UNIVERSITY OF PIRAEUS

DEPARTMENT OF MARITIME STUDIES

MSc SHIPPING

"BIG DATA IN MARITIME INDUSTRY: THE NEXT BIG STEP"

Vandoros Christos

Thesis

Submitted to the Department of Maritime Studies of the University of Piraeus in the framework of the requirements for the award of the Diploma of Postgraduate Studies in MSc Shipping

Piraeus

November 2024

The present Thesis was unanimously approved by the Examination Committee of the Committee appointed by the General Assembly of the Department of Maritime Studies of the University of Piraeus in accordance with the Regulations of the Postgraduate Program in Shipping.

The members of the Committee:

- Alexandros Artikis (Supervisor)
- Dionisis Polemis
- Ioannis Lagoudis

The approval of the thesis by the Department of Maritime Studies of the University of Piraeus does not imply acceptance of the views of the author.

ACKNOWLEDGMENTS

I want to express my very sincere appreciation to all those who helped me and contributed, each in their own way, to completing my studies and more so during the period of this thesis.

First and foremost, I would like to express my deep gratitude to Professor Alexander Artiki, who gave me the opportunity to carry out this thesis under his supervision and guidance. Also, I would like to thank Professor Dionisis Polemis and Proffesor Ioannis Lagoudis for providing me their time, effort and guidance in order to present and finalize my thesis.

In conclusion, I would like to thank my family and friends, my family and friends for their support and encouragement throughout my studies at the University of Piraeus.

I am thankful and lucky to be able to have people like them in my life.

Περίληψη

Η ναυτιλιακή βιομηχανία, ακρογωνιαίος λίθος του παγκόσμιου εμπορίου και των μεταφορών, υφίσταται μια μετασχηματιστική αλλαγή με την έλευση των τεχνολογιών μεγάλων δεδομένων. Η παρούσα διπλωματική εργασία εμβαθύνει στην επιστήμη των μεγάλων δεδομένων στον τομέα της ναυτιλίας, αναδεικνύοντας τις δυνατότητές της να φέρει επανάσταση σε διάφορες πτυχές της βιομηχανίας.

Τα τελευταία χρόνια, η αύξηση των δεδομένων που παράγονται από τα πλοία, τους λιμένες και τα συναφή συστήματα έχει παρουσιάσει τόσο προκλήσεις όσο και ευκαιρίες. Στην παρούσα διπλωματική εργασία, διερευνούμε τις δυνατότητες της ανάλυσης μεγάλων δεδομένων για την ενίσχυση της ασφάλειας, τη βελτιστοποίηση των λειτουργιών και τη βελτίωση της συνολικής αποδοτικότητας στις ναυτιλιακές δραστηριότητες. Με την αξιοποίηση μεγάλων συνόλων δεδομένων από αισθητήρες, δορυφορικές εικόνες και συσκευές IoT (Διαδίκτυο των Πραγμάτων), ο κλάδος μπορεί να αποκτήσει πολύτιμες γνώσεις σχετικά με την απόδοση των πλοίων, την κατανάλωση καυσίμων, τη βελτιστοποίηση διαδρομών και τις περιβαλλοντικές επιπτώσεις.

Η ενσωμάτωση λύσεων μεγάλων δεδομένων διευκολύνει την παρακολούθηση των ναυτιλιακών επιχειρήσεων σε πραγματικό χρόνο, επιτρέποντας την προληπτική αντίδραση σε πιθανά ζητήματα, όπως οι διαταραχές του καιρού, οι βλάβες του εξοπλισμού και οι απειλές ασφαλείας.

Αντιμετωπίζουμε,επίσης,τις προκλήσεις που συνδέονται με την εφαρμογή των μεγάλων δεδομένων στον ναυτιλιακό τομέα, συμπεριλαμβανομένης της ασφάλειας των δεδομένων, της τυποποίησης και της ανάγκης για συνεργασία μεταξύ του κλάδου. Οι ανησυχίες σχετικά με την προστασία της ιδιωτικής ζωής, η κανονιστική συμμόρφωση και η ανάπτυξη ισχυρών πλαισίων διακυβέρνησης δεδομένων αποτελούν βασικά στοιχεία για τη διασφάλιση της υπεύθυνης και ηθικής χρήσης των μεγάλων δεδομένων στον ναυτιλιακό τομέα.

Εν κατακλείδι, η παρούσα διπλωματική εργασία υπογραμμίζει τις μετασχηματιστικές δυνατότητες των μεγάλων δεδομένων στη ναυτιλιακή βιομηχανία, δίνοντας έμφαση στην ικανότητά τους να εγκαινιάσουν μια νέα εποχή αποτελεσματικότητας, ασφάλειας και βιωσιμότητας. Καθώς ο κλάδος συνεχίζει να πλέει στις θάλασσες της ψηφιακής καινοτομίας, η κατανόηση και η αξιοποίηση της επιστήμης των μεγάλων δεδομένων θα είναι καθοριστικής σημασίας για τη ναυτιλιακή βιομηχανία που στοχεύει να παραμείνει μπροστά σε ένα διαρκώς εξελισσόμενο παγκόσμιο τοπίο.

Λέξεις Κλειδιά

Μεγάλα Δεδομένα, Διαδίκτυο των Πραγμάτων, Ναυτιλιακή Βιομηχανία, Ανάλυση Δεδομένων, Ασφάλεια Δεδομένων.

Abstract

The Maritime Industry, a cornerstone of global trade and transportation, is undergoing a transformative shift with the advent of big data technologies. This master thesis delves into the science of big data within the maritime field, highlighting its potential to revolutionize various aspects of industry.

In recent years, the growth of data generated by vessels, ports, and related systems has presented both challenges and opportunities. In this master thesis, we explore the potential of big data analytics to enhance safety, optimize operations, and improve overall efficiency in maritime activities. By harnessing large datasets from sensors, satellite imagery, and IoT devices, the industry can gain valuable insights into vessel performance, fuel consumption, route optimization, and environmental impact.

The integration of big data solutions facilitates real-time monitoring of maritime operations, enabling proactive responses to potential issues such as weather disruptions, equipment failures, and security threats.

We also address the challenges associated with big data implementation in the maritime sector, including data security, standardization, and the need for collaboration among the industry. Privacy concerns, regulatory compliance, and the development of robust data governance frameworks are essential components for ensuring responsible and ethical use of big data in the maritime domain.

In conclusion, this master thesis underscores the transformative potential of big data in the maritime industry, emphasizing its capacity to usher in a new era of efficiency, safety, and sustainability. As industry continues to navigate the seas of digital innovation, understanding and harnessing the science of big data will be pivotal for the maritime industry aiming to stay ahead in an ever-evolving global landscape.

Keywords

Big Data, Internet of Things, Maritime Industry, Data Analysis, Data Security

List of Charts:

List of Tables:

Table of Contexts:

IoT: Internet of Things, devices with sensors that connect and exchange data
FDR: False Discovery Rate
Fintech: Refers to firms using new technology to compete with traditional financial methods
RFID: Radio-Frequency Identification
GFS: Google File System
MapReduce: is a programming model or pattern within the Hadoop framework that is used to access big data stored in the Hadoop File System
AWS: Amazon Web Services
MA: Microsoft Azure, Azure offers a comprehensive set of intelligent solutions for data warehousing, advanced analytics on big data, and real-time streaming
GCP: Google Cloud Platform
GB: Gigabytes
AIS: Automatic Identification Systems
V2V: vehicle-to-vehicle
SRS: Satellite Remote Sensing
AUV: Autonomous Underwater Vehicles
USV: Unmanned Surface Vehicles
ADCP: Acoustic Doppler Current Profilers
CTD: Conductivity-Temperature-Depth
GPS: Global Positioning System
AIS: Automatic Identification Systems
ECDIS: Electronic Chart Display and Information Systems
WRS: Weather Routing Software
NLP: Natural Language Processing
Chat GPT: Chat Generative Pre-Trained Transformer
LCA: Life Cycle Assessment
VDR: Voyage Data Recorder
DPS: Dynamic Positioning System
OTLP: Open Telemetry Protocol
PMIS: Port Management Information Systems
VTS: Vessel Traffic Services
IMO: International Maritime Organization
MARS: Mariners Alerting and Reporting System
EEOI: Energy Efficiency Operational Indicator
LRIT: Long-Range Identification and Tracking
IMSO: International Mobile Satellite Organization
IQRS: Interquartile Range Method
KPI: Key Performance Indicator
OTD: On-Time Delivery
STT: Statistical Time Series
EIA: Environmental Impact Assessment

List of Abbreviations:

1. INTRODUCTION

1.1 BACKGROUND

The application of Big Data in the maritime industry has received a growing interest from academics and professionals in the recent past. There are research papers that discuss Big Data analytics, such as demonstrated by the following works which describe the usage of this technology in routes' optimization, predictive maintenance, and real-time tracking of the ships.

For instance, **M. Al Salim et al**. (2018) analyzed the effects of Big Data in reducing fuel consumption and as a result the cost of consuming fuel and greenhouse gases emissions. **UI Hassan et al**. (2024) investigates predictive maintenance, which uses information gathered from the engine's sensors to forecast when a failure is likely to happen, cuts down on the amount of time and money needed for maintenance.

Furthermore, several difficulties concerning Big Data application in the maritime industry have been also identified by various works, including those ones concerning data integration, data format standardization and cybersecurity risks **(Mirović et al., 2018)**. This work can be seen to offer a helpful background starting point for discussing the present characteristics of big data analytics in the maritime domain and how this may develop in the future.

The need for this research stems from the fact that the application of Big Data analytics can benefit the marine industry. The future challenges of the industry global trade growth, extension of environmental policies, and safety requirements put pressure on the industry to become more efficient and environmentally friendly.

Big Data analytics, which promotes better decision-making, increased operational effectiveness, and the generation of new ideas, thus offers the means necessary to meet these expectations.

Besides, knowing the current condition of Big Data analytics is important to assess future progress and issues amid the industry's tendencies toward automation and digitalization.

Conclusively, by exploring the current development in the application of Big Data analytics, this thesis will provide a rich supply of information useful to the industry stakeholders and policy makers as they seek to tackle challenges arising from digital disruption in logistics, particularly in the maritime industry.

1.2 PURPOSE AND RESEARCH QUESTIONS

The main purpose of this thesis is to identify the current and the new findings in Big Data analytics in the context of the maritime sector as well as assess how these innovations lay groundwork for future progress.

By analyzing the current technologies, methods and applications of Big Data in the maritime business, this study is expected to come up with the current trends and issues that are affecting the digitalization of the maritime industry.

In so doing, this thesis aims to present the future of the maritime industry via the Big Data analytics approach. We also propose some strategies that will help the stakeholders to harness this advancement to sustain their respective organizational competitiveness, sustainability and innovation especially under an increasingly complicated global environment.

Finally, through a step-by-step analysis, this thesis provides a methodology on how raw data, that can be collected from various sources, can also be presented in an organized and clear way to make business decisions easier.

The formulation of this purpose statement has given rise to the following research questions:

Q1. What are the current advancements in Big Data analytics within the maritime industry?

Q2. How are these advancements being applied to improve operational efficiency, safety, and sustainability in maritime operations?

Q3. What are the key challenges associated with the adoption and integration of Big Data analytics in the maritime industry?

These questions will guide the research, providing a structured approach to understanding the role of Big Data analytics in shaping the future of the maritime industry.

1.3 STRUCTURE OF THE THESIS

This master thesis is organized into nine chapters, each delving into essential principles and ideas surrounding big data and its application in the maritime industry. It begins by introducing the overarching concepts and fundamental aspects of big data and subsequently examines these concepts in relation to the field of maritime.

In Chapter 2, readers are introduced to the historical evolution, the fundamental concepts of big data, and its unique characteristics. This chapter provides a comprehensive overview of Big Data Science, detailing its foundational concepts, methodologies, and applications across various industries. It begins by defining Big Data and exploring its characteristics. The chapter delves into the evolution of data science as a discipline, highlighting the key technologies and analytical tools that have emerged to manage and interpret massive datasets. It also mentions the challenges, opportunities and implications of Big Data and the most used technologies.

Chapter 3 reviews the transformative applications of big data in the maritime industry and illustrates how advanced analytics is revolutionizing facets of maritime operations. It points out that big data has been utilized mainly in route optimization, fuel efficiency, predictive maintenance, cargo tracking, and port operations. Each application is accompanied by the description of the technologies and sources of data that relate to it. The chapter also discusses the tangible benefits of big data applications

Chapter 4 identifies emerging trends and prospects for data analytics in the maritime industry, noting how continuous innovation is likely to transform this sector. It begins with an analysis of next-generation technologies that are likely to radically change the maritime data analytics landscape.

The chapter also delves into the potential impact of these innovations on key areas such as autonomous shipping, smart ports, and enhanced supply chain visibility.

Chapter 5 points out numerous challenges, including regulatory and legal challenges, as well as the difficulties of applying big data analytics within maritime companies.

Chapter 6, through case studies and real-world examples, demonstrates how big data is enabling maritime companies to make more informed decisions, optimize their operations, and respond to industry challenges with greater agility.

Chapter 7 provides a detailed, step-by-step guide to conduct data analysis on a maritime dataset, showcasing practical applications of big data techniques within the industry. The chapter begins by introducing the specific maritime dataset used. It outlines the objectives of the analysis, focusing on uncovering patterns and insights that can inform operational improvements and strategic decision-making.

Additionally, it provides instructions for an on-hand data analysis guiding the reader on every step of the analysis (Data Collection, Data Cleaning, Data Analysis, Data Visualization).

Chapter 8 summarizes the key findings of the thesis, highlighting the transformative impact of big data analytics on the maritime industry

2. OVERVIEW OF BIG DATA SCIENCE

2.1 DEFINITION AND CHARACTERISTICS OF BIG DATA

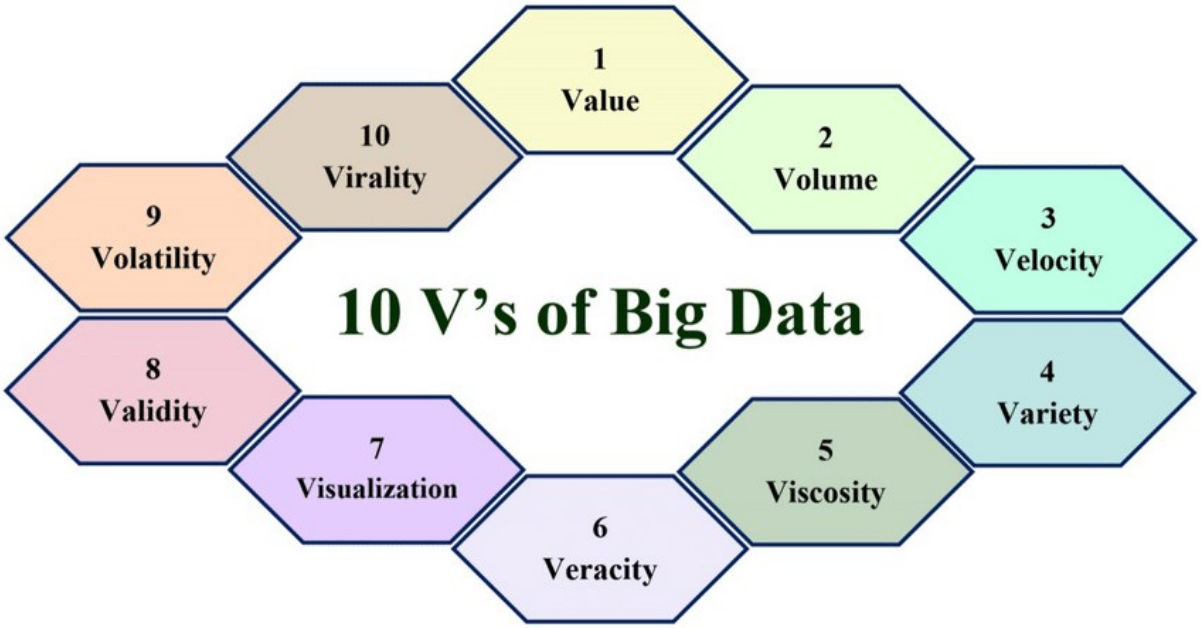
Big data refers to datasets that are too large or complex for traditional data-processing software to handle, often containing a high number of entries or intricate details that can lead to a higher false discovery rate. While there is no exact definition, big data generally represents a vast amount of information that is best utilized in large quantities **(Sreenivasan R., 2017).**

The quantity and scale of accessible data collections have experienced significant increase due to the accumulation of data from various sources such as mobile devices, a multitude of information-gathering IoT devices, drones, software records, cameras, microphones, RFID readers, and wireless sensor networks. Traditional relational database management systems and statistical software designed for desktop use often face challenges when it comes to handling and examining vast volumes of data. To overcome this hurdle, the utilization of massively parallel software that operates across multiple servers can be implemented to efficiently process and analyze large datasets.

Analyzing big data poses numerous challenges, including data collection, storage, analysis, search, transfer, visualization, querying, updating, privacy, and sourcing. there Initially focused on volume, variety, and velocity as the main characteristics but as the field develops, more characteristics have come to represent different Big Data facets. Big data also introduces complexities in terms of observational limitations and sampling. The concept of veracity has been introduced to highlight the importance of data quality and insightfulness in big data analysis. Neglecting to address veracity issues can result in costs and risks that outweigh the benefits of utilizing big data. Beyond just the size of data, big data involves applying advanced analytics methods to extract value from large datasets, uncovering new correlations for various purposes such as business trends, disease prevention, and crime prevention **(Katal A. et al., 2013).**

In Big Data science, the concept of "V's" represents various dimensions that define the challenges and characteristics of Big Data. Initially, there were three core V's, Volume, Variety, and Velocity, but as the field has evolved, more V's have been added to capture additional aspects of Big Data. Figure 1.1 presents the 10 V's of Big Data Science:

Figure 1.1: The 10 V's of Big Data Science



Source: Journal of Ambient Intelligence and Humanized Computing, 2020 [Link]

- **Value**: Is undeniably the most crucial aspect of big data. Without extracting business value from the data, the other attributes of big data hold no significance. Big Data holds immense potential, including the ability to gain a deeper understanding of customers, effectively target them, optimize processes, and enhance machine or business performance. However, it is essential to comprehend the potential and the complexities associated with big data before implementing a big data strategy **(Katal A. et al., 2013)**.

- **Volume**: Refers to the sheer amount of data generated and stored. Big data involves vast amounts of data that cannot be handled by traditional database systems. Social media interactions, sensor data, financial transactions, and

scientific experiments generate massive volumes of data **(Hashem I. A. T., 2015)**. This is not surprising given that over 90% of the data we have today has been generated in the last few years. For example, 300 hours of video are uploaded to YouTube every minute and an estimation in 2016 shows that global mobile traffic reaches 6.2 billion GB every month.

- **Velocity**: Refers to the speed at which data is being generated, produced, created, or refreshed **(Grolinger K. et al., 2016).** For instance, Facebook claims 600 TB of incoming data per day and Google handles over 3.5 billion search queries, equivalent to an average of 40,000 searches every second.

- **Variety**: The different types and sources of data. Big data can come in various formats, including structured, semi-structured, and unstructured data such as text, images, and video, log files, click data, machine, and sensor data **(Katal A. et al., 2013).**

- **Viscosity**: metaphorically describes the friction and inefficiencies encountered when handling large datasets. This includes challenges in data integration, quality, processing, scalability, security, and retrieval. High viscosity can lead to slow processing times and operational bottlenecks. To mitigate these issues, advanced technologies and methodologies such as distributed computing, data lakes, and in-memory processing are employed to streamline data management and enhance efficiency **(Han et al., 2011).**

- **Veracity**: The drop in veracity is an unfortunate characteristic of big data, which occurs as the properties of the data increase. Veracity refers to confidence or trust in the data and it's different from validity or volatility. It is determined by the provenance, reliability, and context of the data source, as well as its meaningfulness to the analysis **(O'Neil C. et al., 2016).**

  For instance, when analyzing a data set, it is important to consider who created the source and the methodology used for the data collection. These factors are crucial in determining the veracity of the information, which in turn helps us

understand the risks associated with making business decisions based on this data set.

- **Visualization**: Big data visualization tools (Power BI, Tableau, QlikView, Targit) are currently encountering technical obstacles due to the constraints of in-memory technology and inadequate scalability, functionality, and response time. Traditionally graphs are insufficient for plotting a billion data points, necessitating alternative methods like data clustering or the utilization of tree maps, sunbursts, parallel coordinates, circular network diagrams, or cone trees to represent the wanted data. Moreover, the numerous variables arising from the variety and velocity of big data, along with their intricate interrelationships, make the development of a meaningful visualization a challenging task **(Healy et al., 2018)**.

- **Validity**: Refers to the accuracy and reliability of data used for analysis. Ensuring validity involves verifying that data accurately represents the real-world phenomena it is intended to model and that it is free from errors or biases that could distort results. Validity is crucial for producing trustworthy insights and requires rigorous data validation processes and quality checks to maintain the integrity of analytical outcomes **(Chen et al., 2014)**.

- **Volatility**: Refers to the rapid and unpredictable changes in data that can complicate analysis and decision-making. This includes the frequent updates and shifts in data patterns, which require systems to adapt quickly to maintain accuracy and relevance. Managing volatility involves implementing agile data management strategies and real-time processing tools to ensure that insights remain current and actionable despite the dynamic nature of the data **(Khan et al., 2017)**.

- **Virality**: refers to the diversity and inconsistency of data types, formats, and sources, which can complicate data management and analysis. This variability requires robust data integration and processing techniques to handle different data structures and ensure coherent analysis. Effective strategies for managing

variability include using flexible data architectures and standardization practices to accommodate the diverse nature of big data and maintain analytical accuracy **(Chen et al., 2014).**

Various industries, including science, business, healthcare, advertising, and government, face challenges when dealing with large datasets in areas like Internet searches, fintech, healthcare analytics, geographic information systems, urban informatics, and business informatics. The use of predictive analytics, user behavior analytics, and other advanced data analytics techniques is essential in unlocking the potential value of big data across different fields and applications **(Khan et al., 2017).**

The concept of "Big Data" is subjective and varies based on the expertise of the analysts and the tools at their disposal, with the definition continuously evolving alongside technological advancements. Depending on the organization, the need to reassess data management strategies may arise when confronted with hundreds of gigabytes of data, while for others, the consideration of data size as a critical factor may only come into play when dealing with tens or hundreds of terabytes of data **(Sagiroglu S. et al., 2013).**

2.2 HISTORICAL EVOLUTION OF BIG DATA SCIENCE

Big data science has undergone a complex and multifaceted evolution throughout several decades, marking significant milestones in its development.

