2024 European Conference on Computing in Construction
Chania, Crete, Greece
July 14-17, 2024

REVOLUTIONISING INDIAN HIGHWAY PROJECTS: UNLEASHING THE POWER OF DATA INTEGRATION WITH CUTTING-EDGE DATA LAKE

Chenchu Murali Krishna[1], Kirti Ruikar[2], and Kumar Neeraj Jha[3]
[1]Indian Institute of Technology Delhi, New Delhi, India
[2]Loughborough University, Loughborough, United Kingdom
[3]Indian Institute of Technology Delhi, New Delhi, India

Abstract

Indian highway projects grapple with multifaceted delays, from contractual hurdles to legal complexities. Addressing this, the National Highway Authority of India (NHAI) embraces a cloud-based Data Lake, ushering in a "Fully Digital" paradigm. This advanced tool, validated through a literature review and NHAI stakeholder interviews, forecasts and manages delays systematically. The study reveals NHAI's adept integration of Data Lake technology across project phases, offering tangible solutions and underlining the transformative potential for data-driven decision-making in highway infrastructure projects. Recommendations advocate wider adoption, continual development, and prioritising education and research in delay management, heralding a new era of operational efficacy.

Keywords: Construction Delays, Data Lake, Highway Projects, Delay Management.

Introduction

In the realm of constructing India's expansive road network, measuring approximately 5.89 million kilometers, the continuous challenge of ensuring the timely completion of highway projects stands as a significant obstacle (Ministry of Road Transport & Highways Year End Review, 2022). This herculean task encounters numerous hurdles ranging from unpredictable weather conditions and complex land acquisition processes to bureaucratic permissions and the looming threat of supply chain disruptions. A study by the National Cooperative Highway Research Program (NCHRP) reveals that a staggering 40 per cent of projects experience delays, leading to escalated costs, compromised quality, and a noticeable decline in public satisfaction. These delays not only inflate project budgets and diminish quality but also erode trust among the public. Traditionally, project managers have relied heavily on their expertise and intuition to anticipate and manage potential delays. Yet, this conventional approach often falls short of identifying hidden delays until they materialise (Tripathi *et al.*, 2023). The imperative arises for a paradigm shift toward a more data-centric approach to delay management.

In this era of data innovation, data lakes have emerged as critical components for managing enormous quantities of structured and unstructured data efficiently. Functioning as a centralised repository, a data lake allows for the storage of data in its original format. It facilitates detailed analysis through an array of data mining and machine-learning techniques (Mathis, 2017). This seismic shift towards data lakes provides a transformative solution for highway project delay management, moving away from the traditional, manual forecasting techniques that are time-consuming data gathering and inherent to inaccuracies.

Data lakes are preferred over traditional approaches like data warehouses and data marts due to their unmatched flexibility, scalability, democratised access to data, and cost-effectiveness (Hai *et al.*, 2023). Unlike rigid data warehouses that require predefined schemas, data lakes store raw data in its original form, accommodating diverse data types seamlessly (Singh *et al.*, 2022). Scalability is another key advantage, enabling data lakes to handle the growing data volumes of highway projects seamlessly. This scalability, coupled with the lower total cost of ownership, makes data lakes an attractive option for organisations and innovation by providing a centralised repository for organisational data, unlike data marts tailored to specific domains (El Haddadi *et al.*, 2020). Additionally, data lakes facilitate advanced analytics techniques like AI and Machine learning, empowering organisations to extract actionable insights from their data. The decision to prioritise data lakes over well-defined data structures and explicit information requirements reflects a strategic choice to embrace technologies such as UML, graph databases, and ontologies (Sawadogo, 2019; Singh and Ahmad, 2019; Nambiar and Mundra, 2022). A comprehensive comparison of these technological options is shown in Table 1, highlighting the unique strengths and advantages that data lakes offer in meeting the evolving demands of construction big data challenges in the digital age.

