PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
الجمهورية الجزائرية الديمقراطية الشعبية
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
وزارة التعليم العالي و البحث العلمي
UNIVERSITY 20 AUGUST 1955 - SKIKDA
FACULTY OF SCIENCES
DEPARTMENT OF COMPUTER SCIENCES

A Thesis

Presented for obtaining the degree of:

DOCTORATE

In: Computer Science

Specialty: Computation and Cognition of Computer Systems

By:

Mohamed Lamine BOUGHOUAS

Theme

Big Data Paradigm: Methods, Tools and Applications, Case of the Internet of Things

---

Thesis defended on 4th July 2024 at Skikda in Front of the Jury Composed of:

Mr Djamel ZEGHIDA	MCA	University 20 August 1955 of Skikda	President
Mr Smaine MAZOUZI	Professor	University 20 August 1955 of Skikda	Supervisor
Mr Yacine KISSOUM	MCA	University 20 August 1955 of Skikda	Co-Supervisor
Mrs Nardjes BOUCHEMAL	MCA	Abdelhafid Boussouf University Center of Mila	Examiner
Mr Soufiane BOULEHOUACH	MCA	University 20 August 1955 of Skikda	Examiner

Academic Year 2023/2024

"Everything has a reality, and the servant will not reach the reality of faith until he knows that what afflicted him could never miss him, and that what missed him could never have afflicted him."

– Prophet Muhammad (PBUH)

Dedication

I proudly dedicate this thesis to:

- My father, my biggest supporter.
- My mum, the loveliest person to me.
- My wife and kids, fuel of my life.
- My lovely brother and sisters.

– Mohamed Lamine

Acknowledgements

Firstly, all praise and thanks go to Allah who gave me the power and courage to conquer life challenges in general and to pursue my higher studies in particular; I am thankful.

I want to express my sincere gratitude to Dr. Yassine Kissoum, who has always been a good listener, for his inspiration and support. This thesis would not have been possible without his constant encouragement and motivation. I also thank Prof. Smaine Mazouzi for his help and advice during my thesis years.

A special note of thanks is also extended to the Examination Committee member. The president of the jury, Dr. Djamel Zeghida, for overseeing this process with such diligence and insight. My examiners, Dr. Nardjes Bouchemal and Dr. Soufiane Boulehouach, for their thorough review, constructive criticism, and thoughtful suggestions, which significantly enhanced the quality of this thesis. Thank you for your support and for being an integral part of this academic endeavor.

I cannot forget my dear parents, wife, children, and brother and sisters for their support and patience.

– Mohamed Lamine Boughouas

Abstract

Big data's distinguishing features include its substantial volume, swift velocity, and wide-ranging variety of data types, which present obstacles for traditional data processing software. The proliferation of big data is propelled by the Internet of Things (IoT), which produces massive data volumes via interconnected devices. The collaboration between IoT and big data is revolutionary, facilitating the development of intelligent applications across various fields, such as healthcare and higher education.

This thesis discusses the challenges and opportunities in managing and extracting insights from Big Data, mainly focusing on IoT-generated Big Data in sectors like higher education and environmental perception and management. It highlights the potential benefits of utilizing Big Data in these sectors and emphasizes the importance of advanced technologies for improving monitoring, analysis, and response mechanisms.

In higher education, the study introduces the concept of Big Data Analytics (BDA) for improving student performance and decision-making, presenting a model for educational supply chain management and predictive analytics for student outcomes. In environmental management, a multi-layered architecture leveraging IoT and fog computing is proposed for forest fire management, emphasizing the importance of fog computing in handling data volume and improving system response time to fire incidents. Implementing advanced technologies like BDA and fog computing can significantly enhance efficiency and decision-making processes in various fields, revolutionizing how data is managed and utilized.

keywords: Big Data, IoT, Data analytics, Higher education, Student performance, Environment, Forest fire detection.

Résumé

Le Big Data se caractérise notamment par son volume considérable, sa vélocité et sa grande variété de types de données, qui constituent autant d'obstacles pour les outils traditionnels de traitement des données. La prolifération des big data est favorisée par l'internet des objets (IdO), qui produit des volumes de données considérables par l'intermédiaire d'appareils interconnectés. La collaboration entre l'IdO et le big data est révolutionnaire, facilitant les applications intelligentes dans divers domaines tels que la santé et l'enseignement supérieur.

Cette thèse aborde les défis et les opportunités liés à la gestion et à l'extraction d'informations à partir des Big Data, en se concentrant particulièrement sur les Big Data générées par l'IdO dans des secteurs tels que l'enseignement supérieur et la perception et la gestion de l'environnement. Elle met en évidence les avantages potentiels de l'utilisation du Big Data dans ces secteurs et souligne l'importance des technologies avancées pour améliorer les mécanismes de surveillance, d'analyse et de réponse.

Dans l'enseignement supérieur, l'étude introduit le concept de Big Data Analytics (BDA) pour améliorer les performances des étudiants et la prise de décision, en présentant un modèle de gestion de la chaîne d'approvisionnement éducationnelle et d'analyse prédictive des résultats des étudiants. Dans le domaine de la gestion de l'environnement, une architecture multicouche tirant parti de l'IdO et du fog computing est proposée pour la gestion des incendies de forêt, soulignant l'importance du fog computing dans le traitement du volume de données et l'amélioration du temps de réponse du système aux incidents d'incendie. La mise en œuvre de technologies avancées telles que le BDA et le fog computing peut considérablement améliorer l'efficacité et les processus de prise de décision dans divers domaines, en révolutionnant la façon dont les données sont gérées et utilisées.

mots-clés: Big Data, IdO, Analyse des données, Enseignement supérieur, Performance des étudiants, Environnement, Détection des feux de forêt.

مُلخّص

تشمل السمات المميزة للبيانات الضخمة حجمها الكبير وسرعة إنتاجها وتعدد أنواع البيانات وتنوعها واسع النطاق، مما يشكل عقبات أمام برامج معالجة البيانات التقليدية. ويُعزى انتشار البيانات الضخمة إلى انتشار إنترنت الأشياء (IoT) ، الذي يُنتج كميات هائلة من البيانات عبر الأجهزة المترابطة. ويُعد التعاون بين إنترنت الأشياء والبيانات الضخمة ثوريًا، مما يسهل تطوير تطبيقات ذكية في مختلف المجالات مثل الرعاية الصحية والتعليم العالي.

تناقش هذه الأطروحة التحديات والفرص في إدارة واستخراج المعلومات من البيانات الضخمة، مع التركيز بشكل خاص على البيانات الضخمة الناتجة عن إنترنت الأشياء في قطاعات مثل التعليم العالي والإدراك البيئي والإدارة البيئية. وتسلط الأطروحة الضوء على الفوائد المحتملة لاستخدام البيانات الضخمة في هذه القطاعات وتؤكد على أهمية التقنيات المتقدمة لتحسين آليات الرصد والتحليل والاستجابة.

في التعليم العالي، تقدم الدراسة مفهوم تحليلات البيانات الضخمة (BDA) لتحسين أداء الطلاب واتخاذ القرارات، وتقدم نموذجًا لإدارة سلسلة التوريد التعليمية والتحليلات التنبؤية لنتائج الطلاب. في مجال الإدارة البيئية، تُقترح بنية متعددة الطبقات تستفيد من إنترنت الأشياء وحوسبة الضباب لإدارة حرائق الغابات، مع التأكيد على أهمية حوسبة الضباب في التعامل مع حجم البيانات وتحسين وقت استجابة النظام لحوادث الحرائق. يمكن أن يؤدي تطبيق التقنيات المتقدمة مثل إنترنت الأشياء وحوسبة الضباب إلى تعزيز الكفاءة وعمليات اتخاذ القرار في مختلف المجالات، مما يُحدث ثورة في طريقة إدارة البيانات واستخدامها.

الكلمات المفتاحيّة: البيانات الضخمة، إنترنت الأشياء، تحليل البيانات، التعليم العالي، أداء الطلاب، البيئة، الكشف عن حرائق الغابات.

Table of Contents

List of Figures

List of Tables

General Introduction

1 Context and motivation

Big data is defined by its high volume, rapid velocity of data processing, and diverse variety of data types. Collectively, these embody the vast and intricate datasets that conventional data processing software cannot handle with optimal efficiency.

The evolution of the Internet of Things (IoT), which encompasses a network of interconnected devices equipped with sensors, software, and other technologies, has significantly contributed to the exponential growth of big data. These IoT devices generate vast amounts of data through their sensors, thereby playing a pivotal role in creating and accumulating big data. The synergy between IoT and big data is foundational and transformative across various sectors, enabling the development of intelligent applications through machine learning and big data analytics (BDA).

The potential applications of combining IoT with big data are vast and varied. In the healthcare sector, for instance, IoT devices can collect heterogeneous data, including electronic health records and laboratory data, which, when analyzed, can lead to improved healthcare solutions, especially in the context of pandemics (Kaur, 2023).

In the higher education sector, by integrating IoT devices and leveraging BDA, educational institutions can enhance teaching, learning, and research processes.

This combination delivers excellent training, personalized learning experiences, and efficient research methodologies (Singh & Madaan, 2022). IoT in higher education can also lead to improved operational efficiency, more engaging learning environments, and enhanced student performance monitoring.

Combining IoT and big data in environmental disaster monitoring offers significant potential applications. By integrating IoT devices with environmental sensors, real-time air, water, and soil quality data can be collected and transmitted to cloud networks for analysis (Fawzy et al., 2022). This integration enables the creation of autonomous monitoring systems that observe environmental parameters dynamically. Such systems enhance decision-making by providing granular and dynamic data for environmental monitoring and decision-making processes, ultimately contributing to more efficient and effective environmental protection strategies.

Managing IoT-generated big data presents significant research challenges, including knowledge discovery and intelligent decision-making. Addressing these challenges requires systematically exploring big data management mechanisms, architectures, and analytics types, underscoring the need for innovative data storage, processing, and security solutions. Thus, IoT integration with big data promises to revolutionize industries by enabling more informed decision-making and offering unprecedented levels of efficiency and customization.

2 Problematic

The principal problematic aspect of Big Data is the challenge of managing and extracting exploitable insights from the vast volumes of diverse and complex data generated in various fields. This involves difficulties in data storage, processing, analysis, and visualization, as well as the need for advanced algorithms and computational tools to manage the complexities of Big Data.

Our research focused on investigating two sectors showing promising potential for reaping significant benefits through data utilization, explicitly focusing on IoT-generated Big Data. These industries include the higher education sector, where data-driven insights can lead to transformative advancements, and the environmental perception and management sector, where leveraging data can bring about crucial improvements in ecological conservation practices and resource management.

In contemporary times, the application of big data analysis is widespread in various fields. Several industries, such as e-commerce, healthcare, technology, cybersecurity, and e-governance, have made notable progress in harnessing multiple data sources. In contrast, sectors like higher education are falling behind in efficiently utilizing data, particularly newly available data sources. Our study demonstrates the encouraging potential of BDA in the realm of higher education. It presents a conceptual framework for effectively incorporating universities' big data decision-support systems.

It is essential to highlight that the environmental perception and management domain plays a vital role in understanding and effectively dealing with various environmental challenges. Utilizing cutting-edge technologies such as the IoT and Big Data is increasingly becoming indispensable. These advanced technologies offer innovative solutions and insights that can significantly enhance the monitoring, analysis, and response mechanisms related to natural calamities like wildfires and floods. By harnessing the power of IoT and Big Data, stakeholders in environmental management can access real-time data and predictive analytics to anticipate, detect, and mitigate the impact of such disasters. This proactive approach not only improves emergency preparedness but also aids in minimizing the socio-economic and environmental consequences associated with these events. Integrating IoT and Big Data technologies in environmental perception and management is paramount for building resilient and sustainable strategies that protect both lives and ecosystems from the devastating effects of wildfires, floods, and other natural disasters.

3 Objectives and contributions

The primary objective of our research is to demonstrate the numerous opportunities and benefits that can be derived from the effective utilization of emerging technologies for processing and analyzing vast amounts of data, with a particular emphasis on their application within the context of IoT applications—our specific area of interest centers on two distinct fields: higher education and environmental perception and management.
The contribution of the present study in these two domains is as follows:

1- *Higher education field*

- First contribution (Boughouas et al., 2022): Our first contribution introduces the concept of BDA in the context of higher education, emphasizing the importance of analyzing large volumes of educational data to improve student performance and institutional decision-making. It discusses the application of three phases of BDA (descriptive, predictive, and prescriptive analytics) through a case study, highlighting the process of using machine learning techniques to predict student outcomes. The study highlights the use of Decision Trees and Random Forest algorithms for building predictive models, providing insights into their effectiveness in forecasting student success rates and offering recommendations for educational improvements.
- Second contribution (Boughouas, Kissoum, & Mazouzi, 2023): Our

second contribution in the higher education field introduces an educational supply chain management (ESCM) model tailored for higher education institutions, highlighting the roles of various actors, types of services, and the importance of decision-making for smooth operations. It explores the potential of BDA in enhancing the decision-making process within the educational supply chain. It proposes a conceptual framework for implementing a big data-driven decision-support solution in universities. Applying BDA in the educational supply chain supports the development of predictive models and recommendations. This guidance guides decision-makers toward more effective and forward-looking decisions, enhancing the overall quality of education and research outcomes. A case study on data analysis is presented to showcase how data can be effectively used in universities to improve quality assurance and policy satisfaction among students, demonstrating the practical benefits of the proposed model.

2- *Environmental perception and monitoring field*

– Our contribution to this field was the introduction of a multi-layered architecture for forest fire management that leverages wireless sensor networks (WSNs), IoT, cloud, and fog computing technologies to enhance wildfire prevention, detection, and intervention (Boughouas, Kissoum, Mazouzi, & Boussouf, 2023). The research emphasizes the significance of fog computing in handling the immense volume of data produced by sensors located in forested regions. It suggests fog computing as a potential resolution for addressing network congestion and delays, which improves the system's response time to fire incidents. The proposed architecture is validated by utilizing the iFogSim simulation tool. It demonstrates that the fog-cloud model significantly reduces latency and bandwidth consumption compared to a cloud-only model, offering a more efficient data processing and management solution for forest fire detection systems (FFDS). By employing drones equipped with fire extinguishing balls, the study suggested an innovative approach to rapidly address fires, enhancing the system's ability to effectively prevent and mitigate forest fires.

4 Thesis Outline

This document will review the studied subjects and aim to place our research within the current literature in this field. As a result, we have divided our manuscript into two main sections.

- ***Part 1:*** *Theoretical background and related works:* This section establishes the foundation of understanding for Big Data and the IoT, exploring their evolution and importance in the modern digital world. The first chapter of this section outlines Big Data's progress and growth, as well as diverse data sources and methods used for processing and analysis to obtain useful insights for decision-making. The second chapter delves into the IoT, clarifying its structure with various components working together to collect, transmit, and analyze data from connected devices. Additionally, it discusses the application of IoT in different sectors like urban planning, healthcare, and industry, demonstrating how IoT technologies can improve efficiency, safety, and service delivery.

  The first part of this thesis concludes with a dedicated related works chapter that discusses studies relevant to the thesis's contributions in the fields of Big Data analytics and IoT, specifically within the contexts of higher education and environmental perception and management. This chapter reviews key research, examining their methodologies, findings, and limitations. By identifying gaps in the existing literature, it highlights areas for further exploration and positions the current study within the broader academic discourse. This comprehensive review sets the stage for the subsequent parts of the thesis, which will detail the research methodologies, experimental setups, and results, ultimately advancing knowledge in Big Data analytics and IoT applications in these two sectors.

- ***Part 2:*** *Scientific contributions:* In this section are presented the contributions of this work to the field of Big Data and IoT summarized into three chapters as follows:

  1. The first chapter of this section deals with improving student performance by exploiting data analysis techniques. A case study is conducted to demonstrate how to extract useful information from massive data sources (big data). Three phases of data analysis were presented (descriptive, predictive, and prescriptive). Suggestions were then made based on the predictive analysis results to help students and teachers improve student performance and the education system as a whole.

  2. The second chapter introduces an ESCM model for higher education institutions, emphasizing the roles of different stakeholders, services, and decision-making processes. It explores using BDA to enhance decision-making within the educational supply chain. A case study on data analysis is presented to demonstrate how data can be effectively utilized in universities to improve quality assurance and student satisfaction, showcasing the practical benefits of the ESCM model.

3. The last chapter of the second part of this manuscript focuses on environmental conservation, particularly in forest areas. This chapter presents an architecture for forest fire detection and management based on fog computing, IoT, and big data analysis technologies. The domain is well-dissected, and some very interesting results have been presented.

PART ONE: BACKGROUND & RELATED WORKS

Chapter I: A Comprehensive Overview of Big Data

Chapter II: The Internet of Things

Chapter III: Related Works

Chapter I:
A Comprehensive Overview of Big Data

1 Introduction

Big data refers to large and complex datasets characterized by their high volume, velocity, and variety. It encompasses massive amounts of information that are beyond the capabilities of traditional data processing tools and techniques to handle effectively. Big data is typically generated from various sources, such as social media, sensors, mobile devices, and online transactions, and it often includes structured, semi-structured, and unstructured data. The distinctiveness of big data lies in its ability to provide insights and value that were previously unattainable. It offers the potential to uncover patterns, trends, and correlations, leading to enhanced decision-making, innovation, and problem-solving across different domains. The challenges associated with big data include data storage, processing, analysis, and privacy concerns. Advanced technologies and analytics approaches, such as machine learning and data mining, are often employed to derive meaningful insights from big data. Today, a growing body of research projects highlights the significance of big data analytics for extracting insights, delves into the challenges and opportunities of processing and analyzing massive datasets, and underscores the transformative influence of big data across various industries (Bhattarai et al., 2019; N. et al., 2022; Dai et al., 2019). In conclusion, big data represents a revolutionary shift in handling and harnessing information. Its defining features distinguish it from traditional data sources, including its sheer volume, velocity, and variety. This vast and diverse collection of data sources, from social media to sensors, offers unprecedented opportunities for uncovering patterns and correlations that were previously inaccessible. Big data's potential for enhancing decision-making, fostering innovation, and solving complex problems is evident across various fields. However, its challenges, from data storage and processing to privacy concerns, require advanced technologies and analytics approaches like machine learning and data mining. The research, exemplified by (M. Chen et al., 2014), underscores the critical role of big data analytics in extracting valuable insights and knowledge from these vast datasets, with applications spanning healthcare, finance, transportation, and more. As we continue to explore and exploit the possibilities of big data, its impact on our world will undoubtedly continue to evolve and expand.

2 Big Data Evolution

Big data has rapidly emerged as a significant field of study and practice, driven by technological advancements and the exponential growth of digital information.
The early 2000s witnessed a significant increase in data generation due to the proliferation of digital technologies and the internet. The International Data Corporation (IDC) predicted a "digital universe" that would grow to 40 zettabytes (1 zettabyte = 1 trillion gigabytes) by 2020, highlighting the need for managing and extracting value from massive datasets (Gantz & Reinsel, 2012).
In 2001, Doug Laney (Laney, 2001) introduced the concept of the 3Vs model, defining big data based on its three main characteristics: volume, velocity, and variety. As data volumes grew exponentially, traditional computing approaches became inadequate. Distributed computing technologies played a crucial role in big data's evolution.
In 2004, Google published a paper on the MapReduce programming model, which formed the foundation for processing large-scale data on distributed systems (Dean & Ghemawat, 2004).
Hadoop, an open-source implementation of MapReduce, was developed by Doug Cutting and Mike Cafarella in 2006. Hadoop provided a scalable and cost-effective framework for processing and storing massive amounts of data (White, 2012).
Big data has gained traction in various industries with scalable computing frameworks like Hadoop. Retailers started using big data analytics to gain insights into customer behavior, optimize inventory management, and enhance personalized marketing campaigns. Healthcare organizations leveraged big data to improve patient outcomes, identify disease patterns, and streamline clinical processes. Financial institutions adopted big data analytics for fraud detection, risk management, and algorithmic trading.
As data volumes grew, new technologies emerged to address the challenges of storing and processing big data. NoSQL databases, such as Apache Cassandra and MongoDB, provide scalable and distributed storage solutions for handling large datasets with high velocity and variety. In-memory computing technologies, like Apache Spark, enabled faster data processing by leveraging distributed computing and caching data in memory.
Machine learning and artificial intelligence (AI) techniques were crucial in deriving meaningful insights from big data. Deep learning algorithms, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), enabled advanced

image and speech recognition capabilities. AI-powered recommendation systems, natural language processing (NLP), and predictive analytics further enhanced the value extraction from big data.
These milestones and technological advancements have collectively shaped big data into a significant field of study and practice, enabling organizations to harness the potential of data for decision-making, innovation, and competitive advantage.

3 Big Data Vs

The three Vs of big data, including volume, velocity, and variety, offer a structured approach that serves as a foundation for comprehending the fundamental traits and obstacles linked to efficient handling and extracting valuable insights from extensive data collections. Figure 1.1 shows the three Vs of big data and their respective specifications.

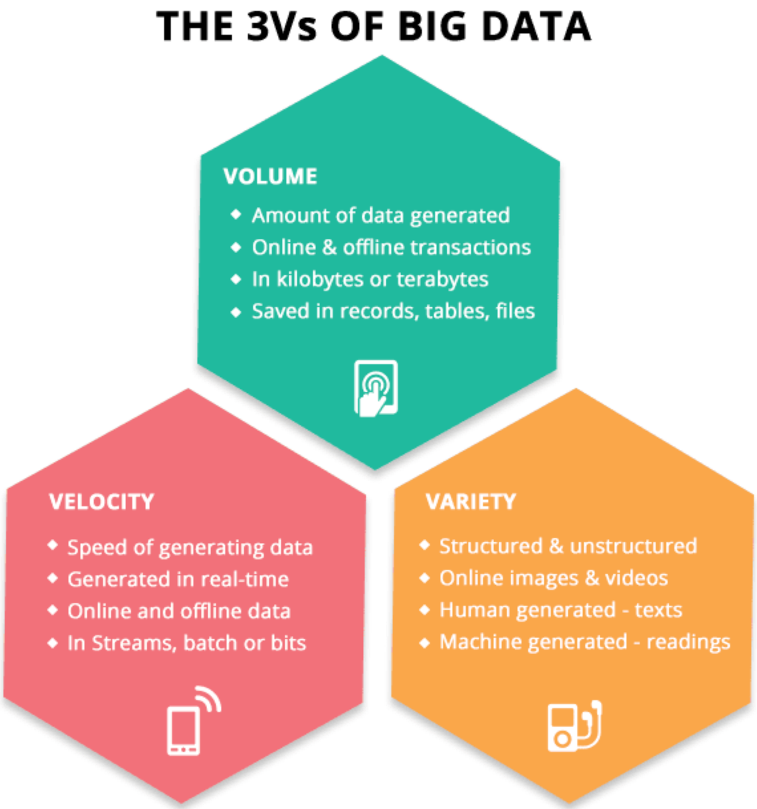

Figure 1.1: The 3Vs of Big Data

3.1 Volume

Volume is a term that denotes the immense quantity of information produced and gathered from various sources, encompassing both the sheer size of data sets and the rapid, exponential expansion of digital content. These data sets can vary anywhere from terabytes to petabytes and even surpass that scale. The rise of digital technologies, social media platforms, IoT devices, and sensor networks has led to a significant surge in the generated data. According to (Reinsel et al., 2017), the overall global data sphere will hit a staggering 163 zettabytes by 2025. The growing volume of data brings various challenges regarding storage capacities, processing capabilities, and the analysis of such vast amounts of information.

3.2 Velocity

Velocity is the rate at which data is created, gathered, and handled, highlighting data flow's immediate or nearly immediate characteristics. Data generation is occurring at a pace never seen before, necessitating that companies possess the capability to efficiently handle and scrutinize it to obtain insights that can be acted upon.
In the context of the IoT, a vast array of devices can produce continuous streams of data in real time, contributing to the ever-growing pool of information. Handling data that is delivered at high speed and volume necessitates the implementation of effective mechanisms for data intake, the processing of streams, as well as the ability to conduct real-time analysis of the data, showcasing the importance of having robust analytics capabilities in place (Perera et al., 2014).

3.3 Variety

Variety represents the diverse types and sources of data. Big data is a term that encompasses a diverse array of data types, which includes structured data, such as relational databases, as well as unstructured data, like text, images, audio, video, social media posts, and sensor data. It encapsulates various information sources and formats that require advanced technologies and techniques for processing, analyzing, and deriving insights. Traditional relational databases are ill-suited to handle the variety of data generated today. Analyzing and deriving insights from this varied data requires flexible data models, storage systems, and analytical tools. In 2001, Laney explained the challenges of managing and analyzing diverse data types. It highlights the importance of flexible data models and analytics techniques capable of handling various data formats and structures (Laney, 2001).

Beyond the original three Vs., some researchers have observed that the framework can be expanded further to encompass additional Vs beyond traditional ones. These additional Vs include veracity, focusing on aspects related to data quality and reliability, as well as value, which pertains to the process of extracting valuable insights and deriving significance from the data at hand. Moreover, another crucial V is variability, which highlights the dynamic nature of data, showcasing how patterns and characteristics can evolve over time. A more comprehensive understanding of the nuances and intricacies associated with big data can be achieved by incorporating these supplementary Vs into the framework. These extra aspects help to understand the diverse characteristics of big data, illuminating its intricacies and providing a comprehensive view of the topic.

Overall, the Vs. of big data underscores and focuses on the extraordinary magnitude, rapid pace, and wide variety of data, highlighting and emphasizing the necessity for pioneering technologies and methodologies to manage, analyze, and derive value from these vast data collections.

4 Big Data Sources

Big data sources are diverse and encompass various data-generating systems, technologies, and platforms. Here are some common sources of big data.

4.1 The Internet of Things (IoT)

IoT devices, encompassing various technologies ranging from sensors to wearable connected gadgets and intelligent appliances, can generate significant amounts of data. These devices have been meticulously crafted to actively gather and transmit data instantaneously, thereby offering substantial and beneficial observations that have the potential to be utilized in a wide range of industries such as healthcare, transportation, and environmental monitoring. Operating within the IoT framework, the objects within this interconnected system are linked to the vast network of the internet, resulting in the generation of a substantial volume of data that requires adequate storage, processing, and presentation in a manner that is easy for users to comprehend (Gubbi et al., 2013).

4.2 Social Media and Online Platforms

Social media platforms, including Facebook, Twitter, and Instagram, are widely recognized for their capacity to produce substantial quantities of user-generated content, which includes various forms of data like text, images, videos, and user engagements. Researchers often analyze this diverse data to gain insights into user behavior, perform sentiment analysis, and identify emerging trends in the online space. It is widely acknowledged that social media big data, in conjunction with the continuous advancements in computing tools, plays a pivotal role in providing valuable and essential insights into human behavior. These insights are highly sought after and are extensively utilized by various entities, including corporations, individuals, and governmental bodies, for many purposes, as highlighted by (Ghani et al., 2018).

4.3 Transactional Data and Enterprise Systems

In the realm of big data, transactional data pertains to the intricate and comprehensive documentation of various business operations and transactions, encompassing a wide range of activities including but not limited to purchases, orders, and payments, all of which play a vital role in facilitating the smooth functioning and accuracy of accounting information systems (Murthy & Geerts, 2017). Enterprise systems are critical in managing these transactions by utilizing Big Data technologies like Apache Hadoop and MapReduce to extract relevant information from massive data volumes (Khanra et al., 2020). For example, integrating Big Data systems in economic activities in China has significantly enhanced the effectiveness of commercial decision-making processes, enabling analysis, forecasting, and improved management strategies based on extracted data insights (ęski, 2019).

4.4 Web and Clickstream Data

Web and Clickstream Data are crucial in big data analytics, especially for E-commerce companies. Clickstream data analysis involves collecting and analyzing data on the web pages visitors click, providing insights into customer behavior and preferences. This data is essential for web marketing, customer prediction, and product management. Clickstream data is commonly preserved within access.log files situated on web servers, encompassing details such as IP addresses, referral pages, and timestamps of access (Ranjan et al., 2018). Analyzing this data using techniques like the Apriori algorithm can help understand user behavior and optimize website performance (Lemzin, 2023). Moreover, the visualization of clickstream data can be enhanced by employing advanced tools like Apache Flume, improving accuracy and performance compared to existing models (Supriyadi et al., 2018).

5 Big Data Technologies

Big data technologies are essential components that significantly impact the effective management, processing, and thorough analysis of vast data, greatly influencing various industries and sectors. These technologies play a crucial role in managing the intricacies and hurdles of the vast amount and diverse nature of data, allowing organizations to derive valuable insights and form informed decisions utilizing data-driven approaches. Moving forward, we will explore key big data technologies that foster innovation and productivity in today's data environment.

5.1 Apache Hadoop

Hadoop stands as an open-source distributed computing framework designed for the scalable management and storage of extensive datasets. Its primary objective is to offer a distributed, expandable, and scalable data storage and processing solution. Hadoop accommodates the handling of diverse data types, including unstructured data. It utilizes two distinct systems for data storage: HDFS (Hadoop Distributed File System) and HBase. These two components combine to create a management system for column-oriented databases distributed across server clusters (A. B. Patel et al., 2012).

Hadoop uses the MapReduce programming model to parallelly process large datasets across numerous nodes within a cluster of computers, thereby accelerating computations and hiding the latency of input and output operations. Hadoop's HDFS provides fault-tolerant storage for massive amounts of data through data replication. There are other parts of Hadoop; however, these are Hadoop's kernel capabilities (White, 2012).

5.1.1 Hadoop Distributed File System (HDFS)

HDFS is a distributed and scalable file system designed to store a large amount of data across distributed clusters. HDFS partitions data into small data blocks (the default block size is 64 MB) and distributes the data blocks across its cluster. Figure 1.2 shows the HDFS architecture. HDFS's main characteristics are as follows:

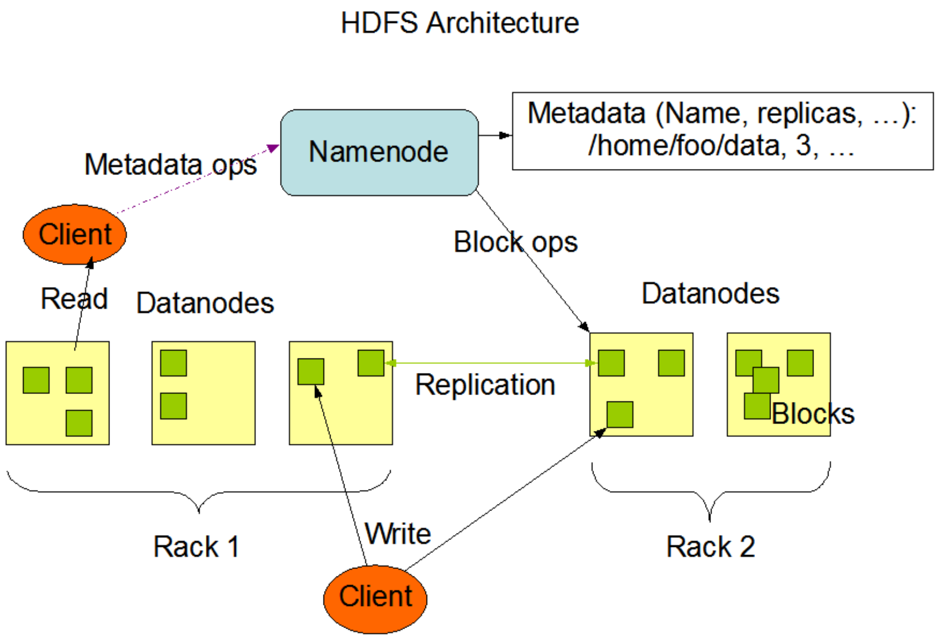

Figure 1.2: HDFS Architecture (Apache Software Foundation, 2023)

- *Fault Tolerance:* HDFS achieves fault tolerance by replicating each data block multiple times and storing them on different machines. If a node fails, HDFS automatically detects the failure and re-replicates the lost blocks to maintain data availability.
- *NameNode and DataNode Architecture:* The HDFS namespace is a hierarchy of files and directories represented on the NameNode. HDFS follows a master-slave architecture. The NameNode is the master node responsible for managing the file system namespace and controlling file access. The NameNode maintains the namespace tree and the mapping of file blocks to Data Nodes. DataNodes are the slave nodes that store and manage the actual data blocks. Data blocks are replicated for fault tolerance and rapid access (Dev & Patgiri, 2014).
- *Data Integrity:* HDFS verifies data integrity using checksums. Each data block is associated with a checksum, and HDFS checks the checksum during read operations to ensure the data's integrity.

5.1.2 MapReduce

MapReduce is a programming model and software framework for processing and analyzing large-scale data sets in a distributed computing environment. Google introduced it in 2004 (Dean & Ghemawat, 2004) to simplify the development of parallel data processing algorithms. In MapReduce, computations are expressed as two functions: a Map function that processes input key-value pairs and produces intermediate key-value pairs and a Reduce function that merges intermediate values associated with the same intermediate key.

The MapReduce framework plays a fundamental role in Hadoop. MapReduce is responsible for the distributed processing and analysis of data in Hadoop.

- *Data Processing Paradigm:* MapReduce provides a programming model and execution framework for processing large-scale datasets in parallel across a cluster of machines. It abstracts the complexity of distributed computing, allowing developers to focus on writing high-level code for data processing tasks.
- *Scalable and Distributed Processing:* MapReduce enables the parallel execution of computations by dividing data into smaller subsets and processing them independently across multiple machines in a Hadoop cluster. It harnesses the computational power of the entire cluster, making it highly scalable and capable of processing massive amounts of data.
- MapReduce optimizes data processing by exploiting data locality. MapReduce tasks are scheduled near the data they need to process, reducing data transfer over the network and improving overall performance.

5.2 Apache Spark

Apache Spark is an open-source big data processing framework that provides fast and in-memory analytics capabilities, Ease of use, and the ability to handle large datasets of various types (text data, graph data, etc.). It handles various data processing tasks, notably batch processing, real-time streaming, machine learning, and graph processing. Spark extends the MapReduce model to efficiently support multiple types of computations, including iterative processing, interactive queries, and stream processing. Spark leverages in-memory computing and a resilient distributed dataset (RDD) abstraction to enable fast and efficient data processing across clusters of machines (Zaharia et al., 2016). It offers a rich set of libraries and APIs for various data processing tasks and has gained significant popularity in the big data analytics community.

