ANALYSIS OF GLOBAL PRACTICES AND DESIGN OF A ROAD SAFETY DATA LAKEHOUSE FOR THE PHILIPPINES

[1]MAYWADEE SOYTONG, [2]ADRIAN ROY L. VALDEZ, [3]SUSAN PANCHO-FESTIN

[1]Scientific Computing Laboratory, Department of Computer Science, the University of the Philippines Diliman, Quezon City, Philippines
[2]Scientific Computing Laboratory, Department of Computer Science, the University of the Philippines Diliman, Quezon City, Philippines
[3]Computer Security Group, Department of Computer Science, the University of the Philippines Diliman, Quezon City, Philippines
E-mail: [1]msoytong1@up.edu.ph, [2]alvaldez@up.edu.ph, [3]spfestin@up.edu.ph

Abstract - There are many ways to store and manage data, depending on the nature and goals of the system. This paper presents a data lakehouse architecture designed to store and integrate road safety data in the Philippines. In the Philippines, road safety information is fragmented among government agencies, including those in charge of the police, health, and transport, and non-system interrelation helps stall an early analysis and coordinated policy response. Here, this paper outlines the research and plan for a unified, scalable data platform using road safety as a case study. Architecture solves issues such as source and format diversity. Specific to the design phase of the architecture, the paper introduces both a conceptual model and an implementation roadmap to guide future development.

Keywords - Data Lakehouse, Road Safety Data, Data Architecture, Data Integration, Data Storage

I. INTRODUCTION

Storing and managing data to make it effective sufficiently to support analysis, integration, and decision-making is critical. In recent years, the data lakehouse has emerged as a powerful architecture that offers the best of both data lakes (unlimited flexibility without constraint) and data warehouses (reliable and governed data). It allows for the ingestion, storage, and processing of large volumes of structured and semi-structured data and can carry out powerful analytics and policy enforcement. This paper applies the data lakehouse to the domain of road safety, focusing on the challenges and opportunities in the Philippines. Road safety data in the country is fragmented across various agencies, including law enforcement, healthcare, transportation, and local governments, which each run their own independent data systems.

The paper is structured into five main parts: the literature review, which covers foundational concepts of data storage and modeling relevant to this study; research on road safety data, which explores how other countries collect and manage such data; and the discussion section, which is divided into two sub-sections, an analysis of similarities and differences across international systems, and an examination of the current fragmented road safety data landscape in the Philippines. Finally, the paper presents a lakehouse architecture proposal for road safety data integration, offering a design approach tailored to the local context.

II. THEORETICAL BACKGROUND

This chapter introduces important theoretical concepts related to data storage, data warehousing and data modeling. The discussion emphasizes the advantages and disadvantages of different data storing approaches and explains traditional and modern data warehousing strategies and data modeling techniques.

A. Data Storage

Modern data storage systems have evolved to handle diverse data types and volumes. Three main models are widely used: data lakes, data warehouses, and the newer data lakehouses. A data lake stores raw, unstructured, or semi-structured data at scale but often lacks governance, risking disorganization [1,2,4]. In contrast, a data warehouse offers structured, governed storage optimized for analytics, but it struggles with flexibility and diverse data formats [3,4]. The data lakehouse combines both approaches, offering the scalability of a data lake with the structure and reliability of a warehouse. It supports ACID compliance, enforces schemas, and is well-suited for managing complex data environments [2,5].

B. Data Modeling

Data modeling defines how data is organized, related, and stored efficiently. A popular approach, dimensional modeling, is used in structured data where it simplifies data into facts and dimensions. However, it is less effective for handling real-time or semi-structured data[8,9]. In contrast, Data Vault modeling offers greater flexibility by separating data into three components: 1) hubs, which represent core business entities; 2) links, which capture relationships between these entities; and 3) satellites, which store descriptive attributes. This structure enables historical tracking and adapts well to changing data sources, making it well-suited for complex and variable

environments such as road safety and public sector systems[10].