In the dynamic landscape of highway construction, where the precision of timing and meticulous management of resources is paramount, the integration of data lakes emerges as a pivotal element for advancing project efficiency (Nargesian *et al.*, 2018). This advanced system empowers project managers to aggregate and scrutinise data from a myriad of sources, encompassing historical project data, predictive weather analytics, and real-time insights from sensors strategically deployed at construction sites. The advent of machine learning algorithms within these data lake frameworks acts as a formidable guard against potential delays. These algorithms are adept at identifying complex patterns, recognizing trends, and discovering connections within

the vast pool of integrated data, offering not only the ability to forecast delays but also the means to proactively prevent them (Gudivada *et al*., 2017). Imagine a scenario in highway construction where delays are not merely managed but actively prevented (Durdyev and Hosseini, 2020). This scenario is made possible through the seamless integration of varied data sources, all housed within the secure confines of a data lake, orchestrated by the strategic application of machine learning algorithms. This not only underscores the role of data lakes as mere storage facilities but elevates them to the role of architects in a data-centric revolution in highway project management. This paradigm shift underscores the narrative that progress in the construction sector transcends the traditional reliance on physical materials, pivoting towards the strategic leverage of data. As the industry leans into this transformative approach, it stands to redefine delays not as mere hurdles but as opportunities to refine resource allocation, boost quality, and uplift public satisfaction.

Table 1: Comparison of the different technologies with Data Lake

S. No.	Database	Stores Raw Data	Structured Data Support	Schema Flexibility	Horizontal Scaling	Querying & Analytics	Why Data Lake is Preferred
1	Data Lake	✓	✓	✓	✓	✓	Flexibility in storing raw data in its original format
2	Data warehouse		✓		✓	✓	Structured storage for consistent querying and reporting
3	NoSQL Databases	✓	✓	✓	✓	✓	Flexibility in schema and scalability for diverse data types
4	Object Storage	✓			✓		Scalability and cost-effectiveness for storing large objects
5	Unified Modeling Language					✓	Conceptual modelling and communication in software development
6	Graph Databases				✓	✓	Ability to model relationships between entities effectively
7	Ontologies					✓	Explicit representation of domain knowledge and relationships

Data Lake Implementation at NHAI

The NHAI has introduced a data lake, marking a significant step forward in managing highway projects. This innovation brings enhanced efficiency, transparency, and the ability to make decisions based on data. The development involved careful planning and implementation, beginning with the initial phase of recognising the essential need for a centralised data repository. This necessity arose from the complexities involved in managing highway projects. Consequently, NHAI decided to establish a cloud-based data lake, laying the foundation for a revolutionary digital infrastructure. In collaboration with technology partners, the NHAI initiated the design of a data lake's architecture, focusing on ensuring scalability, reliability, and the capacity to handle various data types (Khine and Wang, 2018).

The selection of cloud infrastructure from premier providers like AWS or Azure was made to establish a strong foundation for the platform's operations (Sharma, 2018). The development phase involved dedicated teams working diligently to realise the data lake's vision, integrating it effectively with existing systems such as project management and GIS for a smooth transition. Stringent testing measures were put in place to maintain the utmost standards of data integrity, security, and efficiency. This included thorough user acceptance testing with participation from stakeholders across different functions, ensuring it meets the changing requirements of NHAI. Additionally, detailed training programs were organised for end users such as contractors, engineers, project directors, and regional officers, enabling them to utilise the data lake fully in their daily activities.

The data lake was smoothly incorporated into the NHAI's pre-existing workflows, replacing manual procedures with automated systems. This enhancement significantly improved efficiency and accountability across various operations. The integration encompassed project documentation, communications, workflow monitoring, timelines, notifications, and financial transactions related to projects, and all facilitated through the data lake interface. This allowed stakeholders to make well-informed decisions based on up-to-the-minute data insights. Furthermore, the compulsory inclusion of drone surveys for every project enabled thorough monitoring and analysis, substantially strengthening NHAI's capability in managing projects. The data lake serves a broad array of end users, including contractors, engineers, project directors, regional officers, and headquarter staff. Each group of stakeholders enjoys access to customised features and functionalities designed to meet their unique requirements, thereby empowering them to make decisions rooted in data, which in turn promotes project success and organisational development.

NHAI data lake architecture

The data lake architecture employed by the NHAI is designed to efficiently handle large volumes of diverse data, facilitating data-driven decision-making and operational efficiency within the organisation. It consists of key components such as the ingestion layer, storage layer, processing layer, catalogue and metadata management, governance and security, analytics and visualisation, and finally sharing and collaboration components. The simplified breakdown of the key components is as follows:

Data Ingestion Layer: This layer focuses on collecting data from various sources, including internal databases, APIs, and external sources like documents and images. It ensures the seamless ingestion of structured and semi-structured raw data types relevant to NHAI's operations (Singh and Ahmad, 2019).

