

**A
PROJECT REPORT
ON
“The Role of Big Data Analytics in Construction Project
Management”**

**UNDERTAKEN AT
“MIT School of Distance Education”
IN PARTIAL FULFILMENT OF
“Construction & Project Management”
MIT SCHOOL OF DISTANCE EDUCATION, PUNE.**

**GUIDED BY
“Ajay V. Vitalkar”**

**SUBMITTED BY
“Pranay Hanumant Dhawad”**

STUDENT REGISTRATION NO - MIT2021C01716

MIT SCHOOL OF DISTANCE EDUCATION PUNE - 411 038

YEAR 2022-2024

CERTIFICATE

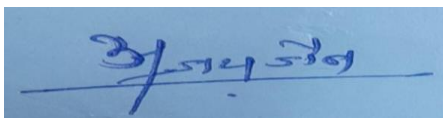
This is to certify that the project report titled “**The Role of Big Data Analytics in Construction Project Management**” has been successfully completed by **Pranay Hanumant Dhawad** as a part of the academic curriculum for **Post Graduate Diploma in Construction & Project Management** at **Maharashtra Institute of Technology, Pune**. The project report was undertaken during the 3rd Semester under the guidance of **Ajay Vitalkar** (Town Planner & Structural Engineer).

The project report has been found to be comprehensive, well-researched, and satisfactorily addresses the objectives outlined. The contents of the report reflect the original work of the author and demonstrate a deep understanding of the subject matter.

This certificate is awarded to acknowledge the successful completion of the project report and to signify the efforts and dedication put forth by the author.

Date: 22/02/2024

Sign –



Project Guide – Ajay V. Vitalkar (Town Planner & Structural Engineer)

Name – Pranay Hanumant Dhawad

Student ID - MIT2021C01716

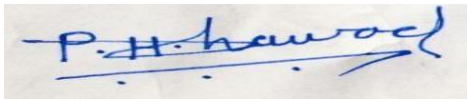
Institution Name – Maharashtra Institute of Technology, Pune

DECLARATION

I hereby declare that this project report entitled “**The Role of Big Data Analytics in Construction Project Management**” bonafide record of the project work carried out by me during the academic year **2022 - 2024**, in fulfillment of the requirements for the award of “**Construction & Project Management**” of MIT School of Distance Education.

This work has not been undertaken or submitted elsewhere in connection with any other academic course.

Sign:-

A handwritten signature in blue ink on a light-colored background. The signature appears to be 'P. H. Dhawad' with a horizontal line underneath.

Name: - Pranay Hanumant Dhawad

Student ID: - MIT2021C01716

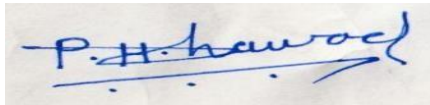
ACKNOWLEDGEMENT

I would like to take this opportunity to express my sincere thanks and gratitude to **“Guided by – Ajay V. Vitalkar”**, Faculty of MIT School of Distance Education, for allowing me to do my project work in your esteemed organization. It has been a great learning and enjoyable experience.

I would like to express my deep sense of gratitude and profound thanks to all staff members of MIT School of Distance Education for their kind support and cooperation which helped me in gaining lots of knowledge and experience to do my project work successfully.

At last but not least, I am thankful to my Family and Friends for their moral support, endurance and encouragement during the course of the project.

Sign:-

A handwritten signature in blue ink on a light-colored background. The signature appears to be 'P. H. Dhawad' with a stylized flourish at the end.

Name: - Pranay Hanumant Dhawad

Student ID: - MIT2021C01716

ABSTRACT

The construction industry, vital to global infrastructure development, faces intricate challenges in project management. Traditional approaches struggle to meet the demands of modern construction projects, necessitating innovative solutions. This research explores "The Role of Big Data Analytics in Construction Project Management," investigating the integration of big data technologies to enhance efficiency, mitigate risks, and optimize decision-making processes.

The study begins with an overview of big data analytics, highlighting its evolution and widespread adoption in various industries. Recognizing the unique challenges within the construction sector, the research explores the limitations of traditional project management methods, setting the stage for the examination of big data's transformative potential.

A comprehensive literature review delves into existing knowledge, examining prior studies on big data applications in construction and traditional project management challenges. Emphasis is placed on successful case studies, illustrating the practical implications and benefits of integrating big data analytics into construction project management.

The methodology section outlines the research design, participant selection, and data analysis methods employed. By investigating the technologies underpinning big data analytics, including Hadoop and Spark, the research aims to establish a clear understanding of their applications in construction project settings.

Case studies further illuminate the practical implementation of big data analytics in construction, presenting real-world scenarios and lessons learned. The findings and discussion section synthesizes research outcomes, emphasizing the impact of big data on project scheduling, cost estimation, and overall project success.

Recommendations for industry practitioners and stakeholders are provided based on the research findings, offering insights into effective strategies for implementing big data analytics in construction project management. The conclusion summarizes key takeaways, underscores the significance of the research, and proposes avenues for future exploration in this dynamic field.

This project contributes to the evolving discourse on construction project management, providing a roadmap for harnessing big data analytics to navigate challenges and optimize outcomes in an increasingly complex and dynamic industry.

TABLE OF CONTENTS

Chapter No.	Title	Page No.
1	Introduction	05
2	Literature Review	07
3	Methodology	09
4	Data Analysis and Interpretation	12
5	Conclusion / Findings	39
6	Suggestions / Recommendations	40

CHAPTER 1: INTRODUCTION

1.1 Background

The construction industry plays a pivotal role in shaping the global landscape, contributing significantly to economic development and societal progress. However, despite its undeniable importance, the industry grapples with numerous challenges in project management. Traditional methodologies, while time-tested, often fall short in addressing the complexities and uncertainties inherent in modern construction projects. This chapter serves as a gateway to understanding the transformative potential of big data analytics in overcoming these challenges and revolutionizing construction project management.

In recent years, the construction landscape has witnessed an unprecedented evolution, marked by increasing project complexities, tighter schedules, and higher stakeholder expectations. Traditional project management approaches, reliant on historical data and conventional methodologies, struggle to keep pace with the dynamic nature of contemporary construction projects. Issues such as cost overruns, delays, and suboptimal resource allocation have become persistent concerns, necessitating a paradigm shift in project management strategies.

1.2 Problem Statement

The limitations of traditional project management methods within the construction industry are pronounced. The reliance on historical data and static planning approaches often leads to inaccurate forecasting and inefficient resource allocation. This misalignment contributes to cost overruns, delays, and suboptimal project outcomes. As construction projects grow in scale and complexity, there is an urgent need for innovative solutions that can enhance the precision, efficiency, and adaptability of project management processes.

1.3 Objectives of the Study

The primary objective of this research is to investigate and analyze the role of big data analytics in construction project management. Specific goals include:

- To examine the limitations of traditional project management in the construction industry.
- To explore the potential benefits and applications of big data analytics in addressing construction project management challenges.
- To identify the key technologies underpinning big data analytics and their relevance to construction projects.
- To analyze case studies illustrating successful implementations of big data analytics in construction project management.

1.4 Significance of the Study

Understanding the role of big data analytics in construction project management is of paramount importance for stakeholders across the construction ecosystem. The findings of this research will contribute valuable insights into how big data can be leveraged to enhance decision-making, optimize resource allocation, and mitigate risks in construction projects. Ultimately, this study aims to provide a foundation for improved project management practices, fostering more successful and sustainable outcomes in the construction industry.

1.5 Scope and Limitations

While this research aims to provide a comprehensive exploration of the role of big data analytics in construction project management, it is essential to acknowledge certain limitations. The scope of the study is focused on the integration of big data analytics technologies in construction, and the findings may not be universally applicable to all construction contexts. Additionally, the rapidly evolving nature of technology and industry practices may pose challenges in capturing the most current developments. Nevertheless, this study strives to offer valuable insights that contribute to the ongoing discourse on enhancing construction project management through innovative technologies.

Chapter 2: Literature Review

2.1 Introduction to Big Data Analytics

The term "big data analytics" refers to the process of examining and interpreting large volumes of complex data to reveal patterns, trends, and associations. In recent years, big data analytics has emerged as a transformative force across various industries. Its ability to extract valuable insights from vast datasets has led to improved decision-making and enhanced operational efficiency. In the context of construction project management, understanding the fundamentals of big data analytics is crucial for appreciating its potential applications and benefits.

2.2 Big Data in Construction

The construction industry, with its intricate network of interconnected processes, generates vast amounts of data at every stage of a project. Big data analytics in construction involves harnessing this data to derive actionable intelligence. Previous studies have explored the applications of big data in construction, showcasing its potential to revolutionize project management. From real-time monitoring of construction sites to predictive analytics for risk management, the literature reveals a diverse array of possibilities for leveraging big data in construction projects.

2.3 Construction Project Management

Traditional project management methodologies in construction often rely on historical data and deterministic planning. This approach, while proven over time, struggles to adapt to the dynamic nature of construction projects. Literature on construction project management highlights the challenges posed by uncertainties, unforeseen events, and the need for real-time decision-making. As construction projects become larger and more complex, there is an increasing awareness of the limitations of traditional project management practices.

2.4 Integration of Big Data Analytics in Construction Project Management

The integration of big data analytics into construction project management holds the promise of addressing longstanding challenges. The literature emphasizes the potential benefits, such as improved project scheduling, cost estimation accuracy, and enhanced risk management. Researchers and industry experts have explored the theoretical frameworks and practical applications of big data analytics in optimizing resource allocation, streamlining communication, and ultimately improving project outcomes. However, challenges and risks associated with the integration process are also acknowledged, underscoring the need for a nuanced understanding of the complexities involved.

2.5 Benefits of Big Data Analytics in Construction

Studies investigating the benefits of big data analytics in construction project management reveal a multitude of advantages. Real-time data analysis allows for proactive decision-making, enabling project managers to respond promptly to emerging challenges. Predictive analytics

aids in risk mitigation by identifying potential issues before they escalate. Additionally, improved accuracy in cost estimation and resource allocation contributes to overall project efficiency. Case studies illustrating successful implementations further emphasize the tangible positive impact of big data analytics on construction projects.

2.6 Challenges and Risks

While the potential benefits are substantial, the literature also acknowledges challenges and risks associated with integrating big data analytics into construction project management. Issues such as data privacy, security concerns, and the complexity of implementing new technologies within existing project management frameworks are explored. Understanding these challenges is essential for developing strategies to mitigate risks and ensure a smooth transition to a data-driven approach in construction project management.

2.7 Frameworks and Models

Various theoretical frameworks and models have been proposed to guide the integration of big data analytics into construction project management. These frameworks provide a structured approach for leveraging big data technologies, emphasizing factors such as data collection, processing, analysis, and decision-making. Reviewing these frameworks aids in synthesizing existing knowledge and understanding the key components of successful big data integration in construction projects.

2.8 Case Studies: Applications in Construction Project Management

Case studies represent a critical component of the literature, offering tangible examples of successful big data analytics applications in construction project management. Examining specific projects provides insights into the challenges faced, the strategies employed, and the outcomes achieved. Case studies serve as practical illustrations of the transformative potential of big data analytics, offering valuable lessons for industry practitioners and researchers alike.