The period from the 1960s to the 1980s marked the Early Computing Era, characterized by the introduction of mainframe computers that revolutionized the way organizations processed large volumes of data. Despite this advancement, the concept of "big data" had not yet taken shape, highlighting the limitations in data processing capabilities during this time **(Codd E. F. 1970).**

In the 1990s, Data Warehousing emerged as a prominent trend as organizations began to focus on storing and managing large amounts of structured data efficiently. Technologies such as relational databases and data warehouses played a crucial role in enabling organizations to streamline data storage and retrieval processes, paving the way for more sophisticated data management practices **(Manyika et al., 2011).**

The 2000s witnessed the Rise of the Internet and Unstructured Data, as the proliferation of the Internet led to an exponential growth in unstructured data types such as text, images, and videos. Companies like Google and Yahoo faced significant challenges in managing vast amounts of data and responded by developing scalable solutions like GFS and MapReduce to address the complexities of handling unstructured data effectively **(Dean J. et al., 2004).**

The year 2008 marked a significant milestone with the Introduction of Hadoop, an open-source framework designed for distributed storage and processing of large datasets. Hadoop played a pivotal role in the big data ecosystem by offering a scalable and cost-effective solution for managing massive amounts of data, thereby revolutionizing the way organizations approached data processing and analysis **(Borthakur D. 2007).**

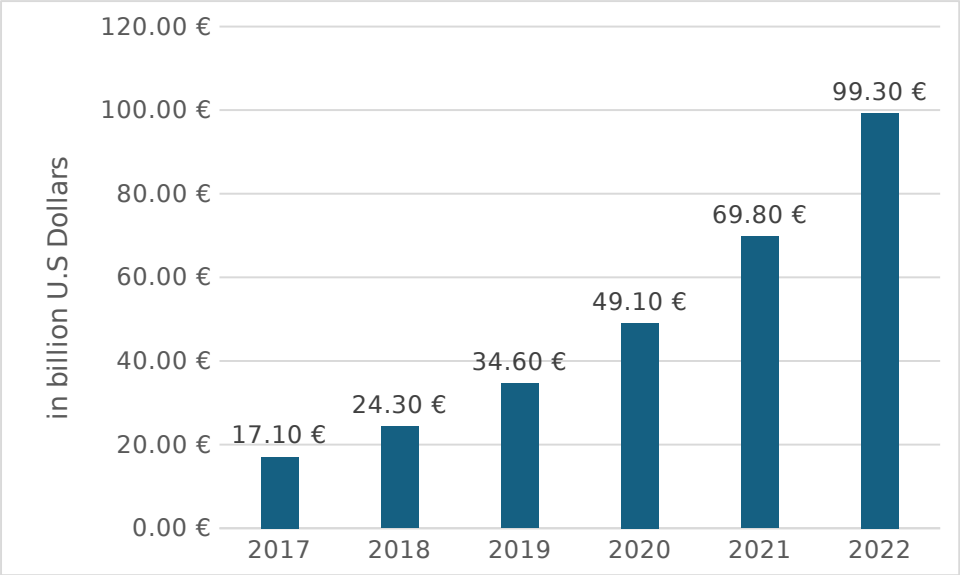


Chart 1.1 Forecasting Size of Hadoop and Big Data Market Worldwide

Source: Statista Research Department, 2016 [Link] **Created by**: Author

The 2010s witnessed the maturation of Big Data technologies, with the development of various tools and frameworks to tackle different aspects of the big data pipeline.

One notable advancement was the emergence of Apache Spark as a faster and more flexible alternative to MapReduce, providing enhanced capabilities for processing and

analyzing large datasets. Cloud computing services, such as AWS, MA, and GCP, played a significant role in the big data landscape during this decade. These services offered scalable infrastructure for processing big data, enabling organizations to leverage the power of the cloud for their data processing needs.

Additionally, data lakes gained popularity as repositories for storing diverse and voluminous raw data, providing a centralized and scalable solution for managing big data.

The integration of machine learning into big data analytics was a significant trend in the 2010s. This integration allowed organizations to extract valuable insights and predictions from vast datasets, paving the way for the emergence of the field of data science. By leveraging machine learning algorithms and techniques, organizations were able to uncover patterns, trends, and correlations within their big data, enabling them to make data-driven decisions and gain a competitive edge in various industries **(Domingos P. 2012).**

In 2012, NoSQL databases gained traction as a popular choice for handling unstructured and semi-structured data. These databases offered greater flexibility and scalability compared to traditional relational databases, making them well-suited for the dynamic and ever-growing nature of big data. The introduction of NoSQL databases provided organizations with more options for efficiently managing and analyzing their data.

Current State of Big Data, the field of big data science is experiencing rapid evolution, largely driven by advancements in artificial intelligence and machine learning. The integration of large-scale data analytics with sophisticated models, like GPT-4 and its successors, has significantly enhanced capabilities in natural language processing (NLP), predictive analytics, and decision-making systems. ChatGPT and similar models are being increasingly utilized for complex data interpretation, automating insights generation, and facilitating human-computer interactions. These AI systems leverage vast datasets to provide more nuanced and contextually aware responses, improving applications in customer service, content generation, and data-driven decision support **(Brown et al., 2020)**.

The growing emphasis on ethical AI and data privacy also underscores current challenges, as researchers and practitioners work to balance innovation with responsible data stewardship **(Binns et al., 2018)**.

Looking ahead, the future of Big Data in the 2020s is expected to be shaped by several emerging trends and technological advancements. Quantum Computing is anticipated to revolutionize data processing, allowing for the analysis of massive datasets at unprecedented speeds, which could open new frontiers in areas like cryptography and complex system modeling **(Preskill 2018)**. Ethical considerations surrounding AI and data usage will become increasingly important, focusing on mitigating biases, ensuring transparency, and safeguarding privacy **(Floridi et al., 2016)**.

Additionally, the development of advanced autonomous systems will rely on real-time Big Data analytics, further embedding AI into various operational contexts. Sustainability will also become a crucial focus, with efforts to minimize the environmental impact of data centers through energy-efficient practices and the adoption of renewable energy sources **(Nader N. et al., 2020)**. Lastly, the democratization of data is expected to make analytics more accessible, empowering a broader range of users within organizations to harness data-driven insights for decision-making **(Bean R. et al., 2018)**.

The historical evolution of big data science reflects the continuous need for innovative solutions to handle the ever-increasing volume, velocity, and variety of data generated in the modern era.

2.3 BIG DATA PLATFORMS

The current era of digital transformation has led to a significant increase in data generation, requiring the creation of specialized platforms to manage and analyze this large influx of information. Big data platforms serve as all-encompassing frameworks that empower businesses to store, process, and analyze extensive quantities of both structured and unstructured data. These platforms are specifically designed to address the challenges posed by the volume, velocity, and variety of data, allowing

organizations to extract valuable insights and improve their decision-making processes through advanced analytics methods **(Stobierski T. 2021).**

The different phases of a Big Data platform workflow can be categorized as follows:

1. **Data Collection**

   Big Data platforms gather information from a multitude of origins, including sensors, weblogs, social media, web scraping, data feeds, APIs, and data integration tools and different databases.

2. **Data Storage**

   The collected data is stored in repositories like Hadoop Distributed File System (HDFS), Amazon S3, or Google Cloud Storage, ensuring high availability, fault tolerance, and scalability through this distributed storage architecture.

3. **Data Processing**

   Data Processing encompasses operations like filtering, altering, and summarizing data to derive important conclusions. This process can utilize distributed processing systems like Apache Spark, Apache Flink, or Apache Storm.

4. **Data Analytics**

   Data analysis encompasses the exploration and interpretation of extensive data sets to derive valuable insights and detect patterns. This involves employing machine learning algorithms, data mining techniques, or visualization tools to gain a deeper understanding of the information at hand. The outcomes of the

analysis can subsequently inform data-driven decision-making, process optimization, opportunity identification, and resolution of intricate problems.

5. **Data Governance**

   The accuracy, consistency, integrity, relevance, and security of data are ensured during this stage. To implement data quality and governance effectively, organizations can utilize techniques such as data quality management, lineage tracking, and cataloging. By implementing strong measures for data quality assurance, organizations can trust the data they rely on for making decisions.

6. **Data Management**

   Efficient data management is essential for big data platforms, encompassing tasks like organization, storage, and retrieval of vast amounts of data. Various strategies like backup, recovery, and archiving are utilized to ensure effective data management, implementing fault tolerance and optimizing data retrieval for diverse use cases.

7. **Data interpretation and visualization**

   Now, with the data analysis and data management part done, it must be now interpreted and thereafter presented in easy-to-understand formats. This could be done in formats such as charts, graphs, or forms of visual aids. Data visualization with noted difficulty information helps in providing accessibility and plainly viewed results.

8. **Data story telling**

   The last step in the data analysis process is data storytelling, presenting the outcomes in an easy-to-understand, engaging way that is helpful to explain the analysis to nontechnical people. This approach is pivotal for the analysis to be communicated to non-technical audiences and to inform better data-driven decisions.

2.3.1 *COMPLEX CLOUD BIG DATA PLATFORMS*

The cloud-based services provided by Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure are known as Complex Cloud Big Data platforms. These platforms are specifically designed to handle and analyze large and intricate data sets.

- **Amazon Web Services (AWS)**

AWS, launched in 2006, offers a comprehensive range of tools and features that form an extensive ecosystem. These include AWS Lambda microservices, which enable the creation of scalable and serverless applications, Amazon OpenSearch Service for efficient search capabilities, Amazon Cognito for secure user authentication, AWS Glue for seamless data transformation, and Amazon Athena for in-depth data analysis. Additionally, AWS provides Amazon EMR for processing and analyzing large volumes of data, Amazon Kinesis for real-time data processing, and Amazon Redshift for efficient data warehousing. These are just a few examples of the wide array of tools and features available within the AWS ecosystem. With these tools, users can easily build and customize a data lake on the cloud, while benefiting from automatic configuration of core AWS services. Furthermore, AWS offers a user-friendly console that allows users to search and browse through available datasets, making the entire process streamlined and efficient **(Haponik A. 2024)**.

- **The Google Cloud Platform (GCP)**

The Google Cloud Platform (GCP) provides a comprehensive suite of cloud services that are designed to meet the needs of businesses and organizations. These services are modular in nature, allowing users to choose the specific components they require for their computing, data storage, data analytics, and machine learning needs. One of the key features of GCP is its ability to support purpose-built data and analytic open-source software clusters, such as Apache Spark. This means that users can quickly and easily set up and manage these clusters in as little as 90 seconds, according to Google.

In addition to this, GCP also offers a range of services specifically tailored for big data processing. These services include Google Cloud Storage, which provides a

reliable and scalable solution for data storage, Google BigQuery, which enables fast and interactive data analysis, and Google Cloud Dataflow, which allows both batch and real-time data processing.

Furthermore, GCP also provides Google Cloud Dataproc, a service that allows users to process big data using popular open-source technologies like Apache Hadoop, Spark, and BigQuery. Additionally, GCP offers AI Platform Notebooks and GPUs, as well as other analytics accelerators, to further enhance the capabilities of big data processing on the platform **(Knox K. 2024)**.

- **The Azure**

Microsoft's Azure platform offers a comprehensive range of features and tools that cater to the needs of developers, data scientists, and analysts when it comes to data storage. Azure seamlessly integrates with data warehouses, ensuring a high level of security, scalability, and adherence to the open HDFS standard. This means that there are no limitations on the size of data that can be stored, and users can leverage parallel analytics to process and analyze their data efficiently.

Azure goes beyond just providing storage capabilities and offers a suite of big data services to address various data processing needs. For instance, Azure Data Lake Storage is specifically designed for storing large volumes of big data, while Azure HDInsight enables users to process this data using popular frameworks like Apache Hadoop and Spark. Real-time data processing is made possible through Azure Stream Analytics, and for big data warehousing, Azure Synapse Analytics (formerly known as SQL DW) is available, providing a powerful solution for managing and analyzing large-scale datasets **(Collier M. 2015)**.

2.3.2 *MOST USED BIG DATA TECHNOLOGIES*

Apache Hadoop: Hadoop is a programming architecture and server software that is open source. Its main purpose is to efficiently store and analyze massive data sets by

utilizing numerous commodity servers in a clustered computing environment. One of its key features is the ability to replicate data, ensuring that no data is lost even in the event of a server or hardware failure.

This powerful platform not only offers essential tools and software for managing big data but also supports various applications that can be run on top of it. While Hadoop is compatible with operating systems like OS X, Linux, and Windows, it is most used on Ubuntu and other Linux variants due to their compatibility and performance advantages **(White T. 2015)**.

Cassandra: Apache Cassandra is a highly scalable, distributed NoSQL System, built for running, fast/complex datasets at huge capacities over numerous ordinary servers, where there is no centralized point for a failure. Due to its distributed nature, it offers very high availability and possesses good fault tolerance capabilities which makes it suitable for use in applications that can have severe consequences if they were not reliable such as mission-critical applications. Cassandra's approach to support flexible schema and its capability to handle the structure of the data with novelty semi structure data helps to manage the query response efficiently as per the data type. Due to the linear scalability of this architecture, it is ideal for heavy duty workloads like real-time analytics, the Internet of Things, and Open Telemetry Protocol (OTLP) **(Hewitt E. et al., 2016)**.

Apache Spark: Apache Spark functions as an open-source data-processing engine that has been meticulously crafted to provide the necessary computational speed and scalability essential for a wide array of applications including streaming data, graph data, machine learning, and artificial intelligence. One of the key distinguishing features of Spark is its ability to process and retain data in memory, eliminating the need for constant read and write operations to disk, thereby significantly enhancing its performance compared to alternatives like Apache Hadoop.

The versatility of Apache Spark extends to its deployment options, offering the flexibility of being deployed either on-premises or on popular cloud platforms such as Amazon Web Services, Google Cloud Platform, and Microsoft Azure. On-premises

deployment grants organizations greater autonomy over their data and computational resources, making it a preferred choice for entities with stringent security and compliance prerequisites. Nonetheless, it is important to note that deploying Spark on-premises necessitates substantial resource allocation in comparison to leveraging cloud-based solutions **(Zaharia M. et al., 2016).**

Databricks: Databricks serves as a cloud-native platform designed for the processing and analysis of large-scale data sets using Apache Spark as its foundation. It fosters a collaborative setting where data scientists, engineers, and business analysts can work together seamlessly, offering a range of functionalities like an interactive workspace, distributed computing capabilities, machine learning tools, and seamless integration with various popular big data solutions.

In addition to its core features, Databricks provides managed Spark clusters and cloud-based infrastructure to facilitate the execution of extensive big data workloads, streamlining the process for organizations looking to handle and dissect vast amounts of data efficiently. This managed service aspect simplifies the setup and maintenance of Spark clusters, enabling smoother operations for data processing and analysis tasks.

While Databricks primarily operates on cloud, it also extends a free community edition tailored for individuals and small teams seeking to explore and experiment with Apache Spark. The Community Edition offers a workspace with restricted computer resources, a subset of the complete Databricks platform features, and access to a limited selection of community-generated content and resources, making it an ideal starting point for those looking to familiarize themselves with the capabilities of Databricks and Apache Spark **(Dacey D. 2024).**

Apache Kafka: Apache Kafka is a distributed event streaming platform which deals with real time data feeds. Originally, Kafka was developed by LinkedIn to enable efficient real time data processing with high throughput and low latency. It is mostly used in constructing real time data ingestion and streaming platforms and the producers publish records to Kafka topics and consumers subscribe to them. Kafka's design can process millions of messages per second; thus, it would be suitable for use in applications such as log data processing, stream processing, and real-time data processing **(Kreps J. 2011).**

Snowflake: Snowflake offers a cloud-native data warehousing solution that encompasses data storage, processing, and analytical functionalities. It caters to both structured and semi-structured data types, offering a SQL interface for efficient data querying and analysis.

The platform operates as a fully managed service, taking care of all infrastructure and management responsibilities such as automatic scaling, backup, and recovery procedures, and ensuring robust security measures are in place to safeguard data integrity. Snowflake facilitates seamless integration with a wide range of data sources, enabling users to consolidate data from various platforms including other cloud-based data services and on-premises databases. This versatility allows for a comprehensive and unified approach to data management and analysis **(Agrawal S. 2024)**.

MongoDB: MongoDB is versatile, a non-relational database, which is easy to use, conventional and highly scalable. This is done through JSON-like documents; its data model is more natural and more flexible than a traditional relational data model. MongoDB is especially good for a high amount of unstructured and semi-structured data and that is why it is effective for content management, IoT and real-time analytics. Its horizontal scaling capability and vast query language help to facilitate complex operations on the data and offer high performance rates at the same time **(Chodorow K. et al., 2019)**.

Apache Storm: Apache Storm is an open-source distributed processing system that is specifically designed to handle and analyze massive amounts of data streams in real-time. It is a highly versatile tool that can be used for various purposes, including real-time analytics, online machine learning, and IoT applications. The way Storm operates is by breaking down data streams into smaller units of work known as "tasks." These tasks are then distributed across a cluster of machines, allowing for parallel processing of data. This distributed approach enables Storm to handle large volumes of data efficiently and effectively, ensuring high performance and scalability.

In addition to its powerful capabilities, Apache Storm is also compatible with popular cloud platforms such as Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure. This means that users have the flexibility to deploy Storm on these cloud platforms or even on-premises, depending on their specific requirements and preferences. This versatility makes Apache Storm a highly

accessible and adaptable solution for organizations of all sizes **(Reddy K. et al., 2014).**

Apache Flink: Apache Flink is an open-source stream-processing framework which in its concurrent is known for stream and batch data processing. It offers accurate and stateful computations of unbounded and bounded streaming datasets with short processing latency and high message processing rates. Some of the advanced features of Flink are complex event processing and Machine Learning and graph processing. The ability to work in large-scale data processing makes it ideal for large-scale data processing applications because it is highly scalable and fault tolerant. While each of the above features is beneficial to Flink, its window combined with various state operations is most helpful when processing stream data **(Tan P. 2020).**

2.4 OPPORTUNITIES AND APPLICATIONS IN BIG DATA UTILIZATION

The application of big data encompasses a wide range of opportunities and applications that are transforming industries and reshaping decision-making processes across various domains in the field of business analytics. Big data enables organizations to make data-driven decisions by processing a vast amount of data, which allows for a comprehensive understanding of market trends, consumer behavior, and operational efficiency.

Predictive analytics, which is a subset of big data applications, empowers businesses to anticipate future trends, optimize strategies, and proactively manage risks. In the healthcare sector, big data plays a pivotal role in clinical analytics. Big data analytics enables remote patient monitoring, which allows healthcare professionals to continuously assess patient data and intervene early with personalized healthcare solutions. The ability to process extensive datasets facilitates a deeper understanding of patient health, aiding in disease prevention and the development of personalized treatment plans **(Rossi et al., 2022)**.

Financial institutions leverage big data for fraud detection and risk management. By conducting sophisticated analysis of financial transactions and market trends, anomalies indicative of fraudulent activities can be identified. Additionally, big data

analytics provides a comprehensive view of risk factors, contributing to more effective risk management practices. In the retail sector, big data applications are beneficial in customer analytics. Retailers can harness data to personalize marketing strategies, enhance customer experience, and optimize inventory management **(Shruti M. 2024)**.

The utilization of big data in the manufacturing sector plays a crucial role in enabling predictive maintenance, which allows businesses to anticipate equipment failures and schedule maintenance activities in advance. By analyzing production data, quality control processes are significantly improved, as issues can be identified and resolved promptly to enhance product quality and reliability **(Choudhary A. K. et al., 2009).**

Telecommunications companies benefit from leveraging big data for network optimization, as it enables the analysis of network performance data to pinpoint congestion areas and improve service quality. Predictive analytics also plays a key role in customer churn prediction, helping companies take proactive measures to retain customers and enhance customer satisfaction **(Gupta J. 2024)**.

In the field of education, big data serves as a fundamental tool for providing personalized learning experiences by analyzing individual student data. This data-driven approach enhances institutional effectiveness by offering insights into student performance, optimizing resource allocation, and improving operational efficiency within educational institutions **(Daniel B. 2015)**.

Smart city initiatives rely on big data for urban planning purposes, allowing for the optimization of traffic flow, management of energy consumption, and enhancement of overall infrastructure planning. Public safety is also significantly improved through big data analytics, as it aids in predicting incidents, planning responses, and preventing potential issues before they occur **(Kitchin R. 2014)**.

Big data also plays a crucial role in marketing and advertising, particularly in facilitating targeted advertising by analyzing customer preferences and behavior. Social media analytics further contribute to understanding customer sentiment,

refining marketing strategies, and ultimately enhancing the overall brand perception in the market **(Wedel M. et al., 2016)**.

The use of big data analytics in environmental monitoring has proven to be highly advantageous. By harnessing the power of big data, climate modeling and prediction have become more accurate and reliable. This has greatly contributed to climate change research, enabling scientists to gain a deeper understanding of the Earth's climate patterns and make more informed decisions. Additionally, big data analytics have facilitated more effective resource management in areas such as water usage, energy consumption, and waste reduction. This has allowed organizations to optimize their resource allocation strategies and minimize their environmental impact **(Kharrazi A. et al., 2016)**.

In the realm of human resources, big data applications have revolutionized talent acquisition processes. Through the analysis of candidate data, organizations are able to make more informed hiring decisions, ensuring that they select the most suitable candidates for their positions. Furthermore, ongoing analysis of employee data has proven invaluable in enhancing retention strategies and improving overall workplace dynamics. By identifying patterns and trends within employee data, organizations can proactively address any issues or concerns, leading to a more engaged and satisfied workforce **(Marler J. H. et al., 2016)**.

The impact of big data extends far beyond specific industries, permeating nearly every sector. As technology and analytics methodologies continue to evolve, the possibilities for leveraging big data are expanding at an unprecedented rate. This offers transformative opportunities for innovation, efficiency, and improved decision-making across various domains. Organizations that embrace big data can tackle complex challenges with greater precision and drive positive outcomes. As a result, big data remains a focal point of ongoing developments, with its role in addressing societal and business challenges becoming increasingly significant **(Ahmed Memon et al., 2017)**.

2.5 CHALLENGES IN BIG DATA MANAGEMENT

The management of Big Data poses various challenges due to the characteristics of data volume, velocity, variety, veracity, and the overall complexity of data ecosystems. The immense amount of data necessitates scalable solutions for storage and processing. Efficiently managing and analyzing large datasets requires infrastructure capable of handling terabytes to petabytes of information. This challenge also involves the requirement for distributed computing frameworks, parallel processing, and horizontally scalable storage systems. The rapid generation of data, often in real-time, introduces an additional level of complexity. Conventional data processing systems may face challenges in handling the fast influx of data, which highlights the need to embrace real-time processing frameworks like Apache Kafka and Apache Flink. It becomes essential to ensure efficient data processing with minimal delay and high capacity, particularly in areas such as financial trading, IoT, and social media analytics **(Alabdullah B. et al., 2018)**.