Data Storage Layer: Once data is ingested, it is stored in a distributed file system, such as Hadoop Distributed File System (HDFS) or cloud-based storage solutions like Amazon S3 or Azure Blob storage. This ensures secure and efficient storage of data, enabling accessibility for further processing and analysis (Inmon, 2016).

Data Processing Layer: In this layer, tools and frameworks like Apache Spark, Apache Flink, or Apache Beam are utilised to process and transform raw data into usable formats. Batch and stream processing techniques enable real-time or near-real-time data processing, enhancing NHAI's analytical capabilities (Sharma, 2018).

Data Catalog and Metadata Management: NHAI's data lake architecture incorporates a metadata management system to catalogue and organise the vast amount of stored data. This metadata provides valuable information about the structure, format, and lineage of the data, facilitating easy discovery and understanding of available datasets (Madera and Laurent, 2016).

Data Governance and Security: Data governance policies and security measures are integral components of NHAI's data lake architecture. Access control mechanisms, encryption, data masking, and compliance with regulatory requirements ensure the privacy and security of sensitive information stored within the data lake (Abraham *et al*., 2019).

Data Analytics and Visualisation: NHAI's data lake enables data analytics and decision-makers to perform advanced analytics and derive insights from the stored data. Tools like Apache Hadoop, Apache Spark, or specialised analytics platforms facilitate data analysis and visualisation through dashboards or reports, empowering informed decision-making (Nagel *et al.*, 2021).

Data Sharing and Collaboration: NHAI's data lake architecture includes features for sharing data and collaborating with internal and external stakeholders. This may involve providing APIs for accessing data, implementing data-sharing agreements, enabling secure data exchange protocols, and fostering collaboration and knowledge sharing within the organisation (Couto and Ruiz, 2022).

Data lake for delay management

In the domain of highway construction, the data lake emerges as a powerful tool for effective delay management. Acting as a centralised repository, it adeptly accommodates structured and unstructured data of any magnitude, enabling the integration and analysis of information from various sources, such as construction sites, traffic sensors, and weather reports. This versatility makes the data lake a crucial component in effective delay management, which is evident in its diverse applications (Edison and Singla, 2020).

Primarily, the real-time monitoring capabilities of data lakes empower project managers to gather instantaneous data from varied sources ranging from sensor metrics and weather projections to traffic dynamics and social media updates (Gorelik, 2019). This comprehensive data snapshot enables proactive decision-making, allowing project managers to discern potential delays and take corrective measures promptly. Secondly, through the lens of predictive analytics, data lakes craft models that discern trends in historical data, translating them into actionable insights for anticipating and mitigating future delays. For instance, a meticulous analysis of past weather patterns facilitates pre-emptive measures against weather-induced schedule disruptions (Nambiar and Mundra, 2022).

Furthermore, data lakes facilitate a detailed optimisation of resource allocation by identifying areas of underutilisation or excess usage. This optimisation strategy not only streamlines project timelines but also acts as a strong deterrent against delays. Additionally, these reservoirs of information play a pivotal role in root cause analysis by storing and examining historical data on previous delays. Armed with this knowledge, organisations can effectively implement measures to minimise and prevent similar delays in future projects. Additionally, data lakes transcend their role as mere repositories by fostering collaboration among project stakeholders. By providing a centralised hub for project-related data, they facilitate seamless communication and collaboration across diverse teams, ensuring alignment toward common project objectives (Hagstroem *et al.*, 2017; Chomo, 2019).

In summary, data lakes serve as invaluable assets in the ongoing effort to minimise delays in highway construction projects. Through the integration and analysis of data from diverse sources, project managers gain deep insights, enabling them to make informed decisions and proactively guide projects toward successful completion. This fusion of big data within the data lake seamlessly addresses the complexities of highway construction delay factors, showcasing a nuanced alignment between the features of the data lake and the diverse challenges encountered in construction projects, as illustrated in Table 2.

Methodology

This study was conducted within the framework of the NHAI and involved a diverse group of stakeholders. The study aimed to understand the subjective experiences and interpretations of these stakeholders' concerning delays in highway projects. By adopting a collaborative approach, the research sought to generate knowledge that reflects the multifaceted perspectives of those involved, thus acknowledging their crucial role in shaping the findings. The methodology of the study was structured into four stages, designed to align with the research objectives. The investigation utilised both secondary and primary data sources. Initially, the research identified the problem of project delays through an extensive literature review, the

author's direct involvement with highway projects, and informed discussions with experts.