2.9 Summary of Key Findings

The literature review concludes by summarizing the key findings from existing studies. Emphasis is placed on the potential benefits, challenges, theoretical frameworks, and practical applications of big data analytics in construction project management. This synthesis of literature sets the stage for the subsequent chapters, providing a comprehensive foundation for the empirical investigation into the role of big data analytics in construction project management.

Chapter 3: Methodology

3.1 Research Design

The research design for this study employs a mixed-methods approach, combining qualitative and quantitative methods to comprehensively investigate the role of big data analytics in construction project management. The qualitative aspect involves an in-depth exploration of existing literature, theoretical frameworks, and case studies, while the quantitative component focuses on gathering empirical data through surveys and interviews.

3.1.1 Literature Review

The qualitative phase begins with an extensive literature review, as detailed in Chapter 2. This involves a systematic examination of peer-reviewed articles, books, and conference proceedings related to big data analytics, construction project management, and their intersection. The synthesis of existing knowledge provides a theoretical foundation for the study, informing the development of research questions and guiding the subsequent empirical investigation.

3.1.2 Theoretical Framework

Building upon the insights gained from the literature review, a theoretical framework is established to guide the empirical investigation. This framework outlines the key variables, relationships, and concepts that will be explored in the study. It serves as a roadmap for the research process, ensuring a structured and systematic approach to data collection and analysis.

3.2 Participants

The study involves participants from diverse backgrounds within the construction industry, including project managers, construction professionals, and stakeholders involved in decision-making processes. A purposive sampling method is employed to ensure that participants possess relevant experience and insights into the use of big data analytics in construction project management.

3.2.1 Survey Participants

A survey is administered to a sample of construction professionals. The survey aims to gather quantitative data on the current state of big data analytics adoption, perceived benefits, and challenges faced in construction project management. The survey questionnaire is designed based on insights gained from the literature review and theoretical framework, ensuring alignment with the research objectives.

3.2.2 Interview Participants

In-depth interviews are conducted with a subset of survey participants to delve deeper into their experiences and perspectives. The semi-structured interviews allow for a nuanced exploration

of specific themes, such as successful implementation strategies, encountered challenges, and recommendations for improving big data analytics integration in construction project management. The selection of interview participants is based on their willingness to provide detailed insights and their varied experiences within the construction industry.

3.3 Data Collection

Data collection involves a combination of surveys, interviews, and the analysis of existing case studies. The survey is distributed electronically to the selected participants, with a focus on gathering quantitative data on the adoption and impact of big data analytics in construction project management. Simultaneously, interviews are conducted either in person or virtually, recorded and transcribed for further analysis. The case studies are examined to supplement the empirical data with real-world examples of successful big data analytics applications.

3.4 Data Analysis

Quantitative data from the surveys are analyzed using statistical tools, such as descriptive statistics and inferential analysis, to identify trends, patterns, and correlations. Qualitative data from interviews are subjected to thematic analysis, allowing for the identification of recurring themes and the generation of in-depth insights. The case studies are qualitatively analyzed to extract lessons learned and best practices in the application of big data analytics in construction project management.

3.5 Validity and Reliability

To ensure the validity of the study, multiple data sources are triangulated, including survey responses, interview transcripts, and case study findings. The use of established theoretical frameworks enhances the reliability of the study by providing a structured and systematic approach to data analysis. Additionally, the inclusion of diverse perspectives from participants with varying roles in construction projects contributes to the robustness of the research findings.

3.6 Ethical Considerations

Ethical considerations include obtaining informed consent from participants, ensuring confidentiality and anonymity, and adhering to ethical guidelines for research involving human subjects. All data are stored securely, and participant identities are protected throughout the research process. Ethical approval for the study is obtained from the relevant institutional review board.

This mixed-methods approach ensures a comprehensive exploration of the role of big data analytics in construction project management, combining quantitative insights with qualitative depth to provide a nuanced understanding of the subject matter. The systematic research design

Enhances the reliability and validity of the study, paving the way for meaningful conclusions and recommendations.

Chapter 4: Big Data Analytics Technologies

4.1 Introduction

This chapter explores the foundational technologies that underpin big data analytics and their specific applications in the context of construction project management. Understanding these technologies is crucial for appreciating how data is collected, processed, and analyzed to derive meaningful insights that enhance decision-making and project outcomes.

4.2 Overview of Big Data Technologies

Big data technologies encompass a diverse set of tools and frameworks designed to handle the volume, velocity, and variety of data generated in construction projects. Key technologies include:

4.2.1 Hadoop

Hadoop is an open-source framework designed for distributed storage and processing of large datasets. It utilizes a distributed file system (HDFS) and a processing engine (Map Reduce) to manage and analyze data across clusters of commodity hardware. In construction project management, Hadoop's scalability and fault tolerance make it valuable for processing large datasets related to project progress, resource utilization, and cost tracking.

4.2.2 Apache Spark

Apache Spark is a fast and general-purpose cluster computing system that enhances data processing speeds compared to traditional MapReduce. Its in-memory processing capability makes it well-suited for iterative algorithms and interactive data analysis. Spark's versatility is applicable in construction for real-time analytics, allowing project managers to make timely decisions based on up-to-date information.

4.2.3 NoSQL Databases

Traditional relational databases may struggle to handle the sheer volume of unstructured data in construction projects. NoSQL databases, such as MongoDB and Cassandra, provide scalable and flexible storage solutions for diverse data types. These databases are instrumental in managing project data that includes sensor readings, images, and other non-tabular information.

4.3 Applications in Construction

The integration of these big data technologies offers a range of applications in construction project management:

4.3.1 Real-Time Monitoring

Big data analytics enables the continuous monitoring of construction sites through the collection and analysis of data from various sensors, wearables, and IoT devices. Real-time insights into equipment usage, worker safety, and project progress empower project managers to proactively address issues and optimize workflows.

4.3.2 Predictive Analytics for Project Scheduling

By analyzing historical project data and external factors, big data analytics facilitates predictive modeling for project schedules. Predictive algorithms can identify potential delays, allowing project managers to adjust timelines and allocate resources more effectively.

4.3.3 Cost Estimation and Resource Allocation

Big data technologies contribute to accurate cost estimation by analyzing historical project costs, market trends, and supplier data. This aids in optimizing resource allocation, preventing overruns, and ensuring efficient utilization of financial resources.

4.3.4 Risk Management

Advanced analytics help identify and assess risks by analyzing historical data and identifying patterns associated with project challenges. This proactive approach allows project managers to implement mitigation strategies before risks escalate.

4.4 Challenges and Considerations

Despite their potential benefits, the implementation of big data technologies in construction project management comes with challenges:

4.4.1 Data Privacy and Security

Construction projects involve sensitive data, and ensuring the privacy and security of this information is paramount. Big data implementations must incorporate robust security measures to safeguard against unauthorized access and data breaches.

4.4.2 Integration with Existing Systems

Integrating big data technologies with existing project management systems can be complex. Compatibility issues and the need for training may pose challenges, emphasizing the importance of a well-planned integration strategy.

4.4.3 Scalability and Infrastructure

Scalability is a critical consideration, especially as construction projects vary in size and complexity. Adequate infrastructure and scalable solutions are required to accommodate the growing volume of project data.

4.5 Case Studies

This chapter includes case studies showcasing successful implementations of big data analytics technologies in construction project management. Examining real-world examples provides

insights into the practical applications, challenges faced, and outcomes achieved, offering valuable lessons for industry practitioners.

4.6 Future Trends

The chapter concludes by exploring emerging trends in big data analytics technologies for construction project management. Innovations such as edge computing, machine learning, and the integration of AI are discussed, providing a glimpse into the evolving landscape of technology applications in the construction industry.

By comprehensively examining these big data technologies and their applications, this chapter lays the groundwork for understanding the practical implications and potential challenges of integrating these tools into construction project management processes.

Materials and Methods

This study follows a multi-stepped approach for reviewing the studies on big data in construction. First, a comprehensive literature retrieval mechanism is adopted from published literature and modified accordingly to retrieve pertinent literature on big data in construction. This is followed by analyses of the retrieved articles in the shape of preliminary analyses, BDE, processing, storage, analytics, and statistical and data mining approaches in relation to the construction industry. These steps are subsequently explained. An extensive literature search was carried out to identify peer-reviewed papers related to big data and construction since 2010, following the approaches adopted in recent studies [31, 32]. This was conducted in order to keep a recent focus and study current articles on big data in construction. Some preliminary analyses, as subsequently discussed, highlighted that big data in construction received more attention in 2010 and onwards; hence, the review period of 2010 and onwards makes sense. A number of scholarly research platforms, including Google Scholar, Scopus, Science Direct, Springer, Elsevier, and IEEE Explore, were consulted for literature search based on the high volume of high-quality research papers available on these platforms following recent studies [33–35]. Once the search engines were selected, a combination of different keywords was developed to identify the most useful publications for this study in the next step. The keyword combinations were developed in a tier-based approach, such that terms related to big data, such as “big data”, “big data analysis”, “big data volume”, and “big data analysis tools” fell into category 1 (S1).

Similarly, all keywords pertaining to construction, such as “construction”, “construction management”, and “construction industry”, were classified into category 2 (S2). Different combinations of keywords from both categories were used to retrieve the most relevant publications. Examples of keyword combinations include big data in construction, big data for construction management, construction management, and big data, etc. Search category was further restricted by including only those papers that were published in 2010 or later years. Since big data technology was used robustly in the

last decade, research publications prior to 2010 were left out. Concept papers, editorials,

notes, perspectives, closures, discussions, conference papers, and others were also excluded from the search to ensure the inclusion of original research papers only. Other publications dealing with classical definitions were also excluded.

Using different combinations of the keywords to identify papers published from 2010 onwards led to a total of more than 10,000 papers being retrieved from the mentioned search engines. The list of articles was narrowed down using the detailed inclusion criteria set for this study. This included removing duplicates and other exclusions, as previously mentioned, which brought the search results down to around 4000 papers. This was further narrowed down in a stepwise manner to ensure that only those papers were included that fit the scope of the current study. In the final step, the content of the papers was analyzed to determine their suitability for this study, resulting in a total of 156 papers.

Figure 1 shows an overview of the different ways in which research studies have addressed the use of big data in construction. There has been a rise in the interest in bigdata usage for the construction industry since 2016. However, the interest has been limited in terms of analyses scope as the trends have remained steady. As shown in Figure 1, the publications on this topic have followed similar terms and research themes over the last few years, leading to gradual evolution. For example, in 2016, most papers related to big data and construction focused on the use of cloud computing, while 2017 saw a trend of developing models and frameworks for implementing big data in the construction industry. Similarly, in 2018 and 2019, researchers have mainly explored how different bigdata models could be implemented within the construction industry. Recently, the research Focus has shifted to using big data in real-time construction projects and identifying how these technologies could be harnessed for developing futuristic construction projects.

