Analyzing Real-World Road Traffic Incident Data Collection and Management

Maywadee SOYTONG [a], Adrian Roy L. VALDEZ [b], Susan PANCHO-FESTIN [c]

[a,b] *Scientific Computing Laboratory, Department of Computer Science, University of the Philippines-Diliman, Quezon City, 1101, Philippines*

[a,b] Intelligent Transportation Systems Laboratory, National Center for Transportation Studies, University of the Philippines-Diliman, Quezon City, 1101, Philippines

[a] *E-mail: msoytong1@up.edu.ph*

[b] *E-mail: alvaldez@up.edu.ph*

[c] *Computer Security Group, Department of Computer Science, University of the Philippines-Diliman, Quezon City, 1101, Philippines*

[c] *E-mail: spfestin@up.edu.ph*

Abstract: Reliable and complete road incident data are essential for effective road safety management and policy development. Traditional data collection systems typically use two-dimensional, flat, fixed-field formats that unable to capture the complexity of real-world road incidents, such as multi-vehicle collisions or cases involving pedestrians. This study analyzes road safety data by classifying the data into distinct entity types and uses this types to design a more flexible data model. The proposes a more flexible data model allows for the recording of multiple related data entity types within a single incident, thereby reducing redundancy and ensuring database integrity. The evaluation of the model highlights its ability to accurately represent complex incidents, minimize data duplication, and maintain integrity. These advancements will improve the quality of data analysis and contribute to the development of more effective road safety policies and interventions worldwide.

Keywords: Road Safety, Data Collection, Crashes, Information Systems, Data Management Systems, Case Study

1. INTRODUCTION

A leading cause of death globally, road traffic crashes demand effective road safety management at the global level. This data is used in the development of policies, allocation of resources. However, in some countries, there are challenges in maintaining high-quality data in incident databases (Bhattacherjee and Tewari, 2015; World Bank, 2011; Al Aamri et al., 2024). Technical obstacles such as non-interconnected reporting systems, inconsistent data fields, and a lack of standardization, as well as system limitations, pose significant challenges.

This research examines actual road safety data collection and management practice in a case-based manner. Through comparisons with a wide range of contexts and practices, it seeks to understand the barriers and opportunities or the what-works and the not-works of documenting, storing, and using crash data. The understanding of these factors is essential for the creation of integrated and robust data systems that can facilitate practical road safety actions in various parts of the world.

Accordingly, this study aims to explore how road safety data are currently collected, managed, and structured, and to identify the factors that limit data quality and interoperability.

The specific objectives are to:

1. Review the existing systems and standards for road safety data management to analyze the structure and limitations of current data models;
2. Define representative data entity types that reflect real-world incident situations; and
3. Propose directions for developing a more efficient and flexible data model.

The study is guided by the following research questions:

1. How effectively do existing data collection and management practices handle real-world road safety data?
2. What types of real-world incident data occur in practice?
3. Which data model can best accommodate the full range of real-world incident data?

The remainder of this paper is structured as follows.

Section 2 provides background information, Section 2.1 describes the organizations and frameworks involved in road safety data management, and Section 2.2 reviews schema design approaches.

Section 3 presents the research methodology, which consists of three stages: (3.1) existing models review, (3.2) classifying of data entity type, and (3.3) development toward a more efficient and flexible data model.

Section 4 discusses the findings, highlighting how the proposed model captures a wider range of data entity type and improves data usability.

Section 5 concludes the paper with key insights and recommendations.

2. BACKGROUND

Effective road safety management relies not only on the collection of accurate road safety data, but also on the proper organization, standardization, and modeling of that data. A clear understanding of existing global and regional standards, the organizations that implement them, and the data structures they use is essential for identifying both strengths and limitations in current practices.

This section provides an overview of the frameworks and institutions that guide road safety data collection and management (Section 2.1), followed by a discussion of the traditional data models and storage practices commonly employed by these organizations (Section 2.2). Understanding these background elements forms the foundation for developing a more efficient and flexible data model, which is the focus of the methodology presented in Section 3.

2.1 GLOBAL AND REGIONAL ROAD SAFETY STANDARDS AND ORGANIZATIONS

There are several international and regional conventions, technical handbooks, and institutional regulations that inform the setting up of adequate road accident data systems. Such references provide typical approaches, goals, and strategies for practices to acquire, analyze, and share information to improve road safety. They have helped in providing insights into how national and local efforts relate to wider regional and global aims.