Spark has significant advantages compared to Hadoop (Zaharia et al., 2016):

- *Speed:* Spark is known for its exceptional speed and performance. It performs in-memory processing to optimize iterative workloads and interactive queries, leading to faster data retrieval and processing than Hadoop's disk-based processing.
- *Advanced Analytics:* Spark offers a rich set of libraries, including MLlib for machine learning, GraphX for graph processing, and Spark SQL for querying structured data. These libraries enable advanced analytics capabilities within the Spark ecosystem, making it suitable for various data processing and analysis tasks.
- *Real-time Processing:*Spark provides built-in support for stream processing, allowing users to process and analyze real-time data streams in a scalable and efficient manner. This is in contrast to Hadoop, which primarily focuses on batch processing.
- *In-Memory Processing:*Spark leverages in-memory computation, which enables faster data processing and iterative operations. It keeps the data in memory, reducing the need for disk I/O and improving overall processing speed.
- Spark integrates seamlessly with other big data tools and frameworks like HDFS, HBase, Cassandra, and more. This allows users to leverage their infrastructure investments and build upon their current data ecosystem.
- *Ease of use:* Spark supports multiple programming languages such as Scala, Java, Python, and R. This makes it easier for developers to write and execute Spark applications, reducing the learning curve.
- *Flexibility:*Spark is designed to cover a wide range of workloads that previously required separate distributed systems, including real-time processing applications, iterative algorithms, interactive queries, and streaming. By supporting these workloads within the same engine, Spark makes it easy and cost-effective to combine different types of processing, which is often necessary for production data analysis pipelines.

5.3 NoSQL Databases

With the growth and multitude of communication and data-sharing tools on the internet, the transferred data has undergone a significant transformation in terms of type, structure, and quantity. This has rendered traditional database management systems inadequate for the new management, processing, and storage requirements. NoSQL databases, such as Apache Cassandra, MongoDB, and Apache HBase, are designed to handle the volume, velocity, and variety of big data. These databases provide flexible and scalable storage solutions for structured, semi-structured, and unstructured data (Han et al., 2011).

The following are some factors that contributed to the emergence of NoSQL:

- *Big Data:* Traditional relational database management systems (RDBMSs) faced challenges handling the sheer volume, velocity, and variety of big data. NoSQL databases were designed to scale horizontally across multiple machines, allowing them to handle massive data sets more effectively.
- *Scalability:* NoSQL databases are built to scale horizontally, meaning they can handle increasing workloads by adding more commodity hardware and distributing data across multiple nodes. This approach contrasts with traditional RDBMSs, which often rely on vertical scaling (adding more resources to a single machine) and can be cost-prohibitive for large datasets or high-traffic loads.
- *Performance:* NoSQL databases are optimized for high-speed read and write operations, making them suitable for applications requiring low-latency data access. They achieve this by sacrificing certain consistency guarantees provided by RDBMSs, favoring availability and partition tolerance (known as the CAP theorem) (Han et al., 2011). This trade-off allows for increased performance and responsiveness in many scenarios.
- *Flexible Data Models:* NoSQL databases support various data models, including key-value, document, columnar, and graph models, which offer flexibility in modeling and querying data. This versatility allows developers to choose the most appropriate data model based on their application's requirements, enabling efficient storage and retrieval of complex, unstructured, or hierarchical data.
- *Agile Development:* NoSQL databases are schema-less or schema-flexible, meaning the data structure can be modified without strict predefined schemas. This flexibility simplifies the development process, allowing for easy iteration and adaptation to changing data requirements. It also facilitates faster application development in scenarios where the data schema evolves rapidly or lacks a fixed structure.

- *Distributed and Fault-Tolerant Architecture:* NoSQL databases are often designed to operate in distributed environments, allowing them to handle large-scale deployments across multiple nodes or data centers. They employ sharding (data partitioning) and replication to ensure data availability and fault tolerance. This architecture improves system resilience and reduces the risk of data loss or downtime.

5.4 Other Big Data Processing Tools

5.4.1 Apache Kafka

Kafka is a distributed messaging system that enables the real-time processing of high-velocity log data streams. It provides high-throughput, fault-tolerant, and scalable messaging capabilities, making it suitable for handling large volumes of data streams and allowing for building real-time data pipelines (Kreps et al., 2011). Kafka follows a publish-subscribe messaging model, where producers publish data to topics in synchronous or asynchronous mode, and consumers subscribe to those topics to receive the data. A topic serves as a classification or label under which messages are published, signifying a flow of data in Kafka. Within Kafka, a topic can be split into numerous partitions, with each partition being a sequentially ordered and unchangeable series of data records. Partitions allow data to be distributed across multiple Kafka brokers, enabling parallel processing and increasing throughput. The data stored in Kafka topics are delivered to consumers. They subscribe to one or more topics and consume records from the assigned partitions. Kafka provides both low-level and high-level consumer APIs for consuming messages. Another building block of Kafka is brokers. Brokers are the servers in a Kafka cluster that handle the storage and replication of data. Their tasks are maintaining and synchronizing partitions, managing producers and consumers, and guaranteeing fault tolerance and high availability. A Kafka cluster comprises numerous brokers who work together to achieve these goals.

5.4.2 Apache Flink

Flink is an open-source stream processing framework that enables high-throughput and low-latency streaming data processing. It supports event time processing, fault tolerance, and stateful computations and provides low-latency, high-throughput processing capabilities (Carbone et al., 2015). Flink's dataflow model allows batch and stream processing, making it suitable for real-time analytics. Flink can also integrate with other tools like Apache Kafka for data ingestion.

Flink follows a distributed processing model and provides several core components and mechanisms to achieve its functionality.

- *Dataflow Model:* Flink uses a directed acyclic graph (DAG) dataflow model to represent and execute data processing tasks. Operators perform transformations on data streams or datasets, allowing for complex processing pipelines.
- *Data Streams and DataSets:* Flink operates on DataStreams for real-time streaming processing and DataSets for batch processing. Both abstractions provide APIs for transformations and computations.
- *Fault Tolerance:* Flink ensures fault tolerance through checkpointing, taking snapshots of the operator state at regular intervals. Flink can restore the state and resume processing from the last successful checkpoint in case of failures.
- *Event Time Processing:* Flink supports event time processing, handling out-of-order events and processing data based on event timestamps.

5.4.3 Apache Pig

Apache Pig is a high-level data flow scripting language and execution framework developed by Yahoo! Research that is designed for processing and analyzing large datasets. It is designed to enable data analysts to efficiently process and analyze large datasets in a distributed computing environment, such as Apache Hadoop. Pig is composed of two main components (Swarna & Ansari, 2017). The first is Pig Latin, a parallel dataflow language that allows users to express data transformations and operations concisely, abstracting the complexities of writing low-level MapReduce programs. Pig Latin scripts are compiled into a series of MapReduce jobs and executed on a Hadoop cluster, providing scalability and parallel processing capabilities for big data analytics tasks. The Apache Pig's second component is the run time environment where Pig Latin scripts are executed.
These technologies represent a subset of the vast array of big data technologies available today. The big data technology landscape continually evolves, and new technologies and frameworks emerge to address evolving challenges and requirements.

6 Big Data Analytics

Big data analytics is a rapidly evolving field that extracts valuable insights and knowledge from vast and complex datasets. With the advancement of technology and the proliferation of digital data, organizations now have access to massive amounts of information from various sources such as social media, sensors, transactions, and more. Big data analytics employs advanced tools, algorithms, and technologies to process, store, and analyze this data, leading to data-driven decision-making and enhanced business outcomes. Big data analytics encompasses various techniques and methodologies, including data mining, machine learning, predictive analytics, and data visualization.

Key Components of Big Data Analytics (Rao et al., 2018):

- *Data Collection:* Big data analytics begins with collecting massive volumes of data from various sources. This data may be structured, semi-structured, or unstructured and come in different formats.
- *Data Storage:* Managing and storing big data efficiently is critical to analytics. Distributed storage systems, data lakes, and cloud-based storage solutions are commonly used to handle enormous data volumes.
- *Data Processing:* Big data analytics involves processing large datasets quickly and efficiently. Technologies such as Hadoop and Spark enable distributed data processing across clusters of interconnected computers.
- *Data Analysis:* Data analysis techniques, such as data mining and machine learning, are applied to identify patterns, trends, and correlations within the data. These analyses help make predictions and gain valuable insights.
- *Data Visualization:* Presenting complex data visually is vital for understanding and communicating insights effectively. Data visualization tools and techniques help to represent data in charts, graphs, and dashboards.

6.1 Big Data Analytics Techniques

6.1.1 Data Mining

Data mining involves discovering patterns, relationships, and insights from large datasets. It utilizes statistical and machine learning techniques to uncover hidden patterns and make informed decisions. Research by (Witten & Frank, 2002) provides an overview of data mining techniques and their applications in different domains.

Applications: Data mining has applications in various industries, including marketing, healthcare, finance, and retail. For example, data mining techniques in marketing can help identify customer segments, predict purchasing behavior, and optimize marketing campaigns.

Benefits: Data mining enables organizations to gain valuable insights from their data, leading to improved decision-making, enhanced customer targeting, cost reduction, fraud detection, and process optimization.

6.1.2 Machine Learning

Machine learning is a complex field that encompasses the process of creating and refining intricate algorithms that have the capability to empower computers with the ability to extract valuable insights from extensive datasets, subsequently utilizing this knowledge to make informed predictions or decisions, all without the need for specific, systematic instructions in the form of traditional programming.

Applications: Machine learning is applied in various industries such as healthcare (diagnosis, personalized medicine), finance (fraud detection, risk assessment), e-commerce (recommendation systems), and manufacturing (predictive maintenance, quality control).

Benefits: Machine learning enables automation, improves decision-making accuracy, enhances productivity, detects anomalies, and enables personalized experiences, increasing efficiency and competitiveness.

6.1.3 Predictive Analytics

Predictive analytics uses historical data and statistical modeling techniques to predict future events or outcomes. Research by (X. Wu et al., 2014) discusses the applications and benefits of predictive analytics in different industries.

Applications: Predictive analytics is applied in sales forecasting, customer churn prediction, risk assessment, demand forecasting, and predictive maintenance.

Benefits: Predictive analytics enables proactive decision-making, improves resource allocation, reduces risks, optimizes operations, and enhances customer satisfaction.

6.1.4 Data Visualization

Data visualization involves representing data visually to facilitate understanding, exploration, and communication of insights. Research by (Heer & Shneiderman, 2012) discusses the principles and techniques of data visualization.

Applications: Data visualization is applied across various domains, including business intelligence, finance, healthcare, and scientific research. It helps in exploring patterns, identifying trends, and communicating insights effectively.

Benefits: Data visualization enables better data comprehension, supports exploratory analysis, enhances decision-making, facilitates communication of complex information, and promotes insights discovery.
The applications and benefits of these big data analytics techniques are not limited to specific industries, as they have broad applicability across various sectors.

7 Big Data Applications

The impact of big data on various sectors, including healthcare, finance, retail, and transportation, has been substantial. In the following, we will discuss the impact of big data in these sectors and the various applications and benefits it brings.

7.1 Healthcare

Big data has revolutionized healthcare by enabling data-driven decision-making, personalized medicine, and improved patient outcomes. (Topol, 2019) highlights the potential of big data in transforming healthcare through advancements in genomics, wearables, electronic health records, and clinical research.
The healthcare industry has witnessed exponential growth in data sources, encompassing hospital records, patient medical records, medical examination results, and data from IoT devices. Furthermore, biomedical research contributes substantially to the public healthcare data repository. However, realizing the potential of this data necessitates efficient management and analysis, as the alternative is akin to searching for a needle in a haystack. Each phase of data handling presents distinct challenges, which require the use of high-end computing solutions for analysis. To address these challenges and enhance public health, healthcare providers must establish robust infrastructure for systematically generating and analyzing big data. Effective management, analysis, and interpretation of big data have the potential to reshape modern healthcare, creating new opportunities. Hence, various industries, including healthcare, actively pursue strategies to translate this potential into improved services and financial benefits. Through the seamless integration of biomedical and healthcare data, contemporary healthcare organizations can revolutionize medical therapies and personalized medicine, ushering in a new era of healthcare delivery (Dash et al., 2019).
Applications: Big data analytics in healthcare includes disease surveillance, early diagnosis, patient monitoring, drug discovery, and precision medicine. It enables population health management, predictive modeling, and healthcare resource

optimization.

Benefits: The use of big data in healthcare leads to improved patient care, better clinical decision-making, reduced healthcare costs, early disease detection, and personalized treatment plans.

7.2 Finance

Big data has transformed the finance industry by providing insights for risk assessment, fraud detection, customer segmentation, and investment strategies.

The financial sector is deeply engaged in big data event calculation, which has led to the generation of extensive volumes of financial transaction data. This data encompasses customer information, logs from various financial products, transaction data utilized for informed decision-making, and external data sources, such as social media and websites. Consequently, professionals and analysts within the financial sector regard this data's effective management and analysis as a burgeoning challenge in financial products and services. Furthermore, it is essential to delve into the specific financial domains where big data exerts a substantial influence, given its profound impacts on various financial products and services. Identifying these areas of financial significance influenced by big data represents a critical area of exploration (Bach et al., 2019; Hasan et al., 2020).

Applications: Big data analytics in finance includes credit scoring, fraud detection, algorithmic trading, customer sentiment analysis, and personalized financial recommendations. It enables real-time risk assessment and enhances regulatory compliance.

Benefits: Using big data in finance leads to improved risk management, increased operational efficiency, enhanced customer experience, fraud prevention, and informed investment decisions.

7.3 Retail

Big data revolutionized the retail sector by delivering insights into customer behavior, preferences, and market trends. Like other sectors, big data in the retail industry brings challenges and opportunities. Presently, scrutinizing data and leaning on its insights has taken on heightened importance in shaping, testing, and designing strategies, marking it as a cornerstone of intelligent retail practices (Lekhwar et al., 2018). Retailers are delving into analytics to achieve a cohesive understanding of their customers and operational activities, encompassing physical stores and online channels.

This endeavor empowers them to make informed, strategic choices that, in turn, foster the advancement of the retail industry (Ying et al., 2020).

Applications: Big data analytics in retail includes customer segmentation, personalized marketing, inventory optimization, demand forecasting, and pricing optimization. It enables the creation of customized shopping experiences and enhances customer loyalty.

Benefits: Using big data in retail leads to improved customer targeting, increased sales and revenue, optimized inventory management, enhanced customer satisfaction, and competitive advantage.

7.4 Transportation

Big data has emerged as a focal point of research within intelligent transportation systems (ITS), a trend evident in numerous projects worldwide. ITS is poised to generate a substantial volume of data. The produced big data will be poised to influence the design and implementation of ITS significantly. As a result, ITS is on track to become safer, more efficient, and economically advantageous. Accurate and efficient data analysis, even on seemingly chaotic data sets, can improve service quality within ITS (Shi & Abdel-Aty, 2015; N. Mohamed & Al-Jaroodi, 2014). The evolution of ITS has led to a shift in data generation from the terabyte level to the petabyte level. Given the vast quantities of data involved, traditional data processing systems prove inefficient and are ill-equipped to meet the demands of data analytics (Zhu et al., 2019).

Applications: Big data analytics in transportation includes traffic prediction, congestion management, public transportation optimization, logistics planning, and vehicle fleet management. It enables real-time monitoring, improves safety, and enhances transportation efficiency.

Benefits: Using big data in transportation leads to reduced traffic congestion, improved route planning, enhanced public transportation services, efficient logistics operations, and better transportation infrastructure planning.

8 Big Data Challenges

The field of big data is constantly evolving, and researchers have identified several challenges and future directions.

8.1 Data Quality and Integration

By rapidly collecting and examining vast datasets from diverse origins and for various purposes, both researchers and decision-makers have come to recognize the manifold advantages of this substantial information resource. These benefits encompass gaining insights into customer requirements, enhancing service excellence, and forecasting and averting potential risks. Nevertheless, it is crucial to emphasize that the effective utilization and examination of big data hinges on the availability of precise and high-quality data, which serves as an indispensable prerequisite for extracting value from this extensive data repository (Cai & Zhu, 2015).

Data Integration involves converting data from its source to a target format. Many data warehousing and management strategies have relied on integration tools to facilitate data migration and transfer, often employing the Extract-Transform-Load (ETL) approach. While these tools excel at handling substantial data volumes, they are less adaptable when dealing with semi-structured or unstructured data. Programmatically driven parallel techniques, like the map-reduce models, have been introduced to address these significant data challenges. Data Integration is a complex and iterative process, particularly when incorporating new data sources. Adding these new sources can be time-consuming, leading to delays, potential data loss, data irrelevance, and the suboptimal utilization of valuable information (Arputhamary & Arockiam, 2015).

Big data integration deviates from conventional data integration in several significant ways (Dong & Srivastava, 2013):

1. The sheer quantity of data sources has expanded to encompass tens of thousands.
2. Most data sources are highly dynamic, constantly generating a vast influx of newly collected data.
3. The structural heterogeneity of these data sources is striking, with notable diversity even among entities that seem quite similar.
4. These data sources exhibit substantial disparities in their quality, encompassing variations in data coverage, accuracy, and the timeliness of the information they furnish.

8.2 Scalability and Performance

Big data scalability is fundamental for organizations dealing with large and complex datasets. It involves managing growing data volumes, diverse data types, and changing

needs while maintaining performance and cost efficiency. Scalable systems often leverage distributed computing frameworks and cloud technologies to achieve these goals, as traditional data processing techniques and infrastructure may not be capable of managing the increasing big data loads. As the volume of data continues to grow, scalability and performance become critical challenges in big data analytics. Research by (Zaharia et al., 2016) discusses the scalability challenges and proposes the Apache Spark framework as a solution for distributed data processing and scalable systems.
Optimizing the scalability of a database involves fine-tuning it with robust distributed frameworks and contemporary programming languages to efficiently process massive data sets in real-time network applications (Sundarakumar et al., 2022). In big data analytics, data processing has seen notable success employing tools like HADOOP and SPARK. Furthermore, emerging programming languages like Python have introduced innovative approaches such as map reduction and erasure coding to address these tasks. Nevertheless, when dealing with extensive datasets on network clusters, Python encounters specific challenges and limitations.

8.3 Data Privacy and Security

Big data raises concerns regarding privacy and security due to the large-scale collection and analysis of sensitive information. The following is an overview of the essential factors concerning privacy and data security in the big data realm.

- *Data Privacy:* Data privacy focuses on protecting individuals' personal information and ensuring its proper handling and use. With the increasing collection and analysis of personal data such as user profiles and log data, there is a growing concern regarding the potential invasion of user privacy when utilizing business analytics on big data (Li, 2014). This issue has gained significant importance.

  To address privacy concerns, researchers have developed various privacy-preserving techniques, such as randomization, k-anonymity, and distributed privacy-preserving, to provide adequate data protection when using business analytics on digital information (Pramanik et al., 2020).

- *Data Governance:* Data governance is critical to managing and protecting big data. Data governance refers to the overall management and control of data assets within an organization. It involves establishing policies, processes, and procedures to ensure data quality, integrity, availability, and security. In big data governance, it is essential to ensure rapid data preparation while prioritizing data consistency and reliability (Al-Badi et al., 2018). It is also necessary to build confidence in the data source and the significance of the results.

- *Ethical considerations:* Working with big data raises important ethical implications, including privacy concerns, potential biases, and issues of consent and transparency.

  Passively collected big data cannot build complex cross-tabulations that link individuals to a diverse range of attributes, family units, residences, neighborhoods, and occupations. Although such data may sometimes be exploited to infer relationships, the inferences made are susceptible to typical statistical problems such as ecological error, non-representative sampling bias, and self-selection bias (Shearmur, 2015).

  Obtaining informed consent from individuals whose data is used in big data analytics is a crucial ethical consideration. Governments and regulatory bodies have introduced privacy regulations to protect individuals' data. The European Union's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) are notable examples. When using data to document certain aspects of human behavior, it is often perceived as a challenge to fundamental values that cover a wide spectrum, from personal autonomy to concepts such as fairness, justice, respect for procedures, property rights, solidarity, and privacy (Barocas & Nissenbaum, 2014).

Big data analytics often involve complex algorithms, making it challenging to understand and explain the decisions based on the data. It relies on using aggregated small-scale data to generate voluminous datasets, which are subsequently subject to analysis in search of valuable information. The data collection process occurs discreetly, and the tools and techniques used are masked by multiple layers of physical, legal, and technical privacy measures deliberately designed to protect sensitive information (Richards & King, 2013). Data collection and processing techniques' transparency is a significant challenge when handling Big Data.

9 Conclusion

This chapter introduces the complex concept of Big Data, highlighting how this advanced technology has developed over time and its profound impact on the contemporary digital landscape. It provides a comprehensive overview of the evolution of Big Data, detailing its remarkable expansion, the diverse origins of the data it encompasses, and the complex processes involved in analyzing and interpreting it to derive valuable insights crucial to informed decision-making processes. Furthermore, it explores the wide range of applications of Big Data across various sectors such as healthcare, finance, transportation, and retail, demonstrating its ability to revolutionize operations, optimize performance, and foster innovation.

Despite these significant advancements and opportunities, Big Data also presents several challenges that must be addressed to fully harness its potential. These challenges include data privacy and security issues, the need for advanced analytical tools and skilled personnel, the handling of large volumes of unstructured data, and ensuring data quality and accuracy. Overcoming these obstacles is crucial for maximizing the benefits of Big Data and ensuring its sustainable and ethical use in the future.

Chapter II: The Internet of Things

1 Introduction

In 1999, Kevin Ashton introduced the IoT concept, describing it as a system of distinct, identifiable, and interconnected objects utilizing radio-frequency identification (RFID) technology. The growth of IoT has been primarily driven by the demands of large companies, who anticipate substantial benefits from the ability to track and monitor all objects within the supply chains in which they are integrated, guaranteeing greater anticipation and predictability. The IoT constitutes a technological paradigm shift that embodies the future of computing and communication, with its progress hinging on continuous technical innovation in various critical domains, ranging from wireless sensors to nanotechnology (Madakam et al., 2015). IoT involves using intelligent physical objects, services, and software systems to manage data (Korte et al., 2021). IoT systems face challenges related to security, privacy, compatibility, and scalability (R. Kumar et al., 2022). IoT has found applications in various fields, including city management, agriculture, remote monitoring, and smart street lighting. This chapter explores the key technologies enabling IoT, the architecture supporting its infrastructure, and the wide-ranging applications demonstrating its potential.

2 IoT Definition

The precise definition of IoT remains a fluid and perspective-dependent concept. In a broader sense, IoT has been described as a globally interconnected network infrastructure with self-configuring capabilities that adhere to established standards and communication protocols (Gokhale et al., 2018).

Following are some IoT definitions from the literature:

Definition 1: *"A worldwide network of interconnected objects uniquely addressable, based on standard communication protocols"* (Atzori et al., 2010).

Definition 2: (K. Patel et al., 2016) defines IoT into three categories: *"Internet of things is an internet of three things: (1). People to people, (2) People to machine /things, (3) Things /machine to things /machine, Interacting through internet."*

Definition 3: *"IoT allows people and things to be connected anytime, anyplace, with anything and anyone, ideally using any path/network and any service"* (Sezer et al., 2018).

3 IoT Enabling Key Technologies

IoT initially exploited existing internet and networking technologies, infrastructure, and conventional computing to suit IoT applications and transform stand-alone objects into intelligent, interconnected devices. However, employing technologies not specifically designed for IoT raised several problems, such as the lack of auto-configuration in the IPv4 protocol. Therefore, there is a growing need to develop novel technologies that tackle critical challenges like effective routing, scalability, and mobility to enhance the simplicity and effectiveness of IoT application development (Alsubaei et al., 2018).

After more than a decade of progress in IoT-enabling technologies, it is evident that efficiency (which necessitates the best use of resources) is one of the most notable trends (e.g., battery and memory). Designers' goals are low device complexity, low energy consumption (extended network lifetime), and a suitable trade-off between communication and signal/data processing power. The lack of interoperability in the IoT is a significant challenge to efficiency (i.e., devices from different brands and models cannot operate well together without a translation layer). Because there are currently no unifying standards for the IoT, interoperability issues remain a significant difficulty in existing IoT technology. Potentially, these issues will become less common as IoT becomes more developed (Alsubaei et al., 2018; Bansal & Kumar, 2020).

In the IoT architecture, several key enabling technologies operate within different layers to ensure the smooth functioning of the entire system. The three main layers explained regarding their associated IoT key enabling technologies are as follows:

3.1 Perception Layer

The perception layer in IoT architectures plays a crucial role in connecting sensor nodes and data acquisition units to capture relevant data from the environment. Various enabling technologies have been proposed to enhance this layer. One approach involves adjusting the time interval between transmissions based on detected events, utilizing communication technologies like LoRaWAN, Wi-Fi, or cellular networks (Dragulinescu et al., 2023). Another innovative technology includes digital tattooing, where a digital signal is embedded into a host signal for security purposes, using techniques like Cyclic Redundancy Check (CRC) for data integrity (Jana et al., 2023). Additionally, an intelligent directional sensitivity-based perception algorithm (DSPA) has been introduced to optimize energy maintenance and perception rate in IoT services, inspired by human visual direction-sensitive systems (Z. Yang et al., 2023). These technologies aim to improve data collection, communication, and security within the perception layer of IoT architectures.

3.2 Network Layer

IoT coordinates different empowering advances at the network layer to guarantee consistent communication and interoperability among an endless cluster of devices. One of the foundational perspectives of IoT is Machine-to-Machine (M2M) communications, which encourages the trade of real-time data between devices, changing them into intelligent systems capable of independent operations (Saponara, 2022). At the network layer, addressing and scalability issues are fundamental, given the endless number of devices associated with the IoT. Using unique identifiers (UIDs) and deploying sensors are critical for data collection and device communication, highlighting the importance of robust protocols and technologies to manage these elements effectively (S. Gupta et al., 2022). Furthermore, the advent of 5G IoT systems introduces a new dimension to network layer technologies, offering enhanced connectivity options and supporting a wide range of applications through improved speed and reliability (Pateromichelakis et al., 2022). Security within the IoT network layer is critical, as devices often gather private and critical data. Ensuring data confidentiality, integrity, and authentication requires robust security solutions and standardized frameworks to protect against potential threats and vulnerabilities (AlAali et al., 2022). Furthermore, the advancement of lightweight encryption algorithms and secure communication protocols, such as the Secure Reliable Message Communication (SEC-RMC) protocol, is crucial in protecting information transmission between IoT devices (Sreekantha et al., 2021). Moreover, the network architecture must be flexible and efficient to accommodate the diversity and volume of IoT services. Network slicing, empowered by Network Functions Virtualization (NFV), offers a promising arrangement by allowing the customization of network slices with diverse QoS to support different IoT applications, guaranteeing ideal execution over the board (Shilpa et al., 2022). In summary, the network layer of IoT encompasses a range of enabling technologies, from communication protocols and security frameworks to advanced network architectures like 5G and NFV-enabled network slicing, all of which are crucial for the seamless, secure, and efficient operation of IoT systems (Alam et al., 2020; Tsai & Lin, 2020).

3.3 Application Layer

IoT speaks to a critical shift in how devices communicate and work, with the application layer significantly empowering these technologies. The application layer protocols are essential for facilitating communication between devices, objects, or things within the IoT framework, ensuring secure, reliable, and efficient data transfer to meet various system requirements (Saponara, 2022). Among the plethora of protocols available, MQTT, XMPP, CoAP, MQTT-SN, STOMP, AMQP, DDS, JMS, LwM2M, REST, HTTP/2, and WebSocket are discussed extensively for their ability to establish trusted connections between objects and things (Win et al., 2023). These protocols are planned to cater to the differing needs of IoT applications, from guaranteeing low latency and power utilization to providing high security and speed (Pateromichelakis et al., 2022). The application layer's significance extends beyond mere communication; it includes integrating vertical IoT applications within future mobile networks through frameworks like vertical application enablement. This integration is facilitated by Network Applications (NetApps) that optimize and translate interactions between IoT apps and mobile networks, highlighting the importance of data management and analytics in enhancing application performance (Choudhary & Tanwar, 2023). Moreover, the application layer's role in IoT is underscored by its capacity to support intelligent decision-making by interfacing physical objects without human intercession, leveraging progressions in RFID, smart sensors, and communication technologies (Naikwadi et al., 2022; Jienan et al., 2021). Security within the application layer is a paramount concern, given the diversity of IoT's application areas. Research focuses on addressing security concerns and mitigating various types of attacks that could compromise the application layer, ensuring the integrity and confidentiality of data being transferred among objects (Sarvaiya, 2022; Lalit et al., 2022). This comprehensive approach to application layer protocols in IoT enables seamless communication among heterogeneous devices. It ensures that these communications are secure and efficient, paving the way for realizing smart application goals (Lalit et al., 2022).

4 IoT Architecture

As examined in different studies, the IoT architectures are fundamental for the successful deployment and operation of IoT systems, with each architecture presenting a unique set of layers that cater to different functionalities and services. A comprehensive investigation reveals that IoT architectures can extend from three to twelve layers, depending on the complexity and requirements of the IoT system in question. A typical beginning point is the three-layer architecture, which includes the perception, network, and application layers. This basic structure is the foundation for understanding IoT systems, focusing on data collection, transmission, and application-specific processing (Fagroud et al., 2023). However, additional layers are often introduced to address more complex needs and functionalities. For instance, a four-layer architecture separates the application from the processing layer, presenting a distinct layer for data processing, which enhances the system's ability to manage and analyze data. Further expanding on this, a seven-layer model has been proposed to address the complexities of IoT communication, ensuring secure and authenticated interactions between devices and cloud computing components (V. Kumar et al., 2023). In more detailed studies, up to twelve layers have been identified. However, after thorough analysis, six key layers were deemed most relevant for constructing a comprehensive IoT system in a cloud environment. These layers include advanced functionalities such as fog computing, which bridges the crevice between cloud and edge devices, and an insight layer that applies machine learning techniques to data (Bouaouad et al., 2020; Isha et al., 2022). Moreover, integrating IoT with Blockchain technology introduces a layered architecture that emphasizes security and immutable data storage, further illustrating the adaptability and diversity of IoT architectures to incorporate emerging technologies (C. K. Wu, 2021; Kakkar et al., 2021). A unified IoT architecture proposal aims to standardize the diverse platforms and layers, suggesting a seven-layer framework that could serve as a defacto standard for IoT applications, thereby simplifying the development process and ensuring consistency across different IoT systems (Sadique & Johannesson, 2021). In summary, IoT architectures are varied and can range from three to twelve layers, each serving specific functions from data collection and processing to security and application-specific services. The choice of architecture and the number of layers depend on the particular necessities of the IoT system, including considerations for security, data processing, and communication needs (Devadas & Subramanian, 2019). The most commonly identified layers are the perception/sensor layer, the network layer, the processing layer, the application layer, and the analysis engine layer, all supported by critical security and management functions.

4.1 Perception Layer

The perception or sensor layer in the IoT architecture serves as the foundational interface between the physical and digital realms. It is primarily responsible for collecting environmental data through sensors and RFID tags. This layer's multifaceted objectives aim to ensure the seamless, secure, and trustworthy interconnection of intelligent sensors to facilitate automated high-level smart applications. Semantic metadata plays a crucial role in enhancing the accessibility and interoperability of these features, making it easier for both machines and humans to process sensory data (Honti & Abonyi, 2019). Effective sensor deployment within this layer is critical for gathering sufficient information to support decision-making processes in IoT-based applications, as demonstrated by the development of algorithms to optimize sensor placement and connectivity (Cheng et al., 2020). Recent advancements have enabled smart sensors to collaborate directly, fostering the development of novel applications without human intervention. This collaboration is pivotal in supporting intelligent and efficient decision-making processes across various technologies and applications interconnected within the IoT (Poongodi et al., 2021). Security within this layer is addressed through lightweight cryptography and single-key management technologies, ensuring the credibility, integrity, and confidentiality of sensed information (Hu et al., 2013). Integrating machine learning algorithms, particularly deep learning techniques, into the perception layer has driven the emergence of Cognitive IoT (CIoT), enhancing the capability for self-diagnosis and decision-making directly at the sensor level (Hussein et al., 2022). Indoor Environmental Quality (IEQ) applications within the IoT leverage sensor networks for user-environment monitoring, dynamically managing indoor conditions to improve safety and health outcomes (Vita et al., 2023). However, the perception layer faces various security threats, necessitating comprehensive countermeasures to protect against attacks and ensure the security of IoT devices (C. K. Wu, 2021). Innovations such as the directional sensitivity-based perception algorithm (DSPA) aim to improve energy maintenance and perception rates by optimizing the sensing direction and region weight, inspired by the human visual direction-sensitive system (Z. Yang et al., 2023; H. Wu, 2022). Additionally, advancements in image processing algorithms for the perception layer have significantly enhanced the accuracy of image search and registration, reducing network load and energy consumption (Z. Yang et al., 2023). Collectively, these research efforts underscore the critical role of the perception layer in the IoT architecture, highlighting its objectives to ensure secure, efficient, and intelligent sensing capabilities across a wide range of applications.

4.2 Network Layer

The network layer in the IoT architecture plays a crucial role as the backbone of IoT systems, facilitating the main communication channel between the application layer and various operational activities within the IoT ecosystem. This layer is responsible for establishing a suitable connection network among nodes, which can be either wire-connected or wireless, depending on the protocol defined by the system designer. It connects IoT devices to the broader internet, making networks vital components for integrating things into the global web (K. S. Mohamed, 2019). Within the setting of a three-layer IoT architecture, the network layer refers explicitly to wide area networks (WANs), encompassing security protocols and techniques based on the TCP/IP network model and mobile communication networks, including industry standards like IPSec, SSL/TLS, and authentication and key agreement (AKA) processes (C. K. Wu, 2021). The objectives of the network layer are multifaceted, aiming to ensure seamless, secure, and efficient communication across IoT devices. Security remains a paramount concern, with the network layer addressing challenges by implementing robust security protocols to safeguard data confidentiality, integrity, and authentication. This is critical as IoT devices often collect sensitive information, making them targets for adversaries (Naing et al., 2023). Moreover, the network layer is designed to handle the high volumes of data generated by IoT devices, employing data compression and packet optimization strategies to mitigate energy consumption and enhance system reliability (Nwadiugwu et al., 2023). Furthermore, the network layer supports the different applications of IoT, from smart cities to healthcare, by empowering the interoperability of a wide range of devices with varying protocols and architectures (AlAali et al., 2022; Mshvidobadze, 2022). It also plays a role in addressing the challenges of heterogeneous IoT (HetIoT) through data integration and processing across different layers of the IoT architecture (Polepaka et al., 2023), while ensuring the delivery of data to end-users through appropriate application layer protocols (Polepaka et al., 2023). Lastly, the network layer contributes to developing standardized security frameworks to address emerging threats and vulnerabilities, fostering the growth and adoption of IoT technologies (Lalit et al., 2022).