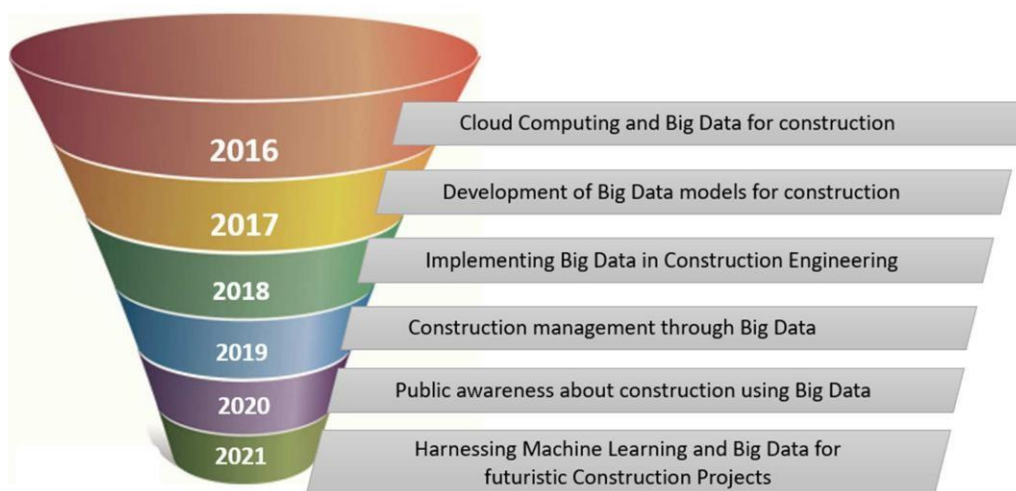


Figure 1. Funnel diagrams showing trends in big data research in construction since 2016.

In addition to big data, some other technologies and methods have been researched in the last couple of years for improving the construction industry. There is a great overlap in the types of technologies studied simultaneously for developing models that could guide future research in the construction industry. Figure 2 shows the overlapping tools and technologies identified from recent literature. It can be observed that big data is not standalone; rather, it depends on other tools and methods, including data analytics, ML, pattern recognition, statistics, deep learning, and artificial intelligence (AI). All these tools and technologies are used in different combinations for developing models that could be used in real time for construction projects. The reliance of all these tools on each other is an important factor to consider when developing construction projects as the computational aspects of the project can only be as good and true and the depth of research is performed for developing and testing the algorithms and frameworks. The construction industry greatly benefits from the overlapping fields of big data technologies. The use of big data requires data mining which generates enormous datasets. The bulk of construction-related data makes the use of statistics inevitable.

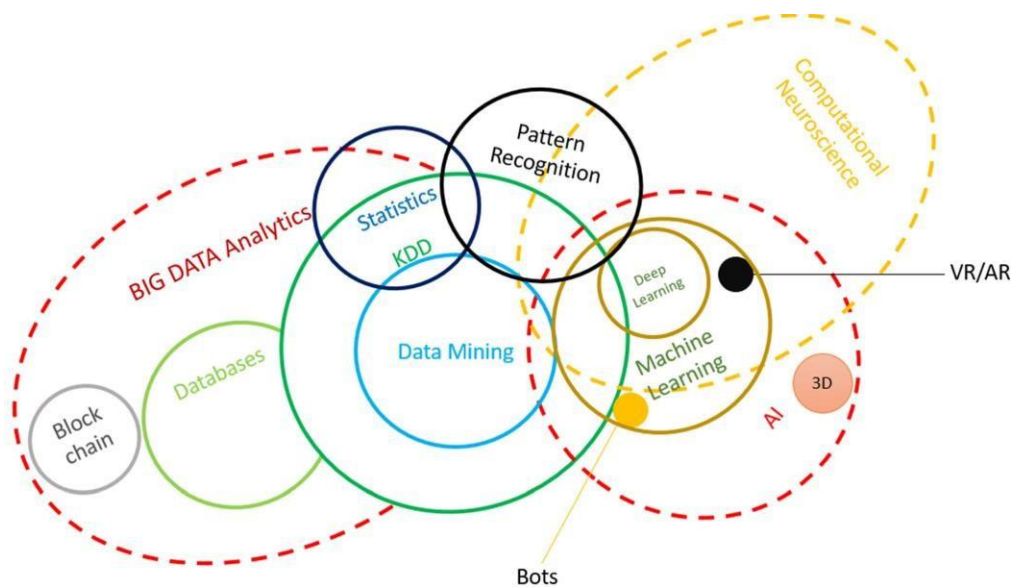


Figure 2. Overlapping fields of research contributing to big data.

Along with data management, statistical analysis, and big data analytics, several different techniques and resources come into use. For example, machine learning tools and artificial intelligence play a crucial role in the construction industry in conjunction with big data. The overlap of all the different fields shown in Figure 2 shows how the field of construction is laden with the use of different technologies, each of which is somehow associated with big data. The use of computational models, databases, deep learning, pattern recognition, virtual reality, bots, and augmented reality contributes to the application of big data in the construction industry. An in-depth analysis of the big data applications and the use of technology in the construction industry results in a much more complex overlap than shown here. However, the core aim of using different technologies is to simplify how datasets can be used to guide future construction projects. Recognizing data patterns and understanding how each dataset fits the needs of a construction project is only possible if the dataset has been analyzed, critically appraised, and classified for its specific usage. The guiding principle here is to use modern technology to upgrade and update the ways in which information could be streamlined for the benefit of different projects. For example, identifying the materials that best suit a particular structure, developing project timelines, and streamlining the resources can become much more straightforward if the construction projects are developed with the help of big data technologies.

As shown in Figure 2, different technologies in the construction industry overlap in different ways. Integrating big data in the construction industry is possible through the combined use of other technologies such as machine learning, AI, VR, AR, pattern recognition, and other such methods.

Preliminary Analyses

As mentioned in the method, some preliminary analyses were conducted on the retrieved articles, including the keywords analysis and the countries of origin of the articles following recently published articles [31,35]. Before this, a basic Google Trend (r) search was conducted using trends.google.com (accessed on 20 November 2021). A comparison was made for three iterations of the keywords previously mentioned. These included construction big data (keyword 1), big data in construction (keyword 2), and big data for construction management (keyword 3). As shown in Figure 3, the earliest attention paid to big data in construction was reported in 2010. This was reported for keyword 1, followed by keyword 2 in 2013 and keyword 3 in 2014. Two clusters are clearly visible from Figure 3. The initial interest cluster showed when big data focused on construction and the spike in interest cluster. The first cluster is evident in 2010–2014, whereas the spike in interest cluster started in 2016. This shows the hotness or relevance of the topic under investigation in the current study.

After the Google Trend analyses, the retrieved articles were analyzed using Vos Viewer[®] tool. The first analysis was that of keywords. The natural distribution of keywords retrieved from the articles shows five distinct clusters: education, city and region, disaster and human interactions, knowledge management, and technology

management in relation to construction, as given in Figure 4. The overall top keywords in

order of priority retrieved from these articles included big data, information management, AI, data mining, internet of things, ML, advanced analytics, data technologies, students, data handling, digital storage, colleges and universities, smart city, decision making, cloud computing, construction industry, and others. These are based on the appearance of the keywords in the titles, abstract, and keywords of a minimum of 30 papers. These keywords are in line with the natural clusters highlighted in Figure 4.

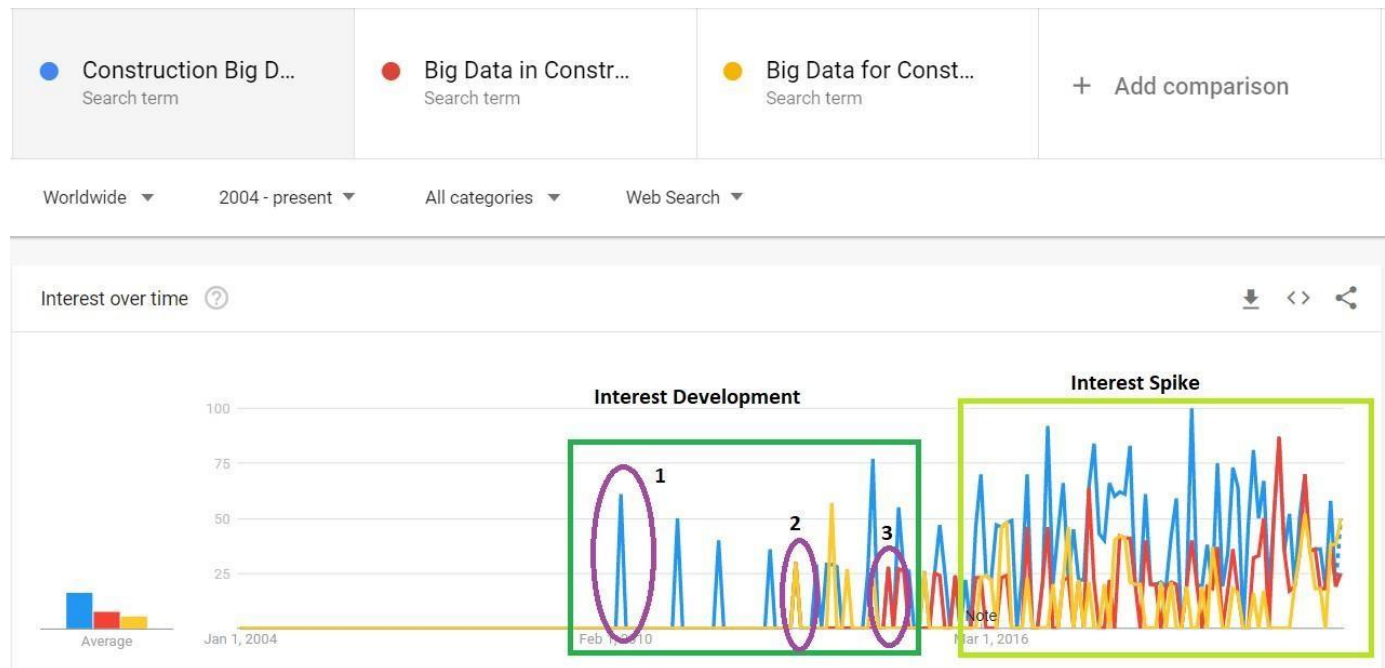


Figure 3. Google Trends analyses for big data in construction, showing interest development and spike in interest.

In another analysis, the top 10 contributing countries to big data research in construction were investigated. These are China, United States, United Kingdom, Russian Federation, Australia, India, South Korea, Germany, Spain, and Italy in terms of the number of contributions as shown in Figure 5. The colors in the country box show the countries with the strongest collaborations, whereas the size of the box refers to the number of papers. For example, most of the papers authored by Chinese authors are in collaboration with authors from Australia, New Zealand, and Indonesia.