2.1.1 GLOBAL PLAN FOR THE DECADE OF ACTION FOR ROAD SAFETY (WHO)

The World Health Organization (WHO) has played a central role in shaping global approaches to road crash data collection. Since 2004, WHO has worked jointly with partners at national and international levels to share knowledge, best practices, and technical guidance for building capacity at the national level for planning, implementing, and monitoring road safety, as revealed in the Global Plan for the Decade of Action 2021–2030 and its objectives (WHO, 2021). Based on the World Health Organization's seminal work, World Report on Road Traffic Injury Prevention (Peden et al., 2004), the establishment of data systems that are useful for identifying risk, targeting interventions, and evaluating outcomes. The report notes that reliable and timely data are vital to support evidence-informed policy actions that enhance road safety performance. It also calls attention to the contrast between detail and feasibility and suggests that the data elements must both be analytically useful and feasible to collect in the field. It specifically recommends that data be of crash analysis usefulness, well-defined, and not excessively difficult to provide.

Building on this foundation, the WHO and World Bank's Data Systems: A Road Safety Manual for Decision-makers and Practitioners (WHO and World Bank, 2010) develops a minimum of the WHO and World Bank's Data set of data elements that countries must collect if they are to undertake a systematic analysis of traffic crashes. These factors are classified into four major categories: (1) crash factors, (2) road factors, (3) vehicle factors, and (4) human factors. It is recommended that one national data system should be tailored locally to reflect local safety conditions and contextual risk factors by the inclusion of further, more context-specific variables. These are offered as a working precedent in developing credible, consistent crash data systems in support of national and even international road safety activities.

2.1.2 DATA FOR ROAD INCIDENT VISUALIZATION, EVALUATION, AND REPORTING (DRIVER)

The Data for Road Incident Visualization, Evaluation, and Reporting (DRIVER) system is an open-source, web-based crash-data management platform developed by the World Bank's Global Road Safety Facility in partnership with the WHO and the Federation Internationale de l'Automobile (FIA) Foundation (World Bank and Global Road Safety Facility, 2018). DRIVER facilitates real-time, geocoded recording and collaborative analysis of road crashes across multiple agencies such as police, health services, and local governments, making it easy to adopt, customize, and deploy at minimal cost. It includes analytical tools for black-spot identification, estimating economic costs of crashes, tracking intervention effectiveness, and integrating seamlessly. Piloted initially in the Philippines in 2013 and later expanded to countries like Brazil, Bangladesh, India, Laos, Thailand, Vietnam, Saudi Arabia, and Kazakhstan (World Bank and Global Road Safety Facility, 2019). DRIVER empowers decision-makers with standardized, accurate, and shareable crash databases crucial for informed, evidence-based road safety strategies.

2.1.3 ASIA-PACIFIC TELECOMMUNITY (APT)

The Asia-Pacific Telecommunity (APT) is an intergovernmental organization that was founded to promote the development of information and communication infrastructure in the region and in the world at large, primarily the economies of Asia and Pacific, promoted by the United Nations Economic and Social Commission for Asia and the Pacific (UNESCAP) and International Telecommunication Union (ITU) and was operational since 1979. The objective of APT is to enhance telecommunication/ICT services and infrastructure in Asia and the

Pacific. According to the Asia-Pacific Telecommunity, traffic accident record systems in Asia are being reformed to improve data standards and analysis methods (APT, 2021), a policy dialogue, information sharing, and coordination forum for telecommunications and ICTs in the region. It supports member countries with respect to ICT standardization, capacity building, and the encouragement of digital inclusion. From 2025, the APT will consist of more than 38 Members, associate members, and a large number of affiliate members, of which the industry representatives will form part.

2.1.4 ASIA-PACIFIC ROAD SAFETY OBSERVATORY (APRSO)

The Asia-Pacific Road Safety Observatory (APRSO) was established in 2009 and makes a significant contribution to road safety developments in the Asia-Pacific region. To standardize and adjust crash data across the member countries, enabling complete regional analysis and evidence-based decision-making. Although the World Health Organisation (WHO) has the foundation for the core crash data elements, research conducted by APRSO with assistance from (APT) (APT, 2021) has identified additional elements that are more specific for the region. The APRSO and APT alliance are focused on using data-centric approaches to lower the number of injuries that happen on the road. The APRSO data fields are in the following categories:

- Crash-related (e.g., date, time, location)
- Road and environment (e.g., road surface, lighting, weather)
- Vehicle-related (e.g., registration, vehicle condition)
- Person-related (e.g., role in the crash, age, gender, injury level)
- Cause factors (e.g., speed, alcohol or drug involvement)
- Post-crash response (e.g., emergency response time, hospitalization details)

Finally, these data sets are intended to fulfill both national reporting requirements and regional integration, as expected to deliver a much more precise and practical picture of road safety trends in Asia and the Pacific countries.