C. Data Warehousing

Data warehousing involves designing systems that collect, integrate, and manage data from various sources to support analysis and decision-making. Three major modeling approaches guide this process: the top-down, bottom-up, and Data Vault models. The top-down approach (Inmon) builds a centralized data warehouse first, then creates departmental data marts [11]. It ensures consistency but requires significant upfront effort. The bottom-up approach (Kimball) starts with data marts and later integrates them into a full warehouse, allowing faster deployment but risking poor integration if not well-managed[11]. Data Vault 2.0 is a full-stack methodology focused on agility, scalability, and traceability. It emphasizes agility, scalability, and traceability across three primary layers: the Raw Data Vault (for storing unmodified, auditable data), the Business Vault (where business rules and logic are applied), and Information Marts (used for reporting and analytics). This separation allows for scalability, historical tracking, and compliance, and it is very suitable for dynamic and complex environments [10,11].

III. THE RESEARCH OF ROAD SAFETY DATA

This section presents a system-level analysis of how various countries collect, manage, and utilize road safety data. The goal is to understand the structure, processes, and technologies adopted internationally in order to inform the design of a suitable architecture for the Philippine context.

1) In Japan, road safety data is collected and coordinated primarily through the National Police Agency (NPA) in collaboration with the Ministry of Land, Infrastructure, Transport and Tourism (MLIT), the Japan Road Traffic Information Center, and local authorities. The system both centrally collects and maintains road traffic information from police accident reports, vehicle registrations, road sensors, weather data, and hospital/clinic records through a standard collection process. Accident reports are primarily data sources filed by police departments and enriched with supplemental data from hospitals, emergency services, vehicle telematics, and traffic monitoring systems.

Centralized storage and integration are handled by institutions such as the Institute for Traffic Accident Research and Data Analysis (ITARDA), which consolidates, cleans, and anonymizes the data for research, policy-making, and urban planning purposes. As highlighted in Kiuchi's study (2018) [12,13], ITARDA's database includes detailed information about accident dynamics, such as the number and types of vehicles involved, time and location of the incident, injury or fatality severity, weather conditions, and road infrastructure characteristics. The system emphasizes centralized data governance, interoperability, and multi-source integration supported by national standards. Local offices also access localized subsets of the data to identify opportunities for intervention in localized traffic patterns or dangerous areas. Japan's approach serves as a reference model for integrated, data-driven traffic safety management [12,13].

2) In India, the technology-based model was used as one of the methods for enhancing road safety through digital data collation and digital reporting by the government. The Ministry of Road Transport and Highways (MoRTH) is responsible for leading national-level efforts through the Road Accident Data Management System (RADMS), a national-level platform designed to capture detailed accident data across states and union territories [14,15]. RADMS collects diverse data sets such as the accident count, fatalities, injuries, details of vehicles, time and location of accidents, road conditions, driver behaviors, and weather conditions. Data are mainly captured by traffic police departments and traffic police units like in other countries, which can guarantee that first-hand information is collected at the scene. The collected data is also processed and analyzed on a national basis for use in analytics, infrastructure design, and decision-making. India has also launched a pilot on AI-based crash detection, with CCTV-based surveillance and vehicle telematics systems to reduce manual data entries and enforce real-time reporting to improve the quality and timeliness of data. RADMS also incorporates dashboards and analytics tools for standardized reporting across geographies to help make rapid decisions and focus safety interventions. Overall, India’s evolving system shows a high level of commitment to data-based road safety management with digital technology tools and centralized governance [14,15].

3) In Nigeria, national road safety data collection and management are the functions of the Federal Road Safety Corps (FRSC). At the heart of its activities is the Road Traffic Crash (RTC) database, which collects comprehensive data on road accidents. Data is collected from a multi-source network, including FRSC field officers, hospitals, police stations, and insurance companies, which feeds into an open and collaborative input system [16,17]. The RTC database captures a number of variables such as crash location, time, vehicle type, casualty details, and suspected causes of incidents. This information is used to generate national crash statistics, support public road safety awareness campaigns, and inform targeted enforcement and policy strategies. While challenges

remain, particularly in achieving full digital integration and standardized reporting across all regions, the current system represents a significant advancement toward structured and evidence-based road safety management in Nigeria [16,17]..