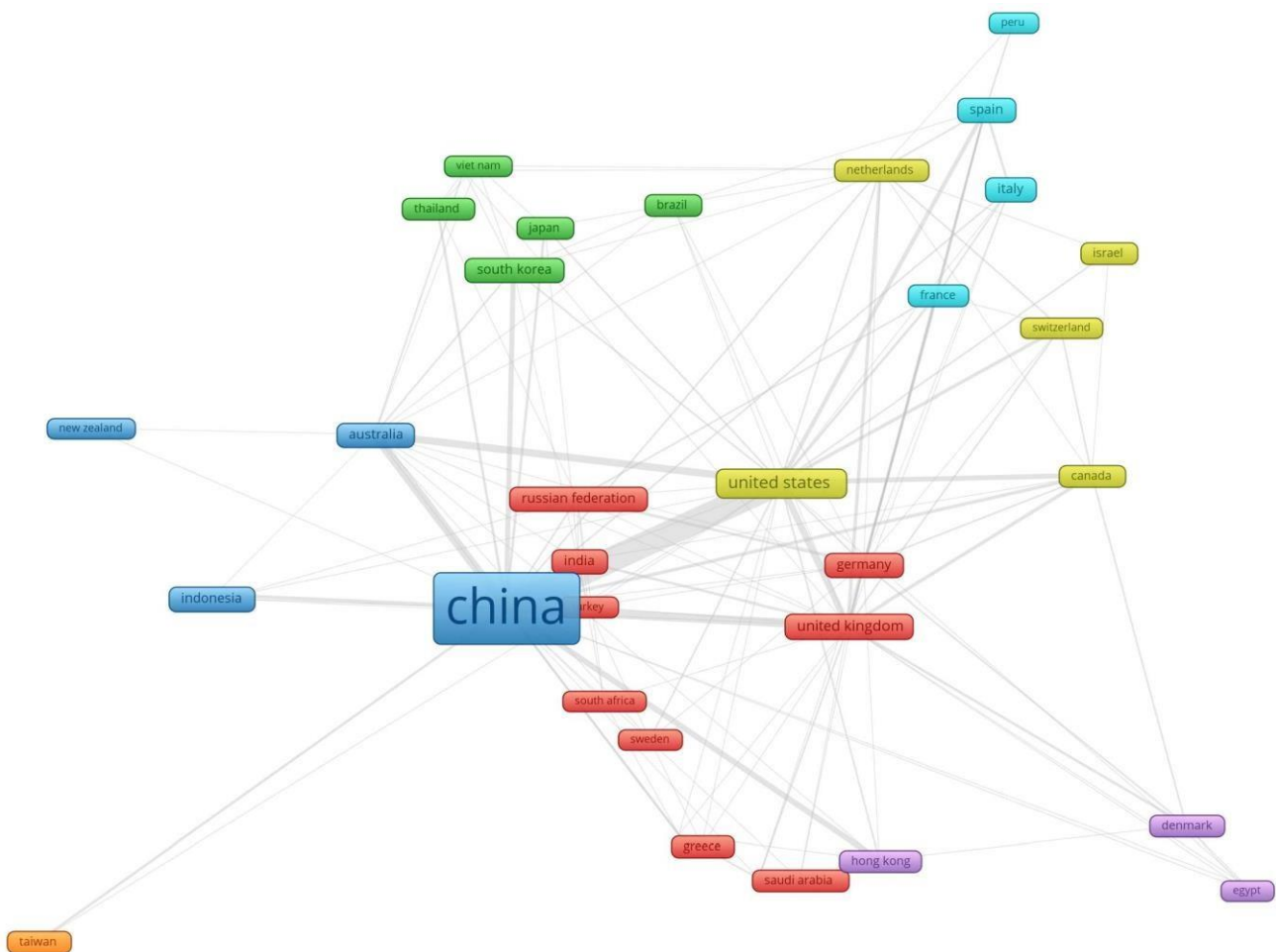


Figure 5. Countries conducting big data research in construction based on reviewed literature.

Big Data Engineering (BDE)

Big data analytics (BDA) is supported by BDE that provides a framework to conduct it. BDE has tremendous applications in construction. It has been used for BIM to improve project management [36]. It has also been used to improve building design and for effective performance monitoring [37], project management, safety, energy management, decision-making design frameworks, resource management [38], quality management, waste management, and others [24].

To understand BDE, it is important to discuss big data platforms. These platforms are

divided into two groups based on variations in their inherent characteristics. These include horizontal scaling platforms (HSP) and vertical scaling platforms (VSP). HSP utilizes multiple servers by distributing processing across them and bringing new machines into the cluster. VSPs are single-server-based configurations that achieve the scaling by up- grading the hardware of the related server. In construction, HSPs have been used for waste management [25], profitability performance [39], smart road construction, and others [40].

Similarly, VSPs have been reported in one-off construction projects [41], transportation [42], and others. This paper focuses on HSPs, particularly Berkeley Data Analytics Stack (BDAS) and Hadoop.

Recently, BDAS has been in the limelight since it has greater performance gains over Hadoop. However, as it is quite recent, it suffers the drawback of limitation in available supporting tools. On the other side, Hadoop has been widely utilized in big data applications. The tools offered by these platforms are useful in the storage and processing of big data. For instance, Bilal et al. [39] investigated the profitability performance of construction projects using big data and used Hadoop Distributed File System (HDFS) for managing the data within the staging area while employing Resource Description Framework (RDF)-enabled Network Data Model (NDM) for storing the persistent data. Similarly, Jun Ying et al. [43] investigated the development and implementation of BDAS by the relevant building authorities in Singapore, which has enhanced knowledge and expertise in build ability. An overview of big data classification into BDE and BDA is shown in Figure 6 and subsequently explained.

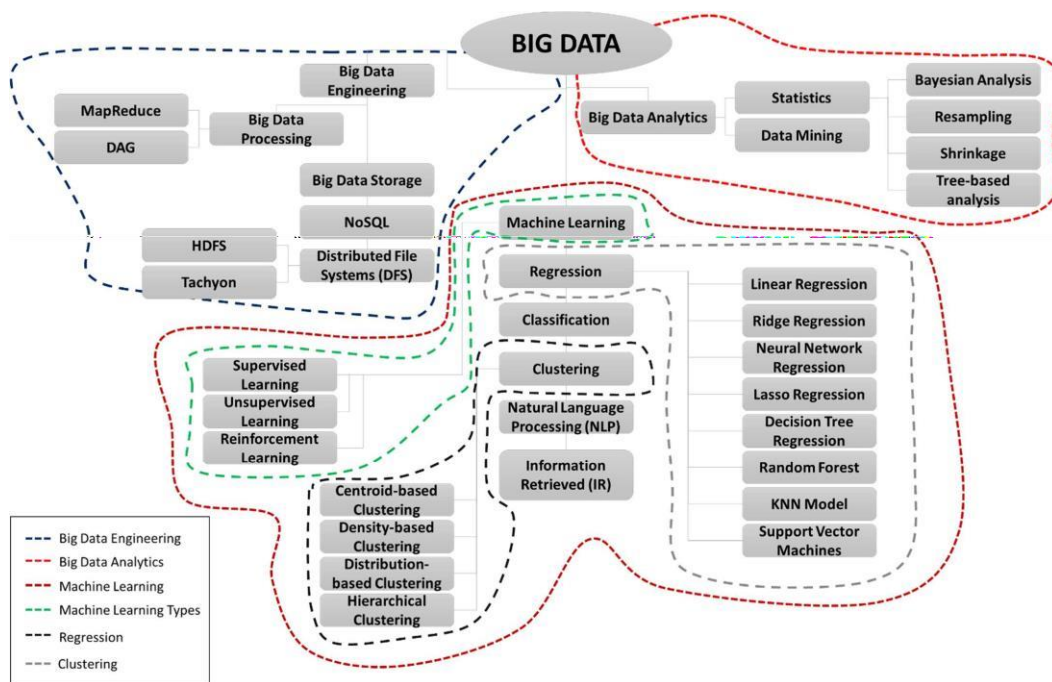


Figure 6. Classification of big data into its key domains.

Big data are classified into two major domains: BDE and BDA. These two main domains are further divided into many classes and subclasses. A third domain that comes under the canopy of big data is ML. The use of ML is inevitable in big data as the data need to be organized, analyzed, and used through ML tools and models such as deep learning and neural networks. Some of the key ML tools and models associated with big data directly or indirectly include regression analysis, clustering, classification, information retrieved (R), and natural language processing (NLP). Some examples of ML in construction include deep-learning-based flood detection and damage assessment [44], projects delay risk prediction [45], construction site safety [46], construction site monitoring [47], neural network models to predict concrete properties [48], and others.

The various algorithms and methods shown in Figure 6 all contribute towards big data in some way. The use of supervised and unsupervised learning approaches is determined depending on the type of datasets available. The major difference between supervised learning and unsupervised learning is that the algorithm for supervised learning utilizes labeled datasets while the unsupervised data do not use labeled data. The supervised and unsupervised algorithms further have different methods and examples. For instance, regression, linear regression, neural network regression, random forest, Naïve Bayes, and lasso regression are examples of supervised learning.

Similarly, clustering, Natural Language Processing (NLP), and KNN are examples of unsupervised learning. The applications of each of these algorithms can differ and hence their integration in the construction industry can vary. The regression models are used in engineering for analyzing trends and correlations between different variables. In the construction industry, these models play a crucial role as the statistical analysis and correlation development between different variables are made easy through linear regression and other similar algorithms. Similarly, machine learning models have made it possible to ensure that construction projects are developed considering safety, time management, and quality.

As shown in Figure 6, the two major big data domains rely on statistics, data processing, and data management. All these features, in turn, are heavily dependent on ML tools and methods. For example, BDE requires data processing and storage, which in turn require regression models, NoSQL, and MapReduce, all of which are different types of computational tools that enable the different applications of big data management. Similarly, BDA heavily depends on ML tools that can use data and statistics to provide organized data solutions. The use of tree-based analysis, Bayesian analysis, and shrinkage are all examples of ML integration in the field of BDA. A wide variety of ML tools have been explored over the years and have been directly or indirectly associated with big data management and analysis. Tools such as linear regression, vector machines, KNN models, clustering, and decision tree regression are among the few examples which enable the use of big data coherently. Furthermore, the classification tree of big data is likely to be further expanded as the ML algorithms are further developed and more analysis methods are added to the list. Therefore, the

constant expansion of the big data analysis tools can enable the use of these tools in the construction industry for improving construction projects in the future. Yang and Yu [49] investigated the application of heterogeneous networks oriented to NoSQL database in optimal post-evaluation indexes of construction projects. NoSQL database is scalable with a powerful and flexible data model and a large amount of data and has increasing application potential in the memory field. Sanni-Anibire et al. [50] investigated the increase in delays and abandonment of tall buildings and developed a machine learning model for delay risk assessment. Methods such as K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Ensemble methods were considered. The model developed for predicting the risk of delay was based on ANN with a classification accuracy of 93.75%. The key components of big data from Figure 6 for its management are discussed below.

Big Data Processing

Distributed and parallel computation is present in the core of BDE. In construction, big data processing has been utilized for waste management [51], prefabricated construction project management [52], profitability analyses, and other construction management applications [39]. For processing information, a considerable number of models are developed. Some of the key big data models are discussed below.

MapReduce (MR)

MapReduce was developed for the handling of big data. It utilizes a distributed processing model in which two functions, as indicated by the name itself, map and reduce, are employed to write analytical tasks. Mappers and reducers are the processes that collect the data from these functions for further processing. Initially, mappers collect and read the input information to process it for subsequent results generation. The output of mappers is used by reducers which give the results that are ultimately stored in the file system. MR has been used by Jiao et al. [53] to develop an augmented framework for BIM. Similarly, it has also been used in construction knowledge maps [54] and other big data applications [54]. The use of MapReduce in the construction industry is inevitable due to the big data applications within the construction industry. The usability of the MapReduce framework in the construction industry relies on the management of big data in a particular way. Accordingly, the datasets are analyzed and divided into categories to reduce clutter and present an easy-to-understand data output. The basic framework of MapReduce includes data input, data chunks, decomposition mappers, decomposed output, linear mappers, linear reducers, and combined output. The exact series and number of components in the framework can vary depending on the version used. However, the overall features and application of MapReduce remain the same, i.e., reduction of data into manageable chunks. The use of MapReduce not only distributes data into smaller chunks but also helps develop datasets that present a more analytic view of big data. Having organized datasets within the construction industry is of key importance as it can greatly increase the efficiency of

data management and decision making based on data analysis.