2.1.5 CENTRAL ASIA REGIONAL ECONOMIC COOPERATION (CAREC)

The CAREC Road Crash Investigation Manual was developed in March 2025 by the Central Asia Regional Economic Cooperation (CAREC) program in collaboration with the Asian Development Bank and aims to provide a standardized approach for road crash data collection, investigation, and analysis (Asian Development Bank, 2025). The manual is intended to enhance the quality of data and ensure consistency across the CAREC member countries to facilitate better road safety interventions in the Central Asia region. It includes road safety data topics such as crashes, persons, vehicles, and roads and includes a multi-faceted array of actions that will be taken to obtain a better understanding of what can be done to prevent road traffic deaths and injuries.

The organizations reviewed in this study contribute substantially to improving data quality in support of road safety enhancement. They provide valuable frameworks for data collection, classification, and reporting, as well as promote collaboration among agencies.

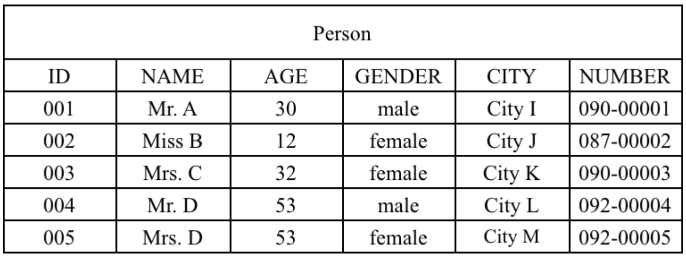
2.2 SCHEMA DESIGN APPROACHES

The organization of information in a database, the so-called schema design, has a strong influence on how efficiently data can be stored, accessed, and queried. Various schema approaches are the result of different trade-offs: simplicity, data richness, flexibility, and analyzability. This section outlines three principal schema design strategies: the flat table approach, the dimensional table approach, and the Data Vault 2.0 approach.

2.2.1 FLAT TABLE MODEL

The flat table model is the most traditional and straightforward way to model data. In this approach, all the data is in a single wide table with many columns as Figure 1. Each record represents one event and includes every possible relevant field. The main advantage of the flat table structure lies in its simplicity. (Rob and Coronel, 2014) It is simple to design, use, and analyze, particularly for small systems or where the data is manually collected.

Flat tables also make it straightforward to export data into Excel or spreadsheet formats or basic statistical tools for routine reporting and descriptive analysis. But usually, this approach often leads to data redundancy and inconsistency, as information may be duplicated across multiple rows. When more attributes are necessary (e.g., new data fields), the schema has to be changed, which can be hard and error-prone. Flat models work well for simple, small datasets, but become inefficient and less flexible. (Connolly and Begg, 2015)