4) In Sweden, there is a Vision Zero strategy that supports road safety data through an extensive, integrated system of data sharing and cooperation among various stakeholders [18,19]. Local police initially collect accident data at the scene and also details of the crash, such as vehicle type, road conditions, and driver behavior. Then, the data is forwarded to national databases managed by the Swedish Transport Administration (Trafikverket) and Statistics Sweden, where it is cleaned, standardized, and matched against other data sources. Moreover, external information, such as traffic sensors, vehicle telematics based on GPS, weather systems, and healthcare provision using Geographic Information Systems (GIS), can be added to the system that makes a comprehensive assessment of traffic incidents. The aggregated dataset is used for real-time monitoring, infrastructure planning, and national traffic safety strategies [18,19].

5) In the US, the Fatality Analysis Reporting System (FARS) is maintained by the National Highway Traffic Safety Administration (NHTSA) as a single source for fatal crash data [20,21]. Law enforcement in every state in the country reports accidents via a common electronic form. These reports are increased by state transportation departments, emergency medical services (EMS), death certificates, and hospitals. All these data are gathered in FARS in a single point, which can be used to explore the trend, the risk factors, and the street safety policy of the country. It also uses information from roadside devices like automatic crash notification systems, vehicle data recorders, and the nation's system of intelligent highways. These information transfers help federal and state agencies share funding, enforcement, and public education efforts in combating the highway death tally [20,21].

6) In South Korea, the government has introduced smart traffic systems powered by state-of-the-art technology for next-generation road safety monitoring and action. The collection of data starts at a street level with a large number of traffic sensors, CCTV cameras, automatic number plate recognition, roadside IoT devices, etc. These systems record data such as vehicle matching sequence times, vehicle speeds, traffic violations, and accident events. Data is sent to the Korea Transportation Safety Authority (KOTSA) and the local government's Integrated Traffic Management Centers when it reacts to an accident [23]. They act as command and control centers in which data from the field is used to feed AI and analysis tools that monitor for unusual traffic events, classify incidents, and fire off alerts. Alongside hardware sensors using infrastructure, mobile GPS phone data and telematics from connected vehicles are also used to increase situational awareness within the system. After they are validated, they are archived in national databases and serve as resources for post-incident analysis, policy development, infrastructure improvements, and predictive traffic risk analysis. The integrated system in South Korea puts advanced traffic management and the first emergency response in place for monitoring traffic and long-term planning to be developed and enables proactive traffic control and prompt emergency response envisioning with smart real-time monitoring, big data analysis, and multi-agency coordination [22,23].

IV. DISCUSSION

The system of every country is a reflection of its legal system, technological capabilities, and inter-agency coordination. But common elements include collecting data in one place, cross-sector cooperation, and using analytics to make decisions before they have to be made. There’s a lesson out there in these cases, an important one for policymakers working toward the establishment of integrated data platforms in countries like the Philippines, where data systems are still fragmented and developing.

THE ANALYSIS OF ROAD SAFETY DATA

A comparison of road safety data systems in several countries (including Japan, India, Nigeria, Sweden, the United States, and Korea) highlights a number of similar features of systems and principles of governance and data management despite variations in technological maturity and resource availability.
One common factor in all of these systems is that police agencies are the collectors of initial crash data. The original point of contact in recording crash scenes is provided in Japan, Sweden, India, the U.S., and Nigeria by police officers or traffic units. Their standardized reports underpin national road safety data systems, supplemented by other contributions from hospitals, emergency services, insurers, and transport authorities. This essential function of the police means that any incident information entered and made available is the quickest, most robotically correct, and consistently structured material possible.