Hadoop was the popular and first big data platform that introduced and made it easy for

people to work on MR by executing its programs successfully. For tasks requiring batch processing, MR proved itself to be an effective tool as a typical cluster contains interlinked mappers and reducers that assist by running MR programs side by side at the same time. Though it has its benefits, these are not devoid of the drawbacks. These drawbacks include running some applications for graph generation and real-time and iterative processing. By dissociating the rest of the ecosystem from the processing of MR, Hadoop's latest versions have tried to sort out the problem. Yet another resource negotiator (YARN) has also been introduced, which functions by providing resource management and scheduling related functions of MR and has made it easy to implement innovative applications by Hadoop. Hadoop models have been used in construction for smart buildings and disaster management [55], failure prediction of construction firms [56], workers' safe behaviors in a metro construction project [57], and other relevant applications. The overall platform design architecture of Hadoop offers high reliability; adopt cluster technology, multi-copy technology, independent backup technology, and other means to reduce the data failure rate effectively and build a reliable data application service platform. First, the processing of big data into batches and simultaneous reduction and refining of the data are carried out using MR. Next, data are batched into similar items to streamline the analyses. This step further reduces noise or datasets that do not align with a particular batch of data. Finally, a dataset is obtained, which is refined and aligned with the original search purpose.

Directed Acyclic Graph

Big data platforms also use Directed Acyclic Graph (DAG) which is an alternative processing model. In comparison with MR, DAG works by relaxing map-then-reduce, the style of MR, which is supported by Spark. Spark is widely accepted for reactive and iterative applications due to its supremacy over MR in high expressiveness and in-memory computation. Disk-resident and memory-resident tasks are conducted ten and one hundred times faster using Spark than MR. DAGs show relationships among variables, making them easier to understand. DAGs provide major advantages that enable experts and researchers to construct complex causal relationships in which nodes represent stochastic variables, and directed edges (arrows) indicate direct probabilistic dependencies among the relevant variables. DAGs are also able to encode deterministic as well as probabilistic relationships among the variables. The usage of Spark and associated DAGs has been reported for construction profitability analysis [39], waste management [25], energy monitoring service on smart campuses [58], and others.

Spark and Hadoop are among the ML tools with enormous potential in construction engineering and management. Figure 7 compares the two tools that can inform research in construction. The speed of both these systems is better than other algorithms and ML tools currently in use in the construction industry. Moreover, fault tolerance in both these

systems is also high and has greater scalability than existing models. The data storage in these systems is slightly different in that Spark uses a memory system while Hadoop

utilizes a disk for data storage. The language for both these tools is also different since Spark is written in Scala while Hadoop has been developed using JavaScript. Despite the slight differences, both these tools provide the opportunity to process data in the form of batches and at a higher speed than previously existing models, making them potential tools for futuristic model developments in construction engineering and management. JavaScript has been used in construction to anticipate building material reuse [59], automated progress control coupled with laser scanning [60], shared virtual reality for design and management [61], construction information mining [62], and others. Similarly, Scala has been used for the process information modeling concept for on-site construction management [63].

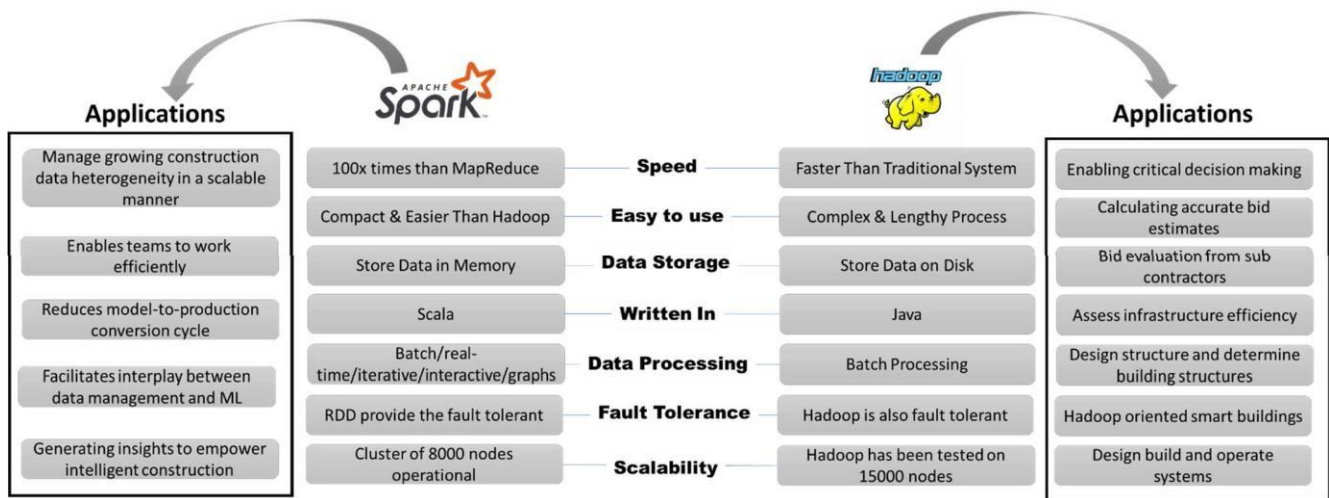


Figure 7. Components of Spark and Hadoop. A side-by-side comparison of Spark and Hadoop provides insights about the usability and applications of each.

Big Data Processing in Construction

Big data processing has been effectively utilized in the construction industry for failure prediction data [56], construction waste analytics [25], profitability data [39], modular and prefabricated construction [52], fire incident management [64], smart campus energy monitoring [58], healthier cities management [28], smart road management [40], and others. Though MR and Spark have their own significance, these are less frequently employed in the construction industry to process big data such as BIM-associated data. Partial BIM models' retrieval was optimized by MR by Bilal, et al. [65] and Chang and Tsai [66]. The authors found a loop in the Hadoop MR logic of data distribution. For overcoming the query problem, a few steps of prepartitioning and processing are introduced for relevant BIM data parts that are later stored in Hadoop clusters. Node multi-threading during data analysis helped by making the CPU work its maximum. This helped in customizing Hadoop for BIM data while the YARN application implemented querying components. YARN applications are further utilized to develop

a BIM system for quantity estimation and clash detection that can execute required tasks with the performance improved many-fold. Another research group worked for naive and

expert BIM users by developing a system for BIM data storage and retrieval [67]. The authors developed a system for cloud BIM to retrieve and represent big data intelligently. This system helped develop an interactive interface to maximize the usability and utility of construction big data. Complex BIM data are retrieved by processing proposed natural languages after reformulating user queries. This data are then visualized by mapping on various visualizations. Before query evaluation, two BIM collections are merged to optimize the process of query execution. Using this technology, a 40% reduction in response time has been witnessed compared to other traditional technologies. Currently, the utilization of BIM is limited across the construction and facilities management stages. The real intent of BIM could only be achieved once applied at each stage of the building lifecycle.

Big Data Storage

Big data storage is also an important aspect of BDE. In construction, big data storage has been explored for forecasting the success of construction projects [68], smart buildings data storage [69], tender price evaluation [70], and others. Despite the availability of BIM data storage, the current applications in construction still require successful implementation. Social BIM, proposed by Das et al. [71], captures building models and the social interactions among the users. The authors developed BIM Cloud based on the distributed BIM framework.

Similarly, a two-tiered hybrid data infrastructure was proposed by Jeong et al. [72] for data management and monitoring of bridges. In this model, the client tier efficiently completes some analytical tasks by storing structured data momentarily using MongoDB, while the central tier stores sensor data permanently using Apache Cassandra. Lin et al. [67] also used MongoDB to store BIM data obtained through building models. Overall big data storage is provided by either emerging NoSQL databases or distributed file systems, as explained subsequently.

Distributed File Systems

The distributed file systems consist of Hadoop Distributed File System (HDFS) and Tachyon. HDFS is designed to deal with large and complex databases such as those related to BIM, waste, and other construction big data sources. It operates with the commodity servers grouped together in a cluster. As it utilizes several servers, the probability of hardware failure also increases. To overcome this problem, HDFS introduces fault tolerance achieved through the distribution of data and their replication. However, in situations where low-latency data access is required, HDFS is not a suitable option as it shows inferior performance. Moreover, it is also troublesome to save many small files due to issues in managing meta-data. Moreover, it is not useful if modifications must be made concurrently at random locations in the data. Nevertheless, HDFS has been utilized by construction

researchers for observing construction workers' behavior [73], improving road performance [39], and investigating profitability performance [39]. Furthermore, based on the distributed input

from HDFS, it facilitates building predictive models for conducting building simulations that give output in a predictive model markup language.

Tachyon is a distributed file system designed to extend HDFS benefits by providing access to the distributed data across the cluster at memory speed. It provides better performance through in-memory data caching and backward compatibility allows MR and Spark tasks to run without changing the codes required in those programs. Tachyon has been utilized in construction for handling unstructured documents [65] and file storage [74]. The Tachyon performs better than HDFS, is backward compatible and can handle the MapReduce jobs without any further modifications.

NoSQL Databases

Relational databases have been common for data management in past decades. However, new applications were designed for better performance, scalability, and flexibility as the technology emerged. Relational databases lag because of their special processing and storage needs. As a result, new systems were devised to fill this technology gap. One such system is the “Not only SQL” system that has optimized data management in several ways. For achieving flexibility, it supports schemeless storage rather than schema-oriented storage. NoSQL has been widely used in different industries, including construction, due to its fragmented nature. Some examples of NoSQL in construction include integration of lessons learned knowledge in BIM [75], web service framework for construction supply chain collaboration and management [76], and Social BIM Cloud implementation [71]. NoSQL systems store schemeless data in a non-relational model. It does not set too many restrictions on value and allows easy product determination. Generally, when NoSQL databases are set to key values, they carry out only specific tasks without evaluating specific values. The key-value database is mainly tailored to the business accessed through the primary key. These systems have four data models that are briefly discussed below.

Key-value

This is the simplest data model used for unstructured data storage. However, the data lack self-description. It has been used for knowledge management in construction [77] and integration of lessons learned knowledge in BIM [75]. BIM provides positive outcomes on project success, such as cost and time reduction, communication and coordination improvement, and increased quality. Big data utilization in BIM can be beneficial to discover root causes of poor building performance, perform real-time data queries, improve the decision-making process, improve productivity, and reveal new designs and services in the construction industry, as is the case in every industry.

Document

This model can store self-describing data. However, this model can lag in terms of efficiency. It has been used for unified lifecycle data management in architecture, engineering, construction, and facilities management through BIM integration [78].

Columnar

Aggregated columns, grouped sub-columns, and sparse data can be stored by using this model. It has been used for integrating digital construction through the internet of things [79] and smart archiving of energy and petroleum construction projects [80].

Graph

This model works well for property-graph-based huge datasets in relationship traversal. It has been used for the 4D construction management information model of prefabricated buildings [81] and the development of a BIM-enabled software tool for facility management [82].