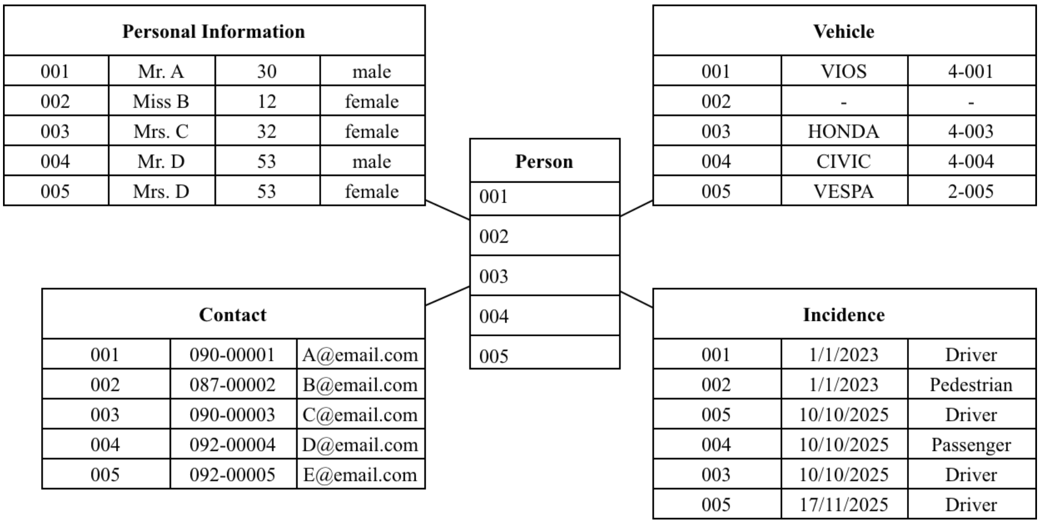


Person					
ID	NAME	AGE	GENDER	CITY	NUMBER
001	Mr. A	30	male	City I	090-00001
002	Miss B	12	female	City J	087-00002
003	Mrs. C	32	female	City K	090-00003
004	Mr. D	53	male	City L	092-00004
005	Mrs. D	53	female	City M	092-00005

Figure 1. Flat table with sample person data

2.2.2 DIMENSION TABLE MODEL

The dimensional table model, by contrast, organizes data into multiple interrelated tables, commonly structured as a star schema or snowflake schema. (Seah, 2014; Nagm, 2020) Instead of keeping all information in one flat table, this approach divides data into fact tables and dimension tables as shown in Figure 2. Every table is connected by unique identifiers, meaning that connections between different data entities can be easily defined. This architecture significantly enhances data structuring, flexibility and analytics potential. It reduces redundancy, makes updating easier, and supports adding new attributes or dimensions without changing the layout. Although such an approach offers several advantages, it also adds a level of complexity in the design and maintenance processes and may be less user-friendly to end users who are unfamiliar with relational databases.



Personal Information			
001	Mr. A	30	male
002	Miss B	12	female
003	Mrs. C	32	female
004	Mr. D	53	male
005	Mrs. D	53	female

Vehicle		
001	VIOS	4-001
002	-	-
003	HONDA	4-003
004	CIVIC	4-004
005	VESPA	2-005

Person
001
002
003
004
005

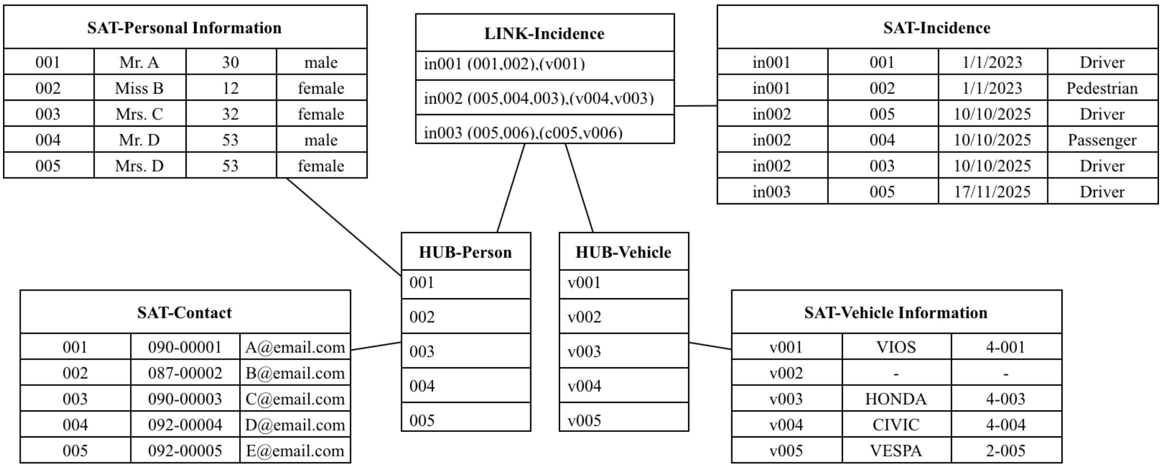
Contact		
001	090-00001	A@email.com
002	087-00002	B@email.com
003	090-00003	C@email.com
004	092-00004	D@email.com
005	092-00005	E@email.com

Incidence		
001	1/1/2023	Driver
002	1/1/2023	Pedestrian
005	10/10/2025	Driver
004	10/10/2025	Passenger
003	10/10/2025	Driver
005	17/11/2025	Driver

Figure 2. Dimension tables (Star schema) with sample person, vehicle and incidence data.