Another trend discovered is the rise of centralized data integration. Be it ITARDA in Japan, RADMS in India, the RTC database of Nigeria, or Vision Zero in Sweden. Each country has created a centralized platform that deploys the multi-source data through a single national framework. These systems typically count on heterogeneous data, including vehicle logs, road environment details, weather conditions, GPS tracking data, and patient records. This multi-agency

effort gives a more complete picture of road safety and accident patterns. In addition, efforts to build digitized infrastructure and advanced technology are taking place in many countries to enhance the speed, accuracy, and availability of road safety data. South Korea is a prime example with real-time intelligent traffic control systems, though India has also launched pilot projects with AI bots for crash detection and telematics. The U.S. and Sweden also integrate with other data sources such as roadside sensors, vehicle black boxes, geographic information systems (GIS), and traffic and weather information for improved situational awareness and decision support.

Finally, all of these systems have a common mission to convert raw data into practical information for infrastructure investment, policy formulation, enforcement tactics, and public communication initiatives. Dashboards, analytics tools, and scheduled reporting are commonly relied upon to facilitate evidence-informed decision-making at national and regional levels.

ROAD SAFETY DATA IN THE PHILIPPINES

In the Philippines. The government has made considerable efforts to manage road safety data through various institutional agencies. Multiple systems have been established by key agencies such as the Metro Manila Accident Reporting and Analysis System (MMARAS), the Department of Transportation's DRIVER system, the Department of Health's Online National Electronic Injury Surveillance System (ONEISS), and the Philippine Network for Injury Data Management System (PNIDMS), and the Philippine Statistics Authority's Vital Statistics Reports (VSR) [24,25]. These sources gather information at various points of the accident path, including incident reporting, vehicle records, injury documentation, and mortality rates. But in many cases, such systems work in standalone mode, causing problems of standardization, integration, and access to data. This fragmentation impedes any integrated analysis across the agencies and stifles the capacity to obtain timely and reliable intelligence for policy and planning.

Compared to international practices, the Philippines demonstrates some alignment with other countries in its multi-agency approach to road safety data. Similar to other countries, the Philippines has a multi-agency approach to road safety data. As in the United States and Sweden, the Philippines also relies on health sector data to supplement traffic and police reports. The presence of these various platforms reflects the decentralized pattern observed in nations like Nigeria and India. Moreover, national statistical systems such as the VSR also play a role similar to Japan and South Korea's ITARDA and statistical reporting on transport-related fatalities, respectively, collecting historical information on road safety incidents. However, significant barriers remain compared with more interconnected global systems. For example, countries such as South Korea and Sweden have established centralized data platforms that collect inputs from different agencies, standardize the formats of reports, and can process the data in an automated or near realtime manner. They allow effective and rapid responses so that governments can rapidly assess lessons learned and maximize road safety.

To keep in line with these global trends, the Philippines may want to consider developing an integrated and scalable standards-based road safety data system. This practice would not only increase the reliability of road traffic data but also would build the country's capability to manage traffic safety proactively in the future by empowering effective coordination between agencies, instituting national data standards, and setting the groundwork for continued access to new and emerging information technologies (e.g., predictive traffic systems, auto incident reporting).

V. A LAKEHOUSE ARCHITECTURE PROPOSAL FOR ROAD SAFETY DATA INTEGRATION IN THE PHILIPPINES

This section explains the implications of the proposed lakehouse design and translates the architectural model into the operational aspects of road safety data in the Philippines. It expects key challenges in data integration, inter-agency coordination, and system implementation. The design leverages a data lakehouse architecture underpinned by the Data Vault 2.0 methodology, which enables unified, historical, and scalable data storage. The Data Vault 2.0 model was selected for its ability to manage dynamic, heterogeneous road incident data across multiple government agencies. In contrast to other data modeling techniques, the Data Vault model offers better flexibility in adapting to the evolving structure of road safety data, such as including new accident types, additional agencies, or contextual metadata.

The lakehouse approach avoids duplication and supports both structured querying and advanced analytics. This design facilitates historical tracking, enables incremental updates, and supports diverse analytical use cases, such as geospatial risk mapping, behavioral analysis, and traffic pattern prediction. While effective in many scenarios, other models often require stable data structures and are less suited to handle highly variable, real-time, or semi-structured data.