Integrating different data formats, such as structured, semi-structured, and unstructured data, can be challenging. Traditional relational databases may not be suitable for handling diverse data types, leading organizations to rely on NoSQL databases, data lakes, and data warehouses to effectively manage various data sources. Ensuring the accuracy and reliability of data is a constant hurdle in the realm of Big Data. Inaccurate, inconsistent, or incomplete data can result in misguided decision-making. To address this, it is crucial to employ data cleansing and validation procedures and implement data quality frameworks to guarantee the dependability of the analyzed information. The challenges in Big Data ecosystems are exacerbated by their complexity. Effective management of data governance, metadata, and data lineage tracking is necessary to integrate technologies, platforms, and data sources. Ensuring interoperability between ecosystem components is vital for gaining valuable insights **(Luo et al., 2017)**.

The inclusion of sensitive information in Big Data environments amplifies security and privacy concerns. It is crucial to prioritize protecting data from unauthorized access, complying with regulations like GDPR or HIPAA, and implementing strong encryption mechanisms to uphold the trust and confidence of stakeholders. Big Data systems must possess scalability as a core necessity. They should be able to scale horizontally, by improving the capabilities of the current infrastructure. This

scalability is crucial to handle the increasing amount of data and ensure optimal performance for data processing and analytics applications **(Dabab et al., 2018)**.

Big Data management presents significant financial challenges. The costs associated with infrastructure development, investment in advanced analytics tools, and hiring skilled professionals can be substantial. Achieving a balance between cost management, performance optimization, and ensuring a return on investment requires meticulous planning and resource allocation **(Almeida 2018)**.

Addressing these challenges necessitates the implementation of efficient data governance. By setting guidelines for data quality, accessibility, and utilization, organizations can ensure data integrity, compliance, and accountability. To successfully manage Big Data, it is crucial to comprehend its intricacies and strategically utilize technologies and practices to unlock its full potential for informed decision-making and business prosperity.

2.6 ETHICAL AND SOCIAL IMPLICATIONS OF BIG DATA

As organizations harness the power of Big Data Analytics to drive decision-making and strategy formulation, the discussions around its ethical implications become increasingly crucial. Addressing issues related to privacy, security, and bias is essential to ensure that the benefits of this invaluable resource are maximized while minimizing any negative consequences that may arise from its misuse or exploitation.

- Ensure data privacy and security measures are in place.

Establishing and maintaining robust data privacy and security measures is paramount amid growing concerns surrounding the gathering and utilization of personal information. This involves implementing advanced encryption, secure authentication, and regular security audits to safeguard data during transmission and storage. Transparent privacy policies, explicit user consent, and adherence to data minimization principles contribute to ethical data handling. Regular software updates and employee training programs are crucial components of a comprehensive approach. Ultimately, a proactive and multifaceted strategy is essential to instill

confidence in users and stakeholders while navigating the complexities of personal data collection and processing **(Dabab et al., 2018)**.

- Transparency in data collection and usage to gain trust.

Transparency in data collection and usage is vital for cultivating trust in the digital age. Clear communication through accessible privacy policies, explicit consent mechanisms, and updates on any changes in data practices empowers users to make informed decisions about their personal information. This transparency not only aligns with privacy regulations but also fosters a sense of user agency and respect. Articulating the specific benefits users gain from data collection further contributes to a positive perception and strengthens the trust relationship between organizations and their users. In essence, transparency is a foundational element for establishing and reinforcing trust in the responsible handling of personal data **(Dabab et al., 2018)**.

- Use of data for societal benefit rather than exploitation.

The ethical use of data emphasizes prioritizing societal benefit over-exploitation, directing data towards initiatives that contribute positively to public interest and welfare. Transparency is crucial, with organizations openly communicating how data is utilized for social good, building trust with users. Striking a balance between the potential benefits of data-driven initiatives and ethical considerations, such as privacy, ensures alignment with societal values. Embracing a perspective where data serves the greater good establishes responsible stewardship, fostering a positive impact on communities and contributing to a sustainable relationship between data practices and societal well-being **(Dabab et al., 2018)**.

- Regular audits and oversight to monitor ethical practices.

Regular audits and oversight are integral for upholding ethical standards in data practices. These systematic reviews serve as proactive measures to identify and address potential lapses, ensuring adherence to ethical, legal, and industry standards. Internal mechanisms, like compliance officers or ethics committees, alongside

external oversight, contribute to accountability. Transparent communication of audit outcomes demonstrates a commitment to continuous improvement and builds trust with users. This ongoing scrutiny fosters a culture of ethical awareness, aligning data-driven practices with established norms and meeting the expectations of users and the broader community **(Dabab et al., 2018)**.

- Engage with stakeholders on ethical dilemmas and decisions.

Engaging with stakeholders on ethical dilemmas and decisions is vital for responsible data management. This collaborative approach ensures diverse perspectives are considered, incorporating input from users, employees, customers, regulators, and the community. By seeking feedback, organizations navigate complex ethical challenges collectively, promoting inclusivity and shared responsibility.

Transparent communication during this process builds trust, as stakeholders feel their perspectives are valued, and decisions are better understood. Overall, stakeholder engagement fosters a comprehensive, informed, and ethical decision-making process, contributing to a trustworthy relationship between organizations and those impacted by their data practices **(Dabab et al., 2018)**.

3. BIG DATA APPLICATIONS IN THE MARITIME INDUSTRY

The maritime industry is keen to operate Big Data applications, in order to focus on optimizing operations, improving safety, and enhancing decision-making through data analysis. The industry generates vast amounts of data from sources like vessel sensors and Automatic Identification Systems (AIS). By leveraging this data, maritime companies can predict equipment failures (predictive maintenance), optimize fuel consumption, improve route efficiency, and ensure compliance with environmental regulations **(Rodseth J. et al, 2016).**

Additionally, Big Data helps mitigate risks by monitoring vessel performance and tracking real-time conditions like sea traffic and weather. These applications reduce operational costs, enhance safety protocols, and contribute to more sustainable shipping practices **(Mirović M. et al., 2018)**.

3.1 DATA COLLECTION AND STORAGE IN MARITIME INDUSTRY

Data collection in the maritime industry is primarily driven by sensors, Automatic Identification Systems (AIS), radar systems, and satellite technologies. These systems capture vast amounts of data in real-time, including vessel location, speed, cargo weight, fuel consumption, weather conditions, and engine performance metrics **(Yang Q. et al., 2020)**. For example, AIS data tracks ships' movements globally, providing critical information for route optimization and collision avoidance **(Barrera S. et al., 2019)**. Additionally, onboard sensors monitor equipment health, enabling predictive maintenance by analyzing performance metrics like vibration levels and engine temperature **(Tinga T. et al., 2017)**.

Storing this data requires advanced infrastructure such as cloud-based platforms and decentralized data storage systems, given the large volume, velocity, and variety of data generated. Cloud storage allows for real-time data processing and analysis, facilitating collaboration between ships, ports, and regulatory bodies **(Hagemann M. 2023)**. The implementation of centralized data lakes ensures data is accessible and can be integrated with machine learning algorithms to derive actionable insights. Secure data storage is also crucial, as maritime operations involve sensitive information, necessitating encryption and compliance with cybersecurity standards **(Akpan F. et al., 2022)**.

3.1.1 TOOLS AND SOURCES FOR DATA COLLECTION

A huge amount of data is collected by ocean-going vessels during every voyage. This data can be used in analyzing and predicting marine traffic, analyzing ship performance for fuel efficiency, and weather routing. There have been several studies carried out on techniques and technologies of data collection, to analyze ship performance, and methods of prediction of the ship's route **(Liu M. et al., 2020).**

3.1.1.1. VESSEL RELATED DATA

These tools provide information related to ship movements, performance, and operations.

Automatic Identification Systems (AIS): It enhances maritime safety by facilitating the tracking and locating of vessels at sea. This technology was developed for the automation of navigation processes in order to improve safety and thus avert

collisions due to direct communication between ships. In addition, AIS provides critical information about the vessel to other ships and interested coastal entities, which is a tool of great importance for traffic management **(Weintrit A. 2009).**

GPS: Is a satellite navigation system that provides location information and accurate timing anywhere in the world under any weather conditions. It can operate at any point on or near the Earth's surface if there is a clear line of sight to at least four GPS satellites **(Hofmann-Wellenhof, B. et al. 1992).**

Onboard Sensors and IoT Devices: Ships are equipped with various sensors that collect data on engine performance, fuel consumption, cargo temperature, and environmental conditions. These sensors are part of the Internet of Things (IoT) network, enabling predictive maintenance and operational efficiency by continuously monitoring equipment **(Aslam S. 2023)**.

Electronic Chart Display and Information Systems (ECDIS)**:** is an innovative version of the navigational chart technology used in naval vessels and a few ships. The system enables the navigation team that navigates the ship to easily access any place and receive accurate directions for their mission **(Weintrit A. 2009)**.

Long-Range Identification and Tracking (LRIT**)** system is a global maritime safety initiative established by the International Maritime Organization (IMO). Its primary purpose is to enhance the tracking and monitoring of ships over long distances, especially for security and safety purposes. The LRIT system enables flag states, coastal states, and port authorities to receive positional data from ships, allowing them to monitor vessels' locations as they transit international waters **(IMSO.com)**.

Marine Radar Systems are used for collecting real-time information on wave heights, directions, and surface currents from installations along shore or onboard vessels to increase navigational safety **(Jerome P.Y. 2005).**

Vessel Traffic Services (VTS)**, l**ike air traffic control, VTS collects and processes data on vessel movements in busy maritime areas to ensure safe navigation and prevent collisions **(IMO)**.

Accident and Incident Databases collect historical data on maritime accidents, near-miss events, and other safety-related incidents to inform risk management and

safety protocols databases like Mariners Alerting and Reporting System (MARS) **(Nautinst).**

3.1.1.2. OCEAN AND ENVIRONMENTAL RELATED DATA

These tools collect data related to oceanographic conditions, weather, and environmental factors.

Satellite Remote Sensing (SRS) has the potential to capture extensive, real-time data regarding sea surface temperature, sea level, ocean color, and wave heights. Hence, satellite remote sensing is essential in weather forecasting, climate studies, and disaster management. Autonomous Underwater Vehicles (AUVs) and Unmanned Surface Vehicles (USVs) accomplish this by unmanned exploration and monitoring. AUVs are specifically useful in deep-sea data gathering surface temperature, salinity, and seafloor mapping, while USVs collect the same data but are more focused on the surface data like meteorological conditions and water quality. In this respect, USVs offer a low-cost solution for coastal area monitoring **(Zwally J. et al., 2001).**

Buoys and Drifters are essential for long-term monitoring. Some buoys are anchored at a fixed point, whereas drifters float freely due to ocean currents. Hence, both types of devices become essential for gathering information on wave conditions, salinity, and sea temperature that is used in climate modeling and for comprehending ocean circulation **(Daidola J. C. et al., 1991).**

Port Weather Stations are installed in port facilities to gather real-time weather data to ensure the safety of ship docking and departure activities.

Towed oceanographic instruments, like Conductivity-Temperature-Depth (CTD) Sensors and plankton nets, are deployed behind research vessels to gain high-resolution data on water properties and marine life along transects **(Dickson A. G. et al., 2007).**

Underwater gliders are self-propelled vehicles that, being controlled by buoyancy adjustment, are very useful in their ability to collect large data sets on water temperature, salinity, and density over great distances and long periods, thus fitting the bill for broad-scale, long-term oceanographic research **(Rudnick D. L. et al., 2004).**

Acoustic Doppler Current Profilers (ADCPs) measure the current velocities against water depth based on the Doppler effect and give valuable information about subsurface currents that form the basis for the study of patterns of ocean circulation and sediment transport **(Gordon R. L., 2011).**

Sonar Technology has many uses, including mapping the seafloor, finding underwater objects, and estimating fish populations. Consequently, it serves as a foundation for fisheries management and marine navigation **(Bollinger M.A. et al., 2017).**

3.1.1.3. PORT AND CARGO RELATED DATA

These tools focus on the operations within ports and the status of cargo.

Port Management Information Systems (PMIS) are digital platforms that manage port operations such as docking, cargo handling, and container movements **(Wartsila.com).**

Blockchain and Digital Shipping Platforms collect and store cargo tracking, customs documentation, and logistics data to improve transparency in maritime supply chains **(Czachorowski K. et al., 2019)**.

Cargo Monitoring Sensors are installed on containers or within cargo holds to monitor conditions such as temperature, humidity, and shocks, especially for sensitive or perishable goods **(Calamp.com).**

3.1.1.4. REGULATORY AND COMPLIENCE RELATED DATA

These sources provide data necessary for regulatory compliance, emissions monitoring, and environmental protection.

Emission Monitoring Systems are sensors that measure greenhouse gas emissions and other pollutants from ships to ensure compliance with International Maritime Organization (IMO) regulations **(Czachorowski, K. et al., 2019)**.

Energy Efficiency Operational Indicator (EEOI) **Data** provides data on a ship's fuel efficiency and carbon emissions, essential for meeting environmental standards **(Wartsila.com).**

The combination of these different sources of data underlies a very comprehensive approach to the monitoring and analysis of the shipping operations regarding critical applications, oceanographic research, maritime safety, and sustainable resource management.

3.1.2 DATA PROCESSING AND ANALYSIS

By implementing automated data gathering and storage technology, maritime stakeholders can build a solid foundation to create and incorporate the functionality needed for further data processing and analysis **(Mirović, M. et al., 2018)**.

The capabilities for processing and analyzing big data are an important step on the way toward the successful implementation of big data technology. After data are gathered from maritime stakeholders, it is prepared for analysis by performing various cleaning and transforming operations such as:

1. **Handling Missing Data in AIS and Sensor Data**

Filling Missing Positions (AIS Data): Ships may lose signal when crossing remote areas or due to bad weather. Imputation methods such as linear interpolation can be applied to fill missing positions based on previous and next positions **(Gao J. et al., 2023)**.

Predictive Maintenance Data Gaps: Missing sensor readings for fuel consumption or engine performance can be filled using historical trends or sensor-specific rules. For critical operations, forward/backward filling is used to maintain data continuity **(Gao J. et al., 2023)**.

2. **Outlier Detection and Treatment in Vessel Performance Data**

Identifying Erroneous Readings: Maritime sensor data (e.g., engine temperature or fuel consumption) can contain outliers due to equipment malfunctions. Z-score, a measure of how many standard deviations a data point is from the mean of a data set,

or Interquartile Range Method (IQR), a measure of the spread of the middle 50% of the dataset, methods that help detect anomalies that could indicate a need for maintenance **(Aggarwal C. 2017)**.

Correcting Navigation Errors: Outliers in AIS data, such as unrealistic ship speeds or abrupt course changes, can be flagged for correction or removed if caused by GPS errors or faulty transmissions **(Aggarwal C. 2017)**.

3. **Data Type Conversion and Parsing**

Date and Time Standardization: Ship logs, AIS, and other time-based data are often reported in different formats. Converting all timestamps into a standardized UTC format ensures consistent tracking and analysis **(Ritu J. 2024).**

Converting Coordinates: Maritime datasets may use different formats for geographic coordinates (degrees, minutes, and seconds vs. decimal). Converting to a standard format enables accurate geospatial analysis **(Ritu J. 2024)**.

4. **Dealing with Duplicate AIS and Sensor Data**

Removing Duplicate Transmissions: AIS data may contain duplicate messages due to multi-channel broadcasting. These duplicates need to be filtered out to avoid bias in vessel tracking and density analysis **(Volosencu C. et al., 2009).**

Duplicate Sensor Data: Ship sensors might transmit data redundantly in high-frequency environments, such as during docking or port operations. Aggregation techniques can be applied to summarize duplicate readings (e.g., averaging overtime intervals) **(Volosencu C. et al., 2009)**.

5. **Binning and Aggregating Geospatial Data**

Binning AIS Data for Route Analysis: Continuous AIS data is often binned into time intervals (e.g., every 10 minutes) or spatial grids (e.g., per nautical mile) to reduce data volume and highlight significant patterns in vessel routes and congestion areas **(Laurent L et al., 2019)**.

Aggregating Weather Data: Weather conditions such as wind speed and wave height can be aggregated over specific time intervals or regions to better understand their impact on maritime operations **(Laurent L et al., 2019)**.

6. **Handling Categorical Variables (Ports, Vessel Types)**

One-Hot Encoding of Vessel Types: Maritime datasets often contain categorical data like vessel types (e.g., cargo ship, tanker, passenger vessel). One-hot encoding these categories allows machine learning models to understand differences between ship classes in predictive tasks like fuel efficiency or arrival time **(Samuels J. A., 2024)**.

Port and Route Encoding: Ports and routes are often represented as categorical variables that need to be encoded for predictive modeling, such as estimating arrival times or traffic congestion at specific ports **(Samuels J. A., 2024).**

7. **Feature Engineering for Maritime Data**

Creating Speed and Acceleration Features: From AIS data, new features like average speed, maximum speed, and acceleration can be calculated to analyze vessel performance under different weather or sea conditions **(Effrosynidis D. et al., 2020)**.

Generating Environmental Impact Metrics: Combining fuel consumption and emission data allows for the creation of environmental impact scores that help analyze compliance with global emission standards and calculate carbon footprints **(Effrosynidis D. et al., 2020)**.

8. **Text Cleaning for Maritime Logs and Reports**

Cleaning Crew Logs: Textual data such as crew logs or incident reports can contain unstructured information. Tokenization (creating a digital representation of a real thing), stop-word removal, and lemmatization (reduce a word to its root form) can help transform this data into structured forms for further analysis of incidents or operational inefficiencies (**Manning C., 2009)**.

Standardizing Incident Reports: Maritime safety incident reports may use inconsistent terminology. Text cleaning and entity recognition help standardize

terminology for analysis of common issues like equipment failure or accident (**Manning C., 2009)**.

9. **Data Filtering and Sampling**

Filtering Out Inactive Vessels: Not all vessels are active in every dataset, so filtering out ships that have not moved in each period helps focus on active maritime traffic analysis **(Kim K. et al., 2018)**.

Sampling High-Frequency Data: In cases where data such as sensor readings or AIS messages are collected at high frequencies, a sampling approach can reduce data size while maintaining the quality of insights **(Kim K. et al., 2018)**.

An example of a Data Processing Analysis flow in the Maritime Sector could be:

Step 1: Use Filling Missing Positions**,** to handle missing AIS positions for vessels crossing remote ocean regions.

Step 2: Detect and remove outliers in fuel consumption data using IQR or Z-score method to correct extreme values due to sensor malfunctions.

Step3: Convert coordinates from degrees, minutes and seconds to decimal format for consistency in geospatial analysis.

Step 4: Normalize fuel consumption data for all vessels to a common scale in order to facilitate comparative analysis between ship classes.

Step 5: One-shot encode vessel types for use in machine learning models to predict fuel efficiency based on ship class and route.

These operations improve quality, consistency, and reliability of maritime data, ensuring that it can be used for insightful analysis, modeling, and decision-making.

3.2 BIG MARITIME DATA APPLICATIONS

3.2.1 PREDICTIVE MAINTENANCE AND CONDITION MONITORING

Predictive maintenance and condition monitoring outline reliability, efficiency, and safety in the vessel business.

Predictive maintenance involves the use of data analysis for forecasting the failure of vessel components. This is especially important for such basic facilities as engines, propulsive mechanisms, and navigation devices. Using the measurements of sensor and monitoring equipment, it is possible to identify precursors of possible problems, and thereby exclude possible failures – which, as a rule, are very costly in terms of repair and the occurrence of other interruptions and accidents at sea **(Mobley R. K. 2022)**.

Condition monitoring is one of the basic tasks of Predictive Maintenance, and it includes continuous assessment of the working machinery or equipment. Vibration, temperature, pressure, and level of lubrication are some of the basic parameters that will be continuously monitored in real-time through sensors installed on board in maritime applications. The data obtained will be analyzed in order to find anomalies, which may mean wear, tear, or other kinds of problems. It ensures the reliability of the vessel regarding critical systems and prolongs its lifespan by addressing equipment condition-related issues proactively **(Lebold M., et al. 1985)**.

Predictive maintenance is also useful to shipping operators as it helps them comply with safety and environmental regulations by preventing equipment failures that could lead to accidents or environmental hazards **(Tinga, T. et al.,2017)**, such as oil spills or engine malfunctions. By maintaining vessels proactively, operators reduce the risk of non-compliance with maritime laws, avoid hefty fines, and ensure safer, more sustainable operations.

The adoption of the above strategies is primarily found on innovative organizational technologies including the IoT sensors, advanced data analytics, and the technology of the remote age monitor that facilitates real-time comprehension of the disturbances. These cutting-edge technologies are quickly finding their way into industry and soon predictive maintenance and condition monitoring will play a very important role in sustainability and efficiency with regard to the operation of maritime industry **(Mobley R. K. 2022).**

3.2.2 ROUTE OPTIMIZATION AND REDUCING EMMISIONS BY USING BIG DATA

Big data has emerged as a transformative force in multiple industries, and one of the key areas where it has significant potential is in optimizing transportation routes to reduce fuel consumption and emissions. By using advanced analytics, machine learning, and statistical methods to process and analyze large datasets, companies can make smarter decisions that lead to more efficient transportation systems. This helps reduce fuel usage, lower emissions, and create more sustainable operations **(Smith, T. W. P., et al., 2014).**

Route optimization refers to finding the most efficient path for a vessel to travel between two points while considering multiple variables such as distance, fuel consumption, traffic, and road conditions **(Roh, M., 2013)**. Traditionally, route planning was based on fixed maps and heuristic approaches. However, with the integration of big data, route optimization has become more dynamic, adaptable, and intelligent.

Big data offers real-time traffic updates, utilizing data from GPS systems, traffic cameras, and crowdsourcing apps like Waze, a GPS-based navigation app and social traffic platform that provides real-time traffic updates and turn-by-turn navigation, to provide information that helps avoid congestion and save fuel. Additionally, weather patterns are analyzed through data on forecasts, allowing for the avoidance of routes prone to delays or increased fuel consumption due to heavy rain, snow, or high winds.

Furthermore, driver behavior is monitored through telematics devices that track performance metrics such as speed, braking, and idling times; adjusting these habits can lead to improved fuel efficiency. Lastly, information about road conditions, including road closures, construction, or accidents, can be integrated into the system to suggest alternative routes that save both time and fuel.

For instance, **UPS's ORION** (On-Road Integrated Optimization and Navigation) **(2016)**, system leverages big data to optimize delivery routes. UPS reports that this system saves millions of miles driven each year, leading to a reduction in fuel consumption by 10 million gallons annually and cutting CO2 emissions by 100,000 metric tons per year.

Maersk, is an another example, who has successfully implemented big data-based route optimization, leading to a 20% reduction in CO2 emissions per container between 2007 and 2021 **(Maersk Sustainability Report, 2021)**. Similarly, **DNV GL (2019)** found that using real-time data for route adjustments reduced emissions by up to 12% on long-distance voyages. In both cases, the how lies in leveraging predictive models and continuous monitoring to ensure that vessels take the most fuel-efficient paths, avoiding rough seas or unfavorable weather that could increase energy consumption.

