 279-284). SPIE.

Sarkar, D., Bali, R. and Sharma, T., 2018. Practical machine learning with Python. Book" Practical Machine Learning with Python, pp.25-30.

Schelter, S., Biessmann, F., Januschowski, T., Salinas, D., Seufert, S. and Szarvas, G., 2015. On challenges in machine learning model management.

Serban, A. and Visser, J., 2022, March. Adapting software architectures to machine learning challenges. In 2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER) (pp. 152-163). IEEE.

Shanahan, J.G. and Dai, L., 2015, August. Large scale distributed data science using apache spark. Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 2323-2324).

Sharma, D. and Kumar, N., 2017. A review on machine learning algorithms, tasks and applications. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 6(10), pp.2278-1323.

Sharma, K., Kaur, A. and Gujral, S., 2014. Brain tumor detection based on machine learning algorithms. International Journal of Computer Applications, 103(1).

Shrestha, A. and Mahmood, A., 2019. Review of deep learning algorithms and architectures. IEEE access, 7, pp.53040-53065.

Smaldone, F., Ippolito, A., Lagger, J. and Pellicano, M., 2022. Employability skills: Profiling data scientists in the digital labour market. European Management Journal, 40(5), pp.671-684.

Stanula, P., Ziegenbein, A. and Metternich, J., 2018. Machine learning algorithms in production: A guideline for efficient data source selection. Procedia CIRP, 78, pp.261-266.

Strickland, E., 2019. IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. IEEE Spectrum, 56(4), pp.24-31.

Sze, V., Chen, Y.H., Emer, J., Suleiman, A. and Zhang, Z., 2017, April. Hardware for machine learning: Challenges and opportunities. In 2017 IEEE custom integrated circuits conference (CICC) (pp. 1-8). IEEE.

Tayefeh Hashemi, S., Ebadati, O.M. and Kaur, H., 2020. Cost estimation and prediction in construction projects: A systematic review on machine learning techniques. SN Applied Sciences, 2(10), p.1703.

Torrance, H., 2012. Triangulation, respondent validation, and democratic participation in mixed methods research. Journal of mixed methods research, 6(2), pp.111-123.

Van de Vijver, F. and Leung, K., 1997. Methods and data analysis of comparative research. Handbook of cross-cultural psychology, 1, pp.257-300.

Walcott, T.H. and Ali, M., 2021, August. Machine Learning for Smaller Firms: Challenges and Opportunities. In 2021 International Conference on Computing, Electronics & Communications Engineering (iCCECE) (pp. 82-86). IEEE.

Wang, J.F., The Impact of Artificial Intelligence (AI) on Customer Relationship Management: A Qualitative Study.

Wang, M. and Deng, W., 2021. Deep face recognition: A survey. Neurocomputing, 429, pp.215-244.

Whang, S.E., Roh, Y., Song, H. and Lee, J.G., 2023. Data collection and quality challenges in deep learning: A data-centric ai perspective. The VLDB Journal, 32(4), pp.791-813.

Wolf, M.J., Miller, K. and Grodzinsky, F.S., 2017. Why we should have seen that coming: comments on Microsoft's tay" experiment," and wider implications. AcmSigcas Computers and Society, 47(3), pp.54-64.

Yudin, A. and Yudin, A., 2021. Data Analysis with Pandas. Basic Python for Data Management, Finance, and Marketing: Advance Your Career by Learning the Most Powerful Analytical Tool, pp.93-150.

Zahid, F., 2023. Security and Compliance Aspects of Data Integrity in Banking AI/ML. INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY, 7(1), pp.323-346.

Zhang, C., Yu, M., Wang, W. and Yan, F., 2020. Enabling cost-effective, slo-aware machine learning inference serving on public cloud. IEEE Transactions on Cloud Computing, 10(3), pp.1765-1779.