The proposed development plan follows the academic stages of Data Vault 2.0 implementation. The data

lakehouse design in this research is presented in four core topics: (1) business processes, (2) business rules, (3) data utilization, and (4) tools and platforms. These research topics reflect the foundational analysis required before implementation.

1. Business Processes

The target business process for this system is the integration of road incident data. Multiple stakeholders are involved in this process, including traffic enforcement units, hospitals, insurance companies, and transport agencies. The design ensures the following:

- Data format independence: Road safety data, whether from police reports (e.g., MMARAS), hospital records (e.g., ONEISS), or surveillance logs, can be ingested in their raw formats without disruption to source systems.
- Non-invasive data collection: Source systems remain operational, while the lakehouse ingests routine or real-time data feeds asynchronously.
- Support for complex scenarios: The model handles simple incidents (e.g., two vehicles colliding) as well as complex multi-party in the same event (e.g., multiple vehicles and pedestrians), enabling seamless data integration across agencies.

2. Business Rules Specification

In Data Vault modeling, identifying the unique business keys can define the hub entity. In the context of Philippine road safety data, the core business events involve vehicle accidents. Business rules are established to handle the following:

- Initially, a basic two-vehicle collision is modeled.
- Then, multi-vehicle collisions are added.
- Additional layers include whether individuals are inside or outside the vehicle during the incident and the presence or absence of vehicle ownership data.

Accordingly, the Hub structure is defined as:

- Hub_Incident: representing the unique identifier for each reported road incident.
- Hub_Vehicle: capturing vehicle identifiers involved in incidents.
- Hub_Person: recording individuals affected by or involved in incidents

These hubs are linked through Link_Incident_Vehicle and Link_Vehicle_Person, while time-variant and other attributes (e.g., crash severity, driver sobriety, time of report) are stored in Satellites for the details. Figure 1 shows the conceptual data model that applied Data Vault modeling designed for road safety data in the Philippines.

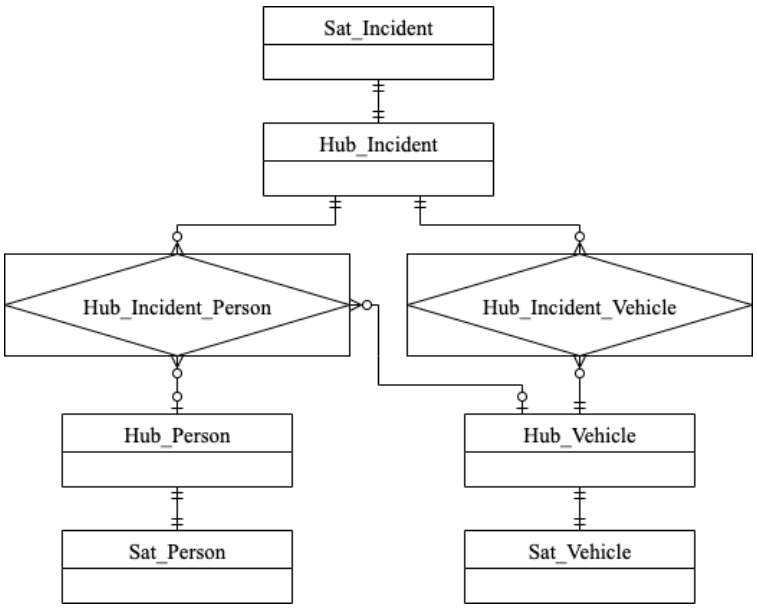


Figure 1 presents a conceptual model for road safety data in the Philippines.

3. Data Usage and Analytical Purpose

The integrated lakehouse is designed for multi-agency collaboration and advanced analytics. The data stored within the lakehouse is expected to support several key applications:

- Predictive modeling using historical and environmental data.
- Real-time dashboards and business intelligence (BI) tools.
- Long-term data retention for regulatory audits and post-crash analytics.

To meet these goals, the platform must be ready for querying, historical traceability and secure sharing with authorized stakeholders.