4.3 Processing Layer

The processing layer within the IoT architecture plays a significant role in managing and interpreting the vast amounts of data generated by connected devices. This layer, often synonymous with cloud computing, ensures data security, facilitates real-time analytics, and supports intelligent decision-making processes (Devadas & Subramanian, 2019). The essential objective of the processing layer is to supply a robust platform for data analysis, empowering the transformation of raw data into significant insights. This is crucial for optimizing operations, enhancing decision-making, and creating value across various IoT applications (Bixio et al., 2020). Recent research highlights integrating machine learning techniques within the processing layer to predict future data requirements and improve system performance through intelligent agents (Mazon-Olivo & Pan, 2022). This is complemented by developing adaptive microservices architectures that support real-time stream processing at IoT platforms' edge and core levels, ensuring flexibility and scalability (Bouaouad et al., 2020). The processing layer's architecture is designed to be comprehensive, incorporating multiple functionalities across different layers to address the diverse needs of IoT systems in a cloud environment (C. K. Wu, 2021). Security within the processing layer is paramount, addressing threats through access control techniques and the security mechanisms of virtual computing (Rago, 2020). Moreover, the application of artificial intelligence (AI) within this layer, specifically through the deployment of lightweight neural network models, enhances the intelligence of edge devices, reducing reliance on centralized servers for data processing (Kabilamani & Gomathy, 2021). The processing layer moreover benefits from progressions in communication technologies, such as Narrowband IoT (NB-IoT), which improve data transmission efficiency and reduce latency, further enhancing the IoT ecosystem's capabilities (Polepaka et al., 2023). Ultimately, the processing layer's objectives are to ensure efficient data management, security, and intelligent processing, supporting the vast array of IoT applications ranging from smart homes to industrial IoT (IIoT) and beyond (Polepaka et al., 2023; Guan & Liang, 2023).

4.4 Application Layer

The application layer in the IoT architecture is pivotal for enabling communication between devices, objects, or things and application interfaces, ensuring the seamless transfer of information within the IoT framework. This layer is designed to meet critical requirements such as security, speed, low latency, less power consumption, reliability, and efficient information transfer, which are essential for the diverse and growing range of IoT applications spanning from industrial processes to smart environments and healthcare systems (P. Gupta & M, 2021; Jienan et al., 2021). The application layer protocols are chosen based on the particular needs and nature of the IoT system, even though there's no standard rule for selecting a specific protocol, making it significant to understand their strengths and weaknesses (Zhang et al., 2022). The objectives of the application layer are multifaceted, aiming to provide a robust framework for real-time communication technologies essential for developing IoT applications, ensuring interoperability among a vast number of devices with varying configurations and operational limitations (Lalit et al., 2022; Fagroud et al., 2023). It seeks to facilitate intelligent services by deploying computing, storage, and other capabilities at the network's edge, near the data source, in this manner tending to the requests for transmission bandwidth, real-time processing, and security in IoT applications (Win et al., 2023). Moreover, the application layer aims to offer a uniform architecture that integrates various components and technologies, supporting the successful development of IoT applications by providing a transparent interaction model among IoT infrastructure components (Sarvaiya, 2022). This layer's protocols also play a crucial role in establishing trusted connections between objects and things, highlighting the importance of selecting the appropriate protocol based on application needs, architecture, and security requirements (Polepaka et al., 2023).

4.5 Fog Layer

The fog layer in the IoT architecture serves as an intermediary computing infrastructure that bridges the gap between end devices and centralized cloud services. Its primary objectives include reducing latency, managing the vast data generated by IoT devices, and enhancing security and privacy measures (Benomar et al., 2021). By processing data closer to its source, the fog layer significantly decreases the response time for IoT applications, facilitating real-time data processing and decision-making (Fazel et al., 2023). One of the key advantages of incorporating the fog layer into IoT architecture is its ability to provide localized data analysis and management, which is crucial for applications requiring low and predictable delay (Rehman et al., 2023). This is particularly important in industrial IoT deployments, where the fog layer hosts data aggregation algorithms to improve network performance metrics such as latency and packet delivery ratio (Javanmardi et al., 2023). Moreover, the fog layer's proximity to IoT devices facilitates more efficient resource utilization. It reduces the bandwidth demand on the network backbone, addressing the challenges of scalability and dynamism in traditional networks (Benomar et al., 2021). Security is another critical objective of the fog layer. Implementing security mechanisms at different levels, including the IoT, fog, and cloud layers, ensures a more secure data processing environment (H. Gupta & Bharti, 2023). The fog layer can also support lightweight and energy-efficient ciphering algorithms for resource-constrained devices, enhancing the overall security posture without significantly impacting device performance (Apat et al., 2023). Furthermore, the fog layer facilitates optimal resource allocation and application placement, addressing the computational limitations of fog devices and ensuring uninterrupted services to end-users (Tselikis, 2023). It also supports the development of intelligent localized applications with strict latency requirements by offering a secure and performance-optimized computing environment (Al-Rubaie et al., 2023). In summary, the fog layer in IoT architecture aims to reduce latency, manage data efficiently, and enhance security, thereby supporting the demands of various IoT applications and improving the quality of service for end-users (Nejad et al., 2023).

4.6 Analysis Engine Layer

The analysis engine layer in IoT architecture serves the purpose of processing and deriving insights from the vast amount of data collected by IoT devices (Muntjir et al., 2017). It integrates analytical mechanisms like predictive analysis and machine learning algorithms to make sense of Big Data gathered from various sources (Chandnani & Khairnar, 2022). This layer plays a significant role in empowering effective data analytics, especially in IIoT applications within the manufacturing industry (Reddy & Sujith, 2018). By leveraging the analysis engine layer, IoT systems can extract valuable information, optimize processes, and enhance decision-making based on the data generated by interconnected devices (Jog & Murugan, 2018). The primary objective of this layer is to facilitate seamless information processing by merging the physical and digital worlds, thereby enabling the ubiquitous use of IoT devices (Zhang et al., 2022).

5 IoT Applications

IoT represents a significant shift in how we interact with the digital and physical worlds, offering various applications across various sectors. At its core, IoT connects devices to the internet and each other, enabling them to collect, share, and analyze data to improve efficiency, enhance safety, and create new services (Ahmed et al., 2022). One of the primary applications of IoT is in smart cities, where it contributes to more efficient and intelligent urban management. This includes everything from traffic control and waste management to energy conservation and public safety, leveraging IoT to create more sustainable and livable environments for their inhabitants (Lampropoulos et al., 2018; Singha, 2018). Similarly, the transportation and logistics sector benefits from IoT through enhanced tracking of goods and optimizing routes, significantly reducing costs and improving delivery times (Vedak & Kapadi, 2023). IoT applications are revolutionizing patient care through remote monitoring and telehealth services in the healthcare sector. By equipping patients with wearable devices that monitor vital signs and other health indicators, healthcare providers can offer personalized care and early intervention for potential health issues, thereby improving patient outcomes (Perwej et al., 2019; Sailaja & Rohitha, 2018). IoT also plays a crucial role in the industrial sector, particularly in the context of Industry 4.0. It enables smart manufacturing processes that can predict maintenance needs, optimize production lines, and ensure the safety of workers through the use of sensors and automation technologies (Gaber et al., 2018; Fawaz, 2022). This not only increases productivity but also reduces operational costs. Moreover, IoT applications extend to the home, where smart home technologies offer homeowners convenience, security, and energy efficiency. From smart thermostats and lighting systems to security cameras and voice-activated assistants, IoT devices are making homes more comfortable and safer (Babu et al., 2017; Shwetha & Suma, 2018). In environmental monitoring, IoT devices are critical in tracking air quality, water levels, and other environmental parameters, providing valuable data that can help conservation efforts and respond to environmental emergencies. However, the widespread adoption of IoT also brings challenges, particularly regarding security and privacy. As more devices connect, the risk of data breaches and other cyber threats increases, necessitating robust security measures to protect sensitive information. In conclusion, the applications of IoT are vast and varied, touching nearly every aspect of modern life. From making cities smarter and improving healthcare to revolutionizing industries and enhancing home life, IoT promises a more connected, efficient, and intelligent world. Yet, as we embrace these advancements, it is crucial to address the accompanying challenges to realize the full potential of IoT.

5.1 IoT in Smart Cities

The IoT is instrumental in developing smart cities by facilitating sophisticated control over infrastructures and services (Whaiduzzaman et al., 2022). IoT enables the connection and collaboration of various devices across distributed environments to provide information and services to urban entities (V. Kumar et al., 2023). When considering incorporating IoT technology within the framework of smart cities, various obstacles and complexities need to be addressed and overcome. These challenges encompass a wide range of issues, including but not limited to the acquisition of data, ensuring robust security measures, the ability to scale effectively as the system grows, and the necessity for different devices and systems to work together seamlessly in a harmonious and interconnected manner (Ali et al., 2023; Haghshenas et al., 2022). These challenges are addressed through middleware solutions like Generic Middleware for Smart City Applications (GMSCA), which integrates IoT and big data applications to support sustainable urban solutions (Zahoor & Mir, 2022). Additionally, it is important to note that integrating and implementing IoT technology within the context of smart cities encompasses the intricate task of effectively overseeing and coordinating urban monitoring systems, which are structured with a three-tier architecture consisting of Endpoint, Edge, and Cloud components. This approach involves managing data and information flow across these layers, ensuring seamless communication and data exchange among devices and systems. Furthermore, it is essential to prioritize mitigating security and privacy risks during this process, including implementing strong data privacy controls and strict network security measures to protect sensitive information and prevent unauthorized access or breaches. Establishing a comprehensive framework that addresses these concerns thoroughly is essential, considering the various potential threats and vulnerabilities that could impact the integrity and confidentiality of data in the IoT ecosystem within smart city environments. Overcoming these challenges is crucial for successfully developing and implementing IoT applications in smart cities.

5.2 IoT in Healthcare Sector

The IoT holds a significant and indispensable position within the healthcare industry, serving a pivotal function in facilitating the monitoring of patients from a distance, offering real-time identification of medical conditions, and enhancing the overall standard and efficiency of healthcare services (Upadhyay et al., 2023). IoT technology enables healthcare providers to effectively monitor patients, particularly in challenging circumstances such as the global COVID-19 pandemic, where remote and real-time patient monitoring is paramount. By leveraging IoT devices and sensors, healthcare professionals can continuously track vital signs and symptoms, enabling early detection of any deterioration in patient health and allowing for timely interventions to be made. This proactive approach helps reduce the risk of virus transmission within healthcare settings. It significantly enhances the overall quality of patient care by providing personalized and data-driven treatment strategies (Al-Atawi et al., 2022). However, IoT-based healthcare systems encounter various obstacles that must be addressed to operate effectively. These obstacles include but are not limited to concerns regarding data security, which involves protecting sensitive information from unauthorized access or breaches. Additionally, privacy issues are a significant challenge as they involve ensuring that personal health data is kept confidential and only accessed by authorized individuals. Interoperability is another hurdle IoT healthcare systems face, referring to the ability of different devices and systems to communicate and exchange data seamlessly. Scalability is also a notable challenge, as healthcare systems must be able to adapt and expand to accommodate growing amounts of data and increasing demands. Furthermore, maintaining data integrity is crucial to ensure IoT devices collect and store accurate, consistent, and reliable information for making informed healthcare decisions (Hegde & Maddikunta, 2023). We can effectively tackle these challenges by integrating Blockchain technology with IoT in healthcare applications. This convergence can potentially significantly improve security measures within the healthcare industry. Also, it offers a promising solution to address the major issues hindering progress in this field (Chakraborty & Kishor, 2022).

5.3 IoT in Industrial Sector

The IoT is being utilized to a great extent within the industrial sector, bringing about a significant transformation in the field of manufacturing through the interconnectedness of various products, processes, and infrastructure. This interconnected network of devices and systems is reshaping how products are produced, monitored, and managed, leading to increased efficiency, automation, and data-driven decision-making in the manufacturing industry (Dallaev et al., 2023). IoT enables the collection and sharing of data through smart devices, which is essential for modern manufacturing systems. It has become integral to manufacturing automation, with PLCs and SCADA facilitating large-scale connectivity within manufacturing plants (Khattar et al., 2023). However, the integration of IoT in various industries is met with numerous obstacles, including but not limited to issues concerning security, privacy, scalability, and efficiency. These challenges arise due to the swift growth and initial phases of development experienced in this field (Khattar et al., 2023; Ullah et al., 2023). Organizations integrating IoT into logistics and supply chain operations also encounter obstacles like lack of comprehensive knowledge and inherent challenges, hindering widespread adoption (Farooq et al., 2022). Combining AI with the IoT is widely regarded as a promising and innovative approach that has the potential to effectively tackle and overcome the various challenges that arise in this context.

5.4 IoT in Smart Homes

IoT applications implemented in smart homes provide the convenience and flexibility of remotely monitoring and controlling various household devices from a distance. This technology enables users to access and manage their appliances, lighting, security systems, and other devices through interconnected networks and smart devices. It offers a seamless and integrated approach to home automation (Muniz & Muniz, 2023; Sissodia et al., 2023). This integration aims to improve various aspects such as automation, convenience, and energy efficiency by leveraging the process of gathering data from a wide range of sensors, including those designed to measure temperature and those that detect motion (Guenfaf & Zafoune, 2023). Challenges related to the field of IoT persist despite advancements, encompassing issues such as ensuring the security of data transmitted and stored by IoT devices, safeguarding the privacy of users' information, as well as addressing the complexities surrounding the automated decision-making processes implemented within these devices (Umer et al., 2023; Vasilescu et al., 2023). Proposed solutions include deep learning-driven systems for automated decision-making and blockchain technology for authentication and identification of devices. Additionally, there is a growing emphasis on ensuring the security of IoT devices in smart homes to prevent vulnerabilities and external attacks. Overall, while IoT applications in smart homes offer numerous benefits, managing security and privacy concerns remains crucial for successful implementation.

5.5 IoT in Environmental Monitoring

The IoT has significantly advanced environmental monitoring by enabling real-time, accurate, and comprehensive data collection and analysis across various ecological parameters (Zhihan, 2022; Zhao et al., 2022). IoT applications in environmental monitoring span air quality assessment, water pollution detection, and radiation pollution monitoring, leveraging modern sensors and IoT devices to facilitate Smart Environment Monitoring (SEM) systems (L. Chen & Zhang, 2022; Ramachandran et al., 2023). These systems utilize wireless sensor networks (WSN) and advanced data analytics, including artificial intelligence (AI), to predict environmental changes and manage pollution effectively (Kaur, 2023; Hulwan et al., 2023). One innovative application of IoT in environmental monitoring is the Multi-Protocol Fusion IoT (MPFIoT), designed for community environmental monitoring. This system integrates various wireless communication protocols to collect and transmit environmental data efficiently, addressing the challenge of communication protocol heterogeneity (Rajbhar, 2022). Similarly, IoT technology's integration with GPS and GIS systems enhances the scope and accuracy of ecological environment monitoring, significantly improving the scientificity of monitoring results (Y. Yang, 2022). However, deploying IoT in environmental monitoring is not without challenges. The heterogeneity of IoT devices and communication protocols can complicate the integration and interoperability of monitoring systems (Anasica, 2022). Security concerns also arise as the extensive deployment of sensor networks increases the risk of data interception and unauthorized access (Lou et al., 2022). Additionally, the complexity and cost of conventional monitoring stations pose significant barriers to widespread adoption, although IoT-based systems offer more portable and cost-effective solutions. Secure IoT-based platforms have been proposed to address these challenges, employing advanced encryption algorithms to protect collected data. Furthermore, the development of IoT technologies, cloud computing, and smart city applications contribute to overcoming obstacles in environmental monitoring, paving the way for more effective management of ecological health. In summary, while IoT applications in environmental monitoring offer promising solutions for real-time and accurate ecological assessments, addressing device heterogeneity, security, and system cost challenges is crucial for successful implementation.

6 Conclusion

the chapter delves deeper into the IoT domain, elucidating its complex architecture by focusing on the seamless collaboration of the components of the IoT ecosystem to facilitate the collection, transmission, and processing of data from interconnected devices. Furthermore, it takes an in-depth look at the various applications of IoT in different sectors, such as smart cities, healthcare, and industry, highlighting the immense potential of IoT technologies to bring about a paradigm shift in these areas through improved operational efficiency, security standards, and service delivery mechanisms.

As we move forward, IoT technologies' continued evolution and integration into various aspects of our daily lives will undoubtedly drive significant advancements. By understanding IoT key components, architecture, and applications, stakeholders can better navigate this rapidly evolving landscape and capitalize on the opportunities it presents.

Chapter III: Related Works

1 Introduction

The purpose of this chapter is to review and analyze the existing literature relevant to the current study, focusing on two key areas: the educational field and the environmental perception and monitoring field. By examining related works, this chapter aims to identify the methodologies, findings, and gaps that have informed current understanding and practices within these domains. This analysis will provide a foundation for positioning the present thesis within the broader academic discourse and highlighting its unique contributions.

In the educational field, the review will focus on two primary subtopics: student performance prediction and academic analysis in higher education. These areas have garnered significant attention due to their potential to enhance educational outcomes and inform policy decisions. The discussion will explore various predictive models, analytical techniques, and their implications for improving student success and institutional effectiveness.

The second part of this chapter will address environmental management, with a specific focus on forest fire detection systems (FFDS). This review will examine key studies that investigate various technologies and methodologies employed for early detection and management of forest fires. It will cover theoretical frameworks, technological advancements, and practical applications that contribute to more effective forest fire prevention and control strategies.

2 Higher education field

2.1 Student Performance Prediction

In the current literature, numerous complementary approaches are available that serve as a fundamental basis for the analysis of predicting student performance. These approaches encompass a wide range of methodologies and techniques aimed at forecasting how well students will perform in their academic endeavors. The utilization of machine learning algorithms to anticipate student performance stands out as a popular subject of exploration in various academic studies. A significant portion of these studies delves into the identification of the most effective predictive models for this purpose. Conversely, other research endeavors are centered around assessing the feasibility and efficacy of employing machine learning in the prediction of student performance outcomes. Moreover, recent studies have also started to investigate the impact of incorporating non-cognitive factors, such as socio-economic background and emotional intelligence, into predictive models to enhance the accuracy of student performance forecasts.

A research investigation was carried out at the Eindhoven University of Technology to assess the efficacy of machine learning in forecasting school dropout (Dekker et al., 2009). The fundamental approach involves constructing multiple prediction models utilizing various machine learning techniques, including classification and regression trees (CART), BayesNet, and Logit. Subsequently, the analysts juxtaposed the predictive outcomes of all the developed models based on their effectiveness.

Researchers from three distinct universities in India carried out a comparable investigation (Yadav et al., 2012). They examined a dataset comprising university students through the utilization of diverse algorithms. Subsequently, an evaluation was performed on the precision and recall metrics of the forecasts. The researchers deduced that the model known as Alternating Decision Tree (ADT) exhibited the highest level of accuracy.

A further investigation was carried out at the University of Jordan (Amrieh et al., 2016) regarding performance forecasting. Alongside the utilization of machine learning techniques, ensemble methods were also implemented by the scholars. Upon evaluation of the utilized approaches, it was determined that Decision Trees yielded the most optimal outcomes. The scholars additionally scrutinized the behavioral characteristics of the students. A model was developed both incorporating and excluding these characteristics. Substantiation was found indicating that the integration of behavioral features enhanced the accuracy of predictions.

2.2 Academic Analysis in Higher Education

The emergence of big data analytics has significantly transformed various industries, including higher education. Universities are now faced with an unprecedented volume, velocity, and variety of data, which presents both opportunities and challenges in optimizing their educational supply chain (ESC) decision-making processes.

The application of big data analytics in higher education can potentially yield a myriad of benefits, such as enhanced student retention, improved resource allocation, and more effective curriculum development (Attaran et al., 2018). Leveraging the wealth of data generated across campus can provide university administrators with a deeper understanding of student behaviors, preferences, and performance, enabling them to make more informed strategic decisions that enhance operational efficiency and deliver a superior educational experience.

The process of Supply Chain Management (SCM) has its roots in the logistics concept and has been integrated into various business sectors for efficient operations. Research and analysis in SCM primarily focus on the manufacturing domain, with limited attention given to the service industry. The application of SCM design in the educational field, categorized as a service industry, has been poorly explored. However, recent studies have started to recognize the potential benefits of implementing SCM principles in educational institutions to enhance resource allocation and improve overall operational effectiveness.

Both the manufacturing and services sectors share a common goal, which is to deliver high-quality outputs throughout the supply chain process, ultimately resulting in finished products that align with the demands of the market and society. In order to realize these objectives, (M. Habib & Jungthirapanich, 2009) introduced a simplified model of SCM tailored for tertiary education. Within this framework, the primary services offered by universities are categorized as Education and Research, with students and research projects serving as the raw materials. The end results manifest as graduates and research findings. Despite its simplified nature and abstract capabilities, the proposed model retains a level of simplicity that overlooks various crucial aspects of the educational framework, including the organizational structure of universities, decision-making processes, information flow among members of the ESC, and their collaborative efforts to attain the ESC's primary goals.

The model proposed by (M. Habib & Jungthirapanich, 2009) has been studied and developed into an Integrated Tertiary Educational Supply Chain Management (ITESCM) model (M. M. Habib & Jungthirapnaich, 2010). This model describes the integrated form of the ESC and educational management for universities that also consists of the ESC and research supply chain. The authors in (Pathik et al., 2012) conducted a case study to evaluate the ITESCM model performance in a private university in Bangladesh. A survey of all stakeholders, including employers, graduates, current students, university administration, and faculty members, was conducted to review the quality outcomes for the final consumer (society). The results of this study show a significant satisfaction of the final consumer toward the outputs of the ITESCM model implementation.

The development of the ESC process and mechanisms is still active through recent studies such as (Saa'da et al., 2022; Waham & Lestari, 2019; Ramzi, 2018; Savignac et al., 2022) with a variety of directions. A few researchers have addressed the adaptation of the ESC in specific higher education systems like Tunisia (Ramzi, 2018) and Jordan (Saa'da et al., 2022). However, the involvement of data analytics in the ESC process as a critical factor in decision-making remains minimal. This thesis studies the adaptation of the ESC model proposed by (M. Habib & Jungthirapanich, 2009) to the most common structure of higher education institutions and investigates the potential of data analysis in the ESC decision-making process.

3 Environmental Perception and Monitoring Field

By exploring the research works related to environmental management and specifically to the fire detection field, we notice two categories of FFDS. The first one is based on data collected from wireless sensor network (WSN) composed of environmental perception sensors like temperature, humidity, wind, and other similar sensor types. In the second category, researchers have leaned towards using unmanned aerial vehicles (UAVs) equipped with cameras as a means of data collection. Image processing techniques were applied to analyze the collected data and retrieve significant information on the state of the forest.

The authors in (Varela et al., 2020) propose a FFDS using WSN for environment perception combined with information fusion techniques. They introduced a low complexity algorithm to detect fire based on only two environmental parameters (Temperature and humidity). Using regression analysis, they created what is called "base function". This function is used to define a base model that is employed as a comparative model when the system detects temperature and humidity that potentially represent a fire. The proposed system showed a 100% fire detection rate when sensor nodes are not exposed to sun's rays. Authors of this FFDS pointed an important issue in the proposed system, which is an incorrect fire detection information when nodes are exposed to sun's rays. This kind of issue can be verified using other type of sensors like flame sensor or by employing drones equipped with cameras to check the real existence of fire. Moreover, the proposed FFDS allows only the fire detection, managing the fire after its outbreak is as important as the fire detection. Employing sensors such as the wind sensor helps to determine the fire propagation direction.

The employment of UAVs in intelligent environmental perception applications is becoming more and more pervasive. Applying such technology in the forest fire fighting field has been widely studied in several research works. UAVs are relatively inexpensive and can cover a large area in different weather conditions with minimal human implication (Kalatzis et al., 2018). Equipping UAVs with cameras for image collection as well as data communication facilities allows good visualization of the area under surveillance. In (Kalatzis et al., 2018), the authors proposed a hierarchical architecture for early forest fire detection by exploiting the benefits of UAVs' sensing abilities and the rich and powerful resources of cloud and fog computing technologies. Several scenarios have been evaluated to demonstrate the usefulness of fog-computing involvement. First, an edge-only scenario is evaluated. In this scenario, the image processing classification tasks are performed at the edge nodes. The results show that edge nodes have difficulty executing such computationally intensive tasks. Moreover, the energy consumption of UAVs is twice that of other scenarios. In another scenario, the authors evaluated the data processing at the cloud level only. All data collected by the UAVs is transmitted to the cloud for data processing purposes, which resulted in a very high data flow between the data source and the cloud. However, this scenario achieves the lowest response time regarding the other scenarios. The last scenario involves the deployment of fog computing as an intermediate layer between the data source and the cloud layer. The fog nodes are deployed close to the data source, which reduces network traffic. Besides, they are in charge of the data processing phase, thus reducing the volume of data transmitted to the cloud and only sending the processing results in case of an unusual situation. The test results show that the last scenario of the proposed FFDS is the most balanced among the three proposed scenarios.

By examining UAV-based solutions, we deduce some relevant challenges that need to be addressed. One of the most critical issues is that the deployment of UAVs in a public environment must be well supervised to avoid privacy-related incidents. In addition, the UAVs' limited battery is a real limitation that must be considered for a more attractive solution.

During the last decade, several researchers studied the implementation of fog computing in time sensitive IoT systems. Fog computing brings the autonomy, computational power, efficiency, and rich resources of cloud computing close to the data sources, which makes it very effective in IoT systems. In (Awaisi et al., 2019), researchers study the efficiency of fog computing implementation compared to cloud computing in a smart car parking system. They proposed a three layer model, where fog computing is applied at the intermediate level between the object layer consisting of cameras and smart LED displays and the cloud layer. The fog nodes are responsible for acquiring and processing the parking slots' images from the cameras and transmitting the analysis results to the LEDs. To evaluate the latency and bandwidth utilization of the proposed system in several scenarios, the authors used the iFogSim simulator. The experimental results show that the scenario based on the fog computing implementation results in low latency and low network usage compared to the cloud computing scenario.

4 Conclusion

This chapter reviewed key studies in the educational field, focusing on student performance prediction and academic analysis in higher education, as well as environmental management, specifically forest fire detection systems. The review highlighted significant methodologies, findings, and gaps in both areas. By identifying these gaps, this chapter positions the present study to address these issues and contribute new insights to the fields of education and environmental management.

PART TWO: SCIENTIFIC CONTRIBUTIONS

Chapter IV: Harnessing Big Data Analytics to Enhance Student Performance in Higher Education

Chapter V: Enhancing University Decision-Making through Big Data Analytics

Chapter VI: An IoT-Fog Based Architecture for Detecting and Managing Forest Fires

Chapter IV: Harnessing Big Data Analytics to Enhance Student Performance in Higher Education

1 Introduction

Data has become so important in practically every area of our life. The emerging amount of data generation peripherals such as computers, smartphones, smart home gadgets, connected vehicles, sensors, and many other devices have introduced the IoT era. Managing such valuable data requires continuously developing new and improved data collection, storage, visualization, and analytics tools and methods.

Extracting valuable information from the massive amount of available data is necessary in all sectors. Some fields, such as marketing, business, and industry, are at the forefront of data mining. While other sectors, such as higher education, are lagging. (Ang et al., 2020) uses the term Big educational data to describe the captured volume, the variety, and the rapid growth of higher education-generated data. The higher education data sources are divided into two types (Vaitsis et al., 2016). The first one is administrative decision-making and learning pattern improvement-related data sources. The other kind are those related to student performance, such as students' interaction with online learning systems, in-classroom student-teacher collaboration, test results, and also the newly available data sources such as social media, personal mobile devices gathered data, campus sensor data, etc.

This chapter focuses on the data analysis process in higher education. The main objective is to demonstrate the benefits of data analytics in predicting students' performance using a machine-learning algorithm and providing recommendations to improve students' results and avoid failure. We present a case study where we have exploited a dataset by applying three analysis phases (Descriptive, predictive, and prescriptive analytics (Morr & Ali-Hassan, 2019)). The dataset used in the current case study cannot be classified as a big dataset. It is a well-structured dataset with a small volume. Processing big educational data sources will significantly and efficiently impact the higher education process and results. However, the lack of such open data sources was a limitation of the current study.

We start our case study by introducing the selected dataset overview. We provide all dataset information beginning with 1) the dataset source, 2) the data collection mechanism, 3) and the dataset's significant features and attributes. After that, we apply descriptive analytics to the dataset. It allows the data to be examined and the identification of relationships among dataset attributes. The next step is predicting student results regarding previously collected data. We start by preprocessing data to eliminate inconsistent or false entries and transform the values into a more suitable format for analysis. We then apply two supervised machine-learning algorithms (Decision Trees and Random Forest) to build a prediction model. By comparing the two algorithms' training results, we concluded that the random forest algorithm gives more consistent and reasonable results than the decision tree algorithm. Finally, regarding the results of the two previous analysis phases, we make some recommendations for higher education institutions and students to improve students' performance.

This project is the first in a series of big educational data analytics studies. Its significant contribution is the application of the three phases of big data analytics (descriptive, predictive, and prescriptive analytics) to a dataset, gathering several characteristics and behaviors of students during their educational process. Moreover, the results of this study demonstrate the potential and benefits of big data analytics for academic institutions.

2 The case study

2.1 Dataset overview

The dataset used in this case study is an educational dataset collected from the Kalboard 360 learning management system (LMS) (Amrieh et al., 2016, 2015). The data was collected using an experienced API (xAPI) learner activity-tracking tool. The xAPI provides the ability to track student-learning actions in the LMS. It records student actions, such as reading an article or watching a training video.

The dataset consists of student learning actions in the Kalboard 360 LMS. It is collected over two semesters of instruction and includes 305 males and 175 females. Figure 4.1 shows the different nationalities of the students in the collected data. In addition to student information and learning activities, the dataset includes an interesting attribute: parent feedback. The feedback from the parents consists of two forms: whether they respond to the learning institution survey or not and whether they are satisfied with the learning institute. Table 4.1 lists and details the dataset's main attributes.

2.2 Descriptive analytics

Descriptive analysis is the foremost approach to understanding existing data and models. It allows us to examine the data, analyze relationships among variables, and identify the relevant characteristics in predicting student performance.

We start by visualizing the dataset attributes in graphical form, trying to identify the

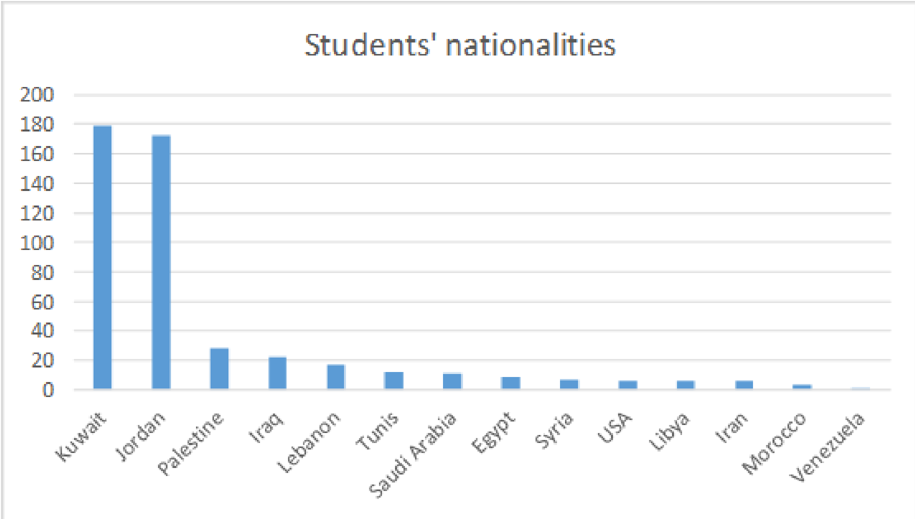

Figure 4.1: The nationalities of the students in the collected data

Attribute	**Description**	**Values**
Gender	student's gender	Male or Female
Nationality	student's nationality	See Figure 4.1
Grade Levels	grade student belongs	From G-01 to G-12
Section ID	the student's classroom	A,B,C
Topic	Course topic	English, Spanish, French, Arabic, IT, Math, Chemistry, Biology, Science, History, Quran, Geology
Semester	school year semester	First, Second
Parent responsible for student		mom, father
Raised hand	how many times the student raises his/her hand on classroom	0-100
Visited resources	how many times the student visits a course content	0-100
Viewing announcements	how many times the student checks the new announcements	0-100
Discussion groups	how many times the student participate on discussion groups	0-100
Parent Answering Survey	parent answered the surveys which are provided by institution or not	Yes, No
Parent School Satisfaction	the degree of parent satisfaction from institution	Good, Bad
Student Absence Days	the number of absence days for each student	above-7, under-7

Table 4.1: The Dataset Main Attributes

essential features for predicting student performance. According to dataset providers (Amrieh et al., 2016; Amrieh et al., 2015), the dataset can be divided into three nominal intervals based on students' total grades/marks.

- Low-level: values between 0 and 69
- Medium-level: values between 70 and 89
- High-level: values between 90 and 100

Figure 4.2 shows the students' classification based on their level. 24.46% of students are classified in low-level, 43.96% in medium-level, and 29.58% in high-level.