Statistical analysis plays a crucial role in analyzing and interpreting the vast amounts of data that are collected. Some of the commonly used methods include:

Regression Analysis: This method helps predict future fuel consumption and emissions based on historical data. By understanding which factors (e.g., speed, traffic density, or vehicle type) most significantly impact fuel consumption, organizations can make informed decisions about optimizing routes.

Cluster Analysis: This is used to group vessels, trips, or routes based on common characteristics (e.g., geographical region, time of day). Identifying clusters helps in designing targeted strategies for reducing fuel consumption for different types of journeys.

Optimization Algorithms: Algorithms like **Dijkstra's shortest path**, which is used to find the shortest path from a given node to every other node, or **Genetic algorithms**, a method for solving constrained and unconstrained optimization problems based on natural selection process, can find the most efficient routes. By analyzing multiple factors like distance, fuel usage, and emissions, these algorithms suggest the optimal path for vehicles.

The primary goal of route optimization through big data is to reduce the environmental footprint of transportation. This is achieved by minimizing unnecessary idling, avoiding traffic jams, and reducing the total distance traveled. Fuel may account for 12-15% of a vessel's operating costs as high as 60% for large container ships.

According to Remoy Shipping (**Recogni, 2022**), the Blue Power EMS reduced fuel consumption in the company's offshore vessels by up to 15%.Blue Power EMS from Recogni is an advanced system that uses big data to enable ships' navigators to operate more efficiently, a visualization and analysis of power consumption is displayed on an intuitive dashboard, allowing the crew to make informed decisions about heading, speed and acceleration.

McKinsey & Company **(Tinnes E. et al., 2020)**, estimates that optimizing truck routes with data analytics could reduce fuel consumption by 10-12%, translating into significant reductions in greenhouse gas emissions.

Another study from **Mao, W. and Larsson, S. in 2022** shows that by using ML Power model (Machine Learning Algorithms) fuel consumption was reduced from 177.9 tons to 154.6 tons (13.1%) overcoming Empirical Power model (used to estimate the relation between speed and power) by 4.9%.

Study	**Reduction in Fuel Consumption**
Remoy Shipping	15%
McKinsey & Company	10-12%
Mao, W. and Larsson, S.	13.1%

Table 3.1. The impact of Big Data on Fuel Consumption **Created by**: Author

Optimized routes and fuel use give more reliable and timely deliveries. Hence, an overall increased effectiveness in maritime logistics. Technological advancements play a crucial role in these processes. Advanced technologies, like:

Weather Routing Software (WRS) further aids in this process by integrating forecasts to help vessels avoid adverse conditions while optimizing fuel use. **(Gong G., 2022)**

Dynamic Positioning System (DPS) automatically maintains a vessel's position and heading by using its own propellers and thrusters. This technology is particularly important for vessels operating in offshore environments, such as oil rigs, or during

port operations. DPS allows for precise control of the vessel's position without the need for anchoring, which reduces fuel consumption during operations that require a stationary position, such as drilling or loading/unloading cargo **(Holvik J. et al., 1998).**

Voyage Data Recorders (VDRs) collect and store critical navigation and operational data, which can be analyzed to optimize future routes and improve fuel efficiency. VDRs are mandatory on large vessels and are used extensively for post-voyage analysis, accident investigations, and improving navigational practices **(Shilavadra B., 2019).**

Trim Optimization Systems: Trim refers to the way a ship sits in the water and optimizing it can reduce drag and fuel consumption. Automated trim optimization systems use real-time data from sensors and analytics algorithms to adjust ballast and cargo distribution for optimal trim. These systems continuously calculate the ideal trim for the ship's current speed, cargo load, and sea conditions **(Reichel M. et al., 2014)**.

Big data and advanced analytics play a pivotal role in optimizing transportation routes and reducing both fuel consumption and emissions. By leveraging real-time data on traffic, weather, and road conditions, companies can significantly improve their efficiency. The use of statistical methods and optimization algorithms enables organizations to make data-driven decisions that lead to sustainable practices. As technology continues to evolve, the role of big data in transportation will only grow, driving further improvements in emissions reduction and fuel efficiency.

One of the vital indicators that helps in monitoring sailing behavior and, at the same time, serves as a basis for future decision-making, is the evaluation of a vessel's voyage performance. A generic voyage performance evaluation can be divided into three domains: overall performance, safety, and operational performance. The results for the overall performance evaluation may yield fuel consumption, cost, and time used for the whole voyage. This information enables management staff to understand the overall performance of the vessel during sailing. The performance also directly affects the operation and safety of the concerned vessel with regards to weather conditions, such as heavy sea, waves, wind, etc. In most cases, performance evaluation will incorporate ship energy efficiency performance and weather routing

performance to provide basic information on making proper decisions related to ship safety and energy efficiency routing **(Poulsen R., T., 2022).**

3.2.3 KEY PERFORMANCE INDICATORS

Key Performance Indicators (KPIs) play a vital role in leveraging data analytics to enhance fuel efficiency, operational effectiveness, and environmental compliance in the maritime industry. Some of the most critical KPIs include fuel consumption, energy efficiency, speed optimization, engine performance, and carbon emissions. Each of these KPIs is informed by vast datasets gathered from IoT sensors, satellite systems, and operational logs, which are then analyzed to optimize various aspects of maritime operations (**Acomi and Acomi, 2016**) some of the key KPIs are:

Vessel Utilization Rate relies on the optimization of the available carrying size for a given ship. This KPI is essential in the definition of organization profitability, given that it reflects the amount of cargo dispatched per voyage. Cargo load as compared to the available capacity can be tracked on data analytical platforms with strategic information on routes and ways to fully utilize the fleet.

Port Turnaround Time refers to time taken by a ship to complete the process of loading and discharging goods at the harbor. This is an effective measure of the line's operational efficiency since longer port stay times decrease total fleet effectiveness and enhance costs. This system tracks the arrival of ships, as well as, docking, and even departure to aid in the recognition of gaps and optimal timing. The optimization of the ports also ensures that the fleets are efficiently utilized, hence reducing costs such as the time the ships take in the ports rather than at sea.

On-Time Delivery (OTD) **Rate** measures the percentage of voyages that deliver cargo on schedule. In the logistics and shipping industries, on-time delivery is essential for customer satisfaction and operational efficiency. Data analytics platforms use real-time monitoring of ship location, weather conditions, and port traffic to predict and optimize delivery schedules, reducing delays and improving the OTD rate.

Safety Incident Frequency is an important element that most shipping companies track as a measure of compliance with safety standards, the occurrence of safety incidents or accidents on board ships. Data analytics may help to track several

operational characteristics related to both: personnel and machinery and identify the risk factors that may cause an accident. With this kind of KPI, companies will be able to have a safer environment for workers and reduce incidents of safety violations which may result in more loss of time and resources.

Maintenance Costs and Downtime, its general maintenance cost per vessel and downtime resulting from repairs. Through real-time predictive maintenance, data analytics systems keep track of the health of the ship components thus reducing breakdowns. Maritime systems use predictive analytics to help operators to schedule maintenance, reduce maintenance costs, and prevent a loss of business resulting in increased product availability of vessels.

Compliance with Environmental Regulations: with increasing regulations, especially from the International Maritime Organization (IMO), compliance with environmental standards (such as sulfur emissions, CO2 emissions, and ballast water treatment) is a key KPI. Data analytics helps track emissions in real-time, ensuring compliance with global standards. This KPI also overlaps with sustainability goals, allowing companies to monitor and reduce their environmental footprint.

Crew Performance and Efficiency is a KPI which relies on the performance and efficiency of the crew that are vital to safe and effective maritime operations. KPIs like crew overtime, rest hours compliance (as per IMO regulations), and crew productivity can be monitored using analytics tools that track work hours, shift patterns, and operational tasks. By analyzing these data points, ship operators can improve crew scheduling, ensure compliance with international labor regulations, and enhance overall productivity.

Cargo Damage and Loss Rate, this KPI tracks the percentage of cargo damaged or lost during transit. High cargo damage rates negatively impact customer satisfaction and revenue. By monitoring data from sensors that track cargo conditions (e.g., temperature, humidity, handling practices), companies can use analytics to reduce cargo loss or damage, ensuring better service delivery and profitability.

Fleet Operational Efficiency incorporates factors such as fuel usage, cargo carried, and distance covered to measure performance across the fleet. It is a KPI that can be examined at the level of the vessel as well as fleeted over time to check if there are inefficiencies or operational bottlenecks. The data analytics systems consolidate

data from vessels to provide a comprehensive fleet view and help operators to refine voyage management while also reducing energy use and improving asset up time.

Vessel Downtime or Utilization, this KPI measures the percentage of time a vessel is operational (i.e., in use for voyages or maintenance). Minimizing downtime is crucial for profitability in the shipping industry. Data analytics help track unplanned downtime due to mechanical failures or other operational issues and use predictive maintenance to ensure that vessels are operational more often, thereby increasing asset efficiency.

Cost per Mile or Voyage tracking the cost per mile, or voyage is a financial KPI that helps shipping companies manage operational expenses. This metric includes everything from fuel costs to crew wages and port fees. By using data analytics to break down each voyage into cost components, operators can identify inefficiencies or areas for cost-saving, such as optimizing routes, reducing unnecessary fuel consumption, or negotiating better port fees.

Customer Satisfaction for the maritime logistics business could be described by several KPIs, including on-time delivery, cargo condition upon arrival, and communication during the shipping process. Stronger operational performance would be ascertained by a high customer retention rate, with good service delivery. This will indeed be driven by key drivers that provide a clear view of how operations have impacted customer satisfaction: data analytics tools track customer interactions, delivery metrics, and reliability of service.

By focusing on a range of KPIs, maritime companies can enhance not just fuel efficiency but also overall operational performance, safety, environmental sustainability, and customer satisfaction. These KPIs, powered by data analytics, enable better decision-making, improve cost control, and enhance the long-term viability of maritime operations.

3.3 SAFETY AND SECURITY

While the maritime industry remains a very critical part of conducting world trade and transportation, types of safety and security issues related to these industries include many types of accidents, collisions, piracy, and illegal activities. Therefore, one of the

potential candidates in the line of providing solutions for better safety and security within the maritime domain must be data analytics.

Data analytics can be applied to a variety of data ranging from AIS data, radar data, satellite data, and social media data, by monitoring social media platforms, maritime authorities can track potential security threats, such as piracy, smuggling, or illegal fishing activities. Real-time data analysis helps identify emerging risks through posts, geotagged images, and communication patterns that could signal illegal activities or safety concerns **(Yaghoubi S. et al., 2015)**.

One of the major applications of data analytics in the maritime industry concerns its ability to detect and monitor risks and hazards **(Huang J. et al., 2023)**. Out of the analysis of historic and real-time data, one can identify patterns and trends indicative of potential risks, such as high traffic density, adverse weather conditions, or unusual vessel behavior. This information can be adopted by implementing proactive measures on mitigating such risks and preventing accidents. For example, data analytics can be used to develop traffic models that optimize vessel routes and speeds to reduce congestion and emissions while enhancing safety.

Another different key application of data analytics in this context lies in the analysis and prediction of safety and security incidents. Such would be feasible from historical data on incidents, where in an analysis by advanced statistical techniques or machine learning algorithms could identify its major causes and contributing factors **(Tsitsarolis A. et al., 2023).** This knowledge can then be used for developing models of risk assessment that evaluate the likelihood and seriousness of similar incidents in the future.

Further, data analytics can be used in developing predictive models that predict probable safety security breaches or safety-related issues from analyzed network traffic, system logs, or user behavior. As **Mirović (2018)** indicates, using predictive models, threats are already realized and responded to before they materialize, hence enhancing general safety and security within the maritime environment.

There is an appreciable level of interest in data analytics applications to improve prevention, detection, and response protocols as it relates to maritime safety and security. A few practical use cases that apply data analytics, comprising several academic, commercial, and publicly funded projects, are presented below:

1. **Big Data for Maritime Security (BESST)**: A European Union-funded initiative uses data analytics to enhance surveillance of illegal activities, such as drug smuggling and human trafficking, by analyzing ship movements and social media posts **(Gatehouse Maritime)**.
2. **Automatic Identification System (AIS) Data**: Commercial platforms like Windward, a maritime intelligence and analytics platform that leverages artificial intelligence (AI) and big data to provide insights into maritime activity to detect unusual ship behaviors, including piracy risks, by employing predictive analytics **(Yang et al., 2024).**
3. **SENTINEL Project**: A publicly funded project that integrates social media analytics with satellite data to improve the detection of maritime risks like oil spills or distress signals **(Santamaria et al., 2017).**

Many of these efforts could find and analyze patterns in relation to the many years of public geo-located maritime dataset collection (**Schoier G., et al., 2017)**, collectively with respect to any certain safety incidents or security breaches.

Due to a lack of relevant, up-to-date terrestrial information, safety and security threats like hijacking, piracy, rogue sailing, distress calls, and others have historically been challenging to analyze. However, the quick emergence of a variety of public maritime datasets has created new opportunities for maritime intelligence and analyses.

Another way that maritime safety and security can be enhanced is by new advances in data analytics relevant to big datasets. When public maritime geo-located datasets become too large to manage effectively, maritime geospatial analysis could benefit from algorithms such as clustering, tracking, graph analysis, and spatiotemporal analysis, which have been successfully developed for crime analysis and similar tasks **(Schoier G. et al., 2017).**

Finally, there is a chance for feedback about the identified insights, patterns, and trends into the consideration of maritime safety and security monitoring and management protocols, as well as into the relevant legislative and regulatory frameworks.

Datasets containing instances of safety and security threats can be analyzed as discrete events, forming the foundation for event retrospective analysis **(Vessel**

Finder). Once a specific threat event is identified, data analytics can query historical datasets across both time and geographic space to detect similar incidents. This approach enhances maritime safety and security by allowing authorities to immediately test the reliability of the event definition through data analysis, offering insights into patterns or anomalies. Over time, this continuous analysis helps build a clearer context for understanding safety risks, improving detection, prevention, and response strategies across various maritime environments.

3.4 LOGISTICS AND SUPPLY CHAIN

Data Analytics used to be a luxury and now it is almost imperative in driving Digital Transformation that helps businesses take data driven decisioning along with operational efficiency through advanced proactive changes based on market dynamics. By including data analytics, greater end-to-end supply chain visibility for demand forecasting vulnerabilities and risks, better inventory optimization and effective risk management are available. Predictive analytics tools let companies analyze historical data to predict changes in demand, reducing the risk of stockouts or overproduction by avoiding holding costs, and helping ensure that inventory levels match customer needs (**Ivanov et al., 2020).**

One of the major ways in which analytics has contributed a great deal to supply chain management involves the development and enhancement of demand forecasting. Advanced predictive models leverage machine learning algorithms and large volumes of data from historical sales, weather conditions, and trends in consumer behavior in making very accurate forecasts. For example, according to **Tiwari et al. (2021)**, predictive analytics has the potential to enable the firm to achieve an upfront view of future spikes in demand and develop mechanisms for resource allocations. The ability to do so mitigates uncertain demand patterns, optimizes resource utilization, and aligns supply and demand better.

Inventory management is another area where data analytics brings substantial value. The application of machine learning and data mining techniques helps in maintaining optimal stock levels and minimizing excess inventory costs. By analyzing sales patterns, lead times, and supplier performance, companies can better predict

replenishment requirements, reduce overstocking, and minimize waste **(Wardle D. 2024).** Advanced warehouse analytics, which use sensor and barcode data, also provide insights into warehouse efficiency, identifying areas where improvements are needed, such as optimizing picking processes or reorganizing storage layouts.

Moreover, data analytics supports better supplier relationship management and enhances resilience in the supply chain. By analyzing supplier data, including lead times, quality, and reliability, businesses can identify the most dependable suppliers and diversify their sourcing strategies. This ensures reduced dependency on a single source and minimizes vulnerabilities in case of disruptions **(Ivanov et al., 2020).** For instance, scenario modeling allows supply chain managers to simulate potential disruptions and design contingency plans to maintain smooth operations.

In the maritime environment, visibility into the supply chain is crucial due to the vast amounts of data from various sources. This includes shipboard systems, satellite tracking of vessel routes, and real-time weather data. Effective management of these data sources is essential for maintaining operational efficiency and safety **(Wamba S. F. et al., 2019)**. All these make it difficult for data users, who are overwhelmed with increased volume and variety.

Apart from that, the process of decision-making in such a supply chain is strongly oriented to experience-based decision-making, with risks related to delays in cargo delivery and missed deadlines. With mechanisms of data collection and electronic data transmission, together with its processing and analysis by machine learning algorithms in, for example, a Big Data warehouse, users of the supply chain will have an opportunity for wider insight into the supply chain and better management of supply chain-based decision-making **(Darvazeh S. et al., 2020)**. The indicative case study of three-sided data supply chain analysis provides the possibility for a higher degree of efficiency and decision-making quality in logistics operators.

In this regard, the value of IoT in logistics and supply chains can be recognized only when there is a growing demand for sharing real-time data. This will include physical and information flows among various parties involved in logistics and supply tasks that can optimize their related functions and operations by real-time cooperation. Such data-driven decisions can further be strengthened in a sustainable manner by

advanced analytical abilities brought in by AI, machine learning, and deep learning **(Ieromonachou P. et al., 2017).**

3.5 ENVIRONMENTAL IMPACT ASSESMENT

Big data analytics play a transformative role in emission reduction and estimation within the maritime industry, supporting efforts to meet the International Maritime Organization (IMO) target of cutting CO2 emissions by 40% by 2030. This is achieved through various data-driven approaches that optimize operations, predict maintenance needs, and ensure compliance with emissions regulations.

Data analytics can be useful in judging the impact of maritime activities on marine biodiversity through the identification of changes in the population of marine species and habitats. It reanalyzes data drawn from underwater sensors, drone footage, and ecological surveys to detect disruptions caused by shipping traffic, noise pollution, and construction activities **(Holling C. S. et al., 1978).**

Slow steaming, a practice of operating ships at reduced speeds to lower fuel consumption, is another area where big data adds value. By analyzing engine performance data in real-time, big data analytics can determine the optimal speed for balancing fuel efficiency with timely deliveries. The system continuously monitors engine outputs, weather conditions, and cargo weight to recommend speed adjustments that minimize fuel use without significantly affecting schedules. The **International Transport Forum (2018)** showed that when combined with big data analytics, slow steaming reduced emissions by 10-20%.

Big data also plays a vital role in predictive maintenance, which helps shipping companies maintain fuel efficiency by ensuring that engines and other equipment are operating optimally. Sensors on the ship monitor the performance of key components, collecting data on temperature, pressure, and vibrations. This data is then analyzed to detect early signs of wear or malfunction before they lead to inefficiencies. For example, if big data analysis shows that an engine is overheating or vibrating unusually, maintenance can be scheduled to prevent fuel inefficiency. Rolls-Royce Marine's predictive maintenance system, which uses big data, reduced fuel consumption by 5% by identifying potential issues early and ensuring engines operated at peak efficiency **(Rolls-Royce Marine, 2019)**.

Another critical area is real-time emissions monitoring and regulatory compliance. Ships are required to report their emissions of pollutants such as CO2, sulfur oxides (SOx), and nitrogen oxides (NOx) under IMO regulations. Big data systems use onboard sensors to collect and analyze emissions data in real-time, ensuring compliance with these standards. This continuous data stream allows for automated reporting and adjustments to operations, such as altering fuel mixtures or adjusting engine performance to reduce pollutants. A 2020 study by the **International Council on Clean Transportation (ICCT)** showed that big data-driven emissions tracking improved compliance with the IMO's sulfur cap by 15%, by accurately measuring and adjusting sulfur emissions in real-time.

In ports, big data optimizes port operations to reduce emissions during docking and loading activities. Smart port technologies use real-time data from ship arrivals, weather forecasts, and cargo handling systems to minimize the time ships spend idling in port, which is a major source of emissions. By coordinating vessel arrivals with port schedules and optimizing loading/unloading processes, ships can reduce the time spent burning fuel while stationary. The Port of Rotterdam, for example, reduced emissions in its vicinity by 15% through such smart port optimizations **(Port of Rotterdam Authority, 2020)**.

Finally, big data also plays a role in carbon footprint calculation across the entire shipping supply chain. By collecting data from various points in the supply chain such as fuel usage, port emissions, and last-mile deliveries, big data platforms can calculate the total carbon footprint of shipping operations. This enables companies to identify areas where emissions can be reduced, such as through more efficient logistics or adopting alternative fuels like LNG. The EU's Horizon 2020 study **(European Commission, 2021)** found that emissions across the supply chain could be cut by 10% when big data was used to optimize logistics and cargo management.

In conclusion, big data analytics is transforming the maritime industry's approach to emission reduction and estimation by providing real-time insights, predictive capabilities, and optimization tools. These technologies are helping industry meet regulatory requirements and achieve significant reductions in fuel consumption and CO2 emissions. Studies predict that big data applications could help the maritime

sector reduce emissions by 20% by 2030 **(IMO, 2021)**, making it an essential tool in the global fight against climate change.

4. THE FUTURE OF MARITIME DATA ANALYTICS AND INNOVATIONS

The maritime industry is undergoing a profound transformation driven by rapid advancements in data analytics. This has led to the need for efficient, safe, and sustainable maritime operations, especially due to the increasing importance of international trade and strengthening environmental standards globally emerging and edge technologies are the new data-driven solution that is frontier to revolutionize the business in every possible aspect from operation of fleets and supply chain. This chapter is a deeper insight into major trends and innovations that are causing prospects in Maritime Data Analysis.

1) Advanced AI and Machine Learning Integration for Predictive Analysis

Currently, AI and Machine Learning are stated to be some of the key drivers for Maritime related data. With the help of this technology, we are able to filter large volumes of data and find such patterns and tendencies that a human might overlook. The utilization of artificial intelligence in the transportation domain has yielded positive results. In addition to real-time data on load, fuel consumption, and freight conditions, it can also obtain route planning and fuel efficiency **(Yadav K. M. et al., 2024)**.

Application of artificial intelligence in predictive analysis will optimize the shipping process, but it's also useful for sanctions compliance, enhanced due diligence, and environmental protection. Moreover, AI will decrease risks and costly accidents or damaging environmental occurrences because it will be able to foresee such situations and prevent them. These applications can help shipping companies optimize operations, enhance safety and security, reduce costs, and minimize their environmental footprint. Some of the most notable cases of maritime predictive intelligence include:

Supply chain efficiency: improve container tracking capabilities to drastically upgrade efficiency. With ocean freight visibility, instantly identify shipments that may be delayed, discover the reason for those delays, and track shipments as they reach milestones along the voyage.