Table 2: Matrix that maps delay management features for highway projects against common causes of delays

Cause of delay	Delay management features			
	Real-time monitoring	***Predictive analytics***	***Resource Allocation & optimisation***	***Collaboration***
Poor planning and scheduling of the project	✓			✓
Site clearances	✓			✓
Land acquisition and rehabilitation issues	✓	✓		✓
Weather and force majeure-related delays	✓	✓	✓	✓
Project changes and redesigns	✓	✓	✓	✓
Permit and approval delays	✓	✓		✓
Stakeholder conflicts	✓	✓	✓	✓
Poor resource management	✓		✓	✓
Inadequate project monitoring	✓	✓		✓
Poor communication between construction parties	✓			✓

This phase of the research identified 182 publications related to project delays and data lakes, with 56 unique articles remaining after duplicates were removed. These articles, dating from 2006, were sourced from databases such as Scopus, Elsevier, and Web of Science. Following the identification of delays, theoretical propositions were formulated to address these delays, based on analysis of the initial data. Qualitative data was collected through semi-structured interviews with six stakeholders, including deputy managers experienced in data lake-integrated projects, two project managers and two senior research followers actively involved in data lake-related projects. Table 3 shows the details of the interviewers. These interviews aimed to gather insights on the identified causes of delays, the theoretical propositions, and the impact of data lake processes and technologies in mitigating these issues. Interviewees are chosen for their efficacy in eliciting information about non-observable phenomena. The final stage involves formulating recommendations and conclusions derived from the content analysis of the collected interview data. This comprehensive approach, integrating data, information, and knowledge through the prism of a data lake, has the potential to revolutionise the landscape of effective data utilisation in highway construction projects, leveraging the vast reservoir of big data available, paving the way for a revolution in the field.

Table 3: Details of the Interviewees

Parameters	Stakeholders					
	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6
Designation	Deputy Manager	Deputy Manager	Project Manager	Project Manager	Senior Researcher	Senior Researcher
Experience (Years)	18	20	15	15	4	3

Propositions

The theoretical propositions of this study are drawn from a thorough review of the literature, focusing on the attributes of a data lake anticipated to contribute significantly to the mitigation of project delays. The study primarily explores the implementation of a data lake in the context of Indian highways, aiming to quantify its advantages vis-à-vis the prominent causes of delays. However, the study's scope is limited to examining a data lake's potential to improve communication, information flow, and coordination among project stakeholders.

The study proposes further investigation into how a data lake can enhance communication and collaboration in highway projects (Nargesian *et al.*, 2018; Giebler *et al.*, 2019). The existing literature research lays a robust foundation for the formulation of key propositions, integral to guiding interview questions design and subsequent data collection:

1. *Proposition 1*: Implementation of a Data Lake in highway projects enhances data visualisation and construction process comprehension, thereby mitigating delays in progress payments and expediting client decision-making.
2. *Proposition 2:* Adoption of a data lake in NHAI projects facilitates effective delay management while fostering improved communication and coordination between the client and other stakeholders.

These propositions serve as the compass for the study, directing the formulation of insightful interview questions and the meticulous collection of data from pertinent stakeholders. This approach allows for a detailed exploration of the benefits and challenges inherent in implementing a data lake within the context of Indian highways, unravelling the untapped potential of data integration for practical usage in highway construction projects.

Interview Analysis

Proposition 1: Implementation of a Data Lake in highway projects enhances data visualisation and construction process comprehension, thereby

mitigating delays in progress payments and expediting client decision-making.

The efficacy of a data lake in highway construction projects is evident in its ability in visualising diverse datasets and operational aspects. Consolidating all data into a singular repository provides a comprehensive overview of project status, progress, and performance, enabling stakeholders to make data-driven decisions. Addressing concerns outlined in prior research such as Keane and Caletka (2015) and Du *et al.* (2018) pertaining to insufficient information about construction progress and communication gaps among stakeholders causing delays in progress payments and decision-making, the content analysis of interviews underscores the substantial benefits that implementing a data lake brings to the fore in managing delays and enhancing decision-making processes.