Based on these intervals, we will try to identify the features that affect the

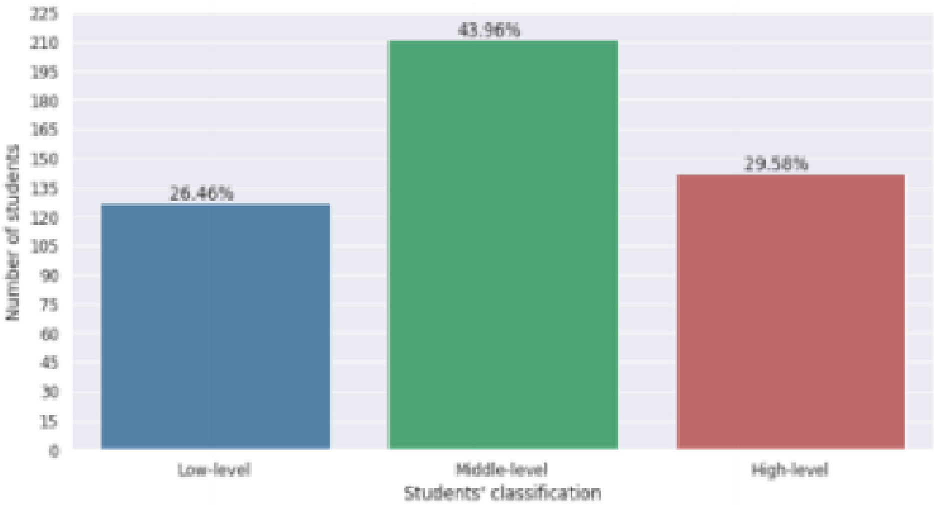

Figure 4.2: Students' classification based on their grade/marks

classification of students by visualizing students' learning records. For this purpose, we cross-reference each dataset attribute with students' level intervals.

1. *Gender:* Figure 4.3 illustrate the students' classification based on their gender. Male students have a lower level than females. Approximatively, the same observation was at the middle level, but the good outcomes (high level) were convergent between male and female students, with a slight superiority for females.

2. *Students absence days:* Figure 4.4 shows the impact of the number of days absent (under/above seven days) on the student's level. We can notice that the number of days of absence directly affects the student's level.

3. *Tracked student learning actions:* From the results in Figure 4.5, there do not

appear to be any specific relationships or patterns in the numerical data. However, we can visualize a high correlation between these characteristics and student outcomes. Combining students' learning actions with students' classification shows that students with high grades are more likely to raise their hands, check for new announcements, view their course content, and participate in discussion groups than students with low or average grades.

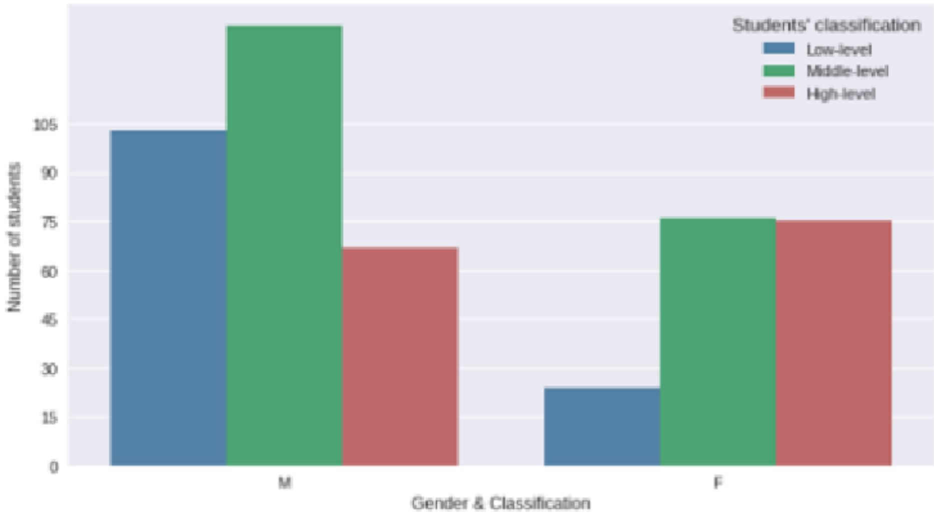

Figure 4.3: Students' gender combined with their classification

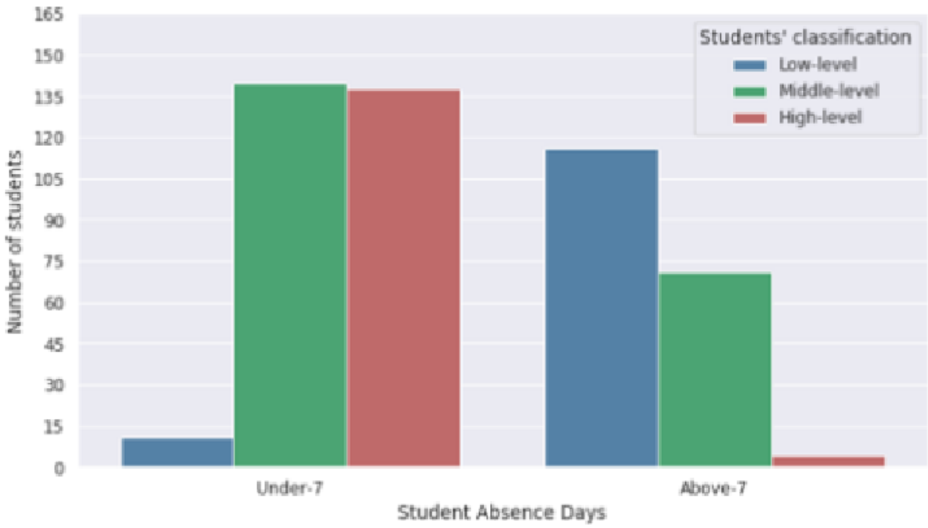

Figure 4.4: Students' absence days combined with their classification

The descriptive phases allowed us to understand the impact of the dataset on students' academic performance and the correlation of its results, which facilitates the application of the predictive phases in the next section.

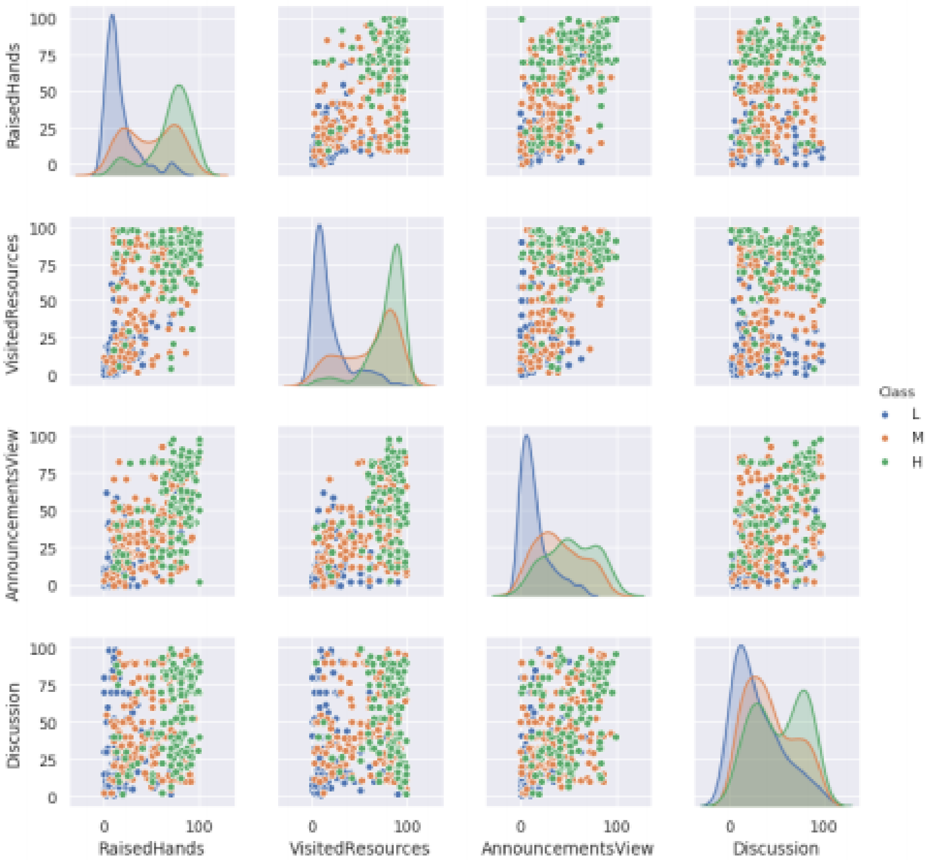

Figure 4.5: Students' absence days combined with their classification

2.3 Predictive analytics

As the name suggests, predictive analytics processes data to predict future events. Predictive analytics uses statistics, data mining, machine learning, and artificial intelligence to build a predictive model around previously collected data.

2.3.1 Preprocessing and feature selection

In the previous section (Descriptive Analysis), we notice that the features in our dataset have different correlations with the student's performance. Now, we will select only those features with a high correlation to the student's performance. To do this, we start by preprocessing the data in our dataset. First, we notice that the data in our dataset is in two different formats (numeric and string). To facilitate the application of the machine learning algorithms, we have transformed the values in string format into numerical values. For example: Gender = ('M' = 1, 'F' = 2), GradeID = ('G-04' = 4, 'G-12' = 12), and so on (see Figure 4.6).

After preprocessing the data and changing all values to numerical values, we proceed

	Gender	Nationality	PlaceofBirth	StageID	GradeID	SectionID	Topic	Semester	Relation	RaisedHands	VisitedResources
0	1	12	12	2	4	1	6	2	1	15	16
1	1	12	12	2	4	1	6	2	1	20	20
2	1	12	12	2	4	1	6	2	1	10	7
3	1	12	12	2	4	1	6	2	1	30	25
4	1	12	12	2	4	1	6	2	1	40	50
5	2	12	12	2	4	1	6	2	1	42	30
6	1	12	12	3	7	1	11	2	1	35	12
7	1	12	12	3	7	1	11	2	1	50	10

AnnouncementsView	Discussion	ParentAnsweringSurvey	ParentschoolSatisfaction	StudentAbsenceDays	Class
2	20	1	1	0	2
3	25	1	1	0	2
0	30	0	0	1	1
5	35	0	0	1	1
12	50	0	0	1	2
13	70	1	0	1	2
0	17	0	0	1	1
15	22	1	1	0	2

Figure 4.6: Dataset after pre-processing

to feature selection. The main objective is to select only those features that strongly correlate with the student's performance.

Feature selection, attribute selection, or variable selection are three terms for one definition. Essentially, it consists of selecting the most relevant features in a dataset. The importance of feature selection is as follows (Sayak, 2020):

- It allows the machine learning algorithm to train faster.

- It reduces the complexity of the model and facilitates its interpretation.
- Overfitting reduction.

There are several general methods of feature selection, such as Filtering methods, Wrapping methods, and Embedded methods (Chandrashekar & Sahin, 2014). Here, we will use the Filter method.
We chose the Filter methods because they utilize variable ranking techniques as the principle criteria for variable selection by ordering. Ranking methods are simple and have success for practical applications (Chandrashekar & Sahin, 2014).
The filter method relies on evaluating the general uniqueness of the data and selecting a subset of features without including any mining algorithm. The filter method uses the exact evaluation criterion, which includes distance, information, dependency, and consistency (Sayak, 2020).
The filtering in the current case study is most commonly done using the Pearson correlation matrix. We plot the Pearson correlation heatmap and explore the independent variables' correlation with the output variable "Class" (students' outcome). Only features with a correlation above 0.1 (taking absolute value) with the output variable are selected. The correlation coefficient has values between -1 and 1.

- A 0 value implies no correlation.
- A value closer to 0 indicates a weaker correlation.
- A value closer to 1 indicates a stronger positive correlation.
- A value closer to -1 indicates a stronger negative correlation.

By inspecting the correlation of the multiple independent variables with the output variable Class, we notice that the features Visited Resources, Student Absence Days, Raised Hands, Announcement Views, Survey Answered, Relation, Parent Satisfaction, Discussion, Gender, and Semester are highly correlated with the output variable 'Class.' Hence, all other features are considered uncorrelated with the output variable and will be ignored. Figure 4.7 shows the Matrix correlation heatmap with only the most correlated features.

2.3.2 Building a Prediction Model:

After preprocessing the data and selecting the most relevant features, we can now construct a predictive model. For this purpose, there are many existing algorithms.

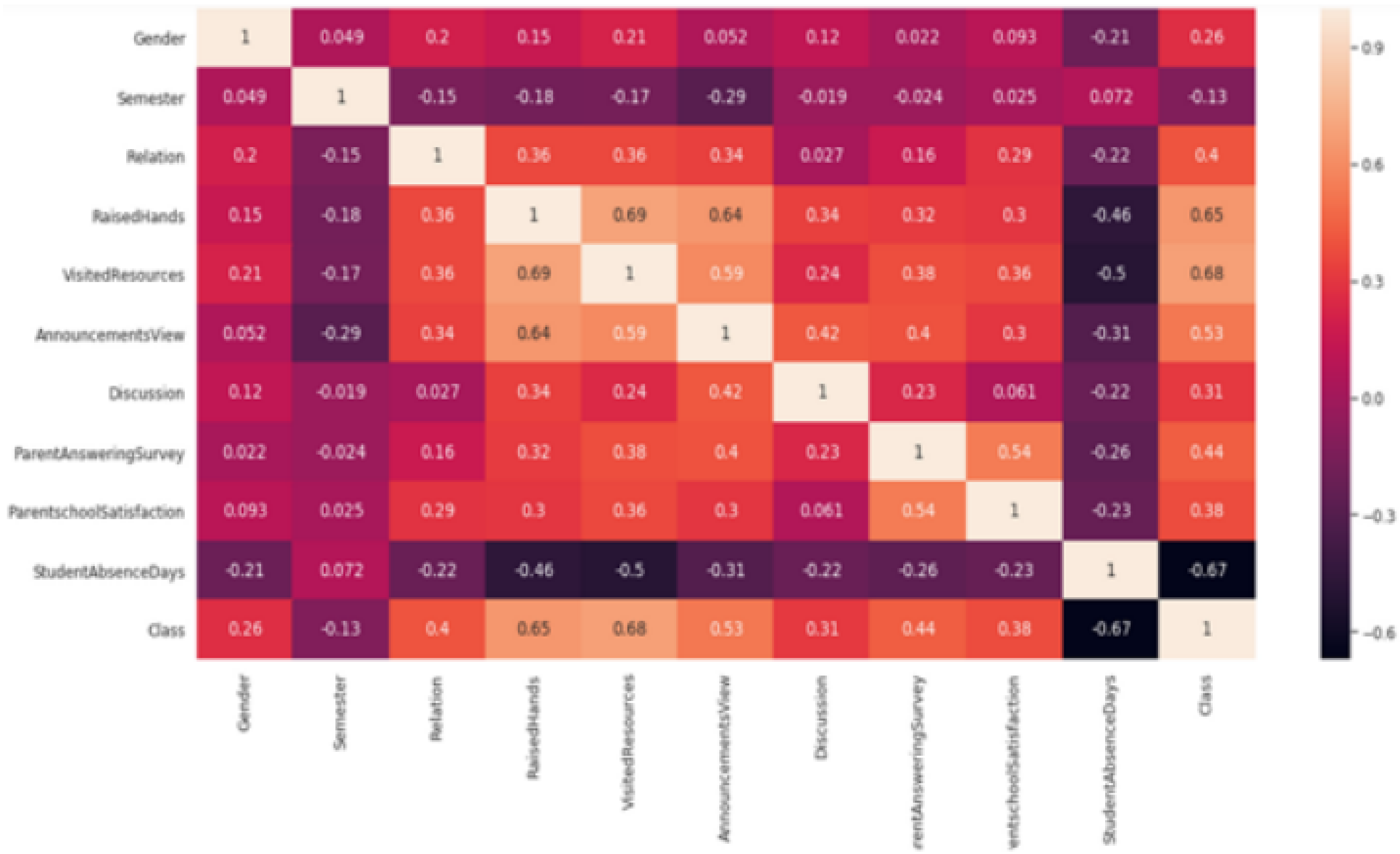

Figure 4.7: Matrix correlation (Pearson Correlation) after features selection

This study uses two different algorithms: Decision Trees and Random Forest. Although these two algorithms have the same objective, predicting a dependent variable from independent variables, they are based on different mathematical methods.

(a) **Decision Tree :** (F.-J. Yang, 2019) defines the decision tree algorithm as follows: "A decision tree is a tree-based method in which each path starting from the root is representing a sequence of data splitting until a Boolean outcome is reached at the leaf node".
After selecting the features best correlated with the output variable, the next step was to divide the data into training and test sets. The training set (70% of the data) was used to build the prediction model, and the test set (30% of the data) was used to test the model. We create the prediction model based on the training dataset using the decision tree method. We then apply the prediction model to the testing dataset and categorize the students according to their educational level. Table 4.2 shows the ratio of correct predictions (Accuracy), which was 72.22%.

	Training dataset	**Testing dataset**
Size	70% (336 students)	30% (144 students)
Accuracy	100%	72.22%

Table 4.2: The Decision tree model implementation results

(b) **Random Forest :** According to (Tang et al., 2018), "A random forest is an ensemble of trees, where the construction of each tree is random. After building an ensemble of trees, the random forest makes predictions by averaging the predictions of individual trees. Random forests often make accurate and robust predictions".
To create the prediction model based on the random forest algorithm, as we did with the decision tree method, we divided the dataset into a training dataset to train our model and a test dataset to test the model and extract the accuracy. Table 4.3 shows the results of implementing the random forest model. The result shows an accuracy of 77.08%.

	Training dataset	**Testing dataset**
Size	70% (336 students)	30% (144 students)
Accuracy	99.44%	77.08%

Table 4.3: The Random forest model implementation results

2.3.3 Testing Classification and Confusion Matrix Comparison

Comparing the two prediction model results for the test dataset, we can see that the results of the second model that relies on the random forest algorithm are better than the results of the first model based on the decision tree algorithm.
Figures 4.8 and 4.9 shows the confusion matrix for the first model based on decision tree and the second model based on random forest.
The confusion matrix of the first prediction model based on the decision tree algorithm

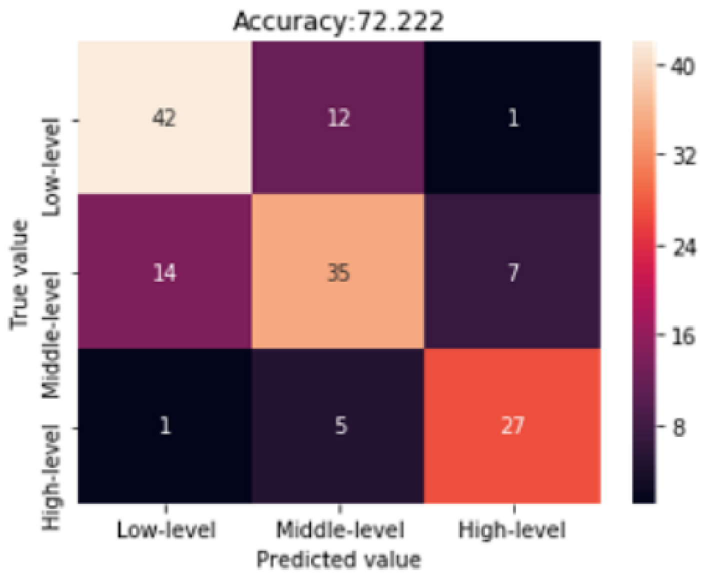

Figure 4.8: Confusion matrix for the decision tree model

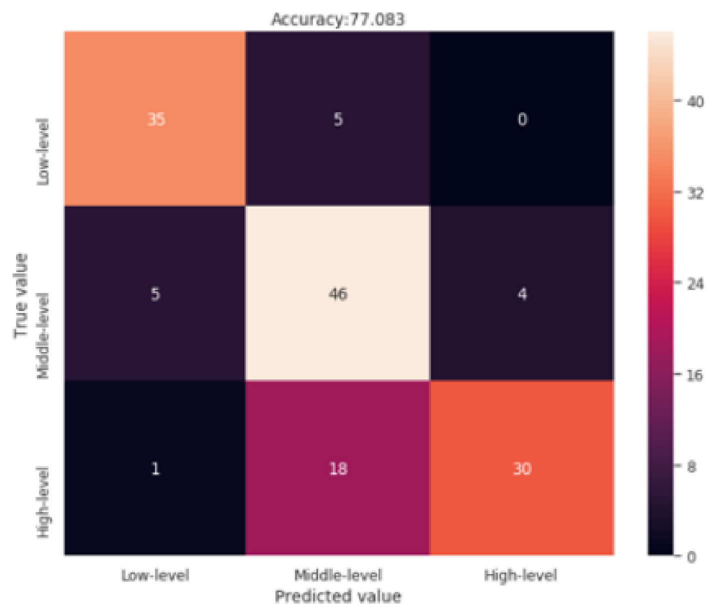

Figure 4.9: Confusion matrix for the random forest model

shows that the model performs well with the low-level class where only 13 students are misclassified out of 55. We can see that 21 out of 56 students are misclassified for the medium level, with 14 classified in the low level and 7 in the high level. Finally, for the high-level class, we can see that 27 out of 33 students are correctly classified in the right class, and 6 out of 33 students are misclassified.
The confusion matrix of the second model based on the random forest algorithm shows that the model works perfectly with the low-level class with only five misclassified students out of 40. We can see that 9 out of 55 students are misclassified for the medium level, 5 of which were classified in the low level and 4 in the high level. Finally, 30 out of 49 students are correctly classified in the correct class at the high level.
Finally, by comparing the confusion matrix of the two models, we conclude that the random forest algorithm applied to the dataset has better prediction results than the decision tree algorithm.

2.4 Prescriptive analytics

After starting with descriptive analytics by visualizing the data set used in the current case study, we built two models to predict future outcomes in predictive analytics. Finally, here we are with the third step of the student performance analysis, prescriptive analytics. The main objective of this phase is to make recommendations for students and educational institutions to improve students' performance.
As discussed in the previous sections, not all dataset features affect student outcomes. Therefore, we will give recommendations based only on the features that affect student performance.

2.4.1 Recommendations for students to improve their performance

(a) **Visited Ressources :** Visiting resources is the most important feature affecting student performance; we recommend students pay more attention to the resources provided by the institution and use different resources, which is significantly more effective than using a single resource.

(b) **Students absence days :** Student absence days are the characteristic that most negatively affects student performance. We have seen that no student with more than seven days of absence can achieve a high level. Therefore, we advise students to attend all classes and establish a culture in which all classes are important.

(c) **Raised hands :** Raised hands are the next important feature that affects student performance. We recommend that students raise their hands and participate in class, ask questions about points they did not understand in class, and ask their teacher for more detailed and specific explanations.

(d) **Announcements View :** Viewing the announcements has a good impact on students' performance. Many students who view the new announcements have a good level of results. Therefore, we recommend that students view the announcements daily. It is good to point out that teachers and institutes, in general, should also facilitate the task by posting new announcements on the institution's website or social media.

(e) **Parent answering survey, Parent school satisfaction and relation :** We can see that family support has a good impact on student outcomes. Here, we recommend parents get involved in their children's learning process. We have seen that students with a mother as their primary caregiver are more likely to succeed because they have more support in their family.

(f) **Discussion :** Discussion also has a good impact on the students' performance. Students who participate in discussion groups get better results. So we recommend students participate in these discussion groups because we learn more when we learn together. Also, we suggest that the teachers create an appropriate environment for the students to organize these groups.

2.4.2 Recommendations for learning institutions

Higher education institutions can use prescriptive analytics at all levels and functional groups within the organization. In addition, it may be used by the administration or faculty. Here, we recommend that institutions use prescriptive analytics to improve the relationship between the student and the institution and change outcomes. To do this, the institution needs to gather student data and analyze it to understand what drives each student's behavior and then offer ongoing, real-time guidance to each student on what they need (specific courses, additional language courses, orientation, etc.). In addition, the institution can engage students by sharing their data with them to improve the relationship between the student and the institution, making the environment more conducive to learning and improving student performance.

3 Conclusion

All educational institutions have the same goal, which is to improve student performance. Big data analytics can be instrumental in achieving this goal in many ways. Predicting student performance is one of the most effective ways, as it allows administration and teachers to identify at-risk groups in their classes early in the study year. It will help them adjust their lessons to help the weaker ones and improve the performance of the stronger ones. We started with descriptive analytics to improve student performance in this case study. This allowed us to understand the dataset and to have a better view of its features. Then, moving into predictive analytics and build two predictive models based on two different machine learning algorithms (Decision Trees and Random Forest) to predict student performance and study the impact of each feature on student outcomes. Based on the results of the predictive analytics, we provided recommendations and advice for students and teachers to help them improve student performance and the educational system.
This work has some limitations that should be noted. The dataset used was small, with only 480 student records. Therefore, we were not able to offer many recommendations.

Chapter V: Enhancing University Decision-Making through Big Data Analytics

1 Introduction

In 2001, (Mentzer et al., 2001) defined the supply chain as *"a set of three or more entities (organizations or individuals) directly involved in the upstream and downstream flows of products, services, finances, and/or information from a source to a customer"*. Entities involved in supply chain processes collaborate in processing inputs to deliver the right products or services in the proper quantity, to the right place, at the right time, and with the lowest cost possible. Supply chain management (SCM) is crucial, primarily for the flow and proper sequencing of supply chain tasks, then for improved operations, better outsourcing, and increased profits.

For decades, researchers and companies have been developing and improving SCM techniques and processes, primarily for the welfare of the manufacturing industry. However, there is a lack of SCM studies in the non-manufacturing or service industry. Possible reasons include the difficulty of measuring how services are performed, the impact of human behavior on service outcomes, challenges of the services' intangible aspects, and the lack of a clear description of the service process (Antony et al., 2020). Educational services are considered intangible actions directed at people's minds (Lovelock & Wirtz, 2004). By projecting the supply chain concept onto an educational ecosystem, we deduce that the Student is the primary customer of academic services and a key supplier. This particular customer particularity is known as the "customer-supplier duality" (Sampson, 2000) in the service industry. In the higher education supply chain, the university is considered the service provider, the students, and research projects are the raw materials and the university's service customers, and graduates and research results are the supply chain outcomes, which the final customer, Society, consume.

Ecosystem, mechanisms, and the organization of higher education institutions differ from one country to another. In this chapter, we present an educational supply chain management (ESCM) model adapted to the most common structure of higher education institutions. This model defines the multiple actors in the educational supply chain (ESC), the types of services provided (primary and support services), the collaboration between members, and the impact of decision-making on the smooth functioning of the ESC.

Our world is currently in a phase known as the Big Data era. The excessive use of the Internet and connected devices from the IoT have led to the appearance of new data types. These data are different from the traditional well-structured data.

Consisting of videos, images, text, signals, and other types, these data have several natures (structured, semi-structured, and unstructured), an exponential generation rate, an immense quantity, and a wide variety. Properly extracting information from this type of data has required the development of new methods and tools.
Nowadays, BDA is applied in several fields. Some areas, such as e-commerce, health, technology, security, and e-government, are well advanced in exploiting the available data in all its forms. Other fields, like higher education, seem to be outdated in the use of data, especially newly available data. In this research, we will demonstrate the potential of BDA in the ESC process and provide a conceptual framework for successfully implementing the big data decision support solution in the university. In addition, we will present a data analysis case study to illustrate the benefits of exploiting data in the university.

2 Educational Supply Chain in the university

Both manufacturing (tangible outputs) and services (intangible outputs) domains have the same objectives: high-quality outputs at each stage of the supply chain and finished products that meet the market and Society's requirements. To achieve these goals, Habib and Jungthirapanich (M. Habib & Jungthirapanich, 2009) proposed a simplified form of SCM for tertiary education (see Figure 5.1). Two universities' main

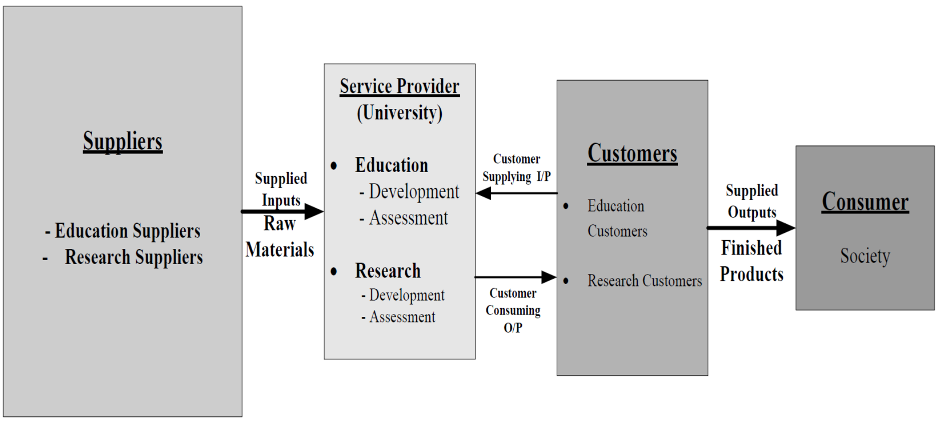

Figure 5.1: Simplified educational supply chain (M. Habib & Jungthirapanich, 2009)

services are identified: Education and Research. Raw materials are students and research projects, whether internal or external. The output products are graduates and research results. Despite its highly simplified form and power of abstraction, the proposed model remains very simple and neglects several significant points of the educational system, such as the structure of the university organization, the decisions and information exchange between the members of the ESC, and their collaboration to achieve the main objectives of the ESC.

Based on the ESC model illustrated in Figure 5.1, we derived an extended model (see Figure 5.2) where the decision-making structure of the universities is presented along with the data exploitation. The essential change is at the service provider level. Three hierarchical levels demonstrate the most applied structure in the universities (top, middle, and bottom levels). Decision-making often follows the top-down approach (blue arrows). However, this does not preclude feedback from the lower to the higher levels (red arrows). The data resulting from the guidelines and decisions made by the multiple hierarchical levels with the help of relevant external data should be stored and analyzed for possible exploitation in similar future situations (dashed black arrows).

(Lambert & Cooper, 2000) divide supply chain members into two categories: primary and support members. Primary members are all those *"who carry out value-adding activities in the business processes"* (Lambert & Cooper, 2000). For example, in the case of the university sector, primary members are faculty and department members, including students, teachers, and administrative staff. These members collaborate to produce better graduates. For the research to impact Society, better members are the researchers, internal and external collaborators, and laboratory staff. Their main objective is to reach research results that have a more critical impact on Society. Supporting members are those who *"simply provide resources, knowledge, utilities, or assets for the primary members of the supply chain"* (Lambert & Cooper, 2000), such as furniture suppliers, catering service providers, teaching and research equipment providers, or cultural and sports activity organizers.

Many research studies in the educational field often aim to improve the learning process and develop research procedures to impact society better. Based on the developed ESC model shown in Figure 5.2, we propose a new ESCM model for the universities illustrated in Figure 5.3.