Due diligence: by understanding the upcoming movements of the vessel, activities in the port and the prevailing weather condition, predictive intelligence offers maritime stakeholders an effective risk management strategy plan. Based on this information, shipping companies can well anticipate the risks and make contingency measures to minimize risks and safeguard assets.

Sanctions compliance: Finding potentially high-risk cargo, ships, or trading partners makes it easier to ensure compliance with national and international sanctions. By gaining access to databases like LexisNexis and real-time behavioral data to identify risk exposure, predictive intelligence technology that is effective can deliver alerts in real-time to pertinent stakeholders. A business can adjust to making sure they're compliant once they have this information.

For instance, **Maritime AI™, Windward's** industry-leading predictive intelligence platform, analyzes billions of data points, including data from over 100 carriers, more than 1,400 ports and terminals, 5,500 container vessels, comprehensive sailing schedules, real-time weather updates and advanced forecasts and live location tracking **(WindWard)**.

2) Digital Twin Technology

Digital twins are usually defined as the virtual representations of physical assets embedded in digital environments such as existing or proposed ships and ports infrastructure. These models are constantly fed with raw data from monitoring instruments such as sensors and other similar equipment. Operators are also able to model various conditions and potential situations using digital twins which may involve an actual asset but are not at risk **(Lv Z. et al., 2023)**.

For example, **Kongsberg Digital**, a Norwegian maritime technology company, has developed a digital twin for a large container ship as part of its "Vessel Insight" platform. The digital twin collects real-time data from the ship's sensors, monitoring systems, and equipment to simulate the vessel's behavior under different conditions. This provides insights into fuel consumption, emissions, and maintenance needs. By running simulations, ship operators can optimize routes, improve energy efficiency, and predict maintenance schedules, reducing downtime and costs **(Kongsberg Digital, 2024)**.

In the future, Digital Twin technology will significantly enhance operational efficiency by enabling real-time optimization of routes, fuel consumption, and logistics. It will play a key role in predictive maintenance, reducing unexpected breakdowns and extending the lifespan of equipment. Technology will also help the maritime industry meet stricter environmental regulations by optimizing energy use and cutting emissions. Additionally, shipbuilders will use digital twins to improve ship design and construction, allowing for more efficient, eco-friendly vessels to be developed and tested virtually before physical implementation.

3) Big Data and Advanced Analytics

The maritime industry creates huge data flow, ranging from signaling via AIS (Automatic Identification System) to weather conditions and shipping contents. Big data analytics is the procedure of collection, storing and analyzing of these extensive and complicated data in order to obtain important data intelligence. Hiring also structured and unstructured data (such as ship's log or post in the social network or weather forecast), advanced analytics help define tendencies and identify correlation leading to better decision making **(Dalakakis D. et al., 2021).** According to a report published by **McKinsley in 2016**"Companies using customer analytics report 115% higher ROI and 93% higher profits than the ones who do not".

Maersk Line has implemented advanced analytics to enhance their predictive maintenance capabilities. By leveraging data from sensors installed on their vessels and integrating it with historical maintenance records and operational data, Maersk can predict when and where equipment failures might occur. This approach helps in

scheduling maintenance activities more effectively, reducing unexpected breakdowns, and minimizing operational downtime **(Landry H. 2024).**

It is therefore evident that the application of big data analytics shall increase the decision-making abilities in the maritime sector at all possible segments. For instance, shipping firms will be able to improve on the effective management of their fleets through understanding of fuel usage, route productivity, and the use of vessels. The big data can be collected and utilized to help ports manage their operations, and thus decrease the amount of congestion and increase throughput. Additionally, the use of advanced analytics can improve the general tracking of cargo and inventory in a firm, hence improving the entire supply chain cost.

4) Blockchain for Supply Chain Transparency

Simple, digital transaction records and tracking of the goods through the chain via constructions known as block chains. The application of block chain in the maritime industry is that one can be able to develop a chain of shipment that cannot be altered, thus enabling all the stake holders in the shipment to have trust in each other. This technology is especially useful in those cases where a supply chain is a bit more convoluted and is at a global level, where different entities are participating and where there are typical issues of fraud or of dispute **(IBM, 2021).**

Global Shipping Business Network (GSBN), a consortium of major shipping companies and port operators, has developed a blockchain-based platform to improve the transparency and efficiency of the global supply chain **(Carranza A. 2023)**. The platform eliminates several inconsistencies by giving all the stakeholders, ranging from the shipping lines, port operators, customs authorities to the freight forwarders access to a single and accurate data version. It replaces the conventional manual process of documents that are usually used in the documentation of containers whereby time and costs are minimized, and handling of cargos is not delayed. Also, the fact that information is fixed and can be changed only with a subsequent consensus in the blockchain produces a positive effect on combating fraudulent actions and documents' forgery, thus strengthening the supply chain.

5) Internet of Things (IoT) and Real-Time Monitoring

Internet of Things (IoT) therefore means the connection of any device, vehicle or object with a communication capability with other objects and devices. IoT devices are subject to the maritime industry to measure real-time vessel, cargo, and environmental conditions. These sensors are designed to constantly generate data which can then be used to fine-tune the processes and check adherence to the rules and regulations **(Aslam S. et al., 2023)**.

Maersk Line has implemented IoT into their fleet to improve its real-time monitoring and control of their ships. The company has an elaborate IoT system through which the organization gathers data from different sensors on its ships, which includes engine data, fuel consumption, and cargo data. This data is sent in real-time to Maersk's central operations and control centers for real-time assessment and subsequent action **(O'Marah K. 2023)**.

The introduction of Internet-of-Things (IoT) in the maritime industry is bound to enhance its operational efficiency and safety. Real-time monitoring tools that are likely to be used by operators can help them identify any problems arising in engines' malfunctioning or deviation from best fuel consumption practices. Moreover, IoT devices will play an essential role in monitoring the environment by providing information regarding emissions, and the quality of ballast water among other things. Thus, shipping industries would easily comply with strict environmental laws and lessen their carbon footprints.

6) Sustainability and Environmental Analytics

With the major intensifying environmental issues, the maritime industry has become a noticeable target to achieve a reduced adverse impact on the environment. Data analytics will be essential for monitoring and improving environmental performance. Predictive models can process information that is coming from multiple sources, for example, from emissions sensors and weather data, to get to the point where they can

predict the environmental impact of different operations and tell which is the best mitigation strategies (**Munim H. Z. et al., 2023).**

Det Norske Veritas (DNV), a leading provider of risk management and quality assurance services, has developed the Veracity data platform, which offers comprehensive environmental analytics to support sustainability efforts in the maritime sector. The platform aggregates data from various sources, including ship sensors, weather conditions, and operational data, to provide actionable insights on fuel consumption, emissions, and overall environmental impact **(Veracity, 2024)**.

Environmental analytics will be the keystone for maritime companies in their quest toward attesting sustainability goals. State-of-the-art fuel consumption optimization and route planning can support companies tremendously in reducing GHG emissions. Data-driven insight will also help to comply with various environmental regulations, which include the sulfur cap by the International Maritime Organization and carbon intensity targets. Finally, environment-related analytics is backing green technology development, including alternative fuels and energy-efficient ship designs.

7) Smart and Autonomous Vessels

Among the unlimited developments in the maritime industry stands out the introduction of autonomous vessels that function on their own without the participation of a human factor, which is one of the most incredible. They use AI, IoT, and Data Analytics to perform different operations such as navigating, avoiding obstacles, and optimizing routes on their own. The smart ship provides the latest technologies such as sensors and systems that are used for steady monitoring of the environment and conditions of the vessel. Consequently, smart ships are being enabled to have commands for action in real-time mode, instead of being always connected to the network communication. The main difference between a smart and an autonomous vessel is that smart vessel despite its enhanced technology still relies on humans for navigation and control **(Wasilewski W. et al., 2021)**.

The **Mayflower Autonomous Vessel (MAS)** is an unmanned, fully autonomous ship that was developed to cross the Atlantic Ocean using AI, machine learning, and advanced technology to navigate and conduct research. It is a modern reimagining of

the historic 17th-century Mayflower ship, which carried the Pilgrims from England to America. MAS was built to commemorate the 400th anniversary of the original Mayflower's voyage **(IBM, 2020).**

The **MV Prism Courage**, an LNG carrier owned by Hyundai LNG Shipping, is a leading example of a smart vessel equipped with Hyundai's HiNAS 2.0 autonomous navigation system. Developed by Avikus, this AI-driven system optimizes routes, monitors real-time data, and enhances fuel efficiency by analyzing sea conditions, weather, and maritime traffic. In May 2022, the ship completed a 12,000-kilometer trans-Pacific voyage, during which the system autonomously controlled half of the journey. This resulted in a 7% reduction in fuel consumption and a 5% decrease in emissions, showcasing the potential of smart ship technology to improve safety, efficiency, and environmental performance **(Szondy, 2022).**

One of the imminent benefits of fully automated ships is that will short cut the labor workforce's cost for the maritime industry as well as the ships being always safe and an invaluable Improvement to the service. Such ships will be self-running 24/7, meaning that they will not have to wait for the crew to rest, enabling quicker and more flexible shipments. Furthermore, the fact that ships can execute the process of navigation independently means that it will be less likely for a human error to take place.

Thus, fewer accidents and environmental incidents will happen. It is predicted that the smart ship will, in fact, be the ship of the future and will connect the rest of the smart fleet as well as shore-based operations centers, leading to creating a more connected and efficient maritime ecosystem.

8) Enhanced Cybersecurity

With digitalization taking more and more dominance in the industry, the likelihood of cyberattacks also grows. Cybersecurity analytics are seen as indispensable in the protection of security in maritime operations against these violating cyber threats. Innovative analytics tools will be employed in order to uncover cyber threats and

provide real-time responses by the use of machine learning to recognize abnormal patterns and vulnerabilities (**IronHack, 2023).**

For example, Deloitte's "**Cyber Risk Services for Shipping and Maritime**" Deloitte offers a suite of cybersecurity-related services in the shipping and maritime sector, underpinning advanced analytics in data to protect ships and onshore shipping operations against cyber threats **(ekathimerini, 2024)**.

With Deloitte's cybersecurity approach, the analysis and monitoring of cybersecurity data are performed continuously for the detection of unusual activities and various forms of cyber-attacks in real time. Using machine learning algorithms along with big data analytics, the company analyzes network traffic and system logs and other sources of data. This helps identify patterns indicating cyber-attacks, like malware infection or phishing attempts, and proactive measures to mitigate the same.

Maritime cybersecurity is an example of how vital it is to secure maritime operations that should be guaranteed to the safety and continuance of the same. The more the vessels and port facilities are interconnected, the more they become potential targets of cyber threats that could paralyze the proper performance of operations, either hack the information they need or even breach the most critical cybersecurity endpoints. Through the deployment of cybersecurity programs and data analytics, shipping companies can “shield” their assets and predict imminent threats from pirates and hackers who can steal such resources or impede shipping schedules through attacks.

Data analysis is a very fast-growing and futuristic field which is being expanded every day. If AI, IoT, and blockchain together with other tools like these undergo further development, they will be responsible for driving serious improvements on how various shipping operations are performed **(SBN Technologies, 2024)**. These advanced products will guarantee that the sector is provided with the potential to perform better than other sectors would have, thus there will be the achievement of efficiency, safety, and sustainability.

5. CHALLENGES AND BARRIERS OF THE DATA ANALYTICS

Integrating data analytics into the shipping business is challenging due to the numerous obstacles and barriers that firms must cope with to get full utilization of these technologies.

5.1 REGULATORY AND COMPLIANCE CHALLENGES

Regulatory and compliance challenges are usually hurdles or the problems that the companies meet while they try to comply with the government's rules and orders. In the perspective of the implementation of data analysis, it concerns mostly the regulatory frameworks and the compliance requirements that are the obstacles to the effective use of data **(Brandon R. 2024).** It is a blockage to a high degree of complexity and contradictory rules and regulations from countries and areas which the organizations do business with. Organizations are obligated to develop their data analytics constructions to meet those rules. Not only should organizations narrow down their focus to specific regulatory and compliance challenges, but they should also investigate different dimensions of regulations and compliance requirements **(Bassi A. C. et al., 2023).**

Given that companies need to follow new regulations imposed on the maritime industry they must have complex processes in data management and analytics. One of the main enablers in this adaptation is the central storage system that gathers multiple operational data on a regular basis. The regulatory authorities sometimes demand timely reporting of the performance of the vessel, safety checks and effects on environment, this makes the data that is extracted, transformed and loaded (ETL) into the analytic platform to be processed a real time process **(Palmer M. et al., 2024).**

By implementing live or near real-time data upload systems, maritime companies can not only streamline operations but also ensure they meet compliance standards set by the International Maritime Organization (IMO) and other regulatory authorities. These tools enable timely and transparent reporting, helping organizations keep up with regulations that demand increased data accuracy, efficiency, and the ability to scale as compliance requirements evolve. In this way, staying up to date with both regulatory mandates and operational performance becomes an integrated, data-driven process **(Bassi A. C. et al., 2023).**

5.2 DATA QUALITY AND SECURITY CONCERNS

The primary issue with data quality and security problems is that even at the time of a successful data analytics implementation, they can create significant obstacles in the analysis attempts of various business sectors. The shipping industry relies heavily on IT-based systems and the data they generate for key areas like fuel efficiency, energy management, safety, and environmental protection. These systems play a crucial role in optimizing operations and ensuring compliance with industry standards.

Furthermore, data quality and security problems have become a big concern for the shipping sector **(Vouros A. et al., 2017).**

Moreover, the transportation industry strongly relies on data, which, in turn, promotes the sensitivity of data quality and data security issues. The snowball effect of inaccurate data after being processed and used for decision-making in load optimization can result in inefficient ship performance and consumption profiling. The stakes could hardly be higher if we encounter accidents or extreme weather forecasts that will severely affect the safety and life of the crew **(Vouros A. et al., 2017)**.

As was mentioned by **ABS (American Bureau of shipping) in 2016** it is necessary to demonstrate clear quality and relevant standards, as well as accountability and mechanisms for compliance with rules for offshore personnel. These include such things as truth, integrity, reliability, criticism, genuineness, safety, accuracy, and timely reporting. Safety violations in fuel shipments may result in fraudulence, manipulation, or any other type of information that might be entered in a manner that interferes with the estimates, for example, position, vessel speed, and heading. Utilizing such dummy data with an algorithm function, it is possible to segment users into both "planners" and "influencers" and then change estimates of fuel consumption for a fleet while staying within the initial size of the algorithm.

5.3 COSTS AND RETURN ON INVESTMENT (ROI)

There are significant resistances around economic aspects, particularly concerning implementation costs and potential ROI **(Shim J. et al., 2015).** For instance,

Businesses face several challenges in adopting data analysis solutions. There are no established business models, making it difficult to implement consistently. Implementation costs are uncertain and hard to estimate, raising concerns about

whether the return on investment (ROI) will meet expectations. Additionally, there's a risk that costs may exceed potential benefits. While the investment is innovative, its true value remains ambiguous, and current efforts could lead to a situation where implementation costs significantly outweigh the expected advantages.

The issue with all these aspects is that the value of data analytics is often misguided. One cannot expect all the benefits to manifest in quantity in a recession. Often, cost-effectiveness does not appear to be easy to obtain, even when indirect benefits are analyzed. Usually, the board of a company is not willing to submit proposals where the debt exceeds the estimated amount for its own benefit. There is concern that the evidence of unfair interest is not conclusive enough to protect the investment.

5.4 ORGANIZATIONAL CULTURE AND CHANGE MANAGEMENT

In the shipping industry, particular issues become obstacles to the implementation of data analytics, namely organizational culture and change management that shape the ways data analytics can be done. These include the values, beliefs, and norms that guide behaviors, practices, and interactions within the organization.

Denison D. R. & Spreitzer M. G. (1991) argued that over time, the people in maritime organizations tend to develop their culture with its associated ways of doing the job. As a result, according to the classification of the organization’s pattern, it needs to be clear that it can be successful or impossible for the organization to adopt and integrate daily operations with data analytics. Cultural and organizational initiatives are complex and extensive endeavors, which create a lot of expectations and demands on the personnel and consumers. Such changes call for transition with regards to the decision-making process and replacing subjectivity with objectivity in the form of data. This implies that information utilized for decision making process should be factual instead of a tendency or assumed.

The focus groups identified that the desire and need for a data-driven culture relates to everyday practice and movement and allows employees to independently explore data and analytics to support their practice emphasis. A data-driven culture welcomes more open access to the data, transparency in operations, and fact-based decision-making with clear accountability and confidence in the analysis to solve

problems when resources are delivered that do not meet expectations **(St. Louis Junior et al., 2019)**.

The perceived "ship doesn't lie" culture, where information about risk is hidden from the operations, and in which evaluation is framed with suspicion, sets the scene for how things play out. The challenges in implementing cultural and organizational changes to create a supportive and optimistic data analytics culture involve several key aspects. Establishing this culture requires careful planning and integration at both the strategic and operational levels of the business. This includes addressing resistance to change, aligning the new culture with existing business practices, building the necessary skills, and maintaining ongoing support and engagement from all stakeholders. It only needs to adopt practical means and standards for changing management and fostering a culture of data analytics.

5.5 TECHNOLOGICAL EXPERTISE SHORTAGE IN THE MARITIME INDUSTRY

The marine environment faces major challenges and barriers to the successful application of data analytics. The biggest challenge yet to be addressed, according to industry experts, is that the maritime sector lacks skilled and specialized people to implement data analytics.

Data technologists are individuals who can manipulate all aspects of data to generate insights for advanced decision-making and the use of created insights **(Medida L. et al., 2024)**. The talent shortage may be due to the tendency to hire seafarers who are simply maritime savvy but have no technology experience, instead of looking for individuals who can teach how to handle and use this kind of technology **(Gavalas D. et al., 2022).**

This deficiency affects extremely mature organizations because they are unable to conduct data analytics efficiently. The use of data analytics must first become more widespread in the maritime industry, as most organizations are unable to meet this challenge **(Bahrami M. et al., 2021)**. Consulting firms that bring data analytics expertise to shippers can help with this. Organizations that already conduct data

analytics may be able to expand data analytics activities rather than limiting them to one or two people who cannot manage these activities.

In an ideal case, companies interested in data analytics would replicate the expertise of a person who can perform data analytics on their respective vessels. This would mean that the companies would have internal experts who are well-equipped with the necessary expertise to handle occupations where data analytics are to be performed to create insight **(Medida L. H. et al., 2024)**. Hiring individuals with this kind of technological background and maritime knowledge as a secondary qualification does not fully address the company's needs. Such employees still require specialized training to effectively integrate maritime expertise with data analytics **(Oksavik A. et al., 2020)**. If there were a professional organization that provided businesses with skilled individuals, it would significantly help bridge the knowledge gap between maritime expertise and data analytics.

However, this would not solve the lack of general talent. Therefore, a maritime education institution must establish interdisciplinary study programs **(Souppez E. G. et al., 2017)** dealing with data analytics and, at the same time, maritime. It means that students could be educated in such a way as to understand both maritime and data analytics, consequently reducing the lack of expertise. Such degree programs already exist in other industries. It would also be useful to make more people aware of the data analytics opportunities in the maritime industry, since most students are not aware of their possibilities. A preparatory course whereby students would be introduced to the maritime industry could create an impact on the candidate pool of companies when hiring bachelors and experts in maritime industry. Solutions to tackle expertise shortage should be parallel, as companies urgently need help in implementing data analytics.

5.6 DATA ANALYTICS INTEGRATION WITH EXISTING SYSTEMS

Many of the organizations that exist within the maritime industry are decades or centuries old and make use of complex, heavy, and expensive technology in a wide variety of forms: hardware, software, processes, databases, standards, etc. These range from those onboard a ship for navigation and berth to those installed in the data center of a huge port **(Durlik I. et al., 2023)** which, among many other things, store, process, analyze, and explore the historical data from vessels docking on the berthing at the

port. For many, merging data analytics with already running systems is not easy. The existing systems might represent a tight weave of hardware, software, and processes. In such cases, they may well span decades of systems evolution and experience **(Lundh M. et al., 2023).**

Such is the difficulty of changing just a trivial aspect of this system that it might require extensive work upsetting a long chain of dependency, and the organizations might very well underestimate the breadth of the issues there. Closely related, adherence to technological maturity levels and technology standards is essential. Not all equipment may be compatible with or support the functions of others. Hence, early testing for the compatibility of technology is vital to avoid such costly surprises further down the track **(Koga S. et al., 2015).**

Many organizations have systems engineered around some central storage system where all operational data gets uploaded and stored in batches, either daily, weekly, or monthly. To achieve that, the required data points from these existing systems would be copied to this central system. These systems also have to include some online or near real-time processes whereby all data points from an existing system are uploaded to the analytic systems almost immediately after they are created **(Maghoromi E. B. et al., 2023).**

This kind of switch would probably be an important reengineering of the existing systems, the most complex and potentially riskiest part of this entire undertaking. Even without reengineering, the integration task would already be heavy and probably a long-term development project due to many differences between the architecture of existing systems and that required for a properly working data analytics system **(Durlik I. et al., 2023).**

These challenges and barriers underline the complexity of implementing data analytics within the maritime industry. Industry players must therefore make a collective effort toward investing in technology, training, and cultural change, while remaining committed to regulatory and cybersecurity concerns **(Maghoromi E. B. et al., 2023).**

6. CASE STUDIES AND REAL-WORLD EXAMPLES

Case studies of data analytics in the maritime industry showcase the transformative potential of advanced analytics and technology in optimizing operations, enhancing safety, and improving decision-making. Through real-world examples, the impact of data-driven strategies on operational efficiency and profitability becomes evident, highlighting the vital role of analytics in the future of maritime operations. In **Table 6.1**, a summary of the key challenges and benefits associated with these case studies highlights the ongoing need for effective data management and the significant advantages that analytics can bring.

6.1 PORT OF ROTTERDAM: SMART PORT INITIATIVE

The Port of Rotterdam, the world's largest and busiest port, has adopted smart port technologies to enhance its operations through data analytics. By using big data and predictive analytics, the port optimizes vessel traffic, cargo handling, and equipment maintenance. This approach improves scheduling, reduces wait times and congestion, and allows for proactive maintenance, preventing equipment breakdowns and further enhancing operational efficiency (**The Digital Port| Port of Rotterdam, 2024)**.

The Smart Port Initiative at the Port of Rotterdam exemplifies how digitization and intelligent systems can enhance port efficiency, safety, and sustainability. Significant milestones achieved through this initiative include establishing a clear governance structure for the port community, defining specific programs for port functionalities, and fostering close collaboration with the government and research institutions. The initiative has also led to the development of numerous pilot projects focused on data sharing and analysis to optimize port operations. As a result, companies using Rotterdam as a base have reduced vessel idle time by 20%, increased throughput capacity, and lowered operational costs, positioning Rotterdam as one of the most efficient and innovative ports. This competitive edge in managing high traffic volumes makes it a strategic hub for international carriers.