The content analysis of interviews highlights that implementing a data lake in highway projects brings significant advantages for managing delays and improving decision-making. Data visualisation enables quick identification of concerns and prompt corrective actions. Access to centralised data facilitates decision-making based on real-time insights. However, successful implementation requires understanding the data and selecting meaningful metrics. Transparent processes and workflows within the data lake enhance utilisation and support process improvement. It is important to note that implementing a data lake alone is not sufficient; it requires a comprehensive approach to leverage its benefits effectively.

The interviews emphasise that the implementation of a data lake in highway projects significantly bolsters delay management and decision-making processes. The key insights derived from the content analysis illuminate critical considerations:

1. *Strategic Metric Selection for Visualization*: Based on the input from deputy manager-level stakeholders and project managers who have experience with data lake-integrated projects, it is unanimously agreed that selecting meaningful metrics for data visualisation is critical. This strategic metric selection ensures that the visualised data is relevant and resonates with its audience, which is crucial for effective decision-making. Strategic metric selection in data lake is key for enhancing business performance through a data-driven approach. Metrics related to operational efficiency, data integration, and business analytics are essential for effective visualisation (Laurent *et al.*, 2020; Barbierato *et al.*, 2021; Kumar and Chundi, 2023).
2. *Expedited Decision-Making Process:* The consensus among interviewees highlights that the implementation of data lakes significantly accelerates the decision-making process by offering streamlined access to comprehensive and centralised data, fostering a culture of prompt and well-informed decisions. Beyond speeding up decision-making, data lakes bring technological advantages by improving flexibility and scalability, democratising data access. They also transform the business paradigm by addressing big data challenges and driving digital transformation, ultimately enhancing business intelligence. By providing a centralised repository of virtually inexhaustible raw data for analytical activities, data lakes enable enterprises to profoundly improve their decision-making processes and business intelligence, thereby transforming their overall business paradigms (Terrizzano *et al.*, 2015; Johny and Pillai, 2022).
3. *Visualisation of Project Status and Impact of Delays:* The interview findings highlight the significance of visualising data pertaining to schedule overruns and activity delays to enhance decision-makers' understanding of project status, with projection of the completion schedule offering insights into the impacts of delays and enabling proactive mitigation efforts. Data lakes at NHAI play a vital role in democratising data access and supporting divers' analytics tasks within enterprises. The data lake structure comprises five layers, enabling effective visualisation of project status by integrating multiple time series data sets. The adaptability of data lakes is exemplified by the NHAI dashboard, which positions stakeholders of data lakes as invaluable tools for the visualisation and analysis of project statuses (Fang, 2015; Mathis, 2017; Kumar and Chundi, 2023; Schneider *et al.*, 2023).
4. *Process Establishment and Improvement:* Interview data affirm that data lakes play a pivotal role in facilitating the establishment and documentation of processes and workflows within NHAI. This analysis helps in identifying areas that need improvement, resulting in greater efficiency in overall processes. Additionally, data lakes enable the use of advanced data-driven analysis techniques, which significantly aid enterprises in optimising the NHAI business operations. This not only captures key process parameters but also serves as a foundation for rigorous analysis. The insights derived from this analysis become instrumental in identifying areas for improvement, ultimately leading to enhanced overall process efficiency (Nagel *et al.*, 2021; Schneider *et al.*, 2023).

In conjunction with the literature review propositions, the synthesis of interview findings underscores that the implementation of data lakes in highway projects has the potential to be a game-changer in decision-making through data visualisation. However, it is imperative to acknowledge that successful implementation necessitates not only adequate resources for maintenance and development but also the addressing of a discernible skill gap in data lake expertise. The collaboration and knowledge-sharing among industry professionals, as underscored by senior research follows actively engaged in data-lake-related projects, emerge as critical components for establishing best practices and standards in data implementation and data lake utilisation. The call to allocate resources for ongoing support, encompassing

infrastructure, data integration, quality management, and governance, resonates as a strategic imperative for organisations venturing into the realm of data lake implementation in highway construction projects.

Proposition 2: Adoption of a data lake in NHAI projects facilitates effective delay management while fostering improved communication and coordination between the client and other stakeholders.

In delving into the realm of effective data utilisation within highway construction projects, insights obtained from interviews with key stakeholders underscore the crucial role of a data lake in enhancing communication dynamics. The literature, supported by expert opinions, emphasises the importance of seamless communication to prevent delays, with inadequate exchange leading to suboptimal designs and planning (Saini, 2015).