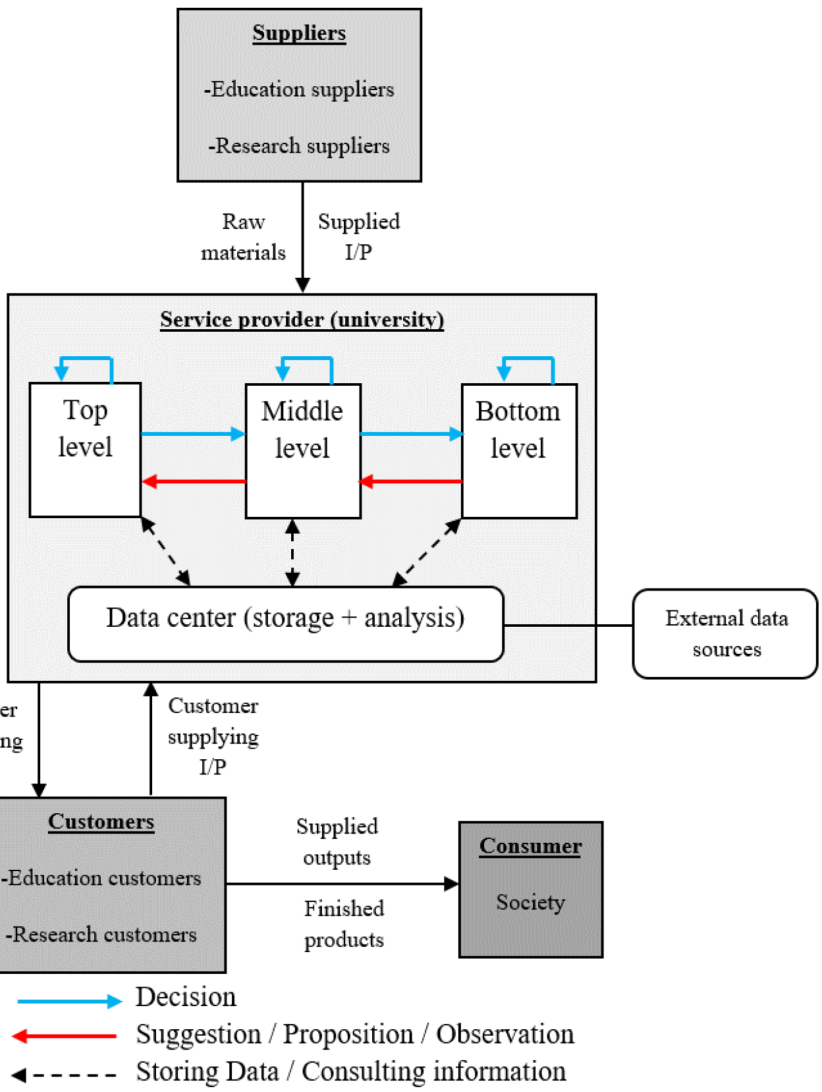

Figure 5.2: Developed educational supply chain

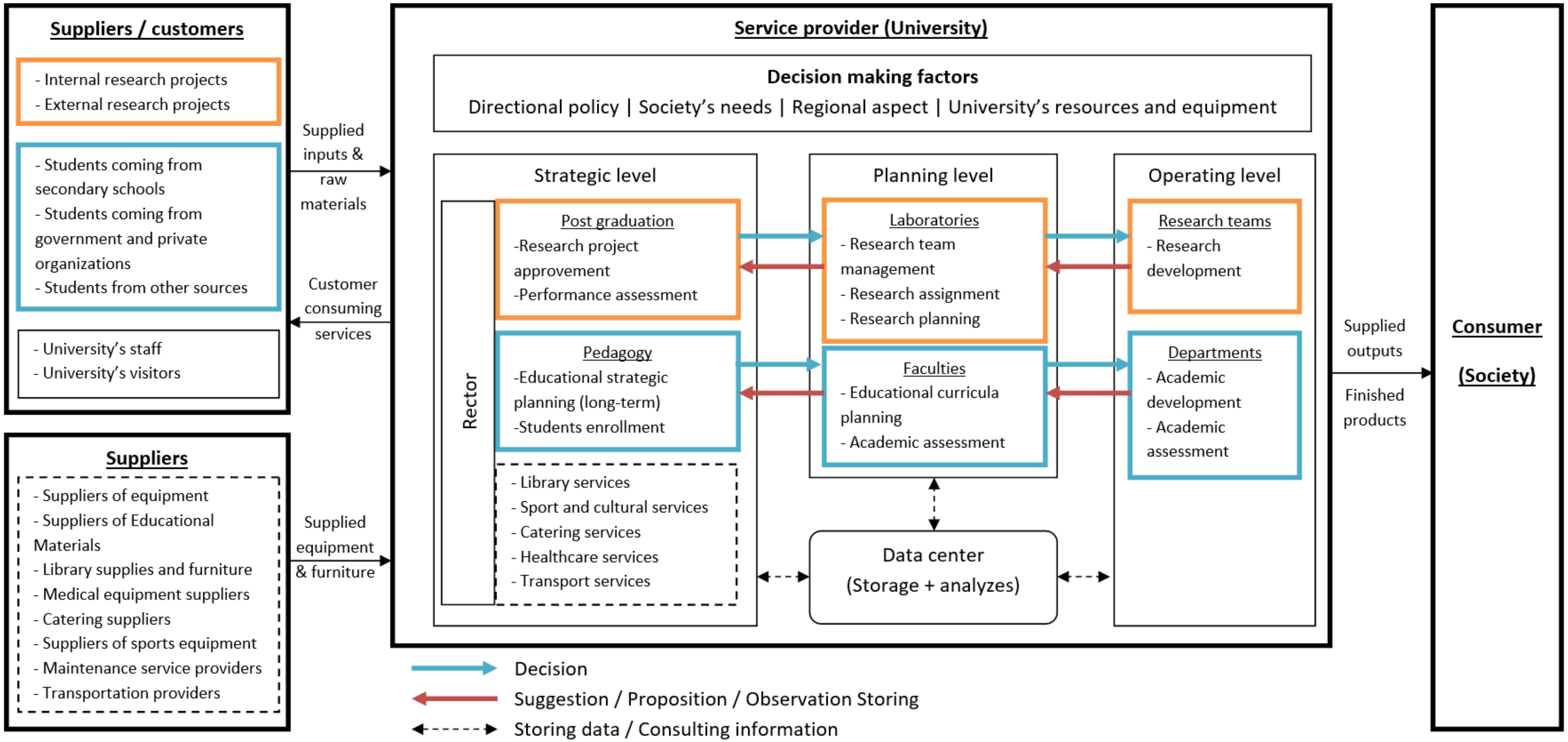

Figure 5.3: Educational supply chain management model for the university

This model identifies the members in each ESC stage on the educational (blue rectangles) and research (orange rectangles) services. In addition to the university's core services, the new ESCM model identifies other services. These services, named support services, are not classified as part of the teaching process or research activities. The support members of the supply chain are often the providers of these services. This type of service, often overlooked, can directly influence the other essential services of the university. For example, providing spaces for and encouraging sports and cultural activities will help the university's students and staff overcome daily stresses and get closer to each other for better collaboration in the educational and research processes. The proposed model illustrates the positioning of each member in this supply chain and the hierarchical decision-making mechanism in all the provided services.
The two main services offered by the university have a top-down decision-making orientation. As illustrated in our model, three decision-making levels follow the structure of the university, namely: strategic level (Top level), planning level (Middle level), and operational level (Bottom level) (Pathik et al., 2012). They are detailed in the following:

- *Strategic level:* At this decision-making level, decisions are made in the context of long-term supply chain planning. Usually, a higher education institution is attached to a ministerial department. The leaders of this department, as well as the rector or director of the university and his staff (Board of Directors), are in charge of using the available information and taking it to this level. The plans from this level are reviewed semi-annually or, sometimes, in annual periods, i.e., quick decision-making is not required at this level. Improving the university's ranking, increasing the number of doctoral students, and restructuring a faculty are examples of these long-term decisions.
- *Planning level:* Decisions at this level have a shorter span of influence than decisions at the top level. These decisions are made in the context of tactical planning. Detailed information is available, and the data are probably very reliable. Decisions at this level are constrained by the strategies developed at the higher level. There is room for modification to account for sudden changes in data. Creating or removing an educational discipline and planning research projects are examples of such decisions.
- *Operating level:* This decision-making level is responsible for implementing the

strategies decided and developed by the higher levels and collecting the resulting data. Decisions made at this level have a short-lived effect. Therefore, a quick response is an absolute necessity. Level three decisions are made regarding equipment and materials management, educational scheduling, educational performance, and scientific research monitoring.

The processes of the ESC and the members of the main and support services are detailed as follows:

2.1 Main services

2.1.1 Supply chain for education

The primary suppliers of this service are the students themselves, who come with their bodies and minds to benefit from the educational services. Students first arrive at the strategic level, where they complete their registration and are enrolled in an educational strategy developed and decided upon by the university rector and the pedagogical services. These strategies are directed to the next decision-making level. At this level, the faculty plays a central role in educational program planning. The departments attached to each faculty, at the operational level, must ensure the smooth running and implementation of the educational plans and strategies decided by the higher levels of decision-making to produce better graduates who will have a vital role in developing and improving our Society.

2.1.2 Supply chain for scientific research

In this category, we focus on research projects proposed to the university internally or by external organizations. Upon entering the university system, research projects must go through the university's decision-making process. They will first be processed and analyzed at the strategic level. After approval and validation, the research projects are forwarded to the next level, planning level. The leading members of the planning level are the laboratory directors. They develop research plans and transmit them to the members of the operational level (research teams). These operational level members will implement the plans and strategies developed by the higher levels to achieve higher quality research results with a higher impact on the end consumer.

2.2 Support services

The university's support services are often under the authority of the university's strategic level. These services must be constantly evaluated, analyzed, and improved to provide a better learning and working environment for all those in the university's ecosystem. Appropriate decisions will directly affect all three hierarchical levels of the university to improve the processes that provide the university's core services. Students, staff, and visitors to the university are the primary beneficiaries of these services. The following factors affect the Student's academic performance and the quality of the university's outcomes, where support services may play a critical role.

- *Student's mental health:* The prevalence of depression, anxiety, and stress among students is a common problem in universities around the world. One very encouraging solution to such a problem is the development of adequate academic support services.

- *Social relationships:* Having good social relations between students, teachers, and all the staff of the main services of the university will make the learning process more pleasant and comfortable. Participating in sports and cultural activities at the university is a positive way to create a better learning atmosphere.

- *Educational and scientific research resources:* Providing the necessary tools, equipment, and resources for learning and scientific inquiry is critical to achieving the desired results. Attempting to achieve primary goals in education and research services does not depend solely on the knowledge and experience of decision-makers. The proposed model identifies four factors influencing the ESC, particularly the decision-making process.

- *Directional policy:* The policy followed by a university's leadership or any decision maker directly affects the conduct and proper functioning of the university's educational process.

- *Society's requirements:* In our model, Society is considered the ultimate client or consumer of the university's outputs. The contemporary needs of Society for high-quality graduates and innovative research is another factor that requires consideration in the university's educational process.

- *Regional aspect:* Many universities consider the regional aspect based on the university's geographic location. The agricultural, economic, commercial, industrial, or other regional aspects are factors that the university must consider when deciding which type of discipline and which scientific orientation are most appropriate to meet regional needs.

- *University's resources and equipment:* The university's availability or lack of resources in terms of equipment, materials, supplies, or human resources can stimulate or hinder the university's educational process.

3 The Role of Big Data Analytics in the Educational Supply Chain and Decision-making Support

3.1 Big data analytics in higher education

In higher education, data analysis typically has two purposes: to predict and anticipate student success rates and to assess organizational performance. Therefore, analytics can be divided into two categories: learning and academic. While learning analytics primarily focuses on improving learner success (Siemens et al., 2011), academic analytics is focused on enhancing resources, processes, and educational plans by leveraging learner and institutional data (Campbell et al., 2007). By exploring the objectives of using data in the three ESC decision levels, we can determine which category of analysis is most appropriate for each level. Table 5.1 shows the suitable type of analysis based on selected objectives at the three ESC decision levels.

University decision levels	Type of analytics	Objective
Strategic level	Academic Analytics	- Revising and enhancing university strategies. - Evaluating the university's reputation. - Assess Student and academic staff satisfaction with the proposed university support services. - Evaluate and improve the Student's academic experience. - Assessing higher research results. - Better and faster consideration of the university service customer's problems and inconveniences.
Planning level	Mainly Learning Analytics and some Academic Analytics	Better management of research projects. Instructor and research project members' evaluation. Predict research project results. Monitoring the university strategies implementation. Revising and enhancing faculty/laboratory plans.
Operating level	Learning Analytics	Earlier detection of students with academic difficulty. Development and enhancement of courses and curricula

Table 5.1: University decision levels and data uses

The following are some examples of university use cases where the exploitation of big data is strongly recommended.

- *The university's reputation and service improvement:* Having a good reputation for the educational and support services offered is significant, especially for attracting excellent students. Using data from social networks to evaluate students' opinions about the university will help to know the university's reputation and thus improve its negative aspects. Applying a social media and text analytics approach to evaluate the customers' experience (students, university administrative staff, teachers, and visitors) will continuously improve the university's multiple services.

- *Student academic experience:* Students with academic difficulty must be identified at the earliest possible stage to react quickly and avoid potential dropout or failure. Students' academic experience should be tracked through data from smart classrooms and online educational systems. Adopting a streaming processing framework is highly recommended to handle such data and extract real-time information.
- *Good management of research projects:* Keeping each research project on time and budget while ensuring good results is critical to avoid wasting valuable time and money in scientific research. Predicting a research project's outcome is now possible by mining historical data from previous research projects.

3.2 Conceptual Big Data-driven Decision-making Framework in the Educational Supply Chain

The managerial experience of the decision-makers allows them to make decisions based on the historical background of their previous experience. Reinforcing this experience with concrete data from real-time situations will help them to make better decisions based on evidence.
This section presents a conceptual framework illustrated in Figure 5.4 based on big data for university decision-making support. This framework outlines the steps and the human, material, and software resources needed to implement the big data decision support solution successfully. The conceptual framework is derived from projecting the business sector's experience applying big data for decision-making. The framework illustrates the constituent elements and the gradual path from big data to decision-making. The framework elements are detailed in the following:

1. *Human resources:* To take full advantage of the benefits of data analysis, the university must develop analytics capabilities. Having competent and professional human resources in data processing is an obligation. The university must

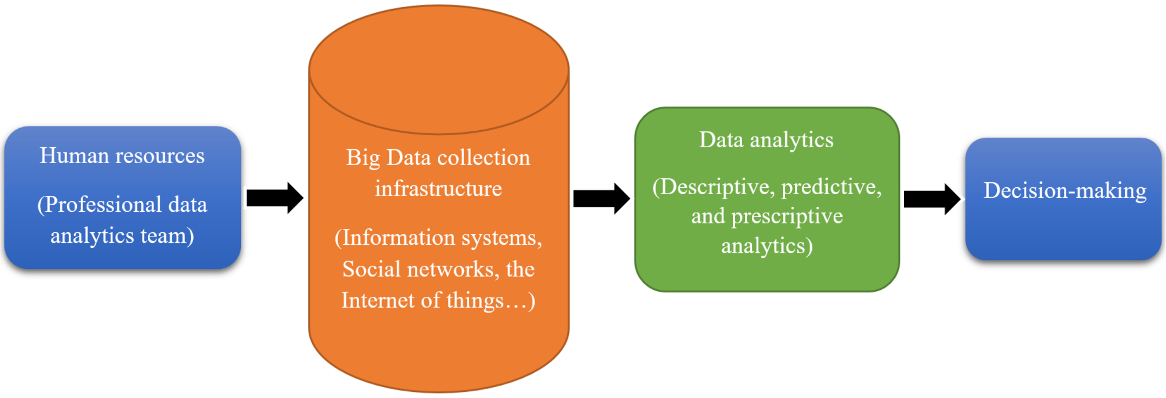

Figure 5.4: Conceptual Big Data-driven Decision-making Framework in the ESC

develop a team that will implement data collection and processing mechanisms for the benefit of each of the university's decision-making levels. This team must have strong skills in data analysis, such as excellent statistics skills, strong data management skills in SQL and NoSql, good data representation skills, and advanced programming knowledge. In addition to technical skills in data analysis, this team must develop knowledge about the academic field and its educational, scientific research, and support processes.

2. *Big Data collection infrastructure:* Several data sources must be exploited to have concrete and reliable data in the data analysis process. In addition to traditional data sources such as the university's information system, it is necessary to develop new data collection infrastructures through several sources like social networks, Internet browsing data, and the IoT (electronic devices, smartphones, sensors, CCTV, and others). For example, it is possible to develop mobile applications that will be deployed on students' smartphones. This will make it easier to communicate information and collect feedback about the services offered by the university. The data collected must be stored according to the requirements of the university's hierarchical decision-making levels.
Cloud computing offers a reliable, secure, and robust solution for implementing such infrastructures. This will avoid the need to procure expensive hardware and thus reduce the cost of deployment and maintenance of such equipment. Cloud computing adoption allows us to focus entirely on data analysis and interpretation tasks. The installation and maintenance of the data analysis tools and the provision of the best hardware are the responsibility of the cloud provider.

3. *Data analytics:* A common approach to extracting valuable information from the

data is to proceed with the three analytics categories (Descriptive, Predictive, and Prescriptive analytics) (Morr & Ali-Hassan, 2019).

Data mining and statistics techniques are employed to discover patterns and correlations between different data features, allowing us to understand what happened in the past and why it occurred.

Machine learning, deep learning, and artificial intelligence techniques are generally applied to build a predictive model around the collected data and the analysis objectives. Several reliable machine-learning algorithms exist depending on the analysis objectives, such as the random forest, K-means, and gradient-boosted model.

The prescriptive analysis phase aims to provide recommendations to decision-makers based on the predictive model resulting from the previous phases and the analysis objective.

The analysis team's skills are a critical factor in the success of the data analytics task.

4. *Decision-making:* The main objective of the proposed conceptual framework is to provide the maximum amount of useful information to decision-makers. Indeed, making better decisions based on realistic information is much easier and safer. Providing a dashboard containing information from data analysis of multiple sources such as social networks, the university's information system, and other sources maintained by a reliable and professional analytics team will undoubtedly improve the decisions made in the three decision-making levels of the university's ESC.

4 The Case Study

The primary aim of this case study is to demonstrate the potential of data processing in evaluating university services and the steps involved in approaching a specific case of data analysis.

4.1 Dataset overview

The dataset used in this case study is from a survey focused on the viewpoints and encounters of students regarding quality assurance within Vietnamese higher education establishments (Ta et al., 2023). The data were collected from July to September 2020 through an online survey conducted via Google Forms, resulting in 1323 valid responses. The data collection tool was developed in alignment with an international survey overseen by UNESCO. This survey aimed to solicit information regarding students' perspectives on institutional quality policies and frameworks, procedures and methodologies for quality assurance, and student surveys. The current dataset is especially helpful for strategic-level policymakers when making decisions on quality assurance for their institutions.

The adapted questionnaire for the Vietnamese context comprises four segments encompassing 60 closed-ended questions. Table 5.2 shows the dataset sections, sub-sections, and a brief explanation of every section. The first section includes four questions about respondents' details: academic year, gender, age, and city of study. The second part, encompassing 29 questions, is dedicated to quality policy and the quality assurance model. This part is subdivided into six subsections, addressing the significance of educational quality and quality policy, the quality assurance manual, the institutional body responsible for quality assurance, the quality assurance objectives, and focal points. The third section, encompassing 21 questions, focuses on the methods and instruments of quality assurance. It is partitioned into three subsections: quality assurance mechanisms and processes, student support services, and tools and procedures for enhancing graduate employability. Finally, the last part comprises ten questions on the survey and evaluation aspects. This part contains two sub-sections: the students' frequency of participation in such surveys and the modifications resulting from evaluation outcomes.

Sections	Sub-sections	Explanation
Personal information		Student's personal information (year of study, age, gender, and city of studying)
Quality policy and quality assurance model	- Importance of education quality in the university's quality policy - Quality (assurance) policy in the university - Quality assurance handbook - Unit in charge of quality assurance - Purposes of quality assurance - Focus level of quality assurance	29 questions about the quality assurance policy in the university. Questions have two types of answer choices: 1) 0: Do not know; 1: Not important / Not any; 2: Not really important / Not much; 3: Moderately important / Moderate; 4: Important / Quite a lot; 5: Very important / A lot. 2) 0: Do not know; 1: No; 2: Yes.
Quality assurance processes and instruments	- Quality assurance instruments and processes - Student support services - Instruments and processes for graduate employability	21 questions on the methods and instruments of quality assurance. Questions have one type of answer choices: 0: Do not know; 1: No; 2: Yes.
Evaluation survey	- Frequency participating in surveys - Positive changes from evaluation results	10 questions highlighting the student's frequency participation in surveys and their perceptions of the positive changes in the university. Questions have two types of answer choices: 1) 0: Do not know; 1: Never; 2: Rarely; 3: Sometimes; 4: Often; 5: Always. 2) 0: Do not know; 1: No change; 2: Change a little; 3: Change some; 4: Change quite a lot; 5: Change a lot.

Table 5.2: Survey questionnaire dataset overview

4.2 Data processing

Before initiating a data analysis process, it is essential to outline the objectives of this analysis. In the current case, we are focusing on two aspects addressed in the dataset:

1. Students' impressions on the effectiveness of quality assurance at their universities and their satisfaction with the applied policy's results.

2. The level of implementation and use of instruments, processes, and support services to assist the proper application of the university's quality assurance policy (QAP).

4.2.1 Data pre-processing

While analyzing the choices within the questionnaire, we consistently observe the presence of the option "Do not know." By selecting this choice, the Student indicates a lack of information regarding the question's subject, making their response entirely unrelated. Omitting such responses would be more beneficial for the analysis results. To better align with the analysis objectives, we will select questions directly related to these objectives. The first section provides us with students' personal information, allowing us to classify each Student in his university. The second section centers on the QAP, and the third section addresses the instruments and processes of quality assurance. These three sections suffice for the outlined analysis objectives.

4.2.2 Data visualization

Our dataset has 64 columns in total; 3 of them are in the "Object/text" type, and 61 columns have numerical values. To better understand the data, we will start by representing the data in a visual format using Python and some helpful libraries like Pandas, Seaborn, and Matplotlib.
The students selected to answer the questionnaire study in five cities of Vietnam: Thai Nguyen (northeastern part), Hanoi (capital city), Vinh (north central coast), Hue (near south central coast), and Ho Chi Minh City (southeastern part). Table 5.3 shows the number of interviewed students per study city.

City	Number of students
HCMC	263
Hanoi	553
Hue	164
Thai Nguyen	181
Vinh	162

Table 5.3: Number of interviewed students per city of study

4.2.2.1 Quality assurance policy assessment:

To measure the university's level of interest in the QAP and its application, we have selected several questions that we believe are of interest. The first question is Q2.1 (In your opinion, how is education quality important in your university's general policy?). Figure 5.5 shows the students' responses to this question for each city of study. We can note that most students find that their respective universities attach considerable importance to the quality of education in their general policies.

The following questions deal with QAP, a document that defines quality objectives, principles, and regulations to influence current and future decisions on quality issues. Selected questions are Q2.2.1 (Does your university have a quality policy or a QAP?),

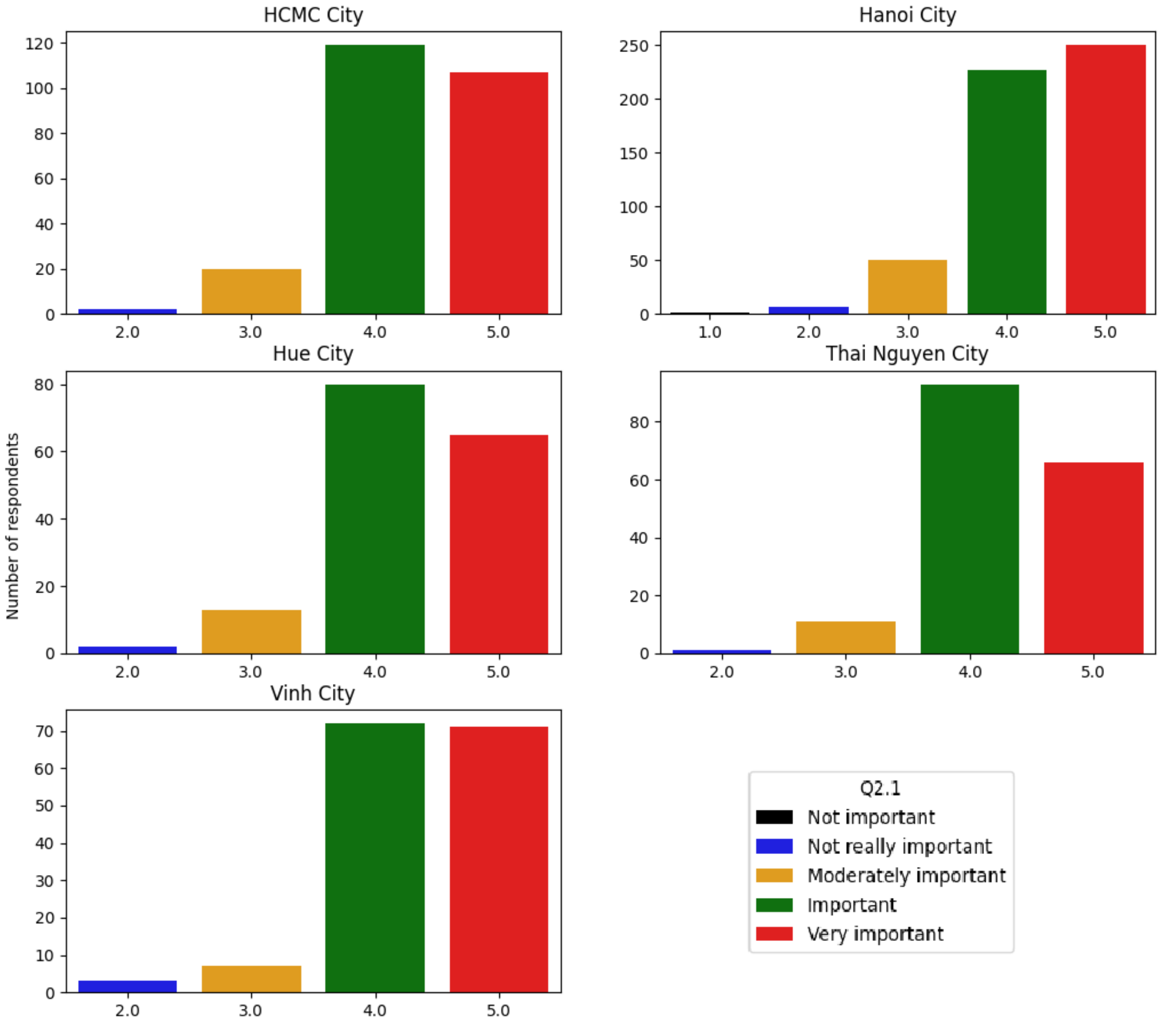

Figure 5.5: Question 2.1: education quality importance in the university's general policy

which will reveal whether or not the university has a QAP. The second question Q2.2.5 (Is the university quality policy / QAP being developed?) will help determine the university's involvement the QAP development and improvement. Figure 5.6 reveals that the universities in the selected cities have a constantly evolving QAP.

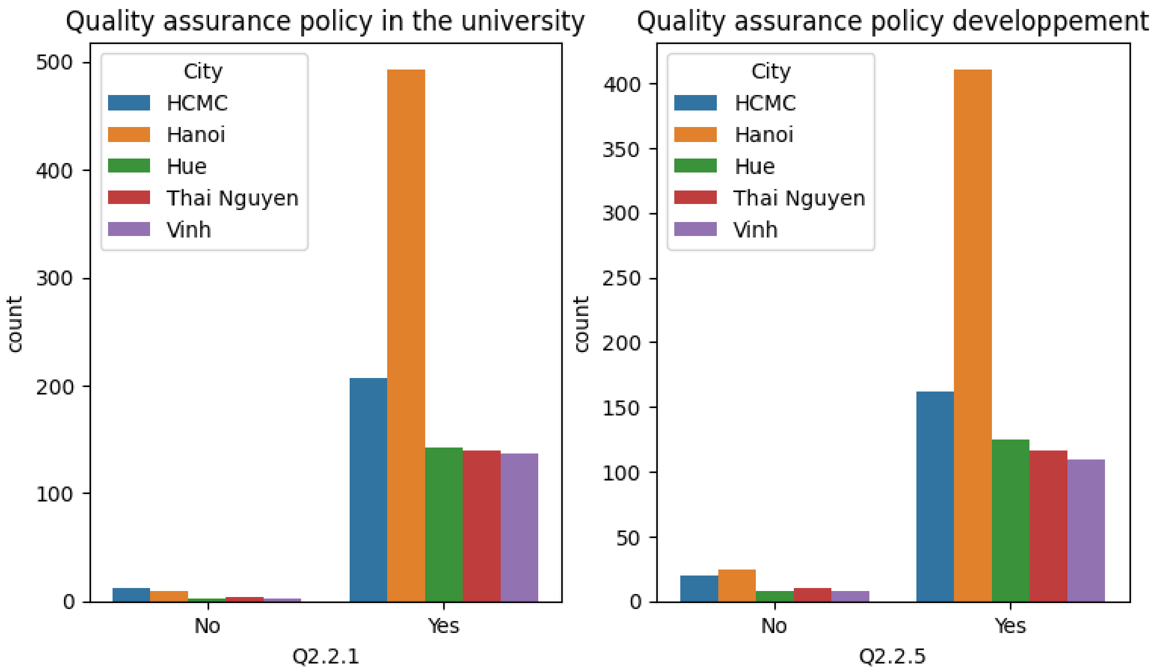

Figure 5.6: Questions 2.2.1 and 2.2.5 on quality assurance policy and its development

To strengthen our findings on quality assurance at the university, we have selected two additional questions (Q2.4.1 and Q2.4.2) on whether or not there is a dedicated unit responsible for quality assurance and the extent of its role in monitoring and improving education quality.
Figure 5.7 confirms that most universities have a dedicated quality assurance assessment and improvement unit.
Next selected questions focus on the importance of various university quality assurance objectives, such as assessing the quality of education at the university (Q2.5.1), improving teaching activities (Q2.5.2), learning activities (Q2.5.3), management activities (Q2.5.4), and support services (Q2.5.5). The answer choices vary from (1: Not Important) to (5: Very important). Figure 5.8 that most respondents consider that their respective universities attach an importance that varies between (3: Moderately important) and (4: Important) for quality assurance concerned in the questions.
Understanding the focus and orientation of the university's QAP is very important and

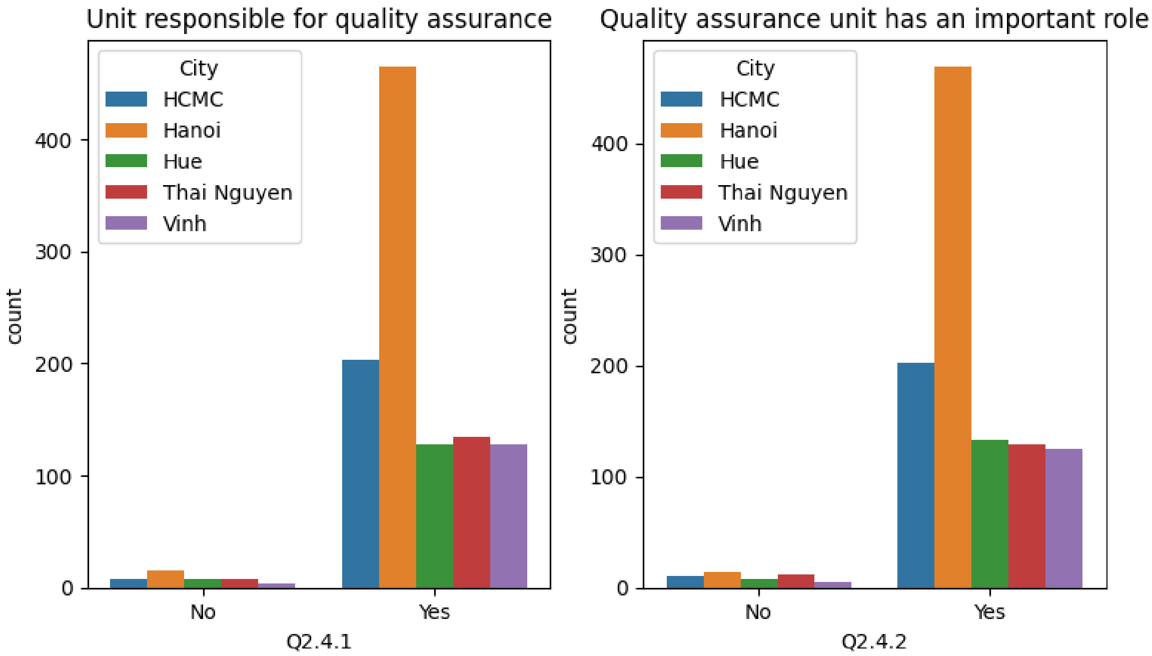

Figure 5.7: Unit responsible for quality assurance

potentially has a major impact on the results of the learning process and the outcomes of scientific research. The sixth sub-section of the QAP model involves questions regarding the focus of the university's policy towards certain main and secondary activities, such as learning and teaching activities (Q2.6.1), graduate employability (Q2.6.2), research (Q2.6.3), management and governance (Q2.6.4), support services (Q2.6.5), and facilities (Q2.6.6). The answer choices in the selected sub-section vary between (1: Not any) and (5: A lot). Figure 5.9 states that the universities in the questionnaire focus on a wide range of activities in their QAP. The answer to the questionnaire is that most respondents rate that their university attaches an importance that varies mainly between (3: Moderate) and (4: Quite a lot).