Despite its successes, the Smart Port Initiative at the Port of Rotterdam faces several challenges, including financial constraints, privacy concerns, and the complexity of

managing numerous innovations. To address these issues, the Port Authority has developed a more structured approach, creating a balanced portfolio that includes various port community programs, partnerships, and pilot projects. This strategy aims to manage these challenges effectively while advancing the initiative. The strategies implemented at the Port of Rotterdam provide valuable insights into other ports looking to establish similar programs.

6.2 MAERSK LINE: FUEL OPTIMIZATION AND EMISSION REDUCTION

This perspective is exemplified by Maersk, which has developed an advanced data analytics system aimed at improving fuel efficiency, consumption, and carbon emissions for its extensive fleet. The system, known as Captain Peter, provides real-time analysis of speed, route, and engine performance data. By continuously receiving data from the ship's equipment such as weather conditions, sea currents, and engine parameters Captain Peter can recommend adjustments to optimize fuel use and reduce emissions **(Maersk Line, 2023)**.

It has realized a total of 4% fuel consumption reduction translating into fuel cost savings and at the same time it has reduced CO2 emission by a wide margin. This is beneficial to Maersk regarding environmental impacts and enables the firm to meet the strict international standards of emissions control. Efficient utilization of fuel has also better stretched Maersk's vessels' useful life since engines do not wear as much as they would if they were constantly on due to inefficiencies.

The two main issues that Maersk faced were bringing together several forms of data into one system and the second one was to ensure that the models had their desired degree of precision. Besides this, another issue was that the system had to be implemented across the company's diverse and global fleet of ships that operate under varying conditions which was to be managed solely by the company. It was also crucial to make sure that crew members were trained sufficiently for them to be able to comprehend and apply the recommendations given by Captain Peter to make the system proper.

6.3 CARNIVAL CORPORATION: PASSENGER EXPERIENCE AND OPERATIONAL EFFICIENCY

With the Ocen Medallion, a wearable device the Carnival Corporation, has developed individual approaches to improve passengers' satisfaction and operation optimization through data analytics. This new generation apparatus records extensive information about the passenger's movement, desires, choices and activities and these are utilized to provide solutions. For instance, information derived from the Ocean Medallion shall allow Carnival to fine-tune the eating offering, and schedule the executions of entertainment and relevant services, while increasing safety standards with real-time crowd densities **(Carnival: Guest Experience Data Platform, 2023).**

The passengers were satisfied with the experience as it was tailor-made according to their preferences, and the onboard spending rose via data analytics. In Carnivals operational context, there is evidence of positive effects of resource positioning where inventory controls, as well as cleaning schedules, were demonstrated to yield positive effects such as cost-cutting. All in all, due to the improved ability to timely anticipate the needs of passengers and respond to them, it is ready to strengthen Carnival's position in the context of competition in the cruise segment.

On the other hand, achieving such a high level of customization requires significant investment in technological and infrastructural resources. This includes expenditure on advanced data analysis tools and wearable devices designed to enhance performance and efficiency. The other two important consideration areas are data management and protection, whereby large amounts of personal data that involved the passengers' details needed to be safeguarded. Also, the company faced issues with the implementation of these technologies in the large and heterogeneous fleet and educating the personnel to use them.

6.4 CARGILL: VOYAGE EFFICIENCY PROGRAM

Cargill, a major multinational in agricultural and food production, utilizes data analytics through its Voyage Efficiency program to enhance shipping operations. By leveraging predictive analytics, Cargill improves voyage planning by considering

factors like weather conditions, fuel costs, and port accessibility. The program integrates historical data with real-time information to determine the optimal route and speed for vessels, aiming to achieve timely delivery with minimal fuel consumption and reduced emissions (**Cargill, 2024)**.

Cargill has been able to reduce fuel consumption and associate CO2 emissions by 5% through the application of data analytics, this contributes not only in terms of environmental sustainability but also results in significant cost savings. Moreover, it improves the company's competence to better schedule and conduct ocean-going voyages with greater predictability which would reduce delays as well as overall operational performance. Cargill has now established its position as a forerunner in sustainable maritime logistics.

One of the major challenges was integrating data from different sources, such as weather data, fuel consumption records, and port information, into one consistent analytics platform. Ensuring that predictive models are accurate and reliable with variables, especially as unstable as the weather, proved quite a challenge. The company also needed to ensure that its operational staff were empowered and well-positioned to interpret and act effectively on these data-driven recommendations.

6.5 SHELL: PREDICTIVE MAINTENANCE FOR OFFSHORE RIGS

By using analytics to drive predictive maintenance strategies on its offshore oil rigs, Shell, a leading global energy company, enhances both safety and operational efficiency. Advanced analysis of sensor data from critical equipment such as pumps, compressors, and turbines allow Shell to anticipate maintenance needs before failures occur. Sensors monitor various metrics, including vibration levels, temperature, and pressure, enabling the company to detect potential issues early and schedule maintenance proactively, thus avoiding unplanned downtime **(Shell, 2024)**.

While deploying such a predictive maintenance solution, unplanned downtime was decreased by 20% and cost savings on maintenance performed were saved at the rate of 10%. Not only will this increase reliability for its offshore endeavours, but it also minimizes the threat of catastrophic equipment failures that could lead to an accident. It enables smarter resource allocation and reduces operational costs at large by aiding in planning regarding equipment scheduling saving the life of important machinery.

Implementing predictive maintenance required enormous investments in sensor technology and data analytics platforms. Shell also had to wrestle with the problems associated with integrating such technologies into existing systems, together with making sure of their continuous accuracy for the predictive models. Another complexity was how to train personnel to act appropriately on predictive maintenance alerts and coordinate corresponding maintenance activities across a multitude of offshore sites.

6.6 WARTSILA: SMART MARINE ECOSYSTEM

Wartsila excels in digital marine technology with the Smart Marine Ecosystem, a compound vehicle that harnesses data analysis to arrive at each sector of the marine domain. The ecosystem employs cutting-edge data analytics to fine-tune energy management, voyage planning, and fleet performance. A fundamental building block of the ecosystem is Wartsila's Eniram platform, which acquires and analyzes the ship's performance data like fuel consumption, engine efficiency, and external variables such as weather and sea setting **(Wartsila, 2023)**.

The Smart Marine Ecosystem is a big part of fuel efficiency has been improved significantly, and costs and environmental impact are reduced by emissions being lower. The results reported by the Eniram Much Meet system show that the operators who can make their decisions based on the data have a better quality of performance for their fleets. This system not only supports more sustainable maritime operations but also helps Wartsila's customers comply with increasingly strict environmental regulations.

To develop and implement the Smart Marine Ecosystem, technical challenges that pertained primarily to data integration from many sources, i.e. onboard sensors, and such external data feeds as weather forecasts were the main area of attention. The comfort of the data in a sprawling network of vehicles and the effectiveness of the analytics models to provide precise advice are also genuine problems. Also, the use of such enhanced technology demanded years of practice besides the bureaus that installed the environment had to transfer their, impressively, changed vision to the whole company.

Case Study	Benefits	Challenges
Port of Rotterdam: Smart Port Initiative	➢ Improved operational efficiency ➢ Reduced Congestion ➢ Enhanced scheduling and maintenance	➢ Financial constraints ➢ Privacy concerns ➢ Managing multiple innovations
Maersk Line: Fuel Optimization and Emission Reduction	➢ Increased fuel efficiency ➢ Reduced emissions ➢ Enhanced voyage planning	➢ High customization costs ➢ Significant investment in technology and infrastructure
Carnival Corporation: Passenger Experience and Operational Efficiency	➢ Enhanced passenger satisfaction ➢ Improved operational efficiency ➢ Better resource management	➢ Balancing personalization with privacy ➢ High costs of implementing new technologies
Cargill: Voyage Efficiency Program	➢ Optimized route and speed ➢ Reduced fuel consumption ➢ Lower emissions	➢ Complexity in integrating real-time and historical data ➢ Dependence on accurate data inputs
Shell: Predictive Maintenance for Offshore Rigs	➢ Prevented unplanned downtime ➢ Increased safety ➢ Improved maintenance scheduling	➢ High costs of sensor installation and data analysis ➢ Managing large volumes of data
Wartsila: Smart Marine Ecosystem	➢ Improved efficiency across marine operations ➢ Enhanced decision-making ➢ Better environmental performance	➢ Integration with existing systems ➢ Initial investment and training costs

Table 6.1: Benefits and Challenges of Data Analytics Case Studies in Maritime Industry **Created by**: Author

7. A STEP-BY STEP DATA ANALYSIS ON A SUPPLY CHAIN DATASET

With the increasing complexity of global logistics, maritime data analysis has appeared to be one of the most important steps in the process of optimizing operations, reducing costs, and improving service delivery. This master thesis aims to conduct a step-by-step data analysis on a comprehensive maritime dataset, focusing on key logistics variables.

Data analysis was conducted on a Supply Chain dataset which was retrieved from Kaggle. The dataset describes the volume of orders that issued, the state and destination of the delivery in the span of three years (2015-2017), with features such as Payment Type, Delivery Status, Customer Segment, Department Name, Market, Order Country, Order Date, Item Quantity, and Order Status can significantly enhance the logistics department's decision-making. By analyzing Delivery Status and Order Status, the department can track delivery performance, identify delays, and optimize routes or processes for timely order fulfillment. Analysis of Item Quantity and Order Date aids to forecast demand, adjust inventory levels, and streamline order processing.

Furthermore, segmenting data by Customer Segment, Market, and Order Country allows the logistics team to tailor strategies based on geographic or demographic trends, enhancing efficiency in distribution and resource allocation. Insights from Department Name and Type can also be used to monitor the performance of various departments or product types, enabling targeted improvements and cost optimization.

7.1. Collect and Store Data

The raw dataset was 96.69 MB and in a CSV format, which consists of 49 **Columns** and 180.509 **Rows** of Raw Data (Total of 8.844.941 values). To manage our data, we inserted the CSV file in EXCEL by using the **Get Data** > **From File** > **From Text/CSV** tool and then we used the **Text to Columns** function in order to make the data more clear and more organized and finally we stored the data in a table by using **Insert>Table** from the Excel Tool Bar.

However, part of the information is unreadable or unwanted, which leads us to the next step.

7.2. Data Description and Preprocessing

Columns	Data Type
Payment Type	Text
Delivery Status	Text
Customer Segment	Text
Department Name	Text
Market	Text
Order Country	Text
Order Date	Date
Item Quantity	Number
Order Status	Text

Table 7.1: Columns and Data Types of Final Dataset **Created by:** Author

Source: Kaggle.com

We filtered every column, and we removed Duplicates, empty cells (NULL cells), rows or columns with no interest for our analysis. A subset of the final dataset is presented in **Table 7.2**.

Payment Type	Delivery Status	Customer Segment	Department Name	Market	Order Country	Order Date	Item Quantity	Order Status
CASH	Shipping on time	Corporate	Apparel	Pacific Asia	Vietnam	2015	1	CLOSED
CASH	Late delivery	Corporate	Apparel	Pacific Asia	Vietnam	2015	1	CLOSED
CASH	Late delivery	Consumer	Apparel	Pacific Asia	Vietnam	2015	1	CLOSED
CASH	Late delivery	Home Office	Apparel	Pacific Asia	Vietnam	2015	1	CLOSED
CASH	Late delivery	Home Office	Apparel	Pacific Asia	Vietnam	2015	2	CLOSED
CASH	Late delivery	Home Office	Apparel	Pacific Asia	Vietnam	2015	4	CLOSED
CASH	Shipping on time	Consumer	Apparel	Pacific Asia	Vietnam	2015	4	CLOSED
CASH	Shipping on	Consumer	Apparel	Pacific	Vietnam	2015	4	CLOSED

	time			Asia				
CASH	Late delivery	Consumer	Apparel	LATAM	Venezuela	2015	1	CLOSED

Table 7.2: Final Data Set Sample **Created by:** Author

Source: Kaggle.com

7.3. Data Analysis

There are multiple tools for analyzing data like EXCEL, SQL, PYTHON, R, POWER BI, TABLEAU etc.

These tools provide in-built functions to perform calculations and create temporary tables. In our case, we decided to use the TABLEAU BI which is more user-friendly and more familiar to us. We visited the Tableau Public, we created a profile and then we created a new viz.

After we connected our EXCEL dataset with TABLEAU, we commenced the analysis by dragging and dropping the data in columns, rows or measures.

Firstly, we created a table to show the number of units that were ordered from every department in the span of three years (2015-2017) including total and percentage and the Grand Total of the 3 years.

A dataset with historical order information across product categories is crucial for logistics, as it enables demand forecasting, helping to optimize inventory and manage peak seasons effectively. By analyzing order trends, the logistics team can coordinate with suppliers, plan resources, and allocate warehouse space efficiently, aligning stock levels with demand to reduce holding costs and prevent last-minute expedited shipping.

Department Name	2015	%	2016	%	2017	%	Grand Total
Apparel	34.687	25,05%	35.324	25,72%	27.754	26,15%	97.765
Book Shop					405	0,38%	405
Discs Shop					1.264	1,19%	1.264
Fan Shop	38.502	27,80%	38.278	27,87%	29.109	27,43%	105.889
Fitness	2.271	1,64%	2.252	1,64%	1.492	1,41%	6.015
Footwear	16.293	11,77%	15.526	11,30%	11.581	10,91%	43.400
Golf	36.613	26,44%	36.071	26,26%	26.613	25,08%	99.297
Health and Beauty					239	0,23%	239
Outdoors	10.114	7,30%	9.901	7,21%	5.948	5,61%	25.963

Pet Shop					246	0,23%	246
Technology					1.463	1,38%	1.463
Total	**138.480**		**1 37.352**		**106.114**		**381.946**

Table 7.3: Orders of Product Categories per Year **Created by:** Author

Source: Kaggle.com

In **Table 7.3** the data indicates that **Apparel**, **Fan Shop**, and **Golf** departments consistently led in sales from 2015 to 2017, accounting for the majority of overall performance, while categories like **Book Shop**, **Health** and **Beauty**, and **Pet Shop** showed minimal activity, only appearing in 2017. Over the three years, there was a general decline in total sales, from 138,480 in 2015 to 106,114 in 2017, possibly due to lower performance in **Apparel**, **Fan Shop**, and **Golf**, as well as decreased interest or seasonal shifts. Potential reasons for the observed trends could include market changes, customer preferences shifting away from lower-selling categories, or strategic reallocation of resources to high-performing departments.

Department Name	2015	2016	2017
Apparel		+1,84%	-21,43%
Book Shop			
Discs Shop			
Fan Shop		-0,58%	-23,95%
Fitness		-0,84%	-33,75%
Footwear		-4,71%	-25,41%
Golf		-1,48%	-26,22%
Health and Beauty			
Outdoors		-2,11%	-39,93%
Pet Shop			
Technology			

Table 7.4: Year to Year Growth **Created by:** Author

Source: Kaggle.com

In **Table 7.4**, several departments show a decline in growth, particularly in 2016-2017. For instance, **Apparel** experienced a mild increase of 1.84% initially, only to face a sharp decline of 21.43% afterward. Similarly, **Fan Shop**, **Fitness**, **Footwear**, and **Outdoors** all display notable declines, with **Outdoors** experiencing the steepest drop at -39.93% between these years.

This trend of reduced growth across multiple categories suggests external challenges or market shifts affecting sales. Potential reasons could include increased

competition, shifts in consumer behavior (such as a preference for online options or new brands), or economic factors limiting consumer spending. Additionally, supply chain issues, particularly if certain products were unavailable or delayed, may have also contributed to these declines. Each category's sharp decline could signal a need for the business to reassess its inventory strategies, explore emerging consumer preferences, or enhance marketing efforts to regain momentum in these departments.

Delivery Status	2015	%	2016	%	2017	%	Grand Total
Shipping on time	10985	17,53%	11281	18,04%	9553	17,96%	31819
Late delivery	34372	54,86%	34446	55,07%	28959	54,45%	97777
Advance shipping	14597	23,30%	14199	22,70%	12326	23,18%	41122
Shipping canceled	2696	4,30%	2624	4,20%	2348	4,41%	7668
Total	**62650**		**62550**		**53186**		**178386**

Table 7.5: Delivery Status per Year **Created by:** Author

Source: Kaggle.com

As is presented on **Table 7.5,** the delivery status data for 2015 to 2017 shows a consistent pattern across the three years, though there are slight variations. The majority of deliveries fell under **Late Delivery**, making up around 55% each year, suggesting a significant and consistent challenge in meeting delivery deadlines. Meanwhile, **On-time Shipping** accounts for about 18% across the years, indicating relatively minor improvement, though it dips slightly in 2017. **Advance shipping** remains stable at around 23%, with minor decreases each year, and **Canceled** shipments are low, at around 4% consistently.

The high rate of late deliveries could stem from several factors, such as supply chain disruptions, logistical inefficiencies, or demand outstripping current shipping capacity. The consistency in these figures suggests systemic issues that could require an overhaul in logistics processes or better forecasting to reduce late shipments. Addressing these late deliveries would be crucial for improving customer satisfaction and operational efficiency, as well as potentially boosting on-time shipping metrics.

Order Status	2015	%	2016	%	2017	%	Grand Total
CLOSED	7067	11,28%	6633	10,60%	5669	10,66 %	19369
COMPLETE	20338	32,46%	2107	33,70%	17361	32,64	58778

			9			%	
ON_HOLD	3398	5,42%	3525	5,64%	2757	5,18%	9680
PENDING_PAYMENT	14003	22,35%	13653	21,83%	11723	22,04%	39379
PAYMENT_REVIEW	636	1,02%	661	1,06%	568	1,07%	1865
PENDING	6967	11,12%	6800	10,87%	6230	11,71%	19997
PROCESSING	7545	12,04%	7575	12,11%	6530	12,28%	21650
SUSPECTED_FRAUD	1385	2,21%	1376	2,20%	1260	2,37%	4021
CANCELED	1311	2,09%	1248	2,00%	1088	2,05%	3647
Total	**62650**		**62550**		**53186**		**178386**

Table 7.6: Order Status per Year **Created by:** Author

Source: Kaggle.com

In **Table 7.6,** the order status data for 2015-2017 reveals consistent trends across various categories, with **Complete** orders making up the largest share at about 32-34% annually. The next largest category is **Pending Payment**, ranging around 22%, indicating a notable portion of orders held due to incomplete payments. **Processing** and **Pending** statuses are also stable, each representing about 11-12% of orders. Notably, **Closed** orders remain around 10%, while **On Hold** and **Suspected Fraud** status stays relatively low but consistent over the years, at approximately 5% and 2%, respectively.

The high percentage of orders stuck in **Pending Payment** suggests a potential area to optimize, possibly by implementing smoother payment processes or clearer communication with customers. The stability in **Complete** orders is positive, but the presence of statuses like **Processing**, **Pending**, and **On Hold** at significant rates points to potential bottlenecks in moving orders through to completion. Addressing these hold-ups could help increase the completion rate, reduce canceled or delayed orders, and ultimately improve both customer satisfaction and operational efficiency.

Market	Order Country	Orders
Africa	Nigeria	2.309
	Egypt	1.189
	Marocco	1.135
	South Africa	1.099
	Republic of Congo	1.010
Europe	France	13.221
	Germany	9.563
	UK	7.301
	Italy	4.988

	Spain	3.868
LATAM	Mexico	13.292
	Brazil	7.987
	El Salvador	3.726
	Republic of Dominican	3.669
	Honduras	3.629
Pacific Asia	Australia	7.954
	China	5.396
	India	4.408
	Indonesia	3.916
	Turkey	3.395
USCA	USA	24.840
	Canada	959
Grand Total		**128.854**

Table 7.7: Top 5 Countries in Orders from every Market **Created by:** Author

Source: Kaggle.com

Lastly, in **Table 7.7**, The order data by market shows strong regional variation, with the **United States** leading with 24,840 orders, followed by **Mexico** in **Latin America** with 13,292 orders and **France** in **Europe** with 13,221 orders. In **Africa**, **Nigeria** has the highest order count (2,309), while in **Pacific Asia**, **Australia** leads with 7,954 orders.

This distribution highlights **North America** and **Europe** as primary markets, suggesting a strong customer base and demand in these regions. **Latin America** also demonstrates notable activity, particularly in **Mexico** and **Brazil**, indicating growth potential in this market. Africa's lower order counts, in comparison, may point to untapped opportunities or logistical and market challenges. Meanwhile**, Pacific Asia** shows moderate activity, with **Australia** and **China** leading, reflecting a possible expansion pathway.

The varied order distribution across regions emphasizes the importance of tailored strategies for each market. For example, the **US** and **European** markets might benefit from enhanced logistics efficiency to maintain customer satisfaction, while the emerging markets in **Africa** and **Latin America** could require market development strategies to drive further engagement and streamline delivery processes.

7.4. Data Visualization

Data visualization is defined as the process of placing data into a visual context. This can be in the form of charts, plots, animations, infographics and the likes. The concept

behind it is to allow human beings to easily single out trends, outgrowths, and other patterns in data. Continuing using the Tableau public we created the below mentioned charts.

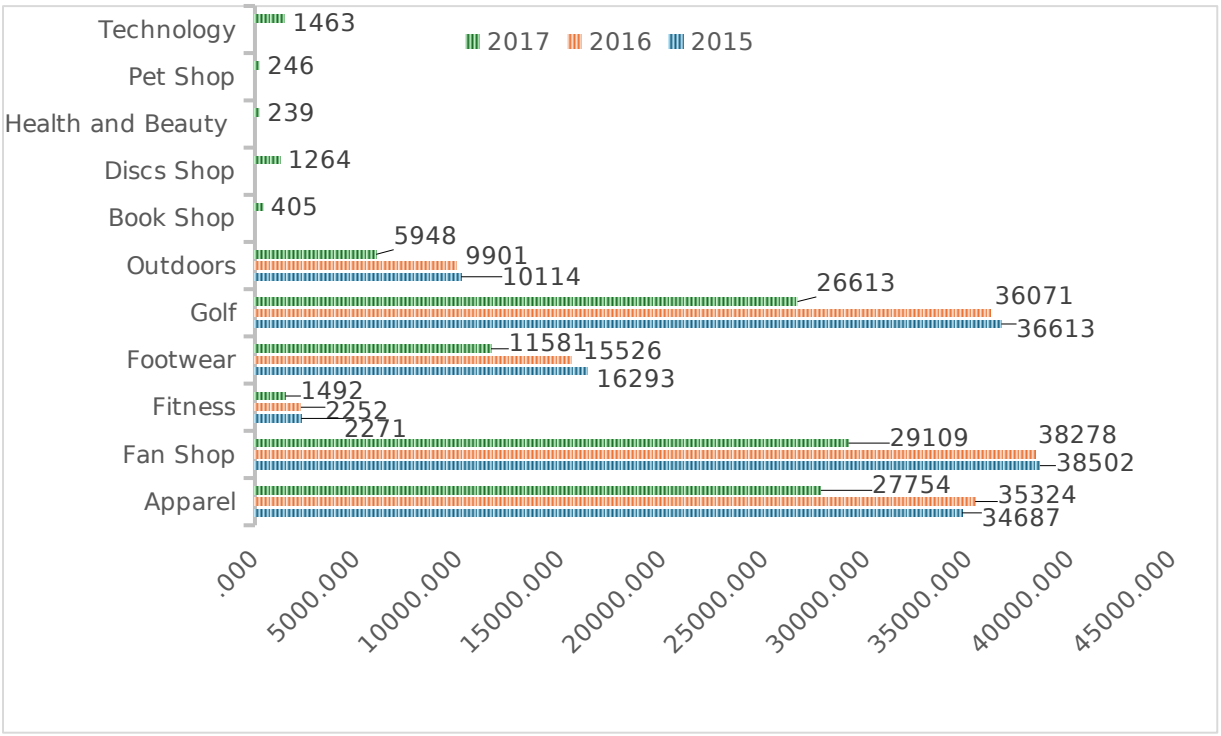


Chart 7.1: Units Sold per Product Category **Created by:** Author

Source: Kaggle.com

Chart 7.1 is a Bar Chart representation of **Table 7.1** where the aforementioned differences are clearly mentioned.