Databases concerning big data storage and management are widely used worldwide for research on various topics. The construction industry also relies on big data sources and databases, observed throughout the last five years to a decade. As shown in Figure 8, the search engine is among the most widely searched database in the last five years, followed by relational and graph DBMS. Until the time of analyzing data for this review, i.e., November 2021, other heavily used databases for extracting and using big data for the construction industry include document stores, native XML, key-value stores, and wide column stores. Object-oriented DBMS and multivalued DBMS search are considerably lower than relational DBMS and graph DBMS, whereas the search engines outperform all other DBMS. These different databases provide data sources for BIM and computational sources for developing structures that could guide larger construction projects. The rising trend in using big data sources shows the increasing interest among the construction industries in big data. For example, exchanging and reusing information is critical for engineering and construction project management. The issues pertaining to data exchange have been minimized with the Extensible Markup Language (XML) application. Such an XML-based Distributed Construction Estimating System (XDCES) has been helpful to reduce the overload of cost-estimating information exchange. Similarly, construction-based DBMS enables all construction companies to build and maintain a database easily. It allows supervisors and workers to capture information using a mobile or tablet device, and then all of that information is stored in the cloud and accessible via a desktop version.

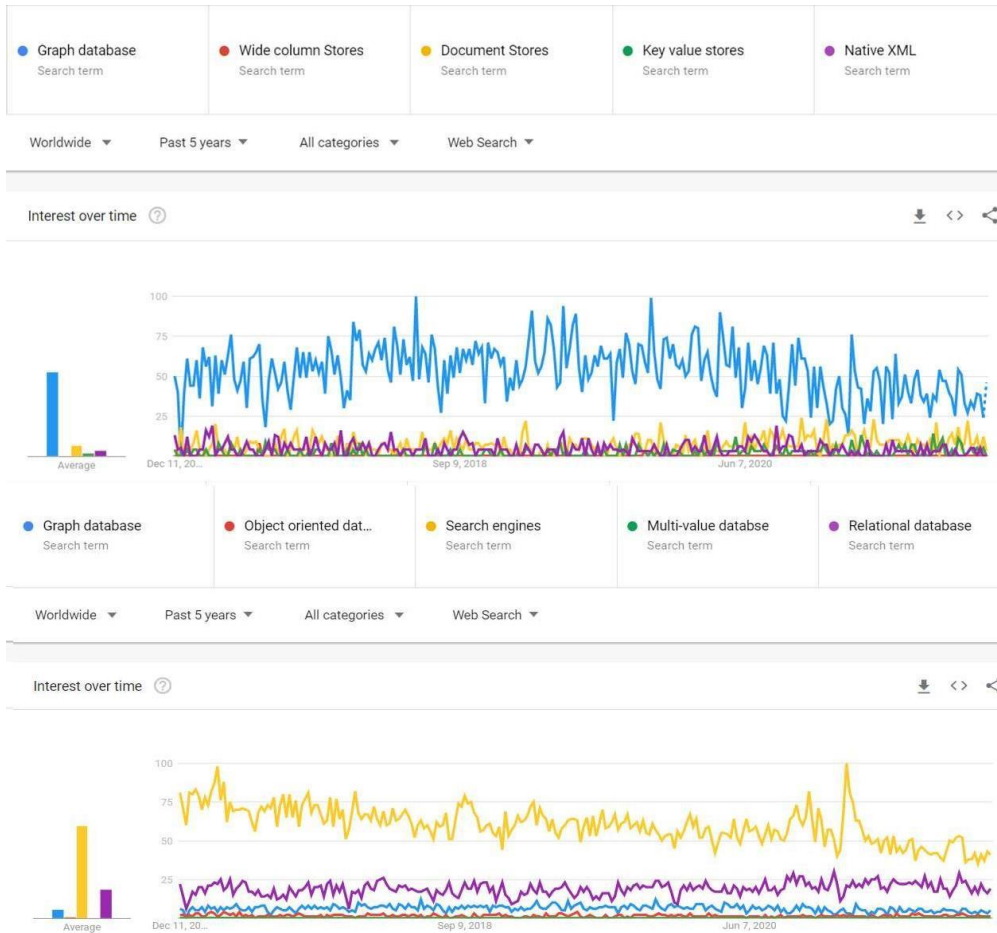


Figure 8. Database popularity in 2016–2021 based on search trends.

Big Data Analytics (BDA)

BDA gathers information from a variety of disciplines. All these disciplines have one thing in common: to find out data patterns. Some of these related disciplines are data mining, statistics, business analytics, predictive analytics, data analytics, knowledge discovery from data, and the most recent one, big data. Big data use the previous techniques to broaden the field of data analytics. For BDA, some of the ML-based tools are developed. In construction projects, BDA has been used for improving building design and effective performance monitoring [37], project safety, energy, resource, overall management and decision-making frameworks [38], and quality and waste management [24]. Big data analytics has been taken a step further by developing predictive analysis techniques. Ngo et al. [83] used a factor-based big data predictive analytic tool for analyzing the capacity of construction industries to deal with big data. This tool was tested and validated on four different construction organizations to ensure that the predictive analytic method could improve how the construction industry can use big data. The integration of big data in the construction industry remains an avenue that requires further research in terms of big data analytics. The gaps in this area were explored by Atuahene et al. [30] and Atuahene et al. [84]. It was identified that the management and processing of data by firms led to the generation of more data, which made data analysis an uphill task. Developing an integrated framework for managing big data and sorting the useful datasets can greatly increase the usability and application of big data in the construction industry. Overall, data analytics is conducted through statistical, data mining, and regression techniques, as explained below.

Statistics

Statistics has wide applications in the construction industry. Statistical techniques including Monte Carlo simulation, Gaussian distribution, non-Bayesian methods, correlation analysis, factor analysis, decision trees, Naïve Bayes, and others have been reported by various studies in construction [85,86]. Some of the areas that benefitted from statistics include learning from post-project reviews, identifying causes of construction delays, analyzing buildings for structural damages, construction litigation, and identifying and recognizing heavy machinery and workers. Other examples of statistics in construction are those of bidding statistics to predict completed construction cost [87], accidents statistics [88], quality control [89], and six sigma for project success [90,91]. From measuring the bid-to-win ratio to how much a project is over budget or schedule, and KPIs, the more numbers you can put behind your work, the better. Data not only allow for more visibility into the state of a particular project, but relevant industry statistics and facts can provide valuable information needed to make important future decisions regarding preconstruction and planning, productivity tools, risk assessment, and workforce and operational efficiency.

Data Mining

Data mining is used to extract meaningful patterns in the data. It has been an integral part of all big data management systems. It employs the techniques used in pattern recognition, ML, and statistics. Several models are assessed, and the ones with the best tolerance and high accuracy are selected and used for obtaining predictive results. In construction, data mining has been reported in waste management [97], BIM-based construction engineering quality management [98], and other relevant areas. Data mining detects useful regularities and information necessary for decision making for construction management projects. A data mining method such as cluster analysis is important for the construction industry, as it combines different construction objects into homogeneous groups and investigates them.

Data mining is supported through data warehousing. Specially structured data is stored in data warehousing for querying and analysis. Extract, transform and load (ETL) is a program that allows the collation of transactional data and operational data. Warehouse data analysis is conducted using Online Analytical Processing (OLAP), which performs better than SQL in computing breakdown and data summaries. OLAP has been used for cost data management in construction cost estimates by Moon et al. OLAP technology deals with the operational data and data obtained using big data technology. OLAP is presented as a multidimensional cube that rapidly processes datasets.

Similarly, different data mining techniques have been used to identify construction delays. For analyzing construction datasets, Kim et al. [12] presented a framework of knowledge discovery in databases (KDD). In the KDD, the most time-consuming and challenging step is data preprocessing. Nevertheless, KDD is a powerful tool for identifying casual relationships in construction projects and reducing construction variability by identifying and eliminating causes for possible deviations. With the application of KDD, randomness of construction projects and novel patterns can be determined. Other techniques include dimensional matrix analysis, link analysis, and text analysis. Other datasets with information related to delay causes, BIM-based knowledge discovery, intelligent learning, and the prevention of occupational injuries can be easily extended in the domain of data mining.

Regression Techniques

Based on an input variable, regression predicts the value of the target variable. It is a supervised ML method. Regression is categorized into simple linear and multiple linear regression based on explanatory variables. In simple linear regression, the relationship between two variables (an explanatory variable x and a dependent variable y) is modeled using ML. While in multiple linear regression, two or more explanatory variables are used and their relationship with the dependent variable is modeled. The more common regression technique is multiple linear regression.

Regression has been extensively used in construction research. For example, it has been used to predict properties of concrete cured under hot weather, predicting final cost for

competitive bids on construction projects, determining contingency in international construction projects, estimating performance time for construction projects, and

others. Moreover, regression has been used for cost estimation, which is a difficult task in the early stages of the project. Adoption of parametric methods such as regression and multiple regression can be applied as both analytical and predictive techniques to estimate the overall reliability of the cost estimation.

The 10 vs. of Big Data

The bulk and variety of big data gathering enormously each day make it virtually impossible to deal with the data sources seamlessly. On the other hand, the enormity of big data gives it many characteristics that further expand the potential of big data and its applications in different research fields. Figure 9 provides an overview of some of the crucial characteristics of big data, also known as the vs. of big data. The 10 vs. of big data have been discussed in Figure 9. Understanding these characteristics of big data enables the identification of opportunities and challenges. The most crucial properties of big data include their value, volume, velocity, variety, veracity, volatility, validity, variability, vulnerability, and visualization, also known as the 10 vs. of big data [104]. These characteristics of vs. are used to guide research in different areas and fields.

In terms of the use of big data in the field of construction, analyzing the vs. can help explore how big data can be used for developing better construction models in the future. Firstly, big data provide great value using various databases and sources that inform the research studies and algorithm developments related to computational models of different building structures. In addition to the value of research, big data also provide a bulk of information needed for research simply through the ever-increasing volume of data that becomes available each day. Furthermore, the velocity with which databases expand each day adds variety to the sort of data available for utilization in fields like construction. The variety of data present is not varying just in terms of the data sources but also the types of data. For example, big data can be present in the form of written text, graphs, pictures, and various other formats to help manage construction project schedules and progress reporting. The increasing amounts of data make the visualization process quite complex. Therefore, it is crucial to develop new ways for data visualization and analysis to keep with the volatility of big data.