4.2.2.2 Quality assurance instruments and support services:

Tools, instruments, and secondary support services majorly impact the university's performance outcomes, i.e., graduates and research results. The university's QAP must take account of the support services aspect carefully and without neglecting them.
To verify the effectiveness of the selected universities' QAPs towards supporting activities and tools, we picked two sub-sections (Q3.1 and Q3.2) of the third section. Sub-section Q3.1 asks students to confirm or deny whether particular processes and tools have been implemented at their respective universities to improve programs. We

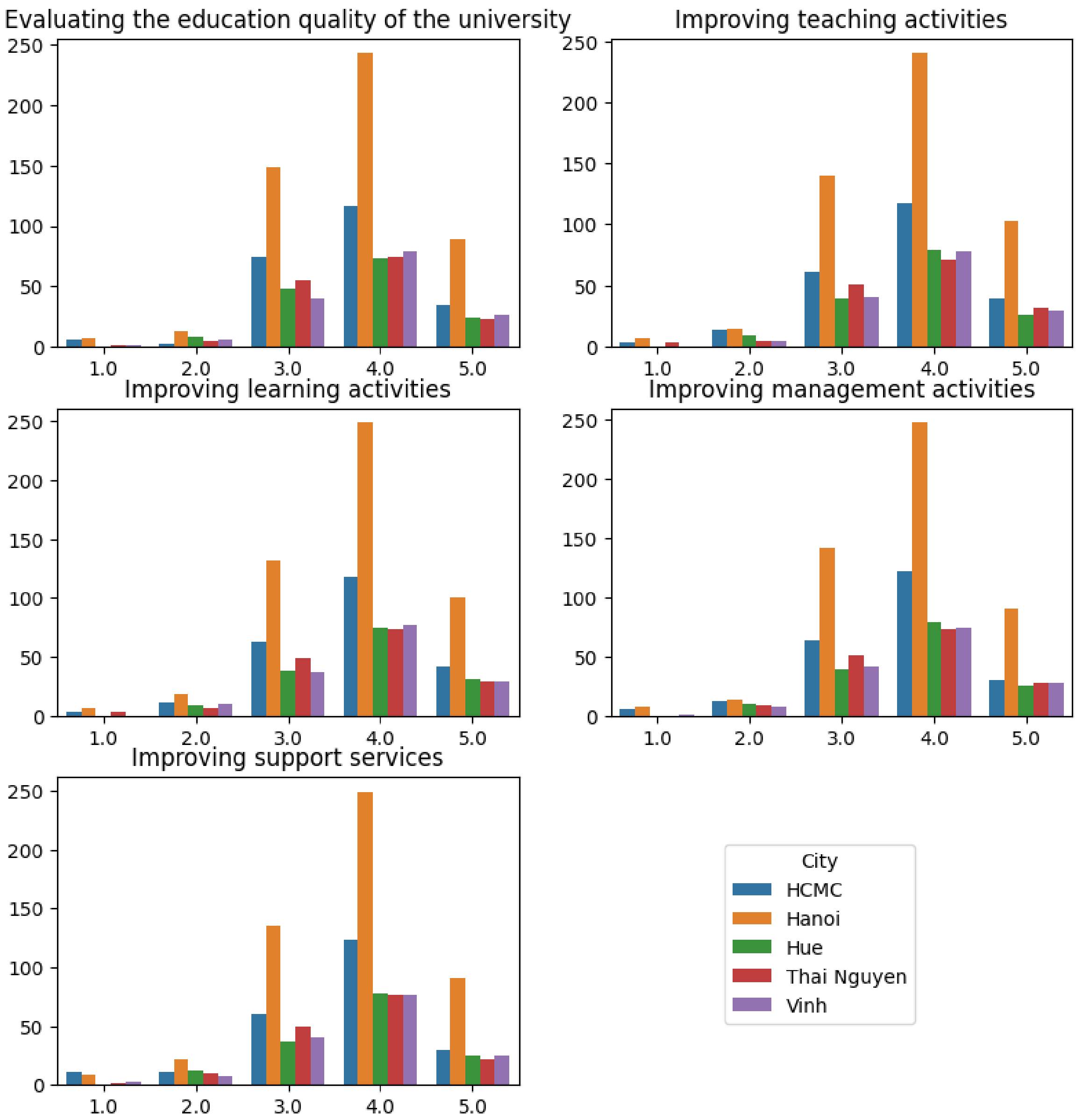

Figure 5.8: Purposes of quality assurance

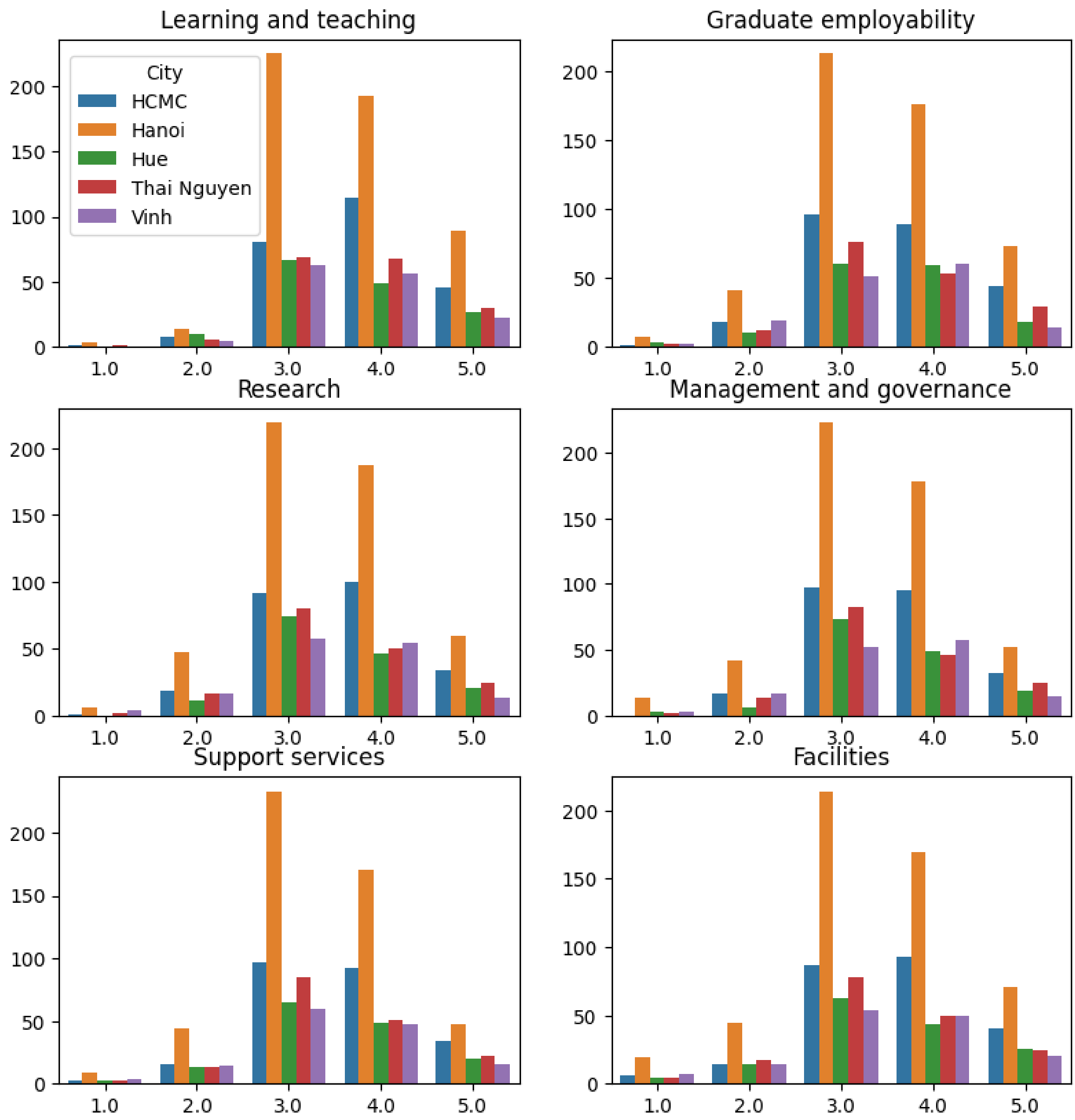

Figure 5.9: University's quality assurance focus level towards some activities

selected one process (Q3.1.5: Student progression monitoring) and one instrument (Q3.1.7: Student satisfaction survey) to verify the real implementation of helpful processes and instruments. Figure 5.10 clearly shows that most of the universities selected attach prime importance to these processes and instruments.

Next, we would like to check the availability and effective use of some support services

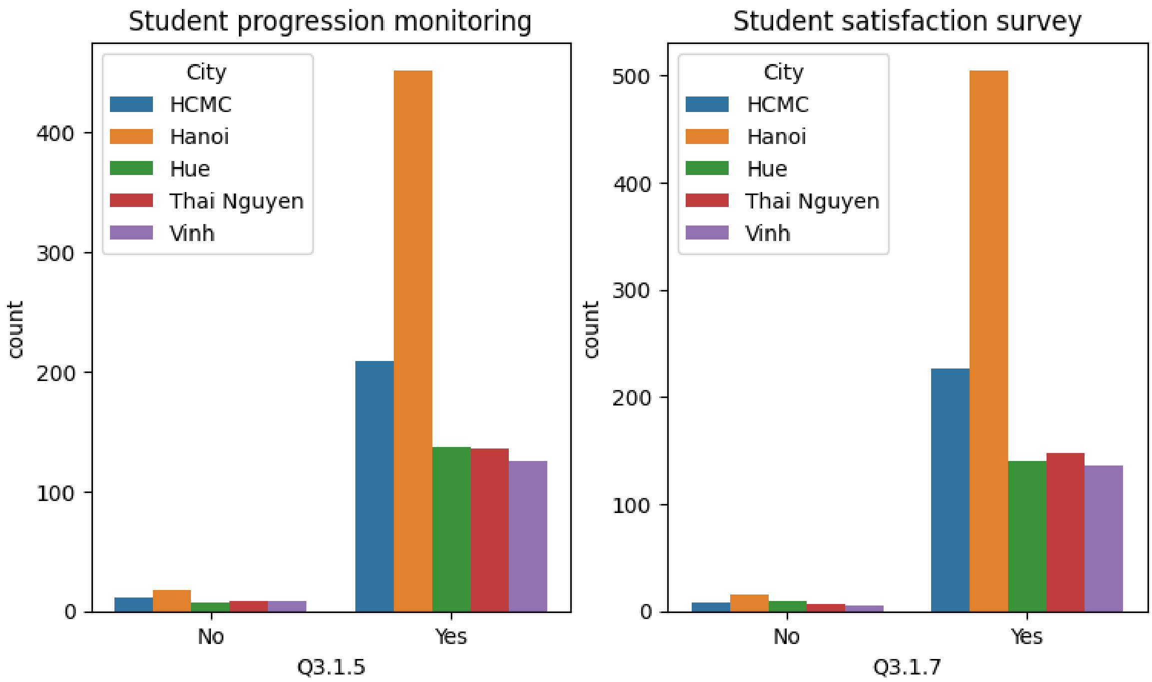

Figure 5.10: Part of the implemented processes and used instruments to improve the university's programs

at students' universities. For this purpose, the second sub-section of the third section (Q3.2) asks students to affirm or not the presence and actual application of certain support services. Figure 5.11 confirms that the most interviewed students' universities provide and use the relevant support services.

4.2.3 Analysis findings

In the present case study, we processed a dataset derived from a survey of students. This survey aimed to evaluate the universities' strategy regarding quality assurance. The survey provides an overview of the effectiveness of the QAPs implemented by universities, their ongoing development, their impact on university outcomes, and the level of student satisfaction with their university's QAP.

Our data consisted of multiple-choice questions, so we opted for a descriptive analysis that visually represents the data, making it easier to understand. The result of

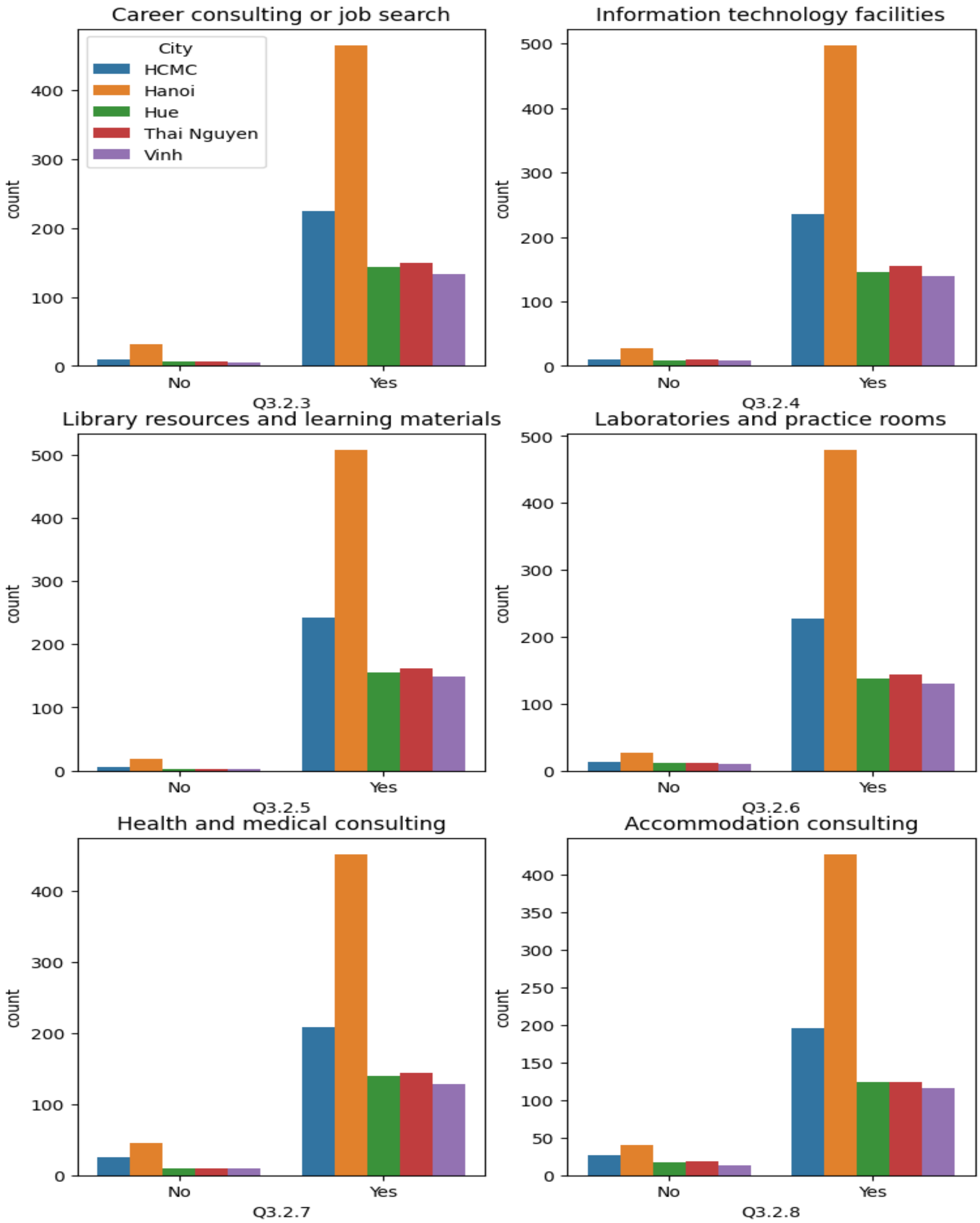

Figure 5.11: Support services

this analysis confirms that most students agree with the fact that their respective universities:

- Have a QAP in constant development.
- Give high priority to the quality of education within their QAP.
- Have a dedicated quality assurance assessment and improvement unit.
- Assign a level of significance ranging from moderately important to important for many quality assurance purposes.
- Have a QAP that focuses on a multitude of activities.
- Attach prime importance towards supporting activities, processes, and instruments.
- Provide and use relevant support services.

5 Conclusion

By analyzing the experience of advanced fields in leveraging supply chain principles, we found that the higher education field could benefit from this experience and apply it to the educational ecosystem. The study proposes an ESCM model to determine the roles of all members of the ESC, specify the decision-making objective of each hierarchical level of the university, strengthen the collaboration between members of the ESC, and outline the essential factors that influence decision-making in the ESC. The involvement and exploitation of data analysis in the ESC decision-making process are paramount to ensure the high quality of supply chain outcomes and rapid responsiveness to changing circumstances. Following the guidelines provided by the conceptual framework for implementing a solution based on big data analysis for decision support in the university will allow taking full advantage of the immense amount of data available today to improve the university's educational services continuously.

Chapter VI: An IoT-Fog Based Architecture for Detecting and Managing Forest Fires

1 Introduction

Forest fires, wildfires, or even firestorms are all terms that describe the existence of a fire in a forested area. Forest fires cause significant losses to environmental systems, infrastructures, biodiversity, and even human lives worldwide yearly. According to the 2020 Global Forest Resources Assessment report (FAO, 2020), an estimated 7.20 billion hectares of land were burned between 2001 and 2018, with an average of about 29% of the area covered by trees.

In Algeria, the year 2021 has seen many devastating wildfires. The most important ones caused the most natural damage, especially human losses during August. The newspaper "le monde" (A, 2021) reports that August's wildfires in Algeria caused at least 90 deaths. The Algerian minister of agriculture and rural development revealed that August's fires destroyed 89,000 hectares of forest in 35 Wilayas (municipalities) (Le Monde, 2021).

Forests occupy a large part of the northern part of Algeria. That makes these places very susceptible to fires. Additionally, the lack of modern fire detection systems and methods and reliance on purely traditional methods significantly impacted the fight against fires and their spread. Therefore, the constant development and implementation of forest fire detection and prevention systems are crucial.

There are two types of forest fire detection systems (FFDS): traditional and modern. Traditional systems encompass old detection methods where human presence in all detection, localization, and extinction operations is indispensable. The total dependence on the human factor may lead to the unreliability of forest fire monitoring due to climatic conditions that hinder surveillance, like clouds, mists, etc. Moreover, human characteristics like fatigue, omissions, and negligence may lead to system disruption (Barmpoutis et al., 2020). Throughout the world's forested areas, implemented fire detection approaches still use traditional methods. Besides, many modern FFDSs based on technological and information systems have been developed. Satellite-based image processing systems are one of these methods. Characterized by their comprehensive geographic coverage, NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) (Barnes et al., 1998) and Visible Infrared Imaging Radiometer Suite (VIIRS) (Murphy et al., 2006) satellites have been widely used to monitor large forest areas. However, their interval time between detecting fire and alarming the concerned departments is relatively high (Grover et al., 2019). In addition, climatic conditions such as the presence of clouds and mists and other obstacles to satellite vision proved the limit of this type of system. An emerging trend in forest fire detection is camera-based image processing systems deployed on unmanned aerial vehicles (UAVs). This type of system is relatively inexpensive and can cover multiple types of terrain (Kalatzis et al., 2018). However, limited resources and processing capabilities, UAV's limited battery life, and privacy preservation are some issues that need to be addressed to achieve better results in such processing systems.

Wireless Sensor Networks (WSNs) have been widely employed for environment perception and monitoring. WSN-based systems are suitable for forest fire detection and have proven their efficiency in detecting, localizing, and tracking forest fires in many applications (Grover et al., 2019; Bouabdellah et al., 2013; Saoudi et al., 2016). The vast geo-distribution of environmental sensor nodes (such as temperature, humidity, precipitation, and wind sensors) in a forest area allows better monitoring and effective detection of forest fires in an early stage. However, WSN-based systems present some limitations. A WSN environmental perception system deploys many sensor types, which generate a vast amount of heterogeneous data in a short lap of time (Smys, 2019). Processing such data very quickly is challenging in traditional WSNs. Sensor node-limited battery life is another limitation every WSN system must overcome.

The paradigm of connected objects or the IoT is a trend that is very appropriate for forest fire detection, prevention, and management. In combination with the cloud computing paradigm, for efficient and fast data processing purposes, several researchers have proposed "real-time" FFDS, such as (Tomar et al., 2019) and (Sungheetha & R, 2020). Data processing and transmission of real-time results are crucial requirements in such systems. Nevertheless, cloud-based forest fire detection implementations have high network usage and latency rates in data acquisition, data processing, and results transmission due to the distance between the source and the data processing center, which is unsuitable for time-sensitive applications.

Existing works in the application of forest fire detection and management have proven their effectiveness by employing technologies such as WSN, IoT, and cloud computing. However, such a system requires deploying a pervasive number of sensors in extravagantly large forest areas, which will generate an excessive amount of data in a short time frame. To address this situation, we suggest implementing an intermediate layer between the data source and the cloud data centers that will be in charge of fast processing of the collected data and minimizing data traffic on the network. This data management mechanism is often called fog computing. The main objective of implementing the fog computing paradigm in the FFDS is to provide fast data processing capability and bandwidth reduction for the network due to the vast number of data-generating devices in such a system.

This study proposes a multi-layered architecture for wildfire prevention, detection, management, and intervention based on WSN, IoT, Cloud, and Fog computing technologies. The architecture is shown in Figure 6.1. The proposed approach has been evaluated by performing simulations in iFogSim (H. Gupta et al., 2017). Two

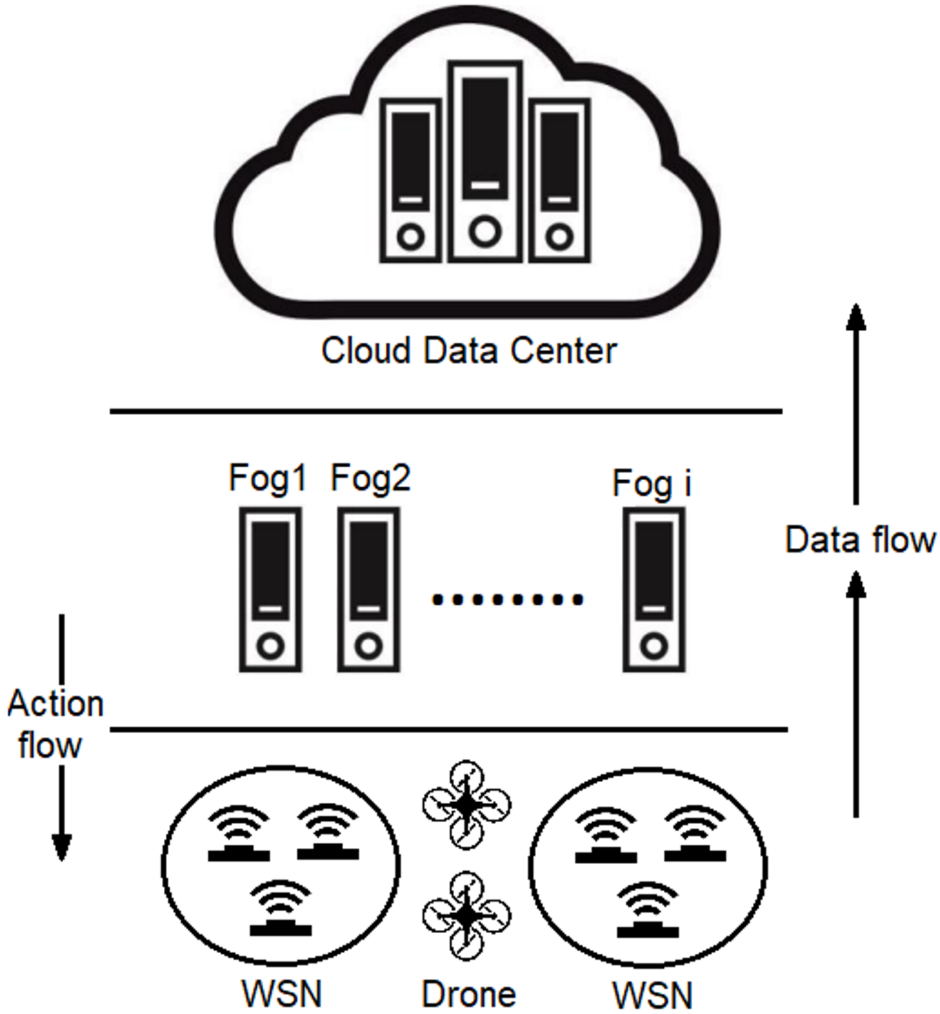

Figure 6.1: Wildfire prevention, detection, and intervention architecture

models have been implemented: cloud-only and fog-cloud models. The experimental results demonstrate that the fog-cloud model minimizes latency in the system and has a considerable reduction of the bandwidth consumption compared to the cloud-only model.

The contribution of this study to the field of forest fire detection is as follows:

1. *Processing performance:* To overcome the processing challenges, fast scalability, and high interoperability of the WSN application, a multi-layer architecture (object, fog computing, cloud computing layer) is presented as a solution. The object layer comprises many environmental parameters sensing objects and others for rapid intervention in case of high fire risk. The fog computing layer brings a good processing ability to the edge of the FFDS. Separating sensing and processing tasks between layers in the proposed architecture also reduces the power consumption of sensors and other devices in the object layer.

2. *Fast intervention and false alarm cost reduction:* A fast fire extinction at its outbreak is crucial in a forest environment that favors rapid-fire spread. Therefore, drones equipped with fire extinguishing balls may be a practical approach for rapid fire extinction. Figure 6.2 shows an example of a drone equipped with fire extinguishing balls. According to (Aydin et al., 2019), a small extinguishing ball of 0.5 kg can deal with a circle of one meter of a short grass fire. Deploying this type of Drone will bring a rapid intervention ability to the presented forest fire fighting system, which is a high requirement to avoid a catastrophic situation.

Figure 6.2: Drone equipped with extinguishing balls

In case of fire risk detection, before deploying human and material resources to control the fire, firstly, an extinguishing drone equipped with a camera is employed to verify the real presence of the fire and to bring a clearer view of the situation to plan better the intervention of the fire-fighting human and material resources if required. Using such an approach will reduce the cost of false alarms and provide a better perception of the area at risk of fire.

3. *Wildfire prediction:* When dealing with a critical situation such as a forest fire, predicting fires and defining a degree or index of fire vulnerability for a forested area is an important capability.
   Cloud computing infrastructure provides powerful and advanced analytical tools. Processing environmental data using advanced data analytics algorithms and methods will provide crucial information to the decision-makers.
   In addition to fog computing's fast data processing ability, using cloud computing as a long-term analysis tool for decision-makers and planners is an interesting approach to providing a fire prediction characteristic to the fire detection system.

2 Proposed System Architecture

This section presents an architecture based on WSN, IoT, fog, and cloud computing for the early detection of wildfires and rapid intervention. The proposed system is a three-layer architecture consisting of an object, fog computing, and cloud computing layers; see Figure 6.1. Each layer has a specific, predefined, and essential functionality.

2.1 System Architecture

2.1.1 Object (Peripheral) Layer

The first layer is a data perception layer, composed of IoT devices with several environmental sensor types like temperature, humidity, precipitation, and wind sensors for fire risk monitoring widely dispersed in a forested area. This kind of sensor has proven: a- their efficiency in sensing fire's causing parameters, b- their reduced fault-tolerance and good sensing abilities, and c- their endurance in challenging wild environments. The IoT sensing nodes are also equipped with a Wi-Fi module to allow the wireless communication of the sensed values. In addition, the GPS service can be easily implemented to provide geo-location capability to every sensing node in the WSN.

This layer is also composed of drones equipped with distinguishing balls. They intervene as early as possible in case of fire risk. These drones receive the exact geo-location provided by the IoT sensing nodes of the area at risk and move to extinguish the fire or at least prevent its spread by the time firefighters arrive. After receiving the exact geo-location of the area at risk of fire, the Drone in the first phase will move automatically to the forest area without any human interaction.

When arriving at the exact location, human interaction may be required to use the extinguishing balls effectively and to avoid possible errors and waste of the limited number of extinguishing balls. These drones are also equipped with cameras to provide a real-time image of the forest area to the fire department to provide the ability to analyze the situation and plan the intervention operation.

2.1.2 Fog Computing Layer

For the low-latency, wide-spread distribution, and the scalable large number of the IoT sensing nodes requirements in an FFDS, deploying the fog computing paradigm is the best choice (Smys, 2019). Fog computing provides easy and efficient connectivity to the object layer. Also, the fog computing paradigm has been proposed to bring the processing power of cloud computing to the edge of the network to minimize bandwidth usage and provide powerful processing units close to the end devices.
The principal function of the fog layer is to acquire, filter, and process the received data from the object layer. The second function of the fog layer is to communicate the data processing results via the Internet to any interested institution or service, such as the fire department. Besides, to intervene as quickly as possible, the fog nodes alert the drones of the object layer with the exact geo-location of the area at risk of fire.

2.1.3 Cloud Computing Layer

Due to the fog layer's reduced storage volume and limited processing capacity, it is difficult to perform an in-depth analysis of the data to predict wildfires. Therefore, a third layer based on the cloud computing paradigm is used. This layer stores data from fog nodes and uses advanced learning tools to improve the fire detection system. The cloud computing layer brings a long-term analysis tool for the FFDS. This analysis will serve as a basis for decision-makers to decide on mechanisms and protocols for forest fire detection and prevention. In addition, such an analysis will identify the areas most vulnerable to fire throughout the year.

2.2 Data Analytics Phases

Fast data analysis is crucial in an FFDS. The data collected by the IoT sensor nodes goes through three processing phases: data collection, data analysis, and decision-making regarding the analysis's results.

2.2.1 Data Collection

The peripheral layer is responsible for the data collection phase. It comprises a network of various IoT sensor nodes deployed in the forest canopy. They collect meteorological data such as temperature, humidity, precipitation, and wind speed and transfer it to the upper layer. The IoT sensing nodes detect, accumulate, and transmit climate data values to the fog nodes at specific intervals. In addition to this, geographic coordinate parameters play a vital role in mitigating the adverse effects of forest fires. The data collected by the IoT sensor nodes are classified into two categories:

1. *Weather data set:* It includes many weather factors such as temperature, humidity, wind speed, and precipitation.
2. *Location parameters:* It includes the geographical coordinates of the monitored area.

The sensors send collected data updates to the appropriate fog node via the Wi-Fi network. After processing this data, the fog node either triggers an action or not.

2.2.2 Data Analysis

The large number of IoT sensor nodes deployed in the target geographical area leads to excessive data generation for analysis. Excessive bandwidth use can clog the network and lead to delayed responses. Hence, the network must be sufficiently scalable and robust to handle this high traffic. The application of the fog computing paradigm is the ideal solution in such systems.
The fog layer lies between the peripheral layer and the cloud layer. It is in charge of acquiring and processing raw and unfiltered data (sent by IoT sensing nodes of the peripheral layer) of a predefined area, then sending orders to the appropriate actuators in this area.
In the presented system, the functions of the fog layer are:

- Data processing and decision making
- Connect the system to the internet
- Transmit data to the upper layer (cloud) for further analysis
- Transmit the decisions resulting from the data analysis to the actuators of the lower layer (peripheral)

To continuously improve the detection system, a second processing of the fog layer analysis results is performed for predictive and prescriptive purposes.
This type of data processing is computationally intensive, which makes cloud computing the appropriate solution. The cloud layer stores the data from the fog nodes and uses advanced learning tools to improve the detection system.
The cloud layer can:

- Store collected data
- Apply an in-depth analysis of the data
- Determine the optimal detection strategy
- Optimize the analysis methods used by the fog layer
- Analyze the received data in the long term to predict forest fires.

This study is not focusing on the data analysis method for forest fire detection. The main interest is in the general architecture and mechanisms that allow better management of a forested area.

2.2.3 Decision Making

Once fog nodes analyze the data collected by the sensor networks, the fog layer sends orders to the actuators located in the peripheral layer in case of a high risk of fire outbreak. The actuators are responsible for applying the instructions of the fog layer. In the proposed system, we suggest combining two types of actuators (traditional and modern) for optimal, functional, and safe control. The human factor is the main component of the traditional forest fire fighting methods. Human resources based methods are still used in all forest areas worldwide, even in countries that have developed other more advanced mechanisms (classified as "current or modern") such as Australia, Canada, and the United States of America.
In addition to traditional methods, the use of fire extinguishing drones is recommended. The main objective of using such types of drones is to intervene when a fire is detected at its early stage. Extinguishing the fire in its first stage will prevent its spread, avoid a probable natural disaster, and limit the exorbitant cost of dealing with such a fire.

2.3 Fire Detection Process Cycle

In the proposed forest fire detection model, the components of the different layers collaborate and communicate to better manage a forest area. Figure 6.3 shows the forest fire detection cycle in the presented architecture. The detection process is data-driven. First, the sensors are responsible for acquiring data on the environmental parameters. This information about the weather conditions is then transmitted to the corresponding fog node in the monitored forest area. Afterward, the fog node performs data analysis to determine the risk rate of fire outbreaks. According to the study, the fog node triggers or does not trigger a fire-extinguishing operation. A drone equipped with fire extinguishing balls and a camera to allow the visualization of the situation in real-time performs this operation. The fog nodes periodically transmit the analysis results to the cloud server for further analysis to determine the rate of fire vulnerability in each forested area.

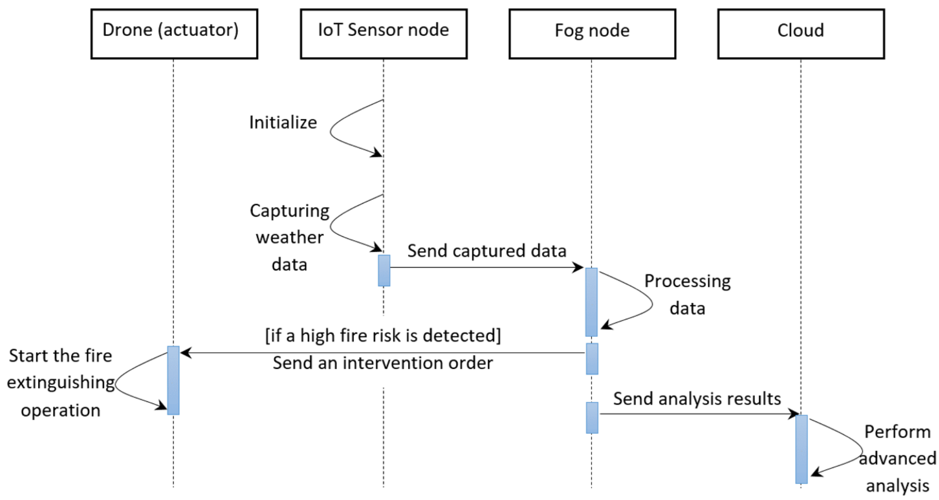

Figure 6.3: Forest fire detection and intervention cycle

3 Experimental Phase

3.1 Simulation

To evaluate the effectiveness and performance of the system outlined in the previous section, two fire detection scenarios are simulated. The first scenario combines cloud and fog computing technologies, while cloud computing is the main data processing engine of the second scenario. The iFogSim toolkit is used to create the network topologies for both scenarios and compare each scenario's results with the other. iFogSim is a simulator that supports evaluating resource management policies by focusing on their impact on latency, energy consumption, network congestion, and operating costs (H. Gupta et al., 2017). It simulates edge devices, cloud data centers, and network links to measure performance metrics. The application model supported by iFogSim is the Sense-Process-Actuate model. In such models, sensors publish data to IoT networks, applications running on fog devices subscribe and process the data from the sensors, and finally, the information obtained is translated into actions transmitted to the actuators.

Components of the simulator are summarized in the subsequent (Mahmud & Buyya, 2019):

- *Physical components:* include fog devices or nodes orchestrated hierarchically. Fog devices act like data centers in a cloud computing paradigm by providing memory, network, and computing resources.
- *Logical components:* In iFogSim, applications are considered a collection of interrelated AppModules. AppEdges' settings define the dependency between two modules. AppModules can be represented by virtual machines, and AppEdges are the logical data flows between two virtual machines.
- *Management component:* Consists of the "Controller" and "Module Mapping" objects. The "Module Mapping" object, according to the requirements of the AppModules, identifies the resources available in the fog devices and places them there. The "Controller" object launches the AppModules on their assigned fog devices following the placement information provided by the "Module Mapping" object. After the simulation, the Controller object collects the results of costs, network usage, and energy consumption during the simulation period from the fog devices.

The interaction between iFogSim components is shown in Figure 6.4.

Figure 6.5 represents the network topology for two monitored forest areas. Our approach to simulating the fire detection scenario starts by creating two forest areas, each represented by a fog node. Every fog node has the function of monitoring and managing all edge devices in its area. In the lower level, several weather parameter sensor types are connected to a microcontroller and represented by a sensor node. The detection frequency is the same for all sensor nodes. The fire-extinguishing Drone is represented by an actuator and connected to the fog node.

The fire detection process is simulated by creating modules for each detection phase as shown in Figure 6.6.