Interview analysis highlights the significant impact of a data lake on communication between clients and stakeholders, serving as a keystone for real-time updates, meticulous tracking of communication flows, and seamless access to relevant information within a centralised platform (Malacarne *et al.*, 2018). This data lake design intricately reflects all communications within stakeholders' dashboards, ensuring immediate action based on document flow, approval status, and remarks of respective officers. This creates a harmonised and transparent communication channel that fosters enhanced coordination, mitigating delays attributed to internal or coordination factors. The implementation of a data lake in highway projects not only improves communication but also streamlines decision-making processes, leading to more efficient project management and ultimately, successful project completion.

In essence, the interview analysis manifests in phrases that unveil the data lake's diverse role:

1. *Improving Communication between Clients and Stakeholders*: The interviews with key stakeholders, including deputy manager-level experts in data lake-integrated projects, project managers, and senior research fellows deeply entrenched in data lake-related endeavors, yielded profound insights into the enhancement of communication dynamics between clients and stakeholders. A pivotal observation emerged: the strategic design of the data lake serves as a communication nexus, seamlessly reflecting all relevant interactions within stakeholders' dashboards. This design ensures not only visibility but catalyses immediate action on requests, leveraging the dynamic flow of documents, approval statuses, and remarks from respective officials.
   The real-time updates from the data lake proved instrumental in transforming project dynamics. Stakeholders, armed with instantaneous information on project status, documentation details, and approvals, navigated a landscape of enhanced coordination. This, in turn, emerged as a powerful countermeasure to delays induced by internal factors or coordination challenges. The interviews underscored a paradigm shift in communication dynamics, where the data lake became a pipeline for real-time collaboration and coordination, moving projects towards streamlined efficiency and reduced delays.
2. *Central Repository for Documents and Information:* In illuminating insights drawn from interviews with key stakeholders, the data lake emerges as a central repository for approved requests, project documents, drawings, notices, and more. According to the perspectives gleaned from two deputy manager-level stakeholders intimately acquainted with data lake integration, two project managers, and two senior researchers who are deeply entrenched in data lake-related projects, the data lake's role as a document repository is pivotal. This repository function, as articulated by the interviewees, not only ensures the organised storage of essential documents but also stands as a safeguard against human errors. The availability of information within this consolidated data lake reduces the likelihood of oversights and inefficiencies, contributing to smoother functioning. Importantly, the improved accessibility to this repository acts as a catalyst, streamlining processes and, consequently, playing a significant role in the reduction of delays within the context of highway construction projects. The consensus among these key stakeholders underscores the transformative impact of a data lake as more than just a repository—it is a dynamic facilitator that not only safeguards against errors but actively contributes to the efficiency and agility of project processes, aligning seamlessly with the overarching goal of minimising delays in highway construction projects.
3. *Facilitating Prompt Resolution of Requests:* In the interview analysis, stakeholders with expertise in data lake-integrated projects unanimously highlighted the data lake's instrumental role in expediting request resolutions and curbing delays. The real-time updates afforded by the data lake emerge as a cornerstone for well-informed decision-making, providing up-to-date insights that prove pivotal in navigating project intricacies. The synergy of improved coordination and communication, facilitated by the data lake, contributes substantially to the efficiency of project management. This integrated approach ensures not only the prompt resolution of requests but also an overarching enhancement of project efficacy through streamlined communication channels and data-driven decision-making.

Conclusions and Recommendations

The research, centered on leveraging data lakes to streamline Indian highway projects, crystallises into two pivotal propositions. The first proposition underscores the transformative impact of visualising data within construction endeavors through data lake implementation. This strategic utilisation addresses delays in progress payments and decision-making with unprecedented efficacy. The second proposition underscores the

instrumental role of data lakes in elevating communication and coordination among project stakeholders, thereby enhancing overall operational efficiency in delay mitigation. The study unequivocally concludes that data lakes represent a potent solution, offering substantial potential to rectify delays by optimising data management, communication, and coordination.

In extrapolating these findings, the imperative for the NHAI becomes evident. Prioritising two key domains data governance and skilled professionals emerges as the strategic pathway to maximise data lake utilisation.