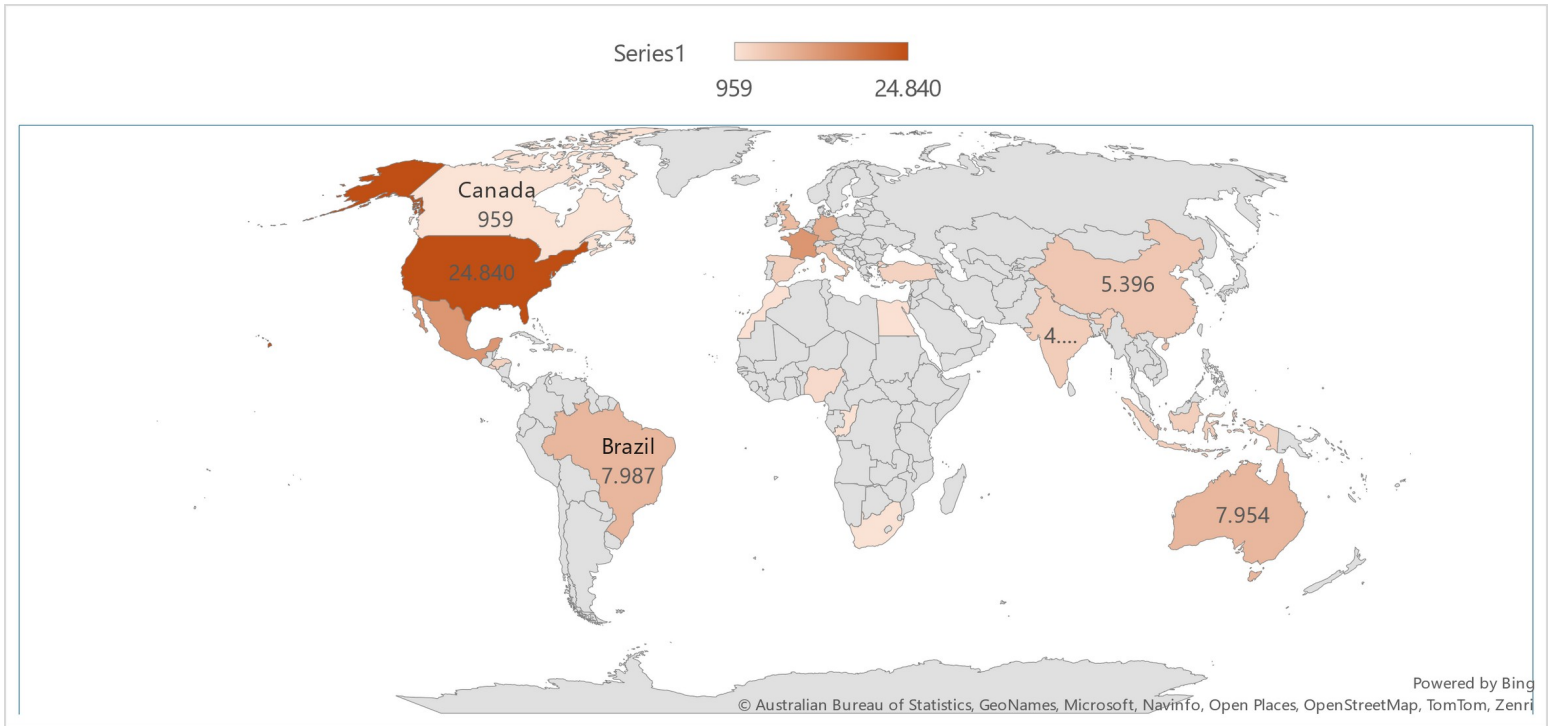


Chart 7.2: Top 5 Countries in Orders from every Market **Created by:** Author

Source: Kaggle.com

Chart 7.2 is a World Map Chart representation of **Table 7.6** in order to visualize the location and maybe the distance between the countries.

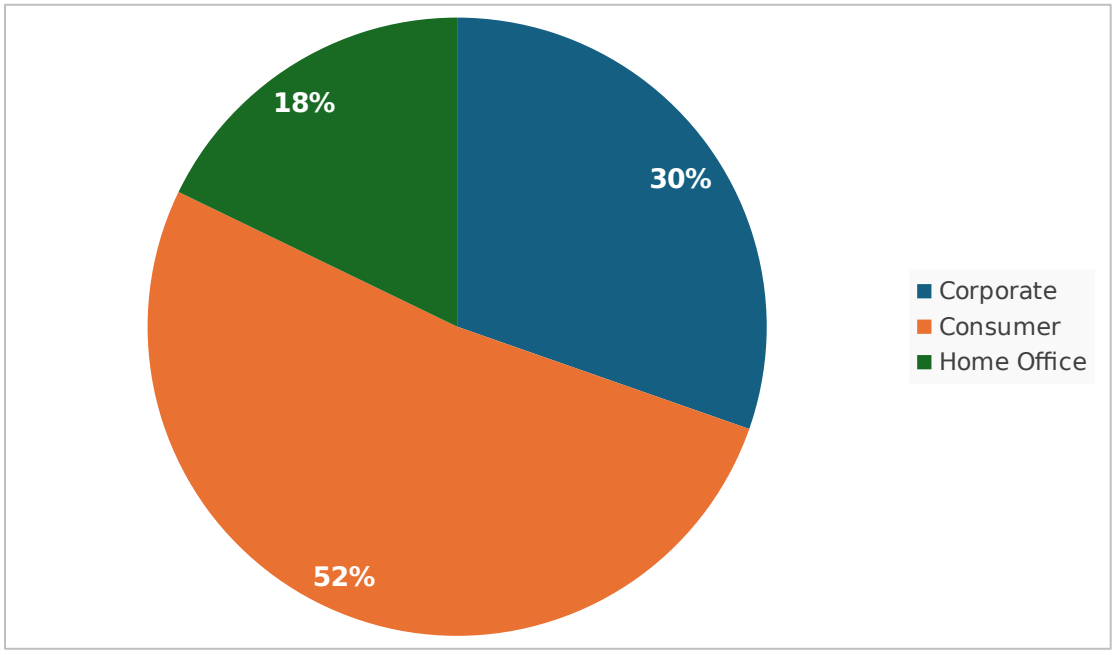


Chart 7.3: Customer Segment **Created by:** Author

Source: Kaggle.com

Chart 7.3 presents the percentage of its customer segment in a 2D Pie Chart where we can see that over half of the orders are made from a Consumer.

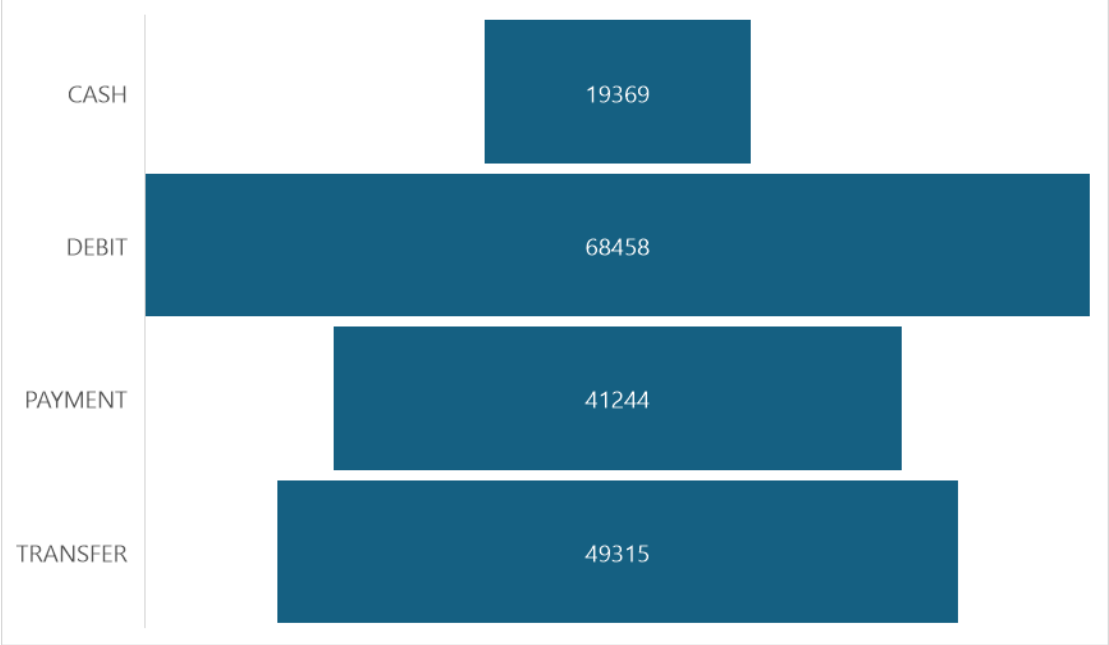


Chart 7.4: Form of Payment **Created by:** Author

Source: Kaggle.com

Chart 7.4 shows the Form of Payment that the Customers prefer in a Funnel Chart. As we can see DEBIT is the most common way of payment with TRANSFER coming in second.

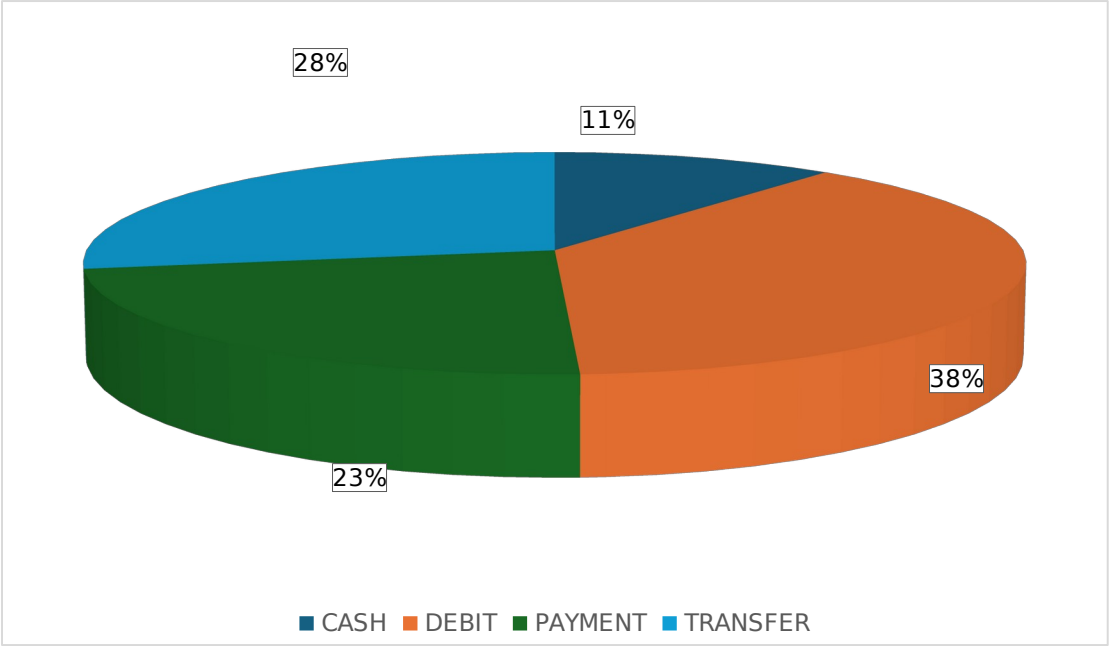


Chart 7.5: Form of Payment (Percentiles). **Created by:** Author

Source: Kaggle.com

Chart 7.5 presents **Chart 7.4** in percentage of the whole in a 3D Pie chart.

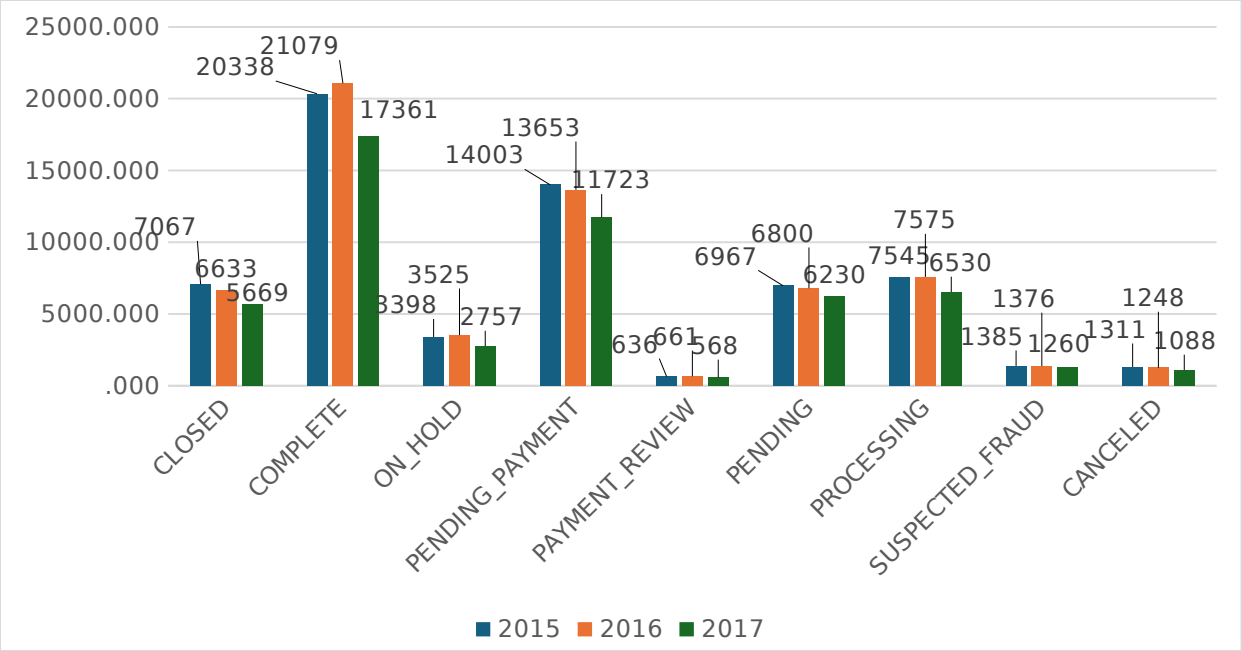


Chart 7.6: Order Status. **Created by:** Author

Source: Kaggle.com

Lastly, **Chart 7.6** visualizes **Table 7.6** in a Bar Chart and **Chart 7.7** presents **Table 7.6** in percentage of the whole and in a Doughnut Chart.

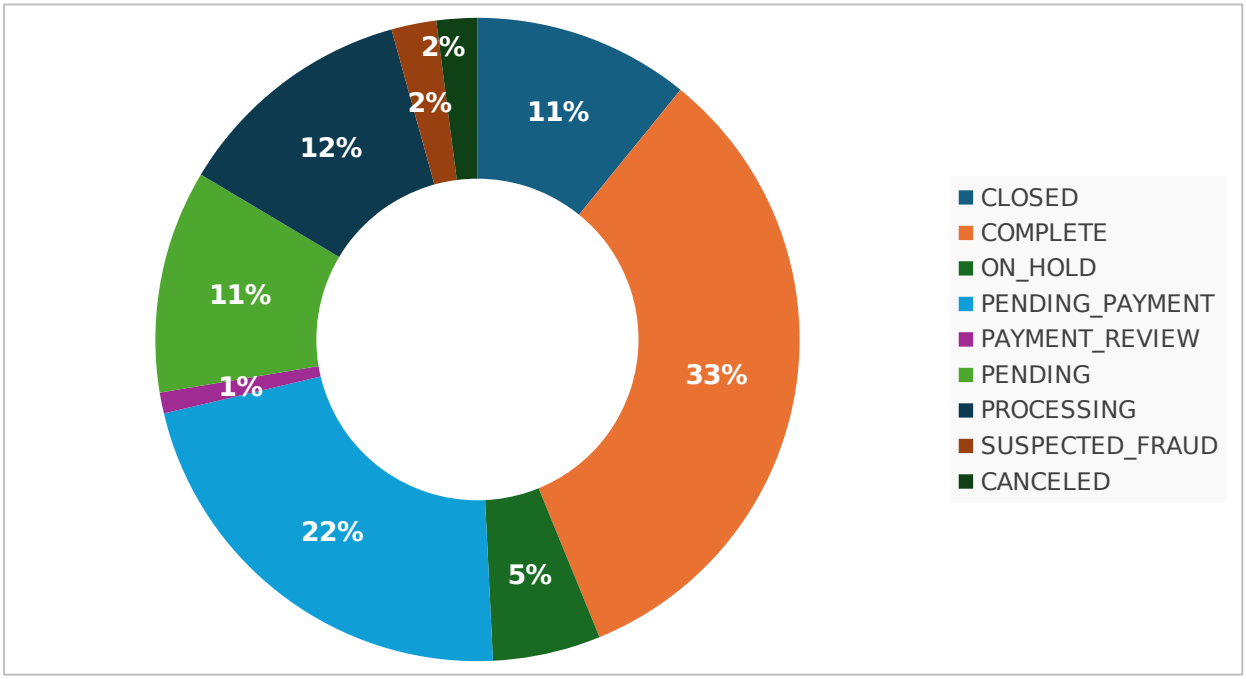


Chart 7.7: Order Status (Percentiles). **Created by:** Author

Source: Kaggle.com

These visuals have been helpful in shaping the analyses and provided consistency for the rest of the analyses' stories. In this way, providing the information in a form of visualizations, we have designed an effective instrument that helps to make a better decision, unite the teams and encourage the key strategic changes in the sphere of maritime.

Finally, the steps of visualization help not only to explain and interpret analyzed data, but also to apply it in practice to the extent that it will be possible to achieve the planned results.

8. CONCLUSION

This thesis has provided an in-depth exploration of Big Data science and its transformative applications within the maritime industry. Beginning with foundational concepts, including the definition, characteristics, and historical evolution of Big Data, the research examined its platforms, prevalent technologies, and the unique opportunities it offers across various sectors. Ethical and social implications of Big Data were also discussed, underscoring the importance of responsible data management practices.

The maritime industry, a critical pillar of global trade, presents a complex yet fertile ground for Big Data applications. This study highlighted various Big Data applications specific to maritime operations, from predictive maintenance and route optimization to environmental impact assessments. Case studies demonstrated the tangible impact of Big Data on maritime giants like Maersk, Shell, and the Port of Rotterdam, illustrating how these organizations are leveraging data for enhanced efficiency, safety, and sustainability. Moreover, a detailed step-by-step data analysis on a supply chain dataset was conducted to provide practical insights into data collection, processing, analysis, and visualization in this context. The analysis provides a short guide on how data analytics can significantly assist the logistics

departments by breaking down complex processes into manageable parts, ensuring well-informed decision-making.

However, the study also revealed substantial challenges hindering widespread adoption. Regulatory concerns, data quality, security issues, and the significant financial and organizational adjustments required to implement data-driven systems were prominent among these challenges. Additionally, the shortage of skilled professionals and the technological integration with existing systems pose significant barriers.

In conclusion, Big Data has transformative potential in the maritime industry. Although challenges persist, organizations investing in data science are poised to unlock unprecedented operational efficiency, cost savings, and environmental benefits. The insights gained in this thesis underline the pivotal role of Big Data in shaping a smarter, safer, and more sustainable maritime sector.

Future research in the maritime industry's Big Data applications should explore advanced predictive analytics, real-time processing via edge computing, and blockchain for enhanced data security. Additionally, there is a need for improved data interoperability standards to enable seamless data integration across platforms, alongside solutions focused on sustainability to reduce the environmental impact. Addressing the skills gap with specialized training programs and conducting economic analyses to understand the cost-benefit dynamics of Big Data in maritime contexts would further support the industry's transition towards data-driven innovation.