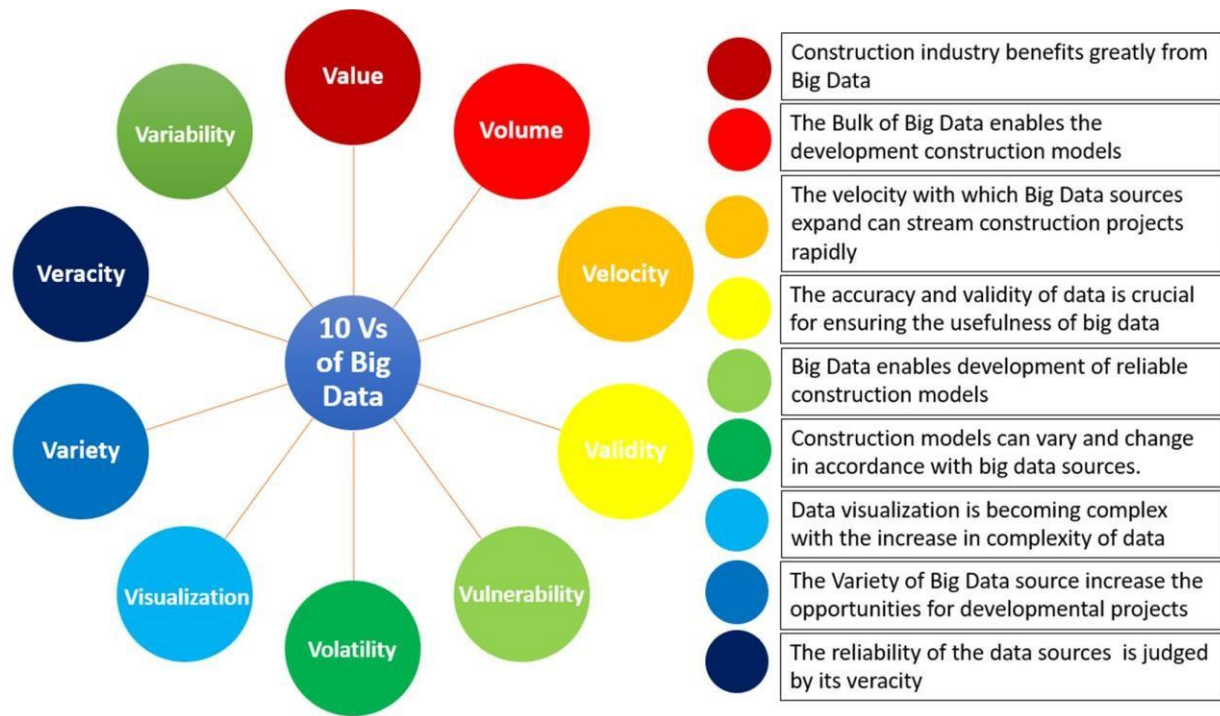


Figure 9. The 10 vs. of big data.

The 10 vs. of big data are among the crucial characteristics representing the true picture of big data as a field of research. The applications of big data in the construction industry are innumerable and they can all be categorized and managed through understanding the characteristic features (or Vs) of big data. The construction industry benefits immensely as a business by integrating big data technologies. The correlation with the business side of the construction industry has been explored in light of the 10 vs. of big data and it has been found that these characteristics provide an immense business growth potential. Starting from the core attributes of volume, variety and velocity, big data have come a long way in terms of their applications and trends. Today, there are 10 characteristics that define big data and are also crucial for implementing big data into different fields. It is crucial to understand that these 10 vs. of big data can be explained in a context-dependent manner considering the field of research. As for the construction industry, the variety and volume of big data are immense, but there is also a great deal of variability in the data present. For example, the choice of building materials and the suitability of the selected materials in different projects depend on several different factors. In this case, analyzing the applicability of big data is possible through data-visualizing techniques that can help deal with the volatility and variability of big data. Similarly, the validity and veracity of big data in construction can be judged only after analyzing the value that the data sources bring and the authenticity that these sources present. Therefore, the increasing velocity of big data is not useful as an independent factor. Instead, the application of big data in the construction industry depends on the 10 different characteristics (Vs) which are associated with big data and are explained in Figure 9

Similarly, these data types can be refined and unstructured, further adding variety to the type of data present for various reporting and research purposes. Veracity refers to the reliability of big data. This is guided by statistics as the enormity of big data makes it hard to identify reliable data sources. Therefore, validating data sources and ensuring that they can be reliably used to guide construction project developments is crucial for research. The veracity of data sources leads to another important characteristic of big data: variability. It is crucial to understand that big data can be highly variable depending on the sources used for extracting the datasets. Understanding these characteristics of big data and analyzing these characteristics given the use of big data in the construction industry can greatly enhance the potential of future construction projects.

Overall, multiple construction-related studies have reported the usage of vs. of big data. For example, velocity has been reported for high-speed construction data processing [105]. Value has been reported for smarter universities and campuses [106]. Volume has been reported for mass level offsite construction material and component production [107]. Variety has been reported for investigating the profitability performance of construction projects [39]. Veracity has been reported for forecasting the success of construction projects [68]. Similarly, variability has been reported for modeling occupational accidents in construction projects [108].

Big data necessitate cost-effective, innovative information processing forms for enhanced insights and decision making. Construction companies can analyze historical datasets and carry out predictive analytics to forecast future events. Data-driven decision making has the potential to reshape the entire business. Together, the 10 attributes or 10 vs. of big data play a crucial role in the construction industry. The volume of data and the velocity through which data are produced at high speed lead to the possibility of validating information related to construction projects. The ability to visualize big data, keep up with the variety of data, and accept the volatility, vulnerability, and variability that come with the veracity of data helps ensure that big data could be truly applicable in the construction industry. Therefore, the value of big data in the construction industry is high and it helps guide future projects.

Machine Learning Techniques

One AI subdomain is ML which can be used to learn from the data using computational systems. The tools used for big data ML are presented in Table 2. ML is further categorized into: (i) supervised learning; (ii) unsupervised learning; (iii) association; and (iv) numeric prediction. ML has several applications in the construction industry. It uses different approaches, including rule-based learning approaches, case-based reasoning techniques, artificial neural networks, and hybrid methodologies.

ML has immense potential as a tool in the field of construction. Over the last two decades, several ML algorithms have been proposed to aid and improve the overall process of construction. For example, ML has been used to predict properties of concrete [48], contract management [109], site safety and injury prediction [46], delay risks management [45], BIM integrated on-demand site monitoring [47], and

other areas of construction engineering and management.

Various ML tools are integrated at different steps along with the construction management processes. Different ML interfaces such as PyTorch and Keras.io help develop computational models based on existing data for building futuristic construction models. BIM can also be improved by using big data and ML tools, as these technologies allow the opportunity to explore how technology could be applied to the construction industry [110]. Over the last few years, different algorithms have been explored to predict various project phases and guide construction projects from inception to closure [111]. Firstly, decision trees and similar tools are used for developing an overall project timeline to predict or determine construction project performance in various phases. Secondly, statistical analysis tools are used for analyzing previous projects and choosing guiding principles for future projects [112]. Finally, design tools are integrated with ML algorithms to build 3D construction models and graphics for building models. These computational models enable analyzing construction projects by planning through look-up schedules and looking for ways to improve buildings and other structures.

The combined use of big data, ML, and AI holds the potential to develop seamless construction projects and enable the development of structures that can withstand severe weather conditions and disasters. For example, one of the key uses of ML tools in futuristic construction projects can be the development of structures that can stand through natural disasters and provide safety nets to communities during floods and other disasters [114]. Similarly, post-disaster evacuation and rescue of individuals can also be carried out more easily if the area contains structures such as roads and buildings built through the use of statistical modeling, thus providing safe routes for people [115]. Although the automation of construction projects remains a future goal, the integration of different ML algorithms is already underway. Managing costs, timelines, and human resources on a construction project are areas guided by various algorithms and computational models [116]. The ML approach can also be applied to develop leading indicators to classify sites according to their safety risk in construction projects.

Table 2. Machine learning tools used for big data.

No.	Tool	Description	Supported Algorithm	Languages	Applications in Construction	Ref.
1	PyTorch	PyTorch is a free tool available for Windows, Mac OS, and Linux for developing ML programs	Regression Classification Clustering Dimensionality reduction Preprocessing	C, C++, Python	Object detection, analyzing buildings and other structures to develop better models	[117]
2	Apache Mahout	An open-source tool that allows high-performing and scalable applications using ML	Distributed Linear Algebra Clustering Regression Preprocessing	Java, Scala	Processing big data for the development of building models and appropriate algorithms	[118,119]
3	Shogun	A diverse ML platform supporting various languages and platforms. Works well with Windows, Linux, and Mac OS	Classification Regression	C++	Provides a platform for analyzing data and developing strategies for construction projects using available information in the form of big data	[120]
4	SciKit Learn		Dimensionality reduction online learning Support vector machines		Enables statistical analysis for construction projects, particularly using existing data for developing suitable construction models	
		A free, machine-learning tool that supports Windows, Mac OS, and Linux	Regression Classification Preprocessing Clustering Model selection	C, C++, Python, Python	Provides training models which can be harnessed for	[121].5.
5.	Keras.io	An ML software that can be used across different platforms	API for neural networks	Python	Improving BIM and creating Confident models for construction projects	[122]

Future Opportunities of Big Data in Construction

There is immense potential for the use of big data in the construction industry. The use of big data and ML can enable construction automation. These tools can also enhance the overall project by removing various hurdles and roadblocks that tend to slow down different projects. The construction industry is quite dynamic and demanding, with the need for labor strength and human resources to ensure the smooth running of projects. The constant challenge of keeping projects on track and ensuring that new buildings and structures are made up to modern standards puts much strain on the project management teams. These roadblocks can greatly be reduced with the use of big data and ML. The core aim of using big data in the construction industry is to enhance the project planning phases and speed up the overall construction process by predicting the possible timelines for particular projects and identifying what factors can be worked on to improve the overall process [123].

The automation of the construction projects will require the combined use of big data, deep learning, and ML tools. One of the major concerns with such projects is ensuring workers' safety and developing strategies for overcoming potential threats to the overall process. Safety of the workers and the structures is essential for the smoother development of construction projects. The use of big data and related tools can ensure that existing data and information can be used for drafting guiding principles and then building computational models accordingly. For example, using sensor-based wearable personal protective equipment, the big data of near misses, onsite accidents, hazards, and other issues can be generated for developing safety plans and management techniques. Similarly, big data, BIM, and cloud-powered simulations can help minimize project waste and help produce superior quality constructed facilities. Further, big data artifacts generated by 3D scanners for as-built drawing development are another key advantage whereby the rehabilitation plans of ancient heritage sites can be developed.

The future holds great potential for the construction industry through big data integration. Some of the key opportunities for the construction industries lie in using big data for business and environmental sustainability. The current roadblocks faced by the construction industry can be overcome in the future through the integration of information extracted through big data. The use of information gathered from past and present projects can help develop sustainable infrastructure in the long term. It is possible to avoid past mistakes and use better quality products guided by the information found through big data in construction. Future research directions in the field of construction rely heavily on big data as the presence of information sources can help in building better infrastructure and greatly improve building designs and the overall construction business. The construction industry must move towards automation and build upon the integration of technology to make the future use of big data seamless and hassle-free. The use of big data tools, BIM, and CAD can only be possible if the relevant support and integration systems are present [107]. Hence, the future of the construction industry depends on upgrading the present environment gradually.

Overall, the role of big data in enabling the entire process of futuristic construction projects is undeniable. Data play a crucial role in developing training models and smoothly enabling the process of

Construction. Future developments in this field will also include the generation and use of more algorithms and models that rely on big data, owing to the need to train the models reliably.

Chapter 5: Conclusion

The construction industry is yet to reap the true benefits of using big data aptly. Over the last two decades, the rapid growth of big data technologies has caused a spike in the number of models and platforms that have been developed for increasing digitalization across different fields. However, the same level of digitalization has not truly been harnessed or integrated by the construction industry. A critical overview of the existing literature points towards the bulk of existing resources and platforms that can easily be applied for construction management. However, the state of implantation of adoption in construction is below par. Therefore, the utilization and commercialization of big data to benefit the construction industry are crucial. An extensive literature review enabled us to identify the potential of big data in construction as the industry generates huge amount of data daily and can greatly improve using the latest technologies. The development of online tools and software which enable infrastructure modeling and CAD is a crucial step in the right direction for futuristic constructions. Having explored the existing ML tools, we found that these tools, coupled with big data, can be applied in the construction industry. In this paper, we have discussed the existing tools used in big data, the use of statistics, big data storage, and BDE. Overlap between these variables further creates complications in that more data are present and the field of big data is ever-expanding.