1. ***Data Governance***: Effective data governance is essential for optimising the use of data lakes. To achieve this, organisations should first establish a clear data governance framework by setting clear goals, such as ensuring data quality, securing sensitive data, complying with regulations, and forming a cross-functional team to oversee policy implementation (Gillan, 2021). It is also important to maintain detailed documentation and manage metadata to keep data easily accessible and understandable. Implementing role-based access controls and maintaining audit trails helps protect sensitive information and monitor data usage. Organisations should classify data based on sensitivity and handle it, accordingly, applying necessary security measures for highly sensitive data. Managing the data lifecycle by setting policies for data retention, archiving, and deletion, and conducting regular audits ensures data remains compliant with legal and business requirements (Brous *et al.*, 2016). Continuous monitoring, measuring effectiveness, and adapting governance practices as necessary will maintain the data lake as a secure, compliant, and efficiently managed resource. This approach empowers organisations to fully leverage their data lakes while ensuring data security and compliance (Duzha *et al.*, 2023).
2. ***Skilled Professionals***: To effectively manage data lake systems in the construction industry, professionals require a combination of technical skills and soft skills. Technical skills encompass knowledge of data platforms like data warehouses and data lakes, proficiency in data management tools such as Delta Lake and Snowflake and understanding of data processing strategies like those in Hadoop (Kaur *et al.*, 2023). Additionally, soft skills like effective communication and presentation are vital for managing both data and human resources within construction teams. Professionals in the construction industry must possess skills in sensor data analytics and data science, including machine learning. Understanding data analytics applications across various construction phases is crucial. A methodology integrating Building Information Modelling (BIM) and Business Intelligence (BI) tools enables collaborative data management. The adoption of information communication technology (ICT) facilitates decision-making and project management (Huang *et al.*, 2012). Managing data lakes requires proficiency in data platforms, management tools, and processing strategies. Investing in skilled professionals is crucial for deriving insights and optimising operations. The NHAI recognises the importance of this and focuses on recruitment, retention, and training programs. Collaborative efforts with academia and industry experts further contribute to bridging the skills gap and fostering innovation within the construction industry. This holistic approach ensures that data lake systems are managed efficiently, driving progress, and enhancing the construction sector's competitiveness.

By focusing on these strategic imperatives of data governance and skilled professionals, NHAI is poised to unlock the full potential of their data lake. This approach not only facilitates effective analysis of highway data but also enhances operational efficiency and safety measures across the expansive highway network.

References

Abraham, R., Schneider, J. and Vom Brocke, J. (2019) 'Data governance: A conceptual framework, structured review, and research agenda', International Journal of Information Management, 49, pp. 424–438.

Barbierato, E., Gribaudo, M., Serazzi, G. and Tanca, L. (2021) 'Performance evaluation of a data lake architecture via modeling techniques', in European Workshop on Performance Engineering. Springer, pp. 115–130.

Brous, P., Herder, P. and Janssen, M. (2016) 'Governing Asset Management Data Infrastructures', in Procedia Computer Science. Elsevier B.V., pp. 303–310.

Chomo, T. (2019) Deploying Data Lake for Big Data Management. Masaryk University.

Couto, J.C. and Ruiz, D.D. (2022) 'An overview about data integration in data lakes', in 2022 17th Iberian Conference on Information Systems and Technologies (CISTI). IEEE, pp. 1–7.

Durdyev, S. and Hosseini, M.R. (2020) 'Causes of delays on construction projects: a comprehensive list', International Journal of Managing Projects in Business, 13(1), pp. 20–46.

Duzha, A., Alexakis, E., Kyriazis, D., Sahi, L.F. and Kandi, MA (2023) 'From Data Governance by design to Data Governance as a Service: A transformative human-centric data governance framework', in Proceedings of the 2023 7th International Conference on Cloud and Big Data Computing, pp. 10–20.

Edison, J.C. and Singla, HK (2020) 'Development of a scale for factors causing delays in infrastructure projects in India', Construction Economics and Building, 20(1), pp. 36–55.

Fang, H. (2015) 'Managing data lakes in big data era: What's a data lake and why has it became popular in

data management ecosystem', in 2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER). IEEE, pp. 820–824.

Giebler, C., Gröger, C., Hoos, E., Schwarz, H. and Mitschang, B. (2019) 'Leveraging the Data Lake: Current State and Challenges', Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11708 LNCS,

Gillan, A. (2021) 'Governance: the key driver for data-driven innovation', Computer Fraud & Security, 2021(4), pp. 10–13.