1. *Filtering Data Module:* A Filtering_data_module is created and assigned to every IoT sensor node. Its function is to validate raw climatic data sensors collect and

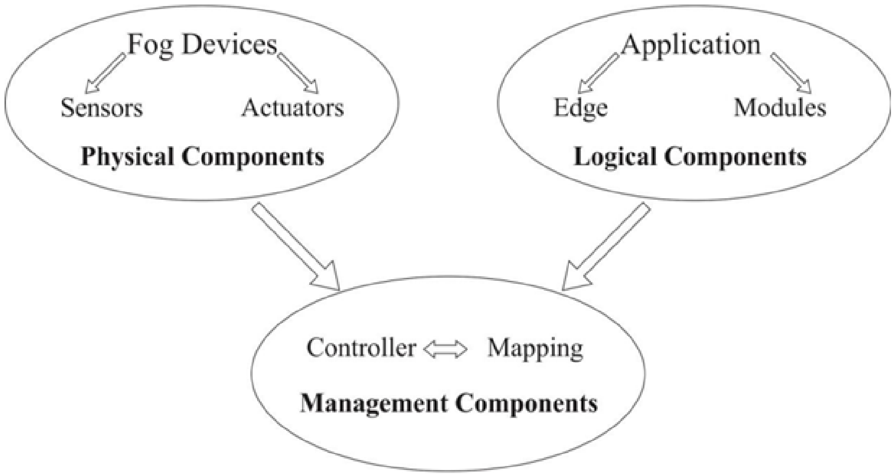

Figure 6.4: The interaction between the components of iFogSim

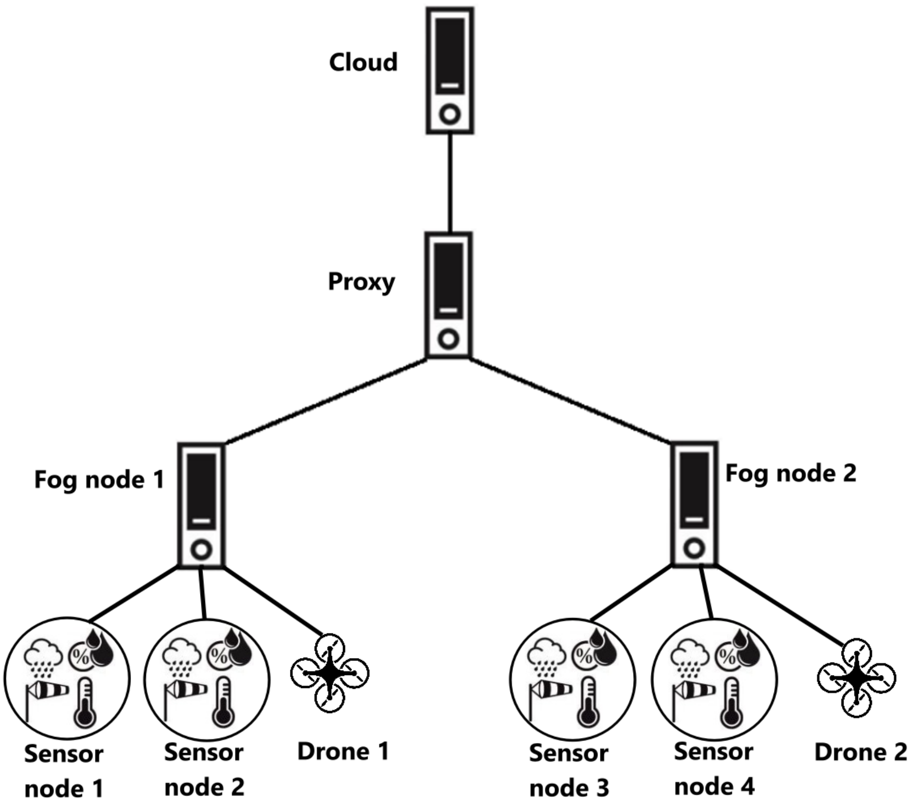

Figure 6.5: Forest fire detection network topology

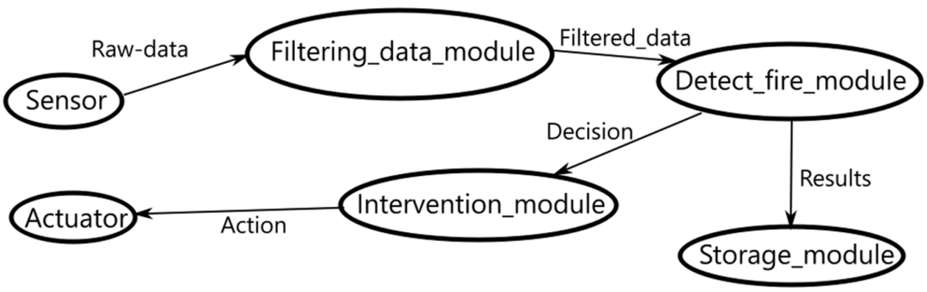

Figure 6.6: Forest fire detection system application modules

transmit them to the next module (Detect_fire_module).

2. *Detect fire module:* After receiving the filtered data, the Detect_fire_module applies a fire detection analysis to these data. Regarding the analysis results, the Detect_fire_module communicates decision information to the Intervention_module. The Detect_fire_module also sends analysis results to the Storage_module located at the cloud level.

3. *Intervention Module:* In case of an intervention decision, the Intervention_module triggers a fire extinguishing action performed by actuators.

4. *Storage Module:* This module is at the cloud level. It stores analytics results that the Detect_fire_module performs for further deep analysis.

Regarding the simulated fire detection scenario (cloud-only or fog-cloud-based scenario) and the available resources of each network component, the iFogSim manages the placement strategy where every module is assigned to the appropriate network component. In the fog-cloud scenario, the Detect_fire_module and Intervention_module are embedded in the fog nodes. Otherwise, these modules are assigned to the cloud server in the cloud-only scenario. The Filtering_data_module is at the sensor node level in the two scenarios. Also, the Storage_module is embedded in the cloud server regardless of the simulated scenario.
Table 6.1 gives the multiple network simulation components characteristics (Cloud server, proxy server, and fog nodes). These parameters are assigned to their respective devices at the network topology initiation time. The simulation parameters include CPU processing capability (million instructions per second), RAM capacity (megabytes), uplink and downlink bandwidth (megabytes), network-level placement of the component, rate per million instructions, and active (busy) and inactive (idle) power.

Parameters	Cloud	Proxy	Fog-node	Sensor-node
CPU (MIPS)	44800	2800	2800	500
RAM (MB)	40000	4000	4000	1000
Uplink BW (MB)	100	10000	10000	10000
Downlink BW (MB)	10000	10000	10000	10000
Level	0	1	2	3
Rate per MIPS	0.01	0.0	0.0	0.0
Busy power	16*103	107.339	107.339	87.53
Idle power	16*83.25	83.4333	83.4333	82.44

Table 6.1: iFogSim Simulation parameters of Cloud, Proxy, Fog-node, and microcontroller sensor-node

To properly evaluate the proposed system, four sensor nodes are assigned per fog node in the first phase. The number of fog nodes progressively increases to assess their impact on latency and network usage. In the second phase, the number of sensor nodes is increased while leaving the number of fog nodes stable. The multiple configurations of the simulated physical topologies are presented in Table 6.2.

Configurations	Phase 1		Phase 2	
	Number of fog nodes (Areas)	Number of sensor nodes per area	Number of fog nodes (Areas)	Number of sensor nodes per area
Config 1	2	4	2	2
Config 2	4	4	2	4
Config 3	8	4	2	8
Config 4	16	4	2	16
Config 5	32	4	2	32

Table 6.2: Configurations of the simulated physical topologies

3.2 Simulation Results

This section discusses the fog computing environment simulation results for the forest fire detection case study. Various metrics reported by iFogSim are collected for the multiple configurations of the physical topology cited in Table 6.2. The latency between system components and bandwidth usage are the parameters to be supervised.

3.2.1 Fire Detection Latency

Figures 6.7 and 6.8 show the average latency of the fire detection process loop for the cloud-only and fog-cloud strategy placements.

First, the phase 1 parameters are applied as cited in Table 6.2. Figure 6.7 shows

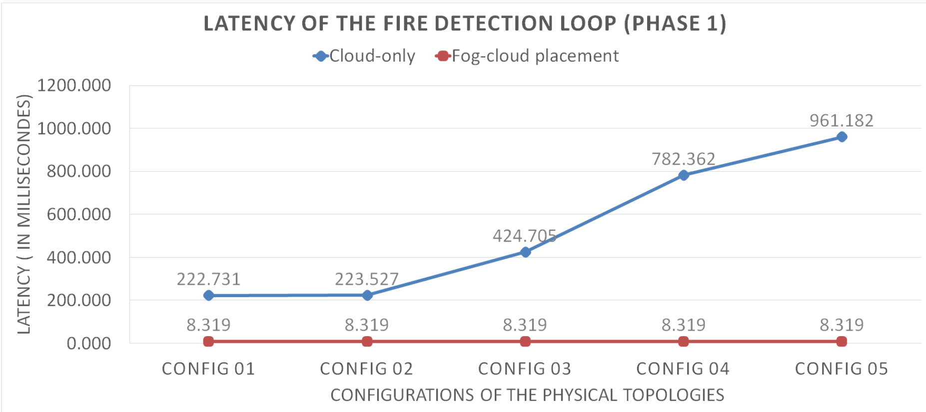

Figure 6.7: Comparing the latency in the cloud-only and the fog-cloud placement strategies (Phase 1 configurations)

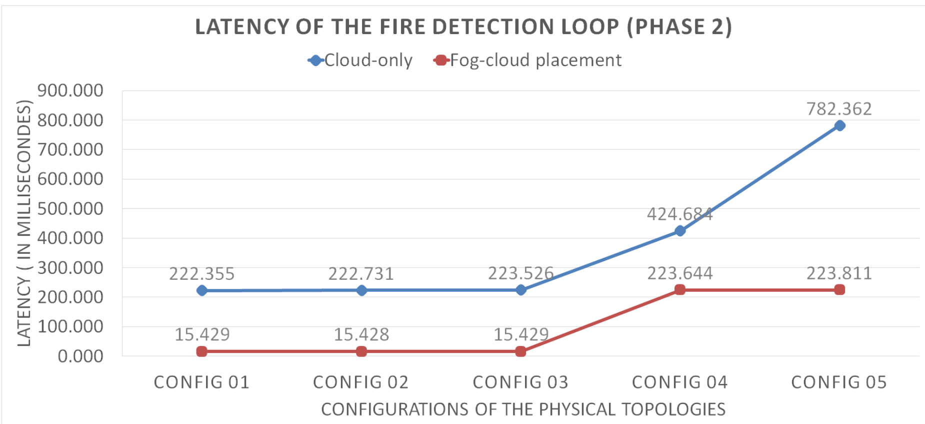

Figure 6.8: Comparing the latency in the cloud-only and the fog-cloud placement strategies (Phase 2 configurations)

the latency simulation results for this phase. We notice that in the case of fog-cloud placement, the latency of the detection loop remains relatively low and stable. The low latency is because, in phase 1, the number of sensor nodes remains the same for each fog node, and all the detection modules are executed in their regular placement. On the other hand, the cloud-only placement strategy results in a much higher latency than the fog-cloud placement. The latency increases considerably as the number of sensor nodes increases. The simulation results conclude that in implementation with a very high number of sensor nodes, the cloud server will have a very high response time to process all the sensor nodes' collected data.
Figure 6.8 shows the latency of the fire detection loop in phase 2. We notice almost the same result as phase 1 for the cloud-only strategy placement. The latency considerably increases as the number of sensor nodes increases. The fog-cloud placement strategy shows some changes regarding the phase 1 results. The latency starts very low for the first three configurations and then starts climbing when dealing with many sensor nodes. The fog nodes have limited resources. By increasing the number of data sources (the sensor nodes), the fog node will not be able to handle all the modules of the detection process. Some modules are then shifted to the cloud server, which explains the increased response time.

3.2.2 Total Network Usage

Figures 6.9 and 6.10 represent the total bandwidth utilization by all system components in the different configurations of the two simulation phases.

We notice that in the case of the cloud-only placement strategy, as the number

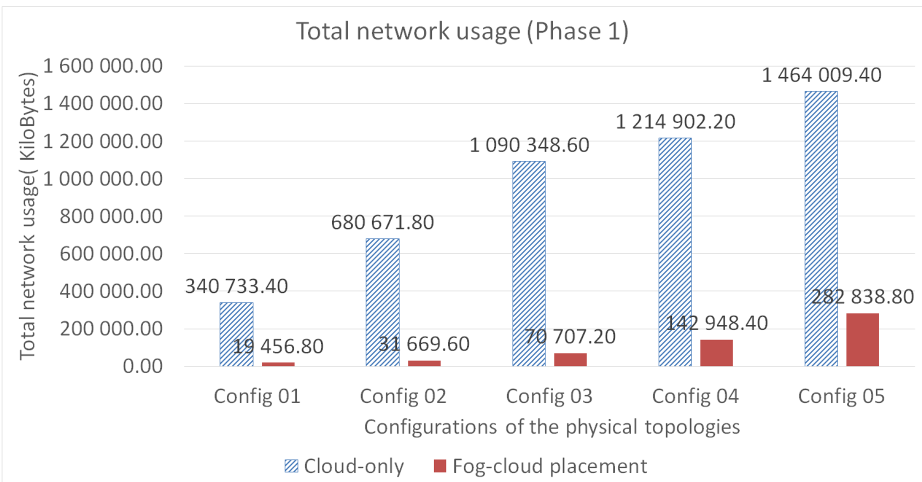

Figure 6.9: Comparing the total network usage in the cloud-only and the fog-cloud placement strategies (Phase 1 configurations)

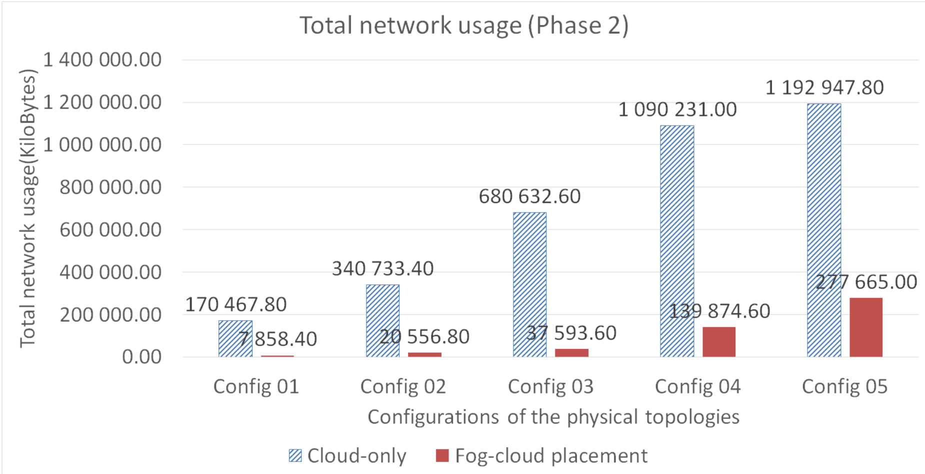

Figure 6.10: Comparing the total network usage in the cloud-only and the fog-cloud placement strategies (Phase 2 configurations)

of sensor nodes increases, the load on the network increases significantly. Also, in the case of fog-cloud placement, the bandwidth usage grows as the number of sensor nodes increases but remains very low compared to the cloud-only placement strategy. The placement of the processing modules is the main factor for network traffic. Indeed, the Detect_fire_module and Intervention_module modules in the fog-cloud placement strategy are located on the fog nodes at the network edge, considerably reducing the data volume sent to the data center. On the other hand, in the case of the cloud-only placement strategy, most of the processing modules are placed on the cloud server, which causes high traffic on the network.

4 Discussion

A forest fire detection and management system must be able to provide some key features to be a fully effective system. Different key features can be considered when classifying an FFDS. In the present research, seven key features are retained: fast data processing, fire spread monitoring, long-term data analysis, fire extinguishing capability, energy efficiency, privacy preservation, and false alarm rate. Details of each feature are given later. Using these features as a classification method will allow us to situate the proposed approach against others. Table 6.3 compares the proposed system in this research study and a selection of recent FFDSs based on the abovementioned features. The values in the table are given according to the consideration or not of each feature within a particular study.

	(Grover et al., 2019)	(Kalatzis et al., 2018)	(Varela et al., 2020)	(Benzekri et al., 2020)	Our Approach
Fast data processing	No	Yes	No	Yes	Yes
Fire spread monitoring	No	Yes	No	Yes	Yes
Long term data analysis	No	Yes	No	No	Yes
Fire extinguishing capability	Yes	No	No	No	Yes
Energy efficiency	No	Yes	Yes	Yes	Yes
Privacy preservation	Yes	No	Yes	Yes	Yes
False alarms rate	Medium	Low	Medium	Medium	Low

Table 6.3: Comparison of our proposed approach with other recent FFDSs

1. *Fast data processing:* Early fire detection is essential in a FFDS. It allows rapid intervention to prevent the spread of the fire and thus avoid a possible disaster.

2. *fire spread monitoring:* Having the ability to monitor the spread of a fire allows for better management. The deployment of a drone equipped with a camera is the ideal solution for this type of mission. In addition, using wind direction sensors will help predict the direction of fire spread.

3. *Long-term data analysis:* After an initial data analysis for fire detection, a second and more sophisticated analysis over the long term is beneficial to designate the most vulnerable zones to fire. Adopting a cloud-based solution is the best practice for such a process.

4. *Fire extinguishing capability:* It is well known that extinguishing a fire within a short period in a forest area will prevent its wide-spread. This feature is not addressed in most studies of FFDSs. The employment of drones equipped with a fire extinguishing solution makes it possible to reach the area at risk quickly and start the fire extinguishing operation.

5. *Energy efficiency:* In an environmental sensing system, reducing the power consumption of the data sensors allows for a long monitoring time without needing maintenance. Therefore, separating the data processing task from the data receiver components will considerably reduce power consumption and prolong the system's lifetime.

6. *Privacy preservation:* UAV-based systems have issues with privacy preservation as they are based on image perception. In contrast, sensor network-based systems have no privacy issues.

7. *False alarms rate:* Sensor network-based systems have a low false alarm rate. In our proposed approach, a second verification of the actual presence of the fire is performed by deploying a drone equipped with a camera. It makes the system

more accurate compared to ordinary sensor network-based systems. In (Kalatzis et al., 2018), when a forest fire is detected, new directives are sent to the UAV to focus the surveillance on the area at risk and thus verify the real presence of fire to avoid the case of a false alarm.

5 Conclusion

IoT applications in time-sensitive areas, such as environmental monitoring, require sophisticated data processing systems. Cloud computing was a suitable solution to meet this requirement. However, the excessive data generation from environmental perception sensors and the high distance between the data source and the data center lead to network congestion and increase the system response time. Fog computing has become necessary to reduce network traffic in such systems while taking advantage of the processing power of the cloud paradigm.

This study investigated the benefits of using the fog computing paradigm to manage a forested area and detect high fire risks. Experimental results showed that the recourse to a three-layer architecture (object, fog, and cloud layer) significantly reduces bandwidth usage and minimizes system response time compared to a cloud-only model.

In future studies, we intend to investigate multiple methods and algorithms for data analysis at fog and cloud levels. The fog level requires fast and efficient data processing methods to detect the fire outbreak immediately. Meanwhile, the cloud level requires more sophisticated data analysis methods to determine the fire vulnerability level of each forest area. Therefore, we can decide on future directives for better preservation of our forests.

The transfer and data exchange in the IoT ecosystem are frequently confronted with security and privacy issues. The case of FFDS is no exception. Through this study, we have evoked several applications where the main concern was privacy preservation and security on the network. Preserving privacy and enhancing the security level on the IoT network is another research guideline to exploit. The authors in (Gadekallu et al., 2022; Prabadevi et al., 2021) discuss this issue and propose the involvement of the Blockchain concept in Edge of Things (EoT) applications. The main reason behind using blockchain in EoT applications is its unique properties. Decentralization in blockchain will multiply the number of access points in the EoT network, thus eliminating the single point of failure of traditional centralized systems and reducing the system's vulnerability to attacks. Due to the blockchain's consensus algorithms, the immutability aspect of data on the network will reinforce the preservation of anonymity in an EoT system. Despite the use of open-source technology in the blockchain, which will bring transparency to the data on the network, critical information is kept secure by complex cryptography algorithms.

Conclusion

This thesis resulted in significant contributions to the field of Big Data with a specific focus on the utilization of the Internet of Things (IoT) by offering a comprehensive study of the field and proposing innovative approaches around the topic of Big Data Analytics for highly significant areas of application.

In the first instance, this work introduced the complex concepts of Big Data and IoT, highlighting how these cutting-edge technologies have developed over time and their profound impact on the contemporary digital environment. It provides a comprehensive overview of the evolution of Big Data, detailing its remarkable expansion, the diverse origins of the data it encompasses, and the complex processes involved in analyzing and interpreting that data to derive valuable insights crucial to informed decision-making processes. In addition, the study delves deeper into the field of IoT, elucidating its complex architecture by focusing on the seamless collaboration of IoT ecosystem components to facilitate the collection, transmission, and processing of data from interconnected devices. Furthermore, the study takes an in-depth view of the various applications of IoT in different sectors such as smart cities, healthcare, and industry, highlighting the immense potential of IoT technologies to bring about a paradigm shift in these areas through improved operational efficiencies, security standards, and service delivery mechanisms.

Following the literature review, this study focused on two application sectors: higher education and environmental perception and management.

1. *Higher education:* The main objective that intrigued us in this field was to show the potential of using data analytics methods to improve students' academic performance and the strategies educational institutions employ to provide optimal services to achieve the most relevant results. As a first step, a case study was carried out using an approach commonly used in the field of Big Data, which includes three phases of investigation: descriptive, predictive, and prescriptive analytics. This produced convincing results regarding student guidance and the supervision of their engagement with the processes implemented to support them in their educational pursuits.

   In addition, having examined the industrial sector and its progress regarding emerging technologies such as Big Data and IoT, we transposed these insights to higher education. We designed an innovative model delineating the educational supply chain, encompassing its many stakeholders, tools, and responsibilities. Furthermore, drawing inspiration from the industrial sector, we have formulated a theoretical framework for providing decision support using Big Data in the context of higher education. This involves identifying the steps required to implement such a framework.

2. *Environmental monitoring:* The use of IoT applications in time-sensitive domains, such as environmental monitoring, requires sophisticated data processing systems. However, the excessive data production by environmental perception sensors and the considerable distance between the data source and the data center lead to network congestion and increased system response times. We studied the adoption of fog computing to alleviate network traffic in such systems while leveraging the processing power of the cloud computing paradigm.

During our study, we investigated the benefits of employing the fog computing paradigm to manage a forest area for detecting high fire risks. Experimental results demonstrated that adopting a three-layer architecture (object, fog, and cloud) significantly reduces bandwidth usage and minimizes system response time compared to a model based solely on cloud computing.

Data transfer and exchange in the IoT ecosystem often encounter challenges related to security and privacy. In our investigation, we explored various use cases that prioritize safeguarding privacy and security within the network. Enhancing the level of security and preserving privacy within the IoT infrastructure stands as a key research focus. Incorporating the Blockchain concept in the Edge of Things (EoT) application is a promising approach to address this issue. The distinctive characteristics of blockchain serve as the primary rationale behind its adoption in EoT applications. The decentralized nature of blockchain leads to increased entry points in the EoT network, consequently eradicating the vulnerabilities associated with traditional centralized systems and enhancing resilience against potential cyber threats.

References

A, I. (2021, 08). *Le ministre de l'agriculture dévoile: plus de 89.000 hectares détruits par les feux de forêts.* Retrieved 2022-06-07, from `https://www.algerie-eco.com/2021/08/24/ministre-agriculture-devoile-plus-de-89-000-hectares-detruits-par-les-feux-de-forets.`

Ahmed, S. A., Alwan, N. F., & Ali, A. M. (2022). Overview for internet of things: Basics, components and applications. *Journal of University of Anbar for Pure Science*, *12*, 47-58. doi: 10.37652/juaps.2022.171846

AlAali, A. M., AlAteeq, A., & Elmedany, W. (2022, 11). Cybersecurity threats and solutions of iot network layer. In (p. 250-257). IEEE. doi: 10.1109/3ICT56508.2022.9990734

Alam, S., Siddiqui, S. T., Ahmad, A., Ahmad, R., & Shuaib, M. (2020). Internet of things (iot) enabling technologies, requirements, and security challenges. *Advances in Data and Information Sciences*, 119-126. doi: 10.1007/978-981-15-0694-9_12

Al-Atawi, A. A., Khan, F., & Kim, C. G. (2022). Application and challenges of iot healthcare system in covid-19. *Sensors*, *22*, 7304. doi: 10.3390/s22197304

Al-Badi, A., Tarhini, A., & Khan, A. I. (2018). Exploring big data governance frameworks. *Procedia Computer Science*, *141*, 271-277. doi: 10.1016/j.procs.2018.10.181

Ali, Z., Mahmood, A., Khatoon, S., Alhakami, W., Ullah, S. S., Iqbal, J., & Hussain, S. (2023). A generic internet of things (iot) middleware for smart city applications. *Sustainability*, *15*, 743. Retrieved from `https://www.mdpi.com/2071-1050/15/1/743` doi: 10.3390/su15010743

Al-Rubaie, N. R. O., Kamel, R. N. N., & Alshemari, R. M. (2023). Simulating fog computing in omnet++. *Bulletin of Electrical Engineering and Informatics*, *12*, 979-986. doi: 10.11591/eei.v12i2.4201

Alsubaei, F., Abuhussein, A., & Shiva, S. (2018). An overview of enabling technologies for the internet of things.
doi: 10.1002/9781119456735.ch3

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2015). Preprocessing and analyzing educational data set using x-api for improving student's performance. *2015 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT)*, 1-5. Retrieved from `https://ieeexplore.ieee.org/document/7360581` doi: 10.1109/AEECT.2015.7360581

Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application*, *9*, 119-136. doi: 10.14257/ijdta.2016.9.8.13

Anasica, S. (2022). Research on smart environment monitoring systems based on secure internet of things (iot). *International Journal on Future Revolution in Computer Science Communication Engineering (IJFRCSCE)*, *8*, 48-55. doi: 10.17762/ijfrcsce.v8i1.2090

Ang, K. L.-M., Ge, F. L., & Seng, K. P. (2020). Big educational data analytics: Survey, architecture and challenges. *IEEE Access*, *8*, 116392-116414. doi: 10.1109/access.2020.2994561

Antony, J., Viles, E., Torres, A. F., de Paula, T. I., Fernandes, M. M., & Cudney, E. A. (2020). Design of experiments in the service industry: a critical literature review and future research directions. *The TQM Journal*, *32*, 1159-1175. doi: 10.1108/tqm-02-2020-0026

Apache Software Foundation. (2023). *Hdfs architecture.* Retrieved from `https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html`

Apat, H. K., Nayak, R., & Sahoo, B. (2023). A comprehensive review on internet of things application placement in fog computing environment. *Internet of Things*, *23*, 100866. Retrieved from `https://www.sciencedirect.com/science/article/abs/pii/S2542660523001890` doi: 10.1016/j.iot.2023.100866

Arputhamary, B., & Arockiam, L. (2015). Data integration in big data environment. *Bonfring International Journal of Data Mining*, *5*, 1-5.

Attaran, M., Stark, J., & Stotler, D. (2018, 04). Opportunities and challenges for big data analytics in us higher education. *Industry and Higher Education*, *32*, 169-182. doi: 10.1177/0950422218770937

Atzori, L., Iera, A., & Morabito, G. (2010). The internet of things: A survey. *Computer Networks*, *54*, 2787-2805. Retrieved from `https://www.sciencedirect.com/science/article/abs/pii/S1389128610001568` doi: 10.1016/j.comnet.2010.05.010

Awaisi, K. S., Abbas, A., Zareei, M., Khattak, H. A., Shahid Khan, M. U., Ali, M., . . . Shah, S. (2019). Towards a fog enabled efficient car parking architecture. *IEEE Access*, *7*, 159100-159111. doi: 10.1109/access.2019.2950950

Aydin, B., Selvi, E., Tao, J., & Starek, M. J. (2019). Use of fire-extinguishing balls for a conceptual system of drone-assisted wildfire fighting. *Drones*, *3*, 17. doi: 10.3390/drones3010017

Babu, B. S., Ramanjaneyulu, T., Narayana, I. L., & Srikanth, K. (2017). Trends of iot. *International Journal of Engineering Trends and Technology*, *43*, 185-188. doi: 10.14445/22315381/ijett-v43p231

Bach, M. P., Živko Krstić, Seljan, S., & Turulja, L. (2019). Text mining for big data analysis in financial sector: A literature review. *Sustainability*, *11*, 1277. doi: 10.3390/su11051277

Bansal, S., & Kumar, D. (2020). Iot ecosystem: A survey on devices, gateways, operating systems, middleware and communication. *International Journal of Wireless Information Networks*, *27*, 340-364. doi: 10.1007/s10776-020-00483-7

Barmpoutis, P., Papaioannou, P., Dimitropoulos, K., & Grammalidis, N. (2020). A review on early forest fire detection systems using optical remote sensing. *Sensors (Basel, Switzerland)*, *20*. Retrieved from `https://pubmed.ncbi.nlm.nih.gov/33187292/` doi: 10.3390/s20226442

Barnes, W. L., Pagano, T. S., & Salomonson, V. V. (1998). Prelaunch characteristics of the moderate resolution imaging spectroradiometer (modis) on eos-am1. *IEEE Transactions on Geoscience and Remote Sensing*, *36*, 1088-1100. doi: 10.1109/36.700993

Barocas, S., & Nissenbaum, H. (2014). Big data's end run around procedural privacy protections. *Communications of the ACM*, *57*, 31-33. doi: 10.1145/2668897

Benomar, Z., Campobello, G., Segreto, A., Battaglia, F., Longo, F., Merlino, G., &

Puliafito, A. (2021). A fog-based architecture for latency-sensitive monitoring applications in industrial internet of things. *IEEE Internet of Things Journal*, *10*, 1908-1918. doi: 10.1109/jiot.2021.3138691

Benzekri, W., El, A., Moussaoui, O., & Berrajaa, M. (2020). Early forest fire detection system using wireless sensor network and deep learning. *International Journal of Advanced Computer Science and Applications*, *11*. doi: 10.14569/ijacsa.2020.0110564

Bhattarai, B. P., Paudyal, S., Luo, Y., Mohanpurkar, M., Cheung, K., Tonkoski, R., ... Zhang, X. (2019). Big data analytics in smart grids: state-of-the-art, challenges, opportunities, and future directions. *IET Smart Grid*, *2*, 141-154. doi: 10.1049/iet-stg.2018.0261

Bixio, L., Delzanno, G., Rebora, S., & Rulli, M. (2020). A flexible iot stream processing architecture based on microservices. *Information*, *11*, 565. doi: 10.3390/info11120565

Bouabdellah, K., Noureddine, H., & Larbi, S. (2013). Using wireless sensor networks for reliable forest fires detection. *Procedia Computer Science*, *19*, 794-801. doi: 10.1016/j.procs.2013.06.104

Bouaouad, A.-E., Cherradi, A., Assoul, S., & Souissi, N. (2020, 11). The key layers of iot architecture. In (p. 1-4). IEEE. doi: 10.1109/CloudTech49835.2020.9365919

Boughouas, M. L., Kissoum, Y., & Mazouzi, S. (2023). An educational supply chain management model as basis for a big data-driven decision support framework for the university. *2023 Third International Conference on Theoretical and Applicative Aspects of Computer Science (ICTAACS), Skikda, Algeria*. doi: 10.1109/ictaacs60400.2023.10449507

Boughouas, M. L., Kissoum, Y., Mazouzi, S., & Boussouf, W. (2023). Towards an efficient fog-based forest fire management architecture. *International Journal of Organizational and Collective Intelligence*, *13*, 1-20. doi: 10.4018/ijoci.317137

Boughouas, M. L., Kissoum, Y., Mouhssen, A., Karek, M. A., & Mazouzi, S. (2022). Towards a big educational data analytics. *2022 International Conference on Advanced Aspects of Software Engineering (ICAASE), Constantine, Algeria*.

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, *14*, 2. Retrieved from `https://datascience.codata.org/articles/10.5334/dsj-2015-002/` doi: 10.5334/dsj-2015-002

Campbell, J. P., DeBlois, P. B., & Oblinger, D. G. (2007). Academic analytics: A new tool for a new era. *EDUCAUSE review*, *42*, 40.

Carbone, P., Katsifodimos, A., Ewen, S., Markl, V., Haridi, S., & Tzoumas, K. (2015). Apache flink™: Stream and batch processing in a single engine. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, *36*, 28-38.