The current study contributes to the body of knowledge by providing a state-of-the-art review of relevant articles focused on big data applications in construction published between 2010 and 2021. It further provides various current applications and future opportunities of big data in the construction industry for practitioners and researchers to ponder upon and initiates the necessary debate around practical implementation and adoption of big data applications in construction.

There are currently various gaps and pitfalls that act as barriers to using big data to its full potential. Firstly, data generation is much faster than the tools available for processing it. Moreover, big data integration into the construction industry is quite an uphill task even with the existing data processing tools.

The current study is limited to the literature published in the last decade and may not include all the available papers due to specific selection criteria developed in this study. Similarly, the search terms may not be holistic and thus not exhaustive; a study conducted in the future with slightly different search strings may produce different results. In the future, the researchers can expand upon and explore the five clusters identified in Figure 4. The individual relations and adoption frameworks for big data in these clusters can be explored.

Author Contributions: Conceptualization, H.S.M. and F.U.; methodology, H.S.M., F.U. and S.Q.; software, H.S.M. and F.U.; validation, H.S.M., F.U., S.Q. and D.S.; formal analysis, H.S.M. and F.U.; investigation, H.S.M., F.U. and S.Q.; resources, H.S.M. and F.U.; data curation, H.S.M., F.U., S.Q. and D.S.; writing—original draft preparation, H.S.M. and F.U.; writing—review and editing, H.S.M., F.U., S.Q. and D.S.; visualization, H.S.M. and F.U.; supervision, F.U.; project administration, H.S.M. and F.U.; funding acquisition, H.S.M. and F.U. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. **Institutional Review**

Board Statement: Not applicable. **Informed Consent Statement:** Not applicable.

Data Availability Statement: Data are available with the first author and can be shared upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

Chapter 6: Suggestions / Recommendations

1. Villars, R.L.; Olofson, C.W.; Eastwood, M. Big data: What it is and why you should care. *White Pap. IDC* **2011**, *14*, 1–14.
2. Siddiqa, A.; Karim, A.; Gani, A. Big data storage technologies: A survey. *Front. Inf. Technol. Electron. Eng.* **2017**, *18*, 1040–1070. [[CrossRef](#)]
3. Phaneendra, S.V.; Reddy, E.M. Big Data-solutions for RDBMS problems-A survey. In Proceedings of the 12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010), Osaka, Japan, 19–23 April 2013.
4. Henry, R.; Venkatraman, S. Big Data Analytics the Next Big Learning Opportunity. *J. Manag. Inf. Decis. Sci.* **2015**, *18*, 17–29.
5. Xu, W.; Sun, J.; Ma, J.; Du, W. A personalized information recommendation system for R&D project opportunity finding in big data contexts. *J. Netw. Comput. Appl.* **2016**, *59*, 362–369.
6. Sepasgozar, S.M.; Davis, S. Construction technology adoption cube: An investigation on process, factors, barriers, drivers and decision makers using NVivo and AHP analysis. *Buildings* **2018**, *8*, 74. [[CrossRef](#)]
7. Ullah, F.; Sepasgozar, S.M.; Wang, C. A systematic review of smart real estate technology: Drivers of, and barriers to, the use of digital disruptive technologies and online platforms. *Sustainability* **2018**, *10*, 3142. [[CrossRef](#)]
8. Kwon, O.; Lee, N.; Shin, B. Data quality management, data usage experience and acquisition intention of big data analytics. *Int. J. Inf. Manag.* **2014**, *34*, 387–394. [[CrossRef](#)]
9. Cui, L.; Yu, F.R.; Yan, Q. When big data meets software-defined networking: SDN for big data and big data for SDN. *IEEE Netw.* **2016**, *30*, 58–65. [[CrossRef](#)]
10. Chaudhary, R.; Aujla, G.S.; Kumar, N.; Rodrigues, J.J. Optimized big data management across multi-cloud data centers: Software- defined-network-based analysis. *IEEE Commun. Mag.* **2018**, *56*, 118–126. [[CrossRef](#)]
11. Simmhan, Y.; Aman, S.; Kumbhare, A.; Liu, R.; Stevens, S.; Zhou, Q.; Prasanna, V. Cloud-based software platform for big data analytics in smart grids. *Comput. Sci. Eng.* **2013**, *15*, 38–47. [[CrossRef](#)]
12. Kim, K.Y. Business intelligence and marketing insights in an era of big data: The q-sorting approach. *KSII Trans. Internet Inf. Syst.(TIIIS)* **2014**, *8*, 567–58
13. Hu, X. Sorting big data by revealed preference with application to college ranking. *J. Big Data* **2020**, *7*, 1–26. [[CrossRef](#)]
14. Custers, B.; Uršič, H. Big data and data reuse: A taxonomy of data reuse for balancing big data benefits and personal data protection. *Int. Data Priv. Law* **2016**, *6*, 4–15. [[CrossRef](#)]
15. Majumdar, J.; Naraseeyappa, S.; Ankalaki, S. Analysis of agriculture data using data mining techniques: Application of big data. *J. Big Data* **2017**, *4*, 1–15. [[CrossRef](#)]
16. Shadroo, S.; Rahmani, A.M. Systematic survey of big data and data mining in internet of things. *Comput. Netw.* **2018**, *139*, 19–47. [[CrossRef](#)]
17. Zhou, R.; Liu, M.; Li, T. Characterizing the efficiency of data deduplication for big data storage management. In Proceedings of the 2013 IEEE international symposium on workload characterization (IISWC), Portland, OR, USA, 22–24 September 2013; pp. 98–108.
18. Petri, I.; Rana, O.; Beach, T.; Rezgui, Y.; Sutton, A. Clouds4Coordination: Managing project collaboration in federated clouds. In Proceedings of the 2015 IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC), Limassol, Cyprus, 7–10 December 2015; pp. 494–499.
19. Hay, B.; Nance, K.; Bishop, M. Storm clouds rising: Security challenges for IaaS cloud computing. In Proceedings of the 2011 44th Hawaii International Conference on System Sciences, Washington, DC, USA, 4–7 January 2011; pp. 1–7.

20. Afolabi, A.; Ojelabi, R.A.; Fagbenle, O.I.; Mosaku, T. The economics of cloud-based computing technologies in construction project delivery. *Int. J. Civ. Eng. Technol. (IJCIET)* **2017**, *8*, 232–242.
21. Moniruzzaman, A.; Hossain, S.A. Nosql database: New era of databases for big data analytics-classification, characteristics and comparison. *arXiv* **2013**, arXiv:1307.0191.
22. Kouanou, A.T.; Tchiotsop, D.; Kengne, R.; Zephirin, D.T.; Armele, N.M.A.; Tchinda, R. An optimal big data workflow for biomedical image analysis. *Inform. Med. Unlocked* **2018**, *11*, 68–74. [[CrossRef](#)]
23. Rodrigues, M.; Santos, M.Y.; Bernardino, J. Big data processing tools: An experimental performance evaluation. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, *9*, e1297. [[CrossRef](#)]
24. Wang, D.; Fan, J.; Fu, H.; Zhang, B. Research on optimization of big data construction engineering quality management based on RNN-LSTM. *Complexity* **2018**, *2018*, 9691868. [[CrossRef](#)]
25. Bilal, M.; Oyedele, L.O.; Akinade, O.O.; Ajayi, S.O.; Alaka, H.A.; Owolabi, H.A.; Qadir, J.; Pasha, M.; Bello, S.A. Big data architecture for construction waste analytics (CWA): A conceptual framework. *J. Build. Eng.* **2016**, *6*, 144–156. [[CrossRef](#)]
26. Munawar, H.S.; Qayyum, S.; Ullah, F.; Sepasgozar, S. Big data and its applications in smart real estate and the disaster management life cycle: A systematic analysis. *Big Data Cogn. Comput.* **2020**, *4*, 4. [[CrossRef](#)]
27. Qadir, Z.; Khan, S.I.; Khalaji, E.; Munawar, H.S.; Al-Turjman, F.; Mahmud, M.P.; Kouzani, A.Z.; Le, K. Predicting the energy output of hybrid PV–wind renewable energy system using feature selection technique for smart grids. *Energy Rep.* **2021**, *7*, 8465–8475. [[CrossRef](#)]
28. Miller, H.J.; Tolle, K. Big data for healthy cities: Using location-aware technologies, open data and 3D urban models to design healthier built environments. *Built Environ.* **2016**, *42*, 441–456. [[CrossRef](#)]
29. Chen, X.; Lu, W. Scenarios for Applying Big Data in Boosting Construction: A Review. In Proceedings of the 21st International Symposium on Advancement of Construction Management and Real Estate, Guiyang, China, 24–27 August 2018; pp. 1299–1306.
30. Atuahene, B.T.; Kanjanabootra, S.; Gajendran, T. Towards an integrated framework of big data capabilities in the construction industry: A systematic literature review. In Proceedings of the 34th Association of Researchers in Construction Management (ARCOM), Belfast, UK, 3–5 September 2018; p. 547.
31. Ullah, F. A beginner’s guide to developing review-based conceptual frameworks in the built environment. *Architecture* **2021**, *1*, 5–24. [[CrossRef](#)]
32. Ullah, F.; Al-Turjman, F. A conceptual framework for blockchain smart contract adoption to manage real estate deals in smart cities. *Neural Comput. Appl.* **2021**, 1–22. [[CrossRef](#)]
33. Ullah, F. *Developing a Novel Technology Adoption Framework for Real Estate Online Platforms: Users’ Perception and Adoption Barriers*; University of New South Wales: Sidney, Australia, 2021.
34. Ullah, F.; Qayyum, S.; Thaheem, M.J.; Al-Turjman, F.; Sepasgozar, S.M. Risk management in sustainable smart cities governance: A TOE framework. *Technol. Forecast. Soc. Change* **2021**, *167*, 120743. [[CrossRef](#)]
35. Qayyum, S.; Ullah, F.; Al-Turjman, F.; Mojtahedi, M. Managing smart cities through six sigma DMADICV method: A review-based conceptual framework. *Sustain. Cities Soc.* **2021**, *72*, 103022. [[CrossRef](#)]
36. Huang, X. Application of BIM Big Data in Construction Engineering Cost. *J. Phys. Conf. Ser.* **2021**, *1865*, 032016. [[CrossRef](#)]
37. Loyola, M. Big data in building design: A review. *J. Inf. Technol. Constr.* **2018**, *23*, 259–284.
38. Ismail, S.A.; Bandi, S.; Maaz, Z.N. An appraisal into the potential application of big data in the construction industry. *Int. J. Built Environ. Sustain.* **2018**, *5*, 145–154. [[CrossRef](#)]
39. Bilal, M.; Oyedele, L.O.; Kusimo, H.O.; Owolabi, H.A.; Akanbi, L.A.; Ajayi, A.O.; Akinade, O.O.; Delgado, J.M.D. Investigating profitability performance of construction projects using big data: A project analytics approach. *J. Build. Eng.* **2019**, *26*, 100850. [[CrossRef](#)]
40. Sharif, M.; Mercelis, S.; Van Den Bergh, W.; Hellinckx, P. Towards real-time smart road construction: Efficient process management through the implementation of internet of things.

END