4. Tools and Implementation Environment

To support the objectives and enable agile integration, this research considers using Dremio, an

open-source data lakehouse query engine. Dremio supports:

- High-performance query execution on large datasets without data duplication
- Semantic layer management for consistent business logic across users
- Native support for lakehouse-style architecture using open formats like Apache Iceberg

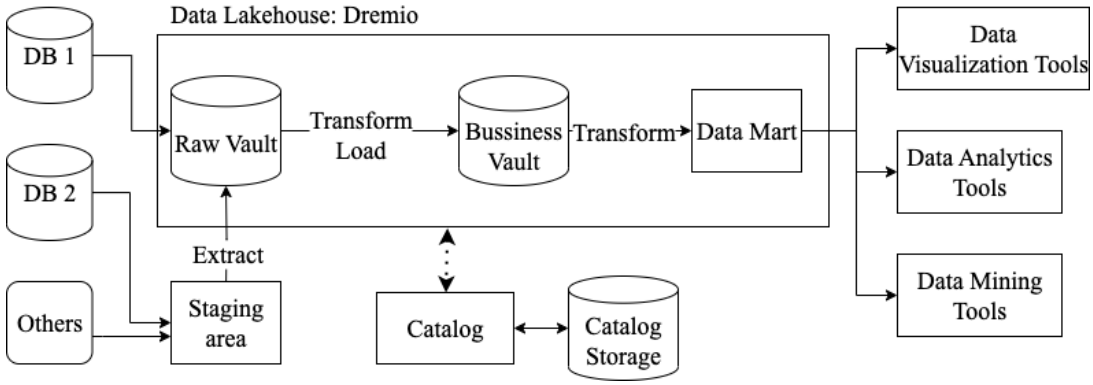


Figure 2 illustrates the proposed system architecture for the Philippine Road Safety Data Lakehouse platform.

Figure 2.0 shows the proposed system, which is a data lakehouse for road safety data in the Philippines. In the ETL process, data is being extracted from databases, and then those databases transformed to obtain the corresponding format to data structured on data lakehouse. These are steps used in the ETL process using Dremio.

VI. CONCLUSION

The growing data on road safety in the Philippines demands an integrated solution that is scalable and data-driven for road incident information. There are lessons to be learned from international models, such as Japan, India, Sweden, and South Korea, which have demonstrated over the years the value of centralized systems, standardized data formats, and analytics. The road safety data system in the Philippines has many core components in place but still lacks this level of integration. The presented design in this paper proposes a foundation for future development, offering a blueprint for how the country can evolve toward a more collaborative road safety data ecosystem.

This paper presents a data lakehouse architecture for consolidated multi-source government data, which scales easily and can be used for both real-time analytics and long-term policy planning. The implementation of the data lakehouse architecture for road incident data integration in the Philippines is a major innovation in the field of public sector data warehousing. By combining the flexibility of the data lakes with the structured reliability of the data warehouses from the old school, this architecture is a modern and scalable approach for handling complex, multi-source data. Utilizing the Data Vault 2.0 approach, the designed warehouse provides flexibility, historical tracking, and audibility, which are appropriate for such a dynamic situation as reporting and responding to road incidents. The next steps will be prototyping and data verification to make sure the system will technically work as well as fit with organizational capabilities and governance structures.

REFERENCES

[1] E. Zagan and M. Danubianu, “Data lake approaches: A survey,” in Proc. 2020 Int. Conf. Development and Application Systems (DAS), 2020, pp. 189–193.

[2] M. Armbrust, A. Ghodsi, R. Xin, and M. Zaharia, “Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics,” [Online]. Available:www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf.

[3] M. B. Shahid, U. Sheikh, B. Raza, and Q. Javaid, “Application of data warehouse in real life: State-of-the-art survey from user preferences’ perspective,” Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 4, 2016.

[4] P. Kuppusamy, “Data lake model to modern educational organizations.”

[5] Databricks, “What is the medallion lakehouse architecture,” [Online].Available:docs.databricks.com/gcp/en/lakehouse/medallion, 2024.

[6] Microsoft, “What is the medallion lakehouse architecture,” [Online].Available:earn.microsoft.com/en-us/azure/databricks/lakehouse/medallion, 2024.