2.2.3 DATA VAULT 2.0 MODEL

The Data Vault 2.0 is combines the scalability of dimensional modeling and the adaptability required for complex and evolving datasets. (Linstedt and Olschimke, 2015; Chaturvedi, 2025) Originally developed by Dan Linstedt in 2015, it emphasizes flexibility, auditability, and long-term data integrity. Primarily built from three main components: hubs, links, and satellites as shown in Figure 3. Hubs represent core business entities. Links capture the relationships or transactions between these entities. Satellites store descriptive attributes and historical changes over time. This architecture makes the model naturally able to express many-to-many relationships without losing data history and contextual information. Furthermore, Data Vault is a modular and scalable structure that can support the addition of new data sources, entities and relationships with little (or no) redesign. (Komninos et al., 2021) Its distinction between structure and content allows parallel data loading and efficient scaling to large, distributed settings. (Yessad et al., 2016) It has nice features that make it suitable for managing data which are heterogeneous, dynamic and continually growing, while ensuring the consistency and traceability of the whole system.



SAT-Personal Information			
001	Mr. A	30	male
002	Miss B	12	female
003	Mrs. C	32	female
004	Mr. D	53	male
005	Mrs. D	53	female

LINK-Incidence
in001 (001,002),(v001)
in002 (005,004,003),(v004,v003)
in003 (005,006),(c005,v006)

SAT-Incidence			
in001	001	1/1/2023	Driver
in001	002	1/1/2023	Pedestrian
in002	005	10/10/2025	Driver
in002	004	10/10/2025	Passenger
in002	003	10/10/2025	Driver
in003	005	17/11/2025	Driver

HUB-Person
001
002
003
004
005

HUB-Vehicle
v001
v002
v003
v004
v005

SAT-Contact		
001	090-00001	A@email.com
002	087-00002	B@email.com
003	090-00003	C@email.com
004	092-00004	D@email.com
005	092-00005	E@email.com

SAT-Vehicle Information		
v001	VIOS	4-001
v002	-	-
v003	HONDA	4-003
v004	CIVIC	4-004
v005	VESPA	2-005

Figure 3. Data Vault 2.0 tables with sample person, vehicle and incidence data.

3. METHODOLOGY

This study takes a quantitative and analytical approach to examine real-world road incident data as structured and managed. The approach of this methodology was applied to find deficiencies in existing databases, and it focused on constructing a more flexible model for accurately capturing the complexity of accident scenarios. The process of the methodology has three principal stages.

3.1 EXISTING MODEL REVIEW

The study began with a review of international and regional frameworks, technical manuals related to road incident data management. These documents offer comprehensive guidance on which data fields should be collected, how they should be defined, and why they are important for analysis and reporting. They also establish standards that help ensure consistency and data quality when implemented in practice. The reviewed documents tend to present variables as sets of tables grouped by domain, typically crash, vehicle, and person. In many systems, crash data are separated into three domain-based tables:

- Crash-level information (date, time, location, severity)
- Vehicle-level information (vehicle type, registration, condition)
- Person-level information (role in crash, age, gender, injury)

As a result, they provide strong foundational guidelines that support more complete and accurate data collection.

However, despite the detailed field specifications, the reviewed documents do not provide examples of database structures or data models. Most frameworks present variables as standalone lists or grouped domains, but they do not illustrate how incidences, vehicles, persons, and contributing factors should be relationally linked within a system. This gap corresponds with the observation that data collected by first responders or frontline data collectors is often stored in a flat table format. Such an approach is convenient for data entry from initial collection tools or sensors, which commonly produce flattened outputs. Yet, significant challenges arise when these data need to be analyzed or integrated with datasets from other agencies.

From an analytical perspective, flat tables require extracting the entire dataset, even when many fields are irrelevant to a given incident. This occurs because flat tables contain numerous context-specific variables. For instance, fields such as injury severity are essential when casualties occur, but in minor cases, such as a single vehicle striking a parked car with no injuries, the same fields become irrelevant and naturally absent. Similarly, certain attributes apply only under specialized conditions (e.g., hazardous material involvement, public-transport passenger counts), making flat structures inefficient and harder to interpret.

Another limitation of flat tables with respect to the analytical query is that they are not able to represent multiple entities, or repeated elements. In case of multiple cars for an accident, however, a simple flat format does not allow for them to be naturally presented as separate records under the same accident. This restriction can reduce analytical depth or, in some cases, render the system impractical for real-world use.