Gorelik, A. (2019) 'The enterprise big data lake: Delivering the promise of big data and data science.' O'Reilly Media.

Gudivada, V., Apon, A. and Ding, J. (2017) 'Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations', International Journal on Advances in Software, 10(1), pp. 1–20.

El Haddadi, O., El Hamlaoui, M., Dkaki, T. and Nassar, M. (2020) 'Data Lake and Digital Enterprise.', in ENASE, pp. 423–429.

Hagstroem, M., Roggendorf, M., Saleh, T. and Sharma, J. (2017) A smarter way to jump into data lakes.

Hai, R., Koutras, C., Quix, C. and Jarke, M. (2023) 'Data lakes: A survey of functions and systems', IEEE Transactions on Knowledge and Data Engineering [Preprint].

Inmon, B. (2016) Data Lake Architecture: Designing the Data Lake and avoiding the garbage dump. Technics publications.

Johny, M.G. and Pillai, S. (2022) 'Analysing the vital role of enterprise data lake in the era of digital transformation', in AIP Conference Proceedings. AIP Publishing.

Kaur, P., Kaushik, A. and Kapoor, A. (2023) 'Skills and Responsibilities of Data Wrangler', Data Wrangling: Concepts, Applications and Tools, p. 19.

Keane, P.J. and Caletka, A.F. (2015) Delay analysis in construction contracts. Wiley Blackwell Oxford, UK.

Khine, P. P. and Wang, Z.S. (2018) 'Data Lake: a new ideology in big data era', ITM Web of Conferences, 17, p. 03025.

Kumar, A. and Chundi, P. (2023) 'Data Lakes', in Encyclopedia of Data Science and Machine Learning. IGI Global, pp. 410–424.

Laurent, A., Libourel, T., Madera, C. and Miralles, A. (2020) 'The Gravity Principle in Data Lakes', Data Lakes, 2, pp. 187–199.

Mathis, C. (2017) 'Data Lakes', Datenbank-Spektrum, 17(3), pp. 289–293.

Ministry of Road Transport & Highways Year End Review-2022: Ministry of Road Transport and Highways.

Nagel, S., Corea, C. and Delfmann, P. (2021) 'Cognitive effects of visualisation techniques for inconsistency metrics on monitoring data-intensive processes', Information Systems Management, 38(4), pp. 342–357.

Nambiar, A. and Mundra, D. (2022) 'An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management', Big Data and Cognitive Computing. MDPI.

Nargesian, F., Zhu, E., Miller, R.J., Pu, K.Q. and Arocena, P.C. (2018) 'Data Lake management: Challenges and opportunities', in Proceedings of the VLDB Endowment. VLDB Endowment, pp. 1986–1989.

Saini, M. (2015) A framework for transferring and sharing tacit knowledge in construction supply chains within lean and agile processes. University of Salford (United Kingdom).

Sawadogo, P.N. (2019) 'Textual Data Analysis from Data Lakes', in New Trends in Databases and Information Systems: ADBIS 2019 Short Papers, Workshops BBIGAP, QAUCA, SemBDM, SIMPDA, M2P, MADEISD, and Doctoral Consortium, Bled, Slovenia, September 8–11, 2019, Proceedings 23. Springer, pp. 558–563.

Schneider, J., Gröger, C., Lutsch, A., Schwarz, H. and Mitschang, B. (2023) 'Assessing the Lakehouse: Analysis, Requirements and Definition.', in ICEIS (1), pp. 44–56.

Sharma, B. (2018) Architecting data lakes: data management architectures for advanced business use cases. O'Reilly Media.

Singh, A. and Ahmad, S. (2019) 'Architecture of data lake', International Journal of Scientific Research in Computer Science, Engineering, and Information Technology, 5(2), p. 4.

Singh, J., Singh, G. and Bhati, B.S. (2022) 'The implication of data lake in enterprises: A deeper analytics', in 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE, pp. 530–534.

Terrizzano, I.G., Schwarz, P.M., Roth, M. and Colino, J.E. (2015) 'Data Wrangling: The Challenging Journey from the Wild to the Lake.', in CIDR. Asilomar.

Tripathi, O.P., Hasan, A., Jha, K.N. and Jain, A.K. (2023) 'Evaluating Government Contracts for Delays, Delay Damages, and Levy of Compensation Provisions', Journal of Legal Affairs and Dispute Resolution in Engineering and Construction, 15(1).