[7] P. Kava and C. Gong, “Build a lake house architecture on AWS,” [Online]. Available:aws.amazon.com/blogs/big-data/build-a-lake-house-architecture-on-aws/, 2021.

[8] B. K. Seah and N. E. Selan, “Design and implementation of data warehouse with data model using survey-based services data,” in Proc. 4th Int. Conf. Innovative Computing Technology (INTECH), 2014, pp. 58–64.

[9] Y. Nagm Aldeen, “Data warehouse – dimensional model vs normalized model,” July 2020.

[10] W. Inmon, L. Daniel, and L. Mary, “Design and build a data vault model in Amazon Redshift from a transactional database.”

[11] L. Yessad and A. Labiod, “Comparative study of data warehouses modeling approaches: Inmon, Kimball and Data Vault,”[Online].Available:www.researchgate.net/publication/312486486_Comparative_study_of_data_warehouses_modeling_approaches_Inmon_Kimball_and_Data_Vault, 2016.

[12] ITARDA, “About ITARDA,” Institute for Traffic Accident Research and Data Analysis (Japan), 1992. Link: https://www.itarda.or.jp/english/

[13] Oguchi, Takashi. (2016). Achieving Safe Road traffic - the experience in Japan. IATSS Research. 39. 10.1016/j.iatssr.2016.01.003.

[14] Bhattacherjee, S., Tewari, V. K. (2015) Road safety management in India: initiatives, issues and challenges. Transportation research procedia, 10, 87-96.

[15] World Bank, "India: Improving Road Safety – Road Accident Data Management System (RADMS) in Tamil Nadu," [Online]. Available: worldbank.org/en/news/feature/2011/12/22/india-improving-road-safety, 2011.

[16] O. T. Shitta-Bey, O. Olatunbosun, F. Bankole, and S. I. Oni, "Road traffic crash data systems in Lagos State, Nigeria: A situational analysis," preprint, 2020.

[17] Federal Road Safety Corps (FRSC), "Evaluation of road traffic crash surveillance system, Federal Capital Territory, Nigeria – 2016," Tephinet, 2016.

[18] M.Å. Belin, P. Tillgren, and E. Vedung, "Vision Zero – a road safety policy innovation," Int. J. Injury Control and Safety Promotion, vol. 19, no. 2, pp. 171–179, 2012. Link: https://doi.org/10.1080/17457300.2011.635213

[19] A.C. Kristianssen, R. Andersson, M.-Å. Belin, and P. Nilsen, "Swedish Vision Zero policies for safety – a comparative policy content analysis," Safety Science, vol. 103, pp. 260–269, 2018. Link: https://doi.org/10.1016/j.ssci.2017.11.005

[20] National Highway Traffic Safety Administration (NHTSA), "Fatality Analysis Reporting System (FARS): census of fatal motor vehicle traffic crashes in the US since 1975," [Online]. Available: nhtsa.gov/crash-data-systems/fatality-analysis-reporting-system.

[21] U.S. Department of Health and Human Services, "Fatality Analysis Reporting System (FARS): data sources and collection methods," [Online]. Available: health.gov data-sources/fatality-analysis-reporting-system-fars.

[22] Korea Transportation Safety Authority, "Smart Traffic Systems and Integrated Traffic Management in South Korea," 2023. [Online]. Available: https://www.kotsa.or.kr/smart-traffic-systems-report.

[23] W. Kang, J. Lee, Y.-S. Jeong, and J. Park, "VCC-SSF: Service-Oriented Security Framework for Vehicular Cloud Computing," Sustainability, vol. 7, no. 2, pp. 2028–2044, 2015. Link: https://doi.org/10.3390/su7022028

[24] Philippine Department of Transportation, "Road Safety Data Systems and Integration Challenges in the Philippines," 2023. [Online]. Available: https://dotr.gov.ph/road-safety-data-systems-report.

[25] Metropolitan Manila Development Authority (MMDA) Traffic Engineering Center – Road Safety Unit, "Metro Manila Accident Reporting and Analysis System (MMARAS) Annual Report 2022," 2022.

★★★