Integration challenges also emerge when attempting to merge flat-structured data with external systems. To consolidate every data field that each organization needed into a single table would build nothing but a single large table, increasing the table size and complexity, making both read and write operations more cumbersome. This leads to inefficiencies, reduced performance, and greater management difficulties in integrated environments.

3.2. DATA ENTITY TYPE CLASSIFICATION

This section classifies road incident data into entity types based on the reviewed frameworks and the challenges in capturing the full complexity of real-world cases. A common feature among these organizations is their focus on three major data factors: incident-related, vehicle-related, and person-related information. This paper also includes a fourth factor: hospital-related data, which represents injuries sustained as a result of the incident. To support this analysis, Figure 4 uses an Euler diagram to illustrate the complex relationships among four major data factors: persons, vehicles, incidents, and hospital data. This classification ensures that no redundant or irrelevant data is captured during the incident reporting process.

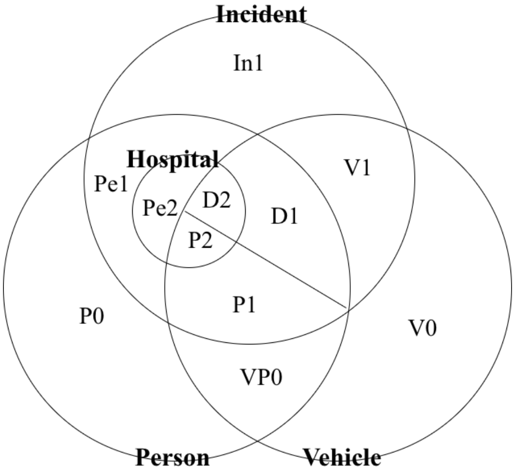


Figure 4. Union of factors illustrating covered road incident cases.

From Figure 4, we examine every possible combination among the four factors. Each region in the diagram represents a different data entity type, corresponding to various combinations of the factors involved, as follows:

People Involved in the Accident
D1 – Driver involved in the accident, in a vehicle, not hospitalized
P1 – Passenger involved in the accident, in a vehicle, not hospitalized
Pe1 – Pedestrian involved in the accident, not in a vehicle, not hospitalized

D2 – Driver involved in the accident, in a vehicle, hospitalized
P2 – Passenger involved in the accident in a vehicle, hospitalized
Pe2 – Pedestrian involved in the accident, not in a vehicle, hospitalized

People Not Involved in the Accident
P0 – Person not involved in the accident (e.g., bystander, witness)

Vehicles Involved or Not Involved in the Accident
V1 – Vehicle involved in the accident, no one inside
VP0 – Vehicle not involved in the accident, but with people inside
V0 – Vehicle not involved in the accident, no relevance to the incident
Other / Special Cases

In1 – Incident not involving a person or a vehicle (e.g., collapsed road, animal obstruction, infrastructure failure)

Figure 4, it can be seen that V0, VP0, and P0 are excluded from the scope of this study, as they represent normal conditions in which no incident has occurred. The table 1 also illustrate from Figure 4 to shows the different levels of specificity for each entity type at a single event.

Entity Type	No Incidence	Incidence							
		No Vehicle			Vehicle				
		No Person	Pedestrian		No Person	Driver		Passenger	
			Not Hospitalized	Hospitalized		Not Hospitalized	Hospitalized	Not Hospitalized	Hospitalized
P0	✓	-	-	-	-	-	-	-	-
VP0	✓	-	-	-	-	-	-	-	-
V0	✓	-	-	-	-	-	-	-	-
In1	-	✓	-	-	-	-	-	-	-
Pe1	-	-	✓	-	-	-	-	-	-
Pe2	-	-	-	✓	-	-	-	-	-
V1	-	-	-	-	✓	-	-	-	-
D1	-	-	-	-	-	✓	-	-	-
D2	-	-	-	-	-	-	✓	-	-
P1	-	-	-	-	-	-	-	✓	-
P2	-	-	-	-	-	-	-	-	✓

Table 1. Shows the different levels of specificity for each dataset
✓ means that the factor applies to the entity type,
While - means that the factor does not apply to the entity type.

Table 2 illustrates that each data domain records only the information relevant to the entities or conditions that actually occur in an incident. For example, if only one driver (Driver 1) is injured while the other (Driver 2) is not, both driver records (D1 and D2) would still exist under the same crash event, but only the injured driver would require hospital-related data. This demonstrates that certain data fields apply only to specific subsets of records, depending on the characteristics of the entities involved. Accordingly, Table 2 highlights that not all fields