Chakraborty, C., & Kishor, A. (2022). Real-time cloud-based patient-centric monitoring using computational health systems. *IEEE Transactions on Computational Social Systems*, *9*, 1613-1623. doi: 10.1109/tcss.2022.3170375

Chandnani, N., & Khairnar, C. N. (2022). An analysis of architecture, framework, security and challenging aspects for data aggregation and routing techniques in iot wsns. *Theoretical Computer Science*, *929*, 95-113. doi: 10.1016/j.tcs.2022.06.032

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers Electrical Engineering*, *40*, 16-28. doi: 10.1016/j.compeleceng.2013.11.024

Chen, L., & Zhang, L. (2022). Design of ecological environment monitoring system based on internet of things technology. *Wireless Communications and Mobile Computing*, *2022*, 1-7. doi: 10.1155/2022/2531425

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and*

Applications, *19*, 171-209. doi: 10.1007/s11036-013-0489-0

Cheng, C.-F., Chen, Y.-C., & Lin, J. C.-W. (2020). A carrier-based sensor deployment algorithm for perception layer in the iot architecture. *IEEE Sensors Journal*, *20*, 10295-10305. doi: 10.1109/JSEN.2020.2989871

Choudhary, V., & Tanwar, S. (2023). A concise review on internet of things: Architecture and its enabling technologies. *Computational Intelligence for Engineering and Management Applications: Select Proceedings of CIEMA 2022*, 443-456. doi: 10.1007/978-981-19-8493-8_34

Dai, H.-N., Wang, H., Xu, G., Wan, J., & Imran, M. (2019). Big data analytics for manufacturing internet of things: opportunities, challenges and enabling technologies. *Enterprise Information Systems*, *14*, 1-25. doi: 10.1080/17517575.2019.1633689

Dallaev, R., Pisarenko, T., Talu, S., Sobola, D., Majzner, J., & Papež, N. (2023). Current applications and challenges of the internet of things. *New Trends in Computer Sciences*, *1*, 51-61. doi: 10.3846/ntcs.2023.17891

Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, *6*, 1-25. doi: 10.1186/s40537-019-0217-0

Dean, J., & Ghemawat, S. (2004). *Mapreduce: Simplified data processing on large clusters.*

Dekker, G., Pechenizkiy, M., & Vleeshouwers, J. (2009). *Predicting students drop out : a case study.* Retrieved from `https://research.tue.nl/en/publications/predicting-students-drop-out-a-case-study`

Dev, D., & Patgiri, R. (2014, 12). Performance evaluation of hdfs in big data management. In (p. 1-7). IEEE. doi: 10.1109/ICHPCA.2014.7045330

Devadas, T. J., & Subramanian, R. R. (2019). Paradigms for intelligent iot architecture. *Principles of Internet of Things (IoT) Ecosystem: Insight Paradigm*, 67-100. doi: 10.1007/978-3-030-33596-0_3

Dong, X. L., & Srivastava, D. (2013, 4). Big data integration. In (p. 1245-1248). IEEE. doi: 10.1109/ICDE.2013.6544914

Dragulinescu, A.-M. C., Marcu, I. M., & Vochin, M. C. (2023). Selecting multiple communication technologies in iot platforms. *Advanced Topics in Optoelectronics, Microelectronics, and Nanotechnologies XI*, *12493*, 232-241. doi: 10.1117/12.2643296

Fagroud, F. Z., Idrissi, N., Lahmar, E. H. B., Zellou, A., Toumi, H., Achtaich, K., & Filali, S. E. (2023). Iot architectures: A brief survey on layers' services. In (p. 641-652). doi: 10.1007/978-3-031-20601-6_52

FAO. (2020). *Global forest resources assessment 2020.* doi: 10.4060/ca9825en

Farooq, S. S., Shah, M. F., Sahar, M. S., Khan, G. M., Kausar, Z., & Farooq, M. U. (2022). Internet of things advancements and challenges in manufacturing industries. *Machine Tools*, 49-56. doi: 10.1201/9781003220985-4

Fawaz, M. M. (2022). Applications and security aspects of internet of things for way of smart world. *i-manager's journal on embedded systems*, *11*, 33. doi: 10.26634/jes.11.1.19136

Fawzy, M., Mahmoud, A. E., & Abdelfatah, A. M. (2022). Environmental and industrial applications using internet of things (iot). *A Roadmap for Enabling Industry 4.0 by Artificial Intelligence*, 141-167. doi: 10.1002/9781119905141.ch9

Fazel, E., Najafabadi, H. E., Rezaei, M., & Leung, H. (2023). Unlocking the power of mist computing through clustering techniques in iot networks. *Internet of Things*, *22*, 100710. doi: 10.1016/j.iot.2023.100710

Gaber, M. M., Aneiba, A., Basurra, S., Batty, O., Elmisery, A. M., Kovalchuk, Y., & Rehman, M. H. U. (2018). Internet of things and data mining: From applications to techniques and systems. *Wiley Interdisciplinary Reviews: Data Mining and*

Knowledge Discovery, *9*, e1292. doi: 10.1002/widm.1292

Gadekallu, T. R., Pham, Q.-V., Nguyen, D. C., Maddikunta, P. K. R., Deepa, N., Prabadevi, B., ... Hwang, W.-J. (2022). Blockchain for edge of things: Applications, opportunities, and challenges. *IEEE Internet of Things Journal*, *9*, 964-988. doi: 10.1109/jiot.2021.3119639

Gantz, J., & Reinsel, D. (2012). *The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east.* IDC iView: IDC Analyze the Future.

Ghani, N. A., Hamid, S., Hashem, I. A. T., & Ahmed, E. (2018). Social media big data analytics: A survey. *Computers in Human Behavior*, *101*, 417-428. doi: 10.1016/j.chb.2018.08.039

Gokhale, P., Bhat, O., & Bhat, S. (2018). Introduction to iot. *International Advanced Research Journal in Science, Engineering and Technology*, *5*, 41-44.

Grover, K., Kahali, D., Verma, S., & Subramanian, B. (2019). Wsn-based system for forest fire detection and mitigation. *Lecture Notes on Multidisciplinary Industrial Engineering*, 249-260. doi: 10.1007/978-981-13-7968-0_19

Guan, J., & Liang, G. (2023, 1). A research of convolutional neural network model deployment in low- to medium-performance microcontrollers. In (p. 44-50). ACM. doi: 10.1145/3585967.3585975

Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of things (iot): A vision, architectural elements, and future directions. *Future Generation Computer Systems*, *29*, 1645-1660. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0167739X13000241` doi: 10.1016/j.future.2013.01.010

Guenfaf, Y., & Zafoune, Y. (2023, 6). An iot ml-based system for energy efficiency in smart homes. In (p. 0198-0203). IEEE. doi: 10.1109/AIIoT58121.2023.10174484

Gupta, H., & Bharti, A. K. (2023). Fog computing iot: Overview, architecture and applications. *arXiv preprint arXiv:2304.08302*. doi: 10.48550/arxiv.2304.08302

Gupta, H., Vahid Dastjerdi, A., Ghosh, S. K., & Buyya, R. (2017). ifogsim: A toolkit for modeling and simulation of resource management techniques in the internet of things, edge and fog computing environments. *Software: Practice and Experience*, *47*, 1275-1296. doi: 10.1002/spe.2509

Gupta, P., & M, I. O. P. (2021, 6). A survey of application layer protocols for internet of things. In (p. 1-6). IEEE. doi: 10.1109/ICCICT50803.2021.9510140

Gupta, S., Nawal, M., Janu, N., & Goyal, D. (2022). Iot, enabling technologies, and sensor node deployment pattern in wsn. *ECS Transactions*, *107*, 7441-7455. doi: 10.1149/10701.7441ecst

Habib, M., & Jungthirapanich, C. (2009, 9). Research framework of educational supply chain management for the universities. In (p. 1-4). IEEE. doi: 10.1109/ICMSS.2009.5303124

Habib, M. M., & Jungthirapnaich, C. (2010, 04). Notice of retraction: An empirical research of integrated educational management for the universities. *LAP Lambert Academic Publishing*, 30–34. doi: 10.1109/ICIME.2010.5477486

Haghshenas, S. S., Guido, G., Vitale, A., & Ghoushchi, S. J. (2022, 9). Quantitative and qualitative analysis of internet of things (iot) in smart cities and its applications. In (p. 1-6). IEEE. doi: 10.1109/DASC/PiCom/CBDCom/Cy55231.2022.9927793

Han, J., E, H., Le, G., & Du, J. (2011, 10). Survey on nosql database. In (p. 363-366). IEEE. doi: 10.1109/ICPCA.2011.6106531

Hasan, M. M., Popp, J., & Oláh, J. (2020). Current landscape and influence of big data on finance. *Journal of Big Data*, *7*, 1-17. doi: 10.1186/s40537-020-00291-z

Heer, J., & Shneiderman, B. (2012). Interactive dynamics for visual analysis.

Communications of the ACM, *55*, 45-54. doi: 10.1145/2133806.2133821

Hegde, P., & Maddikunta, P. K. R. (2023). Amalgamation of blockchain with resource-constrained iot devices for healthcare applications – state of art, challenges and future directions. *International Journal of Cognitive Computing in Engineering*, *4*, 220-239. Retrieved from `https://www.sciencedirect.com/science/article/pii/S2666307423000220` doi: 10.1016/j.ijcce.2023.06.002

Honti, G. M., & Abonyi, J. (2019). A review of semantic sensor technologies in internet of things architectures. *Complexity*, *2019*. doi: 10.1155/2019/6473160

Hu, X., Du, L., & Li, Y. (2013). The perception layer information secu-rity scheme for internet of things based on lightweight cryptography. Atlantis Press. doi: 10.2991/icibet.2013.6

Hulwan, D. B., Kadu, S., Kale, D., Agrawal, K., Kandurke, P., & Rathore, J. S. (2023). Environmental monitoring system. *International Journal for Research in Applied Science Engineering Technology (IJRASET)*, *11*, 3590-3595. doi: 10.22214/ijraset.2023.51826

Hussein, M., Mohammed, Y. S., Galal, A. I., Abd-Elrahman, E., & Zorkany, M. (2022). Smart cognitive iot devices using multi-layer perception neural network on limited microcontroller. *Sensors*, *22*, 5106. doi: 10.3390/s22145106

Isha, Saxena, M., & Jha, C. K. (2022). Multilayered architecture for secure communication and transmission for internet of things. *Soft Computing for Security Applications: Proceedings of ICSCS 2022*, 691-699. doi: 10.1007/978-981-19-3590-9_54

Jana, S. H., Jana, B., & Lu, T. C. (2023). Multimedia authentication through digital tattooing scheme at iot perception layer for smart city. *IETE journal of research*, *69*, 5804-5812. doi: 10.1080/03772063.2023.2220310

Javanmardi, S., Shojafar, M., Mohammadi, R., Alazab, M., & Caruso, A. M. (2023). An sdn perspective iot-fog security: A survey. *Computer Networks*, *229*, 109732. doi: 10.1016/j.comnet.2023.109732

Jienan, D., Xiangning, C., & Shuai, C. (2021, 4). Overview of application layer protocol of internet of things. In (p. 922-926). IEEE. doi: 10.1109/ICCCS52626.2021.9449252

Jog, V. V., & Murugan, T. S. (2018). A critical analysis on the security architectures of internet of things: The road ahead. *Journal of Intelligent Systems*, *27*, 149-162. doi: 10.1515/jisys-2016-0032

Kabilamani, P., & Gomathy, C. (2021). Implementation of downlink physical channel processing architecture for nb-iot using lte/5g networks. *Wireless Personal Communications*, *116*, 3527-3551. doi: 10.1007/s11277-020-07863-5

Kakkar, L., Gupta, D., Saxena, S., & Tanwar, S. (2021). Iot architectures and its security: A review. *Proceedings of the Second International Conference on Information Management and Machine Intelligence: ICIMMI 2020*, 87-94. doi: 10.1007/978-981-15-9689-6_10

Kalatzis, N., Avgeris, M., Dechouniotis, D., Dechouniotis, D., Papadakis-Vlachopapadopoulos, K., Roussaki, I., & Papavassiliou, S. (2018). Edge computing in iot ecosystems for uav-enabled early fire detection. *IEEE International Conference on Smart Computing (SMARTCOMP)*. doi: 10.1109/SMARTCOMP.2018.00080

Kaur, P. (2023). Internet of things (iot) and big data analytics (bda) in healthcare. *Digital Transformation in Healthcare in Post-COVID-19 Times*, 45-57.

Khanra, S., Dhir, A., & Mäntymäki, M. (2020). Big data analytics and enterprises: a bibliometric synthesis of the literature. *Enterprise Information Systems*, *14*, 737-768. Retrieved from `https://www.tandfonline.com/doi/abs/10.1080/`

17517575.2020.1734241?journalCode=teis20 doi: 10.1080/17517575.2020.1734241

Khattar, S., Kaur, R., Verma, T., & Sharma, B. (2023, 3). Investigating challenges to internet of things (iot)applications and technologies. In (p. 1-6). IEEE. doi: 10.1109/ISCON57294.2023.10111967

Korte, A., Tiberius, V., & Brem, A. (2021). Internet of things (iot) technology research in business and management literature: Results from a co-citation analysis. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*, 2073-2090. doi: 10.3390/jtaer16060116

Kreps, J., Corp, L., & Rao, J. (2011). *Kafka: a distributed messaging system for log processing.* Retrieved from http://people.csail.mit.edu/matei/courses/2015/6.S897/readings/kafka.pdf

Kumar, R., ., R., Raj, M., ., V., ., R., & Raj, R. (2022). Internet of things (iot devices)-a literature review. *International Journal of Current Engineering and Technology*, *12*. doi: 10.14741/ijcet/v.12.3.3

Kumar, V., Gunner, S., Spyridopoulos, T., Vafeas, A., Pope, J., Yadav, P., ... Tryfonas, T. (2023). *Challenges in the design and implementation of iot testbeds in smart-cities: A systematic review.*

Lalit, M., Chawla, S. K., Rana, A. K., Nisar, K., Soomro, T. R., & Khan, M. A. (2022, 11). Iot networks: Security vulnerabilities of application layer protocols. In (p. 1-5). IEEE. doi: 10.1109/MACS56771.2022.10022971

Lambert, D. M., & Cooper, M. C. (2000). Issues in supply chain management. *Industrial Marketing Management*, *29*, 65-83. doi: 10.1016/s0019-8501(99)00113-3

Lampropoulos, G., Siakas, K., & Anastasiadis, T. (2018). Internet of things (iot) in industry: Contemporary application domains, innovative technologies and intelligent manufacturing. *International Journal of Advances in Scientific Research and Engineering*, *4*, 109-118. doi: 10.31695/ijasre.2018.32910

Laney, D. (2001). 3d data management: Controlling data volume, velocity and variety. *META group research note*, *6*.

Lekhwar, S., Yadav, S., & Singh, A. (2018). Big data analytics in retail. *Information and Communication Technology for Intelligent Systems*, *107*, 469-477. doi: 10.1007/978-981-13-1747-7_45

Le Monde, A. (Ed.). (2021, 08). En algérie, les incendies qui ont ravagé la kabylie sont éteints. *Le Monde.fr*. Retrieved 2022-06-02, from https://www.lemonde.fr/afrique/article/2021/08/18/en-algerie-les-incendies-qui-ont-ravage-la-kabylie-sont-eteints_6091748_3212.html

Lemzin, A. (2023). Streaming data processing. *Asian Journal of Research in Computer Science*, *15*, 11-21. doi: 10.9734/ajrcos/2023/v15i1311

Li, Y. (2014). The impact of disposition to privacy, website reputation and website familiarity on information privacy concerns. *Decision Support Systems*, *57*, 343-354. doi: 10.1016/j.dss.2013.09.018

Lou, H. H., Mukherjee, R., Wang, Z., Olsen, T., Diwekar, U., & Lin, S. (2022). A new area of utilizing industrial internet of things in environmental monitoring. *Frontiers in chemical engineering*, *4*. doi: 10.3389/fceng.2022.842514

Lovelock, C. H., & Wirtz, J. (2004). Services marketing: People, technology, strategy.

Madakam, S., Ramaswamy, R., & Tripathi, S. (2015). Internet of things (iot): A literature review. *Journal of Computer and Communications*, *03*, 164-173. doi: 10.4236/jcc.2015.35021

Mahmud, R., & Buyya, R. (2019). Modeling and simulation of fog and edge computing environments using ifogsim toolkit. *Fog and Edge Computing: Principles and*

Paradigms, 433-465. doi: 10.1002/9781119525080.ch17

Mazon-Olivo, B., & Pan, A. (2022). Internet of things: State-of-the-art, computing paradigms and reference architectures. *IEEE Latin America Transactions*, *20*, 49-63. doi: 10.1109/tla.2022.9662173

Mentzer, J. T., DeWitt, W., Keebler, J. S., Min, S., Nix, N. W., Smith, C. D., & Zacharia, Z. G. (2001). Defining supply chain management. *Journal of Business Logistics*, *22*, 1-25.

Mohamed, K. S. (2019). Iot networking and communication layer. *The Era of Internet of Things: Towards a Smart World*, 49-70. doi: 10.1007/978-3-030-18133-8_3

Mohamed, N., & Al-Jaroodi, J. (2014, 7). Real-time big data analytics: Applications and challenges. In (p. 305-310). IEEE. doi: 10.1109/HPCSim.2014.6903700

Morr, C. E., & Ali-Hassan, H. (2019). Descriptive, predictive, and prescriptive analytics. *Analytics in healthcare: a practical introduction*, 31-55. doi: 10.1007/978-3-030-04506-7_3

Mshvidobadze, T. (2022). Heterogeneous iot platform in network architecture. *International Journal of Computers*, *16*, 60-65. doi: 10.46300/9108.2022.16.12

Muniz, R. D. F., & Muniz, S. M. (2023). Investigation of iot-integrated smart homes. *Journal of operational and strategic analytics*, *1*, 42-45. doi: 10.56578/josa010106

Muntjir, M., Rahul, M., & Alhumyani, H. A. (2017). An analysis of internet of things (iot): Novel architectures, modern applications, security aspects and future scope with latest case studies. *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH TECHNOLOGY (IJERT)*, *6*.

Murphy, R. E., Ardanuy, P., Deluccia, F. J., Clement, J. E., & Schueler, C. F. (2006). The visible infrared imaging radiometer suite. *Earth Science Satellite Remote Sensing*, 199-223. doi: 10.1007/978-3-540-37293-6_11

Murthy, U. S., & Geerts, G. L. (2017). An rea ontology-based model for mapping big data to accounting information systems elements. *Journal of Information Systems*, *31*, 45-61. doi: 10.2308/isys-51803

N., D., Pham, Q.-V., Nguyen, D. C., Bhattacharya, S., Prabadevi, B., Gadekallu, T. R., ... Pathirana, P. N. (2022). A survey on blockchain for big data: Approaches, opportunities, and future directions. *Future Generation Computer Systems*, *131*. doi: 10.1016/j.future.2022.01.017

Naikwadi, S. N., Lokhande, S., Patil, B. V., & Mudalkar, P. K. (2022). An analysis of enabling technologies for the internet of things. *International Journal of Research Publication and Reviews*, *03*, 884-887. doi: 10.55248/gengpi.2022.31222

Naing, S. K., Oo, S. T., Oo, H. W. Y., Lin, P. Y., Nwe, K. M., & Thwin, M. M. S. (2023, 2). Iot network forensics based on transport layer. In (p. 312-317). IEEE. doi: 10.1109/ICCA51723.2023.10182142

Nejad, S. S., Khademzadeh, A., masoud Rahmani, A., & Broumandnia, A. (2023). Software defined fog platform. *International Journal of Electrical and Computer Engineering (IJECE)*, *13*, 5454. doi: 10.11591/ijece.v13i5.pp5454-5461

Nwadiugwu, W.-P., Ejaz, W., Kaneko, M., & Anpalagan, A. (2023). Neural-network-assisted packet accelerators for internet of things network systems. *IEEE internet of things journal*, *10*, 15238-15251. doi: 10.1109/jiot.2023.3265344

Patel, A. B., Birla, M., & Nair, U. (2012, 12). Addressing big data problem using hadoop and map reduce. In (p. 1-5). IEEE. doi: 10.1109/NUICONE.2012.6493198

Patel, K., Patel, S., & Scholar, P. (2016). Internet of things-iot: Definition, characteristics, architecture, enabling technologies, application future challenges. *International journal of engineering science and computing*, *6*. doi: 10.4010/2016.1482

Pateromichelakis, E., Dimopoulos, D., & Salkintzis, A. (2022). Netapps enabling application-layer analytics for vertical iot industry. *IEEE Internet of Things Magazine*, *5*, 130-135. doi: 10.1109/iotm.001.2200212

Pathik, B. B., Habib, M. M., & Chowdhury, M. T. (2012). Analysis of educational supply chain management model: a case study approach. *Proceedings of the 2012 International Conference on Industrial Engineering and Operations Management*, 9.

Perera, C., Zaslavsky, A., Christen, P., & Georgakopoulos, D. (2014). Context aware computing for the internet of things: A survey. *IEEE Communications Surveys Tutorials*, *16*, 414-454. doi: 10.1109/surv.2013.042313.00197

Perwej, Y., Haq, K., Parwej, F., & M., M. (2019). The internet of things (iot) and its application domains. *International Journal of Computer Applications*, *182*, 36-49. doi: 10.5120/ijca2019918763

Polepaka, S., Kumar, R. P. R., Nandakumar, S. D., Sureshkumar, R., & Vivekanandhan, V. (2023, 1). An insight on iot architecture and its applications. In (p. 1-6). IEEE. doi: 10.1109/ICCCI56745.2023.10128574

Poongodi, T., Gopal, R., & Saini, A. (2021). Iot architecture, communication technologies, and its applications. *Internet of Things, Artificial Intelligence and Blockchain Technology*, 1-24. doi: 10.1007/978-3-030-74150-1_1

Prabadevi, B., Deepa, N., Pham, Q.-V., Nguyen, D. C., M, P. K. R., G, T. R., ... Dobre, O. (2021). Toward blockchain for edge-of-things: A new paradigm, opportunities, and future directions. *IEEE Internet of Things Magazine*, *4*, 102-108. doi: 10.1109/iotm.0001.2000191

Pramanik, M. I., Lau, R. Y. K., Hossain, M. S., Rahoman, M. M., Debnath, S. K., Rashed, M. G., & Uddin, M. Z. (2020). Privacy preserving big data analytics: A critical analysis of state-of-the-art. *WIREs Data Mining and Knowledge Discovery*, *11*. doi: 10.1002/widm.1387

Rago, F. (2020, 7). Phenomenological layer structure of an intelligent agent for iot. In (p. 155-158). IEEE. doi: 10.1109/ICPICS50287.2020.9202220

Ramachandran, A., Gayathri, K., Aswathy, R. H., & Beevi, S. Z. (2023). Modeling of internet of things enabled sustainable environment air pollution monitoring system. *Global NEST*, *25*, 172-179. doi: 10.30955/gnj.004707

Ramzi, S. (2018, 04). Modeling the education supply chain with network dea model: The case of tunisia. *Journal of Quantitative Economics*, *17*, 525-540. doi: 10.1007/s40953-018-0131-x

Ranjan, R., Jayaraman, P. P., Villari, M., & Georgakopoulos, D. (2018). A note on resource management techniques and systems for big data workflow processing. *Computing*, *100*, 1-2. doi: 10.1007/s00607-018-0586-9

Rao, T. R., Mitra, P., Bhatt, R., & Goswami, A. (2018). The big data system, components, tools, and technologies: a survey. *Knowledge and Information Systems*, *60*, 1165-1245. doi: 10.1007/s10115-018-1248-0

Reddy, B. R., & Sujith, A. V. L. N. (2018). A comprehensive literature review on data analytics in iiot (industrial internet of things). *HELIX*, *8*, 2757-2764. doi: 10.29042/2018-2757-2764

Rehman, A., Awan, K. A., Din, I. U., Almogren, A., & Alabdulkareem, M. (2023). Fogtrust: Fog-integrated multi-leveled trust management mechanism for internet of things. *Technologies*, *11*, 27. doi: 10.3390/technologies11010027

Reinsel, D., Gantz, J., & Rydning, J. (2017). *Data age 2025: The evolution of data to life-critical don't focus on big data; focus on the data that's big sponsored by seagate the evolution of data to life-critical don't focus on big data; focus on the data that's big.* Retrieved from `https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf`

Richards, N. M., & King, J. H. (2013). Three paradoxes of big data..

Saa'da, R. J., Nimer, N. A., Altarawneh, A. M., Al-Anber, Z. A., & Al-Kalha, Z. (2022). Enabling a supply chain scor model for strategic decision-making and growth of the higher education sector in jordan. *Journal of Southwest Jiaotong University*, *57*.

Sadique, K. M., & Johannesson, P. (2021). Layered architecture for end-to-end security, trust, and privacy for the internet of things. *Intelligent Computing and Innovation on Data Science: Proceedings of ICTIDS 2021*, 289-298. doi: 10.1007/978-981-16-3153-5_32

Sailaja, D. K., & Rohitha, M. (2018). Literature survey on real world applications using internet of things. *SSRN Electronic Journal*. doi: 10.2139/ssrn.3165327

Sampson, S. E. (2000). Customer-supplier duality and bidirectional supply chains in service organizations. *International journal of service industry management*, *11*, 348-364.

Saoudi, M., Bounceur, A., Euler, R., & Kechadi, T. (2016). Data mining techniques applied to wireless sensor networks for early forest fire detection. *Proceedings of the International Conference on Internet of things and Cloud Computing*. doi: 10.1145/2896387.2900323

Saponara, S. (2022). *Enabling technologies for the internet of things.* CRC Press. doi: 10.1201/9781003338109

Sarvaiya, S. B. (2022). Analysis of iot data transfer messaging protocols on application layer. *International Journal for Research in Applied Science and Engineering Technology*, *10*, 1812-1819. doi: 10.22214/ijraset.2022.45604

Savignac, P., Petitdemange, E., Oger, R., Lauras, M., & Gac, J. (2022, 06). *Toward an educational supply chain operations reference model.* doi: 10.1109/GOL53975.2022.9820327

Sayak, P. (2020, 5). *Feature selection tutorial in python sklearn.*

Sezer, O. B., Dogdu, E., & Ozbayoglu, A. M. (2018). Context-aware computing, learning, and big data in internet of things: A survey. *IEEE Internet of Things Journal*, *5*, 1-27. doi: 10.1109/jiot.2017.2773600

Shearmur, R. (2015). Dazzled by data: Big data, the census and urban geography. *Urban Geography*, *36*, 965-968. doi: 10.1080/02723638.2015.1050922

Shi, Q., & Abdel-Aty, M. (2015). Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. *Transportation Research Part C: Emerging Technologies*, *58*, 380-394. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0968090X15000777` doi: 10.1016/j.trc.2015.02.022

Shilpa, V., Vidya, A., & Pattar, S. (2022). Mqtt based secure transport layer communication for mutual authentication in iot network. *Global Transitions Proceedings*, *3*, 60-66. doi: 10.1016/j.gltp.2022.04.015

Shwetha, C. A., & Suma, M. (2018). Remote management of smart equipments and services in manufacturing industry through iot applications using android compatible devices. *International Journal of Emerging Research in Management and Technology*, *6*, 290. doi: 10.23956/ijermt.v6i7.228

Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S., Shum, S. B., & Baker, R. S. J. D. (2011). Open learning analytics: an integrated modularized platform.

Singh, A., & Madaan, G. (2022). Integration of iot and big data technologies for higher education. *Edutech Enabled Teaching*, 27-44. doi: 10.1201/9781003254942-3

Singha, T. (2018). Iot and its smart applications. *International Journal of Information Science and Computing*, *5*. doi: 10.30954/2348-7437.1.2018.2

Sissodia, R., Rauthan, M. S., & Barthwal, V. (2023). Challenges in various applications using iot. *Handbook of Research on Machine Learning-Enabled IoT for Smart*

Applications Across Industries, 1-17. doi: 10.4018/978-1-6684-8785-3.ch001

Smys, S. (2019). Energy-aware security routing protocol for wsn in big-data applications. *Journal of ISMAC*, *01*, 39-55. doi: 10.36548/jismac.2019.1.004

Sreekantha, D. K., Koujalagi, A., Girish, T. M., & Sairam, K. V. S. S. S. S. (2021). Internet of things (iot) enabling technologies and applications—a study. In (p. 1425-1442). doi: 10.1007/978-981-15-3514-7_107

Sundarakumar, M. R., Mahadevan, G., Natchadalingam, R., Karthikeyan, G., Ashok, J., Manoharan, J. S., ... Velmurugadass, P. (2022). A comprehensive study and review of tuning the performance on database scalability in big data analytics. *Journal of Intelligent Fuzzy Systems*, *44*, 1-25. doi: 10.3233/jifs-223295

Sungheetha, A., & R, R. S. (2020). Real time monitoring and fire detection using internet of things and cloud based drones. *Journal of Soft Computing Paradigm*, *2*, 168-174. doi: 10.36548/jscp.2020.3.004

Supriyadi, S., Nurhadryani, Y., & Suroso, A. I. (2018). Website content analysis using clickstream data and apriori algorithm. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *16*, 2118-2126. doi: 10.12928/telkomnika.v16i5.7573

Swarna, C., & Ansari, Z. (2017). Apache pig-a data flow framework based on hadoop map reduce. *International Journal of Engineering Trends and Technology (IJETT)*, *50*, 271-275.

Ta, H. T. T., Nguyen, C. H., Le, H. T., Pham, N. T. T., Pham, H. T., & Trinh, N. T. (2023). Survey dataset on student perceptions and experiences of quality assurance in vietnamese universities. *Data in Brief*, *49*, 109305. Retrieved from `https://www.sciencedirect.com/science/article/pii/S2352340923004249` doi: 10.1016/j.dib.2023.109305

Tang, C., Garreau, D., & von Luxburg, U. (2018). When do random forests fail? *Advances in neural information processing systems*, *31*.

Tomar, R., Tiwari, R., & Sarishma. (2019). Information delivery system for early forest fire detection using internet of things. *Communications in Computer and Information Science*, 477-486. doi: 10.1007/978-981-13-9939-8_42

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, *25*, 44-56. doi: 10.1038/s41591-018-0300-7

Tsai, T.-A., & Lin, F. J. (2020). Enabling iot network slicing with network function virtualization. *Advances in Internet of Things*, *10*, 17-35. doi: 10.4236/ait.2020.103003

Tselikis, C. (2023). On the software architectures for fog-based secure iot deployments. *Proceedings of the 22nd European conference on information warfare and security*, *22*, 493-499. doi: 10.34190/eccws.22.1.1222

Ullah, I., Shukla, J. V., & Singh, D. K. (2023, 4). The applications, opportunities and challenges of iot in supply chain management: Insights from literature review. In (p. 1-5). IEEE. doi: 10.1109/ICETET-SIP58143.2023.10151507

Umer, M., Sadiq, S., Alhebshi, R. M., Sabir, M. F., Alsubai, S., Hejaili, A. A., ... Mohamed, A. (2023). Iot based smart home automation using blockchain and deep learning models. *PeerJ Computer Science*, *9*, e1332. doi: 10.7717/peerj-cs.1332

Upadhyay, S., Kumar, M., Upadhyay, A., Verma, S., Kavita, Kaur, M., ... Castillo, P. A. (2023). Challenges and limitation analysis of an iot-dependent system for deployment in smart healthcare using communication standards features. *Sensors*, *23*, 5155. doi: 10.3390/s23115155

Vaitsis, C., Hervatis, V., & Zary, N. (2016). Introduction to big data in education and its contribution to the quality improvement processes. *Big Data on Real-World*

Applications, *113*, 58. doi: 10.5772/63896

Varela, N., L, D.-M. J., Ospino, A., & Zelaya, N. A. L. (2020). Wireless sensor network for forest fire detection. *Procedia Computer Science*, *175*, 435-440. doi: 10.1016/j.procs.2020.07.061

Vasilescu, N.-G., Pocatilu, P., & Doinea, M. (2023). Iot security challenges for smart homes. In (p. 41-49). doi: 10.1007/978-981-19-6755-9_4

Vedak, M., & Kapadi, O. (2023). Iot and it's smart applications. *International Journal of Advanced Research in Science, Communication and Technology*, *3*, 349-353. doi: 10.48175/ijarsct-8168

Vita, M. D., Laurini, E., Rotilio, M., Stornelli, V., & Berardinis, P. D. (2023). Improving indoor well-being through iot: A methodology for user safety in confined spaces. *Lecture notes in networks and systems*, 767-779. doi: 10.1007/978-981-19-6634-7_54

Waham, M. A., & Lestari, F. (2019, 12). The integration of supply chain coordination in higher education supply chain framework. *International Business Education Journal*, *12*, 53-61. doi: 10.37134/ibej.vol12.5.2019

Whaiduzzaman, M., Barros, A., Chanda, M., Barman, S., Sultana, T., Rahman, M. S., … Fidge, C. (2022). A review of emerging technologies for iot-based smart cities. *Sensors*, *22*, 9271. doi: 10.3390/s22239271

White, T. E. (2012). *Hadoop : The definitive guide*. O'reilly Media.

Win, S. P., Phyu, W. L. L., & Thaw, T. Z. (2023, 2). Shedding light on technological treasures within iot application layer protocol. In (p. 164-169). IEEE. doi: 10.1109/ICCA51723.2023.10182211

Witten, I. H., & Frank, E. (2002). Data mining: practical machine learning tools and techniques with java implementations. *ACM SIGMOD Record*, *31*, 76-77. doi: 10.1145/507338.507355

Wu, C. K. (2021). Iot perception layer security. *Internet of Things Security. Advances in Computer Science and Technology. Springer, Singapore.*. doi: 10.1007/978-981-16-1372-2_6

Wu, H. (2022). Image self-coding algorithm based on iot perception layer. *Mobile Information Systems*, *2022*, 1-11. doi: 10.1155/2022/9910655

Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, *26*, 97-107. doi: 10.1109/tkde.2013.109

Yadav, S. K., Bharadwaj, B., & Pal, S. (2012, 03). Mining education data to predict student's retention: A comparative study. *arXiv:1203.2987 [cs]*.

Yang, F.-J. (2019). An extended idea about decision trees. *2019 International Conference on Computational Science and Computational Intelligence (CSCI).* doi: 10.1109/csci49370.2019.00068

Yang, Y. (2022). Iot-based air pollution monitoring system. *Highlights in Science, Engineering and Technology*, *17*, 299-307. doi: 10.54097/hset.v17i.2619

Yang, Z., Zhang, J., Jiang, Y., & Jin, Y. (2023). A self-organizing iot service perception algorithm based on human visual direction-sensitive system. *IEEE Internet of Things Journal*, *10*, 6193-6204.

Ying, S., Sindakis, S., Aggarwal, S., Chen, C., & Su, J. (2020). Managing big data in the retail industry of singapore: Examining the impact on customer satisfaction and organizational performance. *European Management Journal*, *39*. doi: 10.1016/j.emj.2020.04.001

Zaharia, M., Franklin, M. J., Ghodsi, A., Gonzalez, J., Shenker, S., Stoica, I., … Venkataraman, S. (2016). Apache spark. *Communications of the ACM*, *59*, 56-65. doi: 10.1145/2934664

Zahoor, S., & Mir, R. N. (2022). An iot enabled smart city: Assessing the applications, resource constraints, existing solutions and research directions. *International Journal of Computing and Digital Systems*, *12*, 279-293. doi: 10.12785/ijcds/120127

Zhang, Y., Yu, H., Zhou, W., & Man, M. (2022). Application and research of iot architecture for end-net-cloud edge computing. *Electronics*, *12*, 1. doi: 10.3390/electronics12010001

Zhao, X., Li, J., Chen, R., Li, C., Chen, Y., Zhang, T., & Zhang, Z. (2022). Design and implementation of environmental monitoring system based on multi-protocol fusion internet of things. *Journal of advanced computational intelligence and intelligent informatics*, *26*, 715-721. doi: 10.20965/jaciii.2022.p0715

Zhihan, L. (2022). Practical application of internet of things in the creation of intelligent services and environments. *Frontiers in the Internet of Things*, *1*, 912388. doi: 10.3389/friot.2022.912388

Zhu, L., Yu, F. R., Wang, Y., Ning, B., & Tang, T. (2019). Big data analytics in intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, *20*, 383-398. doi: 10.1109/tits.2018.2815678

ęski, Z. (2019). The perspective of using big data technology for the purposes of educational transactional analysis. *Edukacyjna Analiza Transakcyjna*, *8*, 81-88. doi: 10.16926/eat.2019.08.06