REFERENCES:

1. **ABS (2016).** “Data Integrity for Marine and Offshore Operations”. ABS Cyber Safety. [PDF]
2. **Agrawal, S., (2024). “**Snowflake Data Warehouse 101: A Comprehensive Guide”. Hevodata. [Link]

3. **Akpan, F., Bendiab, G., Shiaeles, S., Karamperidis, S., Michaloliakos, M. (2022)**. “Cybersecurity Challenges in the Maritime Sector”. Advances on Networks and Cyber Security. [Link],
4. **Alabdullah, B., Beloff, N., & White, M., 2018.** Rise of big data – issues and challenges. [PDF]
5. **Almeida, F., 2018.** Big Data: Concept, Potentialities and Vulnerabilities. [PDF]
6. **Artur Haponik., 2024.** “Introduction to Big Data Platforms” [Link]
7. **Aslam. S. (2023).** “IoT for the Maritime Industry: Challenges and Emerging Applications”. Proceedings of the 18th Conference on Computer Science and Intelligence Systems. [PDF]
8. **Bahrami, M. & Shokouhyar, S. (2021).** “The role of big data analytics capabilities in bolstering supply chain resilience and firm performance: a dynamic capability view”. Information Technology and People. [Link]
9. **Barrera, S., Lopez, L., Cedres, D. (2023).** “Maritime Surveillance by Multiple Data Fusion: An Application Based on Deep Learning Object Detection, AIS Data and Geofencing”. VISIGRAPP 2023 (Lisboa). [Link]
10. **Bassi, A. C. & Alves N. S. (2023).** “Challenges to Implementing Effective Data Governance”. University of Sao Paulo. [PDF]
11. **Bean, R. (2018).** “Big Data and AI Are Driving Business Innovation in 2018”. NewVantage Partners LLC. [Link]
12. **Bedford, T., & Cooke, R. (2001)**. “Probabilistic Risk Analysis: Foundations and Methods”. Cambridge University Press. [PDF]
13. **Binns, R., Veale, M., Van Kleek, M., et al. (2018).** 'A dangerous" black box": A critical review of the role of black box machine learning in consumer credit decision-making.' Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. [Link]
14. **Bollinger, M.A. and Kline, R.J., (2017).** “Validating Sidescan Sonar as a Fish Survey Tool over Artificial Reefs”. Coastal Research. [Link]
15. **Borthakur, D. (2007).** "The Hadoop Distributed File System: Architecture and Design." The Apache Software Foundation. [PDF]
16. **Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008).** “Time Series Analysis: Forecasting and Control” (4th Edition). Wiley. [PDF]
17. **Brandon, B. (2024).** “Navigating Regulatory Compliance with Modern Data Strategies”. IFA. [Link]

18. **Brouer, B. D., & Pisinger, D. (2017).** “Optimization in Container Liner Shipping”. INFORMS Journal on Computing. [PDF]
19. **Brown, T. B., Mann, B., Ryder, N., et al. (2020).** “Language Models are Few-Shot Learners”. arXiv preprint. [Link]
20. **BSR (2016)**. “Looking Under the Hood: ORION Technology Adoption at UPS”. BSR’s Center for Technology and Sustainability [Link]
21. **Calamp.** “Cargo Monitoring” [Link]
22. **Cargill (2024).**“Predictive Analytics Require Good Data”. Cargill [Link]
23. **Carnival: Guest Experience Data Platform** [Link]
24. **Carranza, A. (2023).** “ONE joins GBSN after trial of blockchain platform”. Supply Chain Drive. [Link]
25. **Chaal, M., (2018).** “Ship operational performance modelling for voyage optimization through fuel consumption minimization”. World Maritime University [PDF]
26. **Chachorowski, K., Solesvik, M., Kondratenko, Y. (2019).** “The Application of Blockchain Technology in the Maritime Industry”. Green IT Engineering: Social, Business and Industrial Applications. [Link]
27. **Chen et al., (2014)**. “Big Data: A Survey”. Springer Science Business Media New York 2014.[PDF]
28. **Chodorow, K., & Dirolf, M. (2019).** “MongoDB: The Definitive Guide: Powerful and Scalable Data Storage”. O'Reilly Media. [PDF]
29. **Choudhary, A. K., Harding, J. A., & Tiwari, M. K. (2009)**. "Data mining in manufacturing: a review based on the kind of knowledge." Journal of Intelligent Manufacturing. [Link]
30. **Codd, E. F. (1970).** "A Relational Model of Data for Large Shared Data Banks." Communications of the ACM. [PDF]
31. **Dabab, M., Craven, R., Barham, H., & Gibson, E., 2018**. Exploratory Strategic Roadmapping Framework for Big Data Privacy Issues. [PDF]
32. **Dacey, D. (2024). “**What is Databricks: The Power of Unified Analytics”. Dovetail. [Link]
33. **Daidola, John C.; Basar, Nedret S.; Reyling, Christopher J.; Johnson, Fontain M.; Walker, Richard T. (1991).** “Worldwide Buoy Technology”. U.S Department of Transportation. [PDF]

34. **Dalakakis. D., Vaitsos, G., Nikitakos, N., Papachristos, D., Dalakakis, A., (2021).** "Big data management in the shipping industry: examining strengths vs weaknesses and highlighting relevant business opportunities". International Association of Maritime Universities (IAMU) Conference. [PDF]
35. **Daniel, B. (2015).** "Big Data and analytics in higher education: Opportunities and challenges." British Journal of Educational Technology, [Link]
36. **Daniil, G & Boviatsis, M. (2022).** "Evaluation of Environmental Impact Assessment Factors in Maritime Industry". Journal of Environmental Science and Engineering. [Link]
37. **Darvazeh, S., & Vanani, I. (2020).** "Big Data Analytics and Its Applications in Supply Chain Management". New Trends in the Use of Artificial Intelligence for Industry. [Link]
38. **Dean, J., & Ghemawat, S. (2004)**. "MapReduce: Simplified Data Processing on Large Clusters." Proceedings of the 6th Symposium on Operating Systems Design and Implementation (OSDI). [PDF]
39. **Deane, F. & Nay, Z. (2021).** "Automated Decision-Making and Environmental Impact Assessments: Decisions, Data Analysis and Predictions". Queensland University of Technology. [Link]
40. **Denison, D. R. & Spreitzer, M. G. (1991).** "Organizational Culture and Organizational Development.". JAI Press Inc. [PDF]
41. **Dickson, A. G., et al., (2007)**."Guide to Best Practices for Ocean CO2 Measurements," PICES Special Publication. [Link]
42. **DNV GL (2019)**. "Big data and route optimization in shipping: Emission reduction case study". [PDF]
43. **Domingos, P. (2012).** "A Few Useful Things to Know About Machine Learning." Communications of the ACM [PDF]
44. **Durlik, I. & Miller, T., (2023).** "Navigating the Sea of Data: A Comprehensive Review on Data Analysis in Maritime IoT Applications". MDPI. [Link]
45. **Ehrahardt, m. & Burns, K. (2007)**. "Methods of Seawater Analysis" [Link]
46. **Ekathimerini.com (2024).** "Insights into the future of Shipping – A Deloitte survey". [Link]
47. **European Commission (2021). "**Horizon 2020 program: Optimizing shipping supply chain for emission reduction". [PDF]

48. **European Maritime Safety Agency (2019)** “Report on emission reduction through big data in shipping”. [Link]
49. **Floridi, L., & Taddeo, M. (2016).** “What is Data Ethics? Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences”, Oxford Internet Institute [PDF]
50. **Gatehouse Maritime (2024).** “Advanced Analytics” [Link]
51. **Gavalas, D., Syriopoulos, T. & Roumpis, E., (2022).** “Digital adoption and efficiency in the maritime industry”. Journal of Shipping and Trade. [Link]
52. **Glasson, J., Therivel, R., & Chadwick, A. (2013).** “Introduction to Environmental Impact Assessment”. Routledge. [PDF]
53. **Gong, G., (2022)**. "Weather routing for the refined management of passage," World Maritime University, [Link]
54. **Gordon, R. L., (2011)**, "Acoustic Doppler Current Profiler Principles of Operation: A Practical Primer," Teledyne RD Instruments. [PDF]
55. **Grolinger, K., Higashino, W. A., & Tiwari, A. (2016).** "Data management in cloud environments: NoSQL and NewSQL data stores." Journal of Cloud Computing: Advances, Systems and Applications [PDF]
56. **Gupta Jay et al., (2024)**. “Data Analytics in Telecommunications”. Quantexa. [Link]
57. **Hagemann, M. (2023).** “Cloud technology for the maritime industry”. LinkedIn. [Link]
58. **Han et al., (2011).** “Data Mining: Concepts and Techniques”. Morgan Kaufmann Publishers is an imprint of Elsevier. [PDF]
59. **Hashem, I. A. T., (2015).** "The role of big data in smart city." International Journal of Information Management [PDF]
60. **Healy et al., (2018)**. “Data Visualization: A Practical Introduction”. Princeton University Press. [Link]
61. **Hewitt, E., & Kopp, J. (2016).** “Cassandra: The Definitive Guide”. O'Reilly Media, Inc. [PDF]
62. **Hofmann-Wellenhof, B., et al. (1992).** "Global Positioning System: Theory and Practice," Springer, [PDF]
63. **Holling, C. S. (1978).** “Adaptive Environmental Assessment and Management”. International Institute for Applied Systems Analysis. [PDF]

64. **Holvik, J., Kongsberg, S., (1998).** “Basics of Dynamic Positioning”. DYNAMIC POSITIONING CONFERENCE. [PDF]
65. **Huang, J., & Ung, S. (2023)**. “Risk Assessment and Traffic Behaviour Evaluation of Ships”. Department of Merchant Marine, National Taiwan Ocean University. [Link]
66. **IBM. (2021). “**Blockchain and Maritime: Transforming Supply Chain Management”. [Link]
67. **Ieromonachou, P., & Nguyen, T. (2017).** “Big data analytics in supply chain management: A state-of-the-art literature review”. Computers & Operations Research. [Link]
68. **IMO** (International Maritime Organization). Guidelines for Environmental Impact Assessment in the Marine Environment. [Link]
69. **IMO.** “Vessel Traffic Services” [Link]
70. **International Council on Clean Transportation (2020).** “Big data-driven emissions monitoring: Ensuring compliance with IMO regulations”. [PDF]
71. **International Maritime Organization (2021)**. “IMO 2030 emissions targets: The role of big data in maritime sustainability”. [Link]
72. **International Transport Forum (2018)**. “Slow steaming and big data: The impact on emissions in the shipping industry”. [Link]
73. **IronHack (2023).** “The Crucial Role of Data Analytics in Enhancing Cybersecurity”. [Link]
74. **Ivanov, D., & Dolgui, A. (2020). “**A Digital Supply Chain Twin for Managing the Disruption Risks and Resilience in the Era of Industry”. Transportation Research Part E: Logistics and Transportation Review. [Link]
75. **Jerome P.-Y. (2005).** “X-band Radar Wave Observation System”. Minerals Management Service U. S. Department of the Interior. [PDF]
76. **Kar, U. (2019).** “Application of Artificial Intelligence in Automation of Supply Chain Management”. Journal of Strategic Innovation and Sustainability. [Link]
77. **Karlos Knox., (2024).** “What is Google Cloud Platform (GCP)”. Pluralsight. [Link]
78. **Katal, A., Wazid, M., & Goudar, R. H. (2013).** "Big data: Issues, challenges, Tools and Good Practices." 2013 International Conference on Emerging Trends and Applications in Computer Science [PDF]

79. **Khan et al., (2017)**. “Big Data Analytics: Challenges and Applications”. IEEE International Conference on Communications (ICC).[Link]
80. **Kharrazi, A., Qin, H., & Zhang, Y. (2016).** "Urban big data and sustainable development goals: Challenges and opportunities." Sustainability. [Link]
81. **Kitchin, R. (2014).** "The real-time city? Big data and smart urbanism." GeoJournal. [Link]
82. **Koga, S. (2015).** “Major challenges and solutions for utilizing big data in the Maritime Industry”. World Maritime University. [PDF]
83. **Konsberg Digital (2024).** “Power workflows with digital twin technology to generate more value”. [Link]
84. **Kreps, J., Narkhede, N., & Rao, J. (2011).** “Kafka: The Definitive Guide”. O'Reilly Media.[PDF]
85. **Kshetri, N. (2014).** "Big Data's Impact on Privacy, Security, and Consumer Welfare." Telecommunications Policy [PDF]
86. **Landry, H. (2024).** “Predicting the future with real time data”. Maersk. [Link]
87. **Lebold, M., et al. (1985)**, "Review of Vibration Analysis Methods for Gearbox Diagnostics and Prognostics," Proceedings of the 54th Meeting of the Society for Machinery Failure Prevention Technology. [Link]
88. **Liu, M., & Zhou, Q. (2020).** “Voyage performance evaluation based on a digital twin model”. IOP Conference Series Materials Science and Engineering. [Link]
89. **Liu, Z. & Gao, Z., (2010).** “A data mining method to extract traffic network for maritime transport management”. ScienceDirect. [Link]
90. **Lundh, M. & Huffmeier, J. (2023).** “The Impact of Digitalization on Maritime Safety and the Work Environment of the Crew”. Swedish Shipowners Association. [Link]
91. **Luo, C., 2017.** Survey of Parallel Processing on Big Data. [PDF]
92. **Lv, Z., Lv, H., Fridenfalk, M. (2023).** “Digital Twins in the Marine Industry”. Electronics. [PDF]
93. **M. Reichel, A. Minchev & N.L. Larsen. (2014).** “Trim Optimization - Theory and Practice”. the International Journal on Marine Navigation. [PDF]
94. **M.Collier., (2015).** “Microsoft Azure Essentials: Fundamentals of Azure.”. Microsoft Press. [PDF]
95. **Maersk (2021)**. “Sustainability report: 20% reduction in emissions through big data”. [Link]

96. **Maersk Line (2024).** “Captain Peter”. Maersk. [Link]
97. **Maghoromi, E. B. (2023).** “Impact of emerging technologies on maritime education and training: a phenomenological study”. World Maritime University. [PDF]
98. **Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011)**. "Big Data: The Next Frontier for Innovation, Competition, and Productivity." McKinsey Global Institute. [PDF]
99. **Mao, W. and Larsson, S. (2022).** “Increase shipping efficiency using ship data analytics and AI to assist ship operations”. Lighthouse Reports. [PDF]
100. **Marler, J. H., & Boudreau, J. W. (2016).** "An evidence-based review of HR Analytics." The International Journal of Human Resource Management. [Link]
101. **Matthews, H. S. & Hendrickson, C. (2008).** “The Importance of Carbon Footprint Estimation Boundaries”. Environmental Science and Technology. [Link]
102. **Medida, L. H. & Kumar, (2024).** “Addressing Challenges in Data Analytics: A Comprehensive Review and Proposed Solutions”. Critical Approaches to Data Engineering Systems and Analysis. [Link]
103. **Mirović, M., Milicevic, M. & Obradovic I. (2018).** “Big Data in the Maritime Industry”. University of Dubrovnik. [PDF]
104. **Mobley R. K. (2022)**. “AN INTRODUCTION TO PREDICTIVE MAINTENANCE”. Elsevier Science. [PDF]
105. **Morgan, R. K. (2012). “**Environmental Impact Assessment: The State of the Art”. Impact Assessment and Project Appraisal. [Link]
106. **Munim, H. Z., (2023).** “The Impact of Big Data Analytics Capabilities on the Sustainability of Maritime Firms”. Data Analytics for Supply Chain Networks. [Link]
107. **Nader Nada, Abusfian Elgelany., et al. (2014).** “Green Technology, Cloud Computing and Data Centers: The Need for Integrated Energy Efficiency Framework and Effective Metric”. International Journal of Advanced Computer Science and Applications [PDF]
108. **Nautinst**. “MARS”. [Link]
109. **O’Marah, K. (2023)**. “Container Shipping Gets Digital”. Forbes. [Link]
110. **O’Neil, C. (2016).** “Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy”. Crown Publishing Group [PDF]

111. **Oksavik, H.P. Hildre, Y. Pan, I. Jenkinson, B. Kelly, D. Paraskevadakis, R. Pyne, (2020).** "FUTURE SKILL AND COMPETENCE NEEDS". SkillSea. [PDF]
112. **Paine, L. P. ,2014.** "The Sea and Civilization: A Maritime History of the World". Atlantic Books
113. **Palmer M. (2024).** "Understanding ETL, Data Pipelines for Modern Data Architectures". O'Reilly Media. [PDF]
114. **Poulsen, R., T., (2022).** "Energy efficiency in ship operations - Exploring voyage decisions and decision-makers". Science Direct. [Link]
115. **Predictive Analytics Require Good Data**. [Link]
116. **Predictive Maintenance| New Perspectives**. [Link]
117. **Preskill, J. (2018).** "Quantum Computing in the NISQ era and beyond". Quantum, 2, 79. [PDF]
118. **Rao, R. & Mahesh, B. (2024).** "Challenges and Opportunities of Big Data Analytics for Maritime and Shipping Industry". International Journal of Engineering Technology and Management Sciences. [Link]
119. **Recogni (2022).** "Recogni reduces vessel fuel consumption using data analytics". Business Norway. [Link]
120. **Reddy, K., & Srinivas, K. (2014). "**A Survey on Platforms for Big Data Analytics". Journal of Big Data. [Link]
121. **Remote Container Management** [Link]
122. **Rodseth, J., Perera, P., Mo, B. (2016).** "Big Data in Shipping: Challenges and Opportunities". COMPIT 2016. [Link]
123. **Roh, M., (2013).** "Determination of an economical shipping route considering the effects of sea state for lower fuel consumption". International Journal of Naval Architecture and Ocean Engineering. [Link]
124. **Rolls-Royce Marine (2019)**. "Predictive maintenance and emission reduction in maritime shipping". [Link]
125. **Rossi, R. & Hirama, K., 2022.** "Towards a Conceptual Approach of Analytical Engineering for Big Data". [PDF]
126. **Rudnick, D. L., et al. (2004).** "Underwater Gliders for Ocean Research," Marine Technology Society Journal, [Link]
127. **Sagiroglu, S., & Sinanc, D. (2013).** "Big Data: A Review." 2013 International Conference on Collaboration Technologies and Systems (CTS). [PDF]

128. **Santamaria, C. & Alvarez, M. (2017).** “Mass Processing of Sentinel-1 Images for Maritime Surveillance”. Remote Sensing. [Link]
129. **SBN Technologies (2024).** “Importance of Data Analytics in Ship Management”. [Link]
130. **Schoier, G., & Gregorio, C. (2017).** “Clustering Algorithms for Spatial Big Data”. DEAMS, University of Trieste. [PDF]
131. **Shell (2024).** “Predictive maintenance: Industry buzzword or maintenance must-have?”. Shell [Link]
132. **Shilavadra, B., (2019).** “Voyage Data Recorder (VDR) on a Ship Explained”. Marine Insight. [Link]
133. **Shim, J., French, A., Guo, C. (2015).** “Big Data and Analytics: Issues, Solutions, and ROI”. Communications of the Association for Information Systems. [Link]
134. **Shruti M., 2024.** “Big Data Examples & Applications Across Industries” [Link]
135. **Skredderberget, A. (2024).** “The first ever zero emission, autonomous ship”. Yara. [Link]
136. **Smith, T. W. P., et al., (2014).** "Third IMO GHG Study 2014," International Maritime Organization, [Link]
137. **Souppez, E. G. (2017).** “Fostering Maritime Education Through Interdisciplinary Training”. International Conference on Maritime Policy, Technology and Education. [Link]
138. **Stopford M. (2009).** “Maritime Economics (3rd Edition)”. Routledge [PDF]
139. **Tan, P. (2020).** “Stream Processing with Apache Flink: Fundamentals, Implementation, and Operations”. Packt Publishing. [PDF]
140. **The Digital Port| Port of Rotterdam, (2024)** [Link]
141. **The Energy Dilemma: AI’s Double-Edged Sword** [Link]
142. **Tinga, T. & Tiddens, W. (2017).** “Predictive maintenance of maritime systems: Models and challenges”. 27th European Safety and Reliability Conference. [Link]
143. **Tinnes, E., Perez, F., Kandel, M., Probst, T. (2020).** “Decarbonizing logistics: Charting the path ahead”. McKinsley & Company. [Link]
144. **Tsitsarolis, A. (2023).** “Collision Risk Assessment and Forecasting on Maritime Data”. University of Piraeus. [PDF]

145. **Tu, E., Zhang, G., Rachmawati, L., (2014).** "Exploiting AIS Data for Intelligent Maritime Navigation," Proceedings of the 7th ACM SIGSPATIAL International Workshop on Computational Transportation Science. [Link]
146. **Veracity, (2024).** “Sharing of verified emissions data and ETS statements made easy as Veracity by DNV partners with maritime solution providers”. [Link]
147. **Vessel Finder.** “Historical AIS Data” [Link]
148. **Vouros, G., Doulkeridis, C., Santipantakis, G., and Vlachou, A. (2017)** "Taming big maritime data to support analytics.". University of Piraeus [PDF]
149. **Wamba, S. F., & Akter, S. (2019)**. “Understanding supply chain analytics capabilities and agility for data-rich environments”. International Journal of Operations & Production Management. [Link]
150. **Wartsila (2023).** “The Energy Dilemma: AI’s Double-Edged Sword”. Wartsila. [Link]
151. **Wasilewski, W., Wolak, K., Zaras, M. (2021).** “Autonomous shipping. The future of the maritime industry?”. MWSE. [Link]
152. **Wedel, M., & Kannan, P. K. (2016).** "Marketing Analytics for Data-Rich Environments." Journal of Marketing. [Link]
153. **Weintrit, A. (2009).** "The Electronic Chart Display and Information System (ECDIS): An Operational Handbook," CRC Press, [Link]
154. **White, T. (2015).** “Hadoop: The Definitive Guide. O'Reilly Media”. O’Reilly Media, Inc. [PDF]
155. **WindWard (2024).** “Predictive Intelligence”. [Link]
156. **Yadav, K., M. (2024)**. “Transforming the Shipping Industry: Integrating AI-Powered Virtual Port Operators for End-To-End Optimization”. Journal of Advanced Research Engineering and Technology (JARET). [PDF]
157. **Yan, Q. & Ferlita, A. (2024).** “A framework of a data-driven model for ship performance”. Ocean Engineering. [Link]
158. **Yang, Y. & Liu, Y. (2024).** “Harnessing the power of Machine learning for AIS Data-Driven maritime Research”. Science Direct. [Link]
159. **Zaharia, M., & Chowdhury, M. (2016). “**Learning Spark: Lightning-Fast Data Analytics”. O'Reilly Media. [PDF]

160. **Zhong, Q. & Liu, X.M., (2021). "**Monitoring Methods of Marine Pollution Range Based on Big Data Technology". Nature Environment and Pollution Technology. [Link]
161. **Zwally, J., & Brenner, A. (2001).** " Satellite Altimetry and Earth Sciences", Academic Press. [Link]
162. **Gao J., Cai Z., Sun W., Jiao Y. (2023)**. "A Novel Method for Imputing Missing Values in Ship Static Data Based on Generative Adversarial Networks". Section Ocean Engineering. [Link]
163. **Aggarwal C. (2017).** "Outliners Analysis". Springer [PDF]
164. **Ritu J.** (2024). "Data Parsing Explained: Definition, Benefits, and Techniques". Docsumo. [Link]
165. **Volosencu C., Dan P., Jurca L. (2009).** "Redundancy and Its Applications in Wireless Sensor Networks: A Survey". WSEAS TRANSACTIONS ON COMPUTERS. [Link]
166. **Laurent L., Francisco S. (2019).** "Introduction to Spatial Network Forecast with R". GitHub. [Link]
167. **Samuels J. A., (2024).** "One-Hot Encoding and Two-Hot Encoding: An Introduction". Imperial College. [Link]
168. **Effrosynidis D., Tsikliras A., Arampatzis A. (2020). "**Species Distribution Modelling via Feature Engineering and Machine Learning for Pelagic Fishes in the Mediterranean Sea". Applied Science. [Link]
169. **Manning C., Raghavan P., Schütze H., (2009).** "An Introduction to Information Retrieval". Standford University. [PDF]
170. **Kim K., Wang Z. (2018).** "Sampling Techniques for Big Data Analysis". International Statistical Review. [Link]
171. **IBM (2020).** "Sea change setting a new course for ocean research". IBM. [Link]
172. **Szondy (2022).** "Massive LNG tanker sails itself across the Pacific in shipping world first". New Atlas. [Link]
173. **Wardle D. (2024).** "The Role of Data and Analytics in Supply Chain Management". Global Partner Solutions. [Link]
174. **Tiwari, S., Wee, H. M., & Daryanto, Y. (2021)**. "Big data analytics in supply chain management: Current trends and future perspectives" Journal of Business Research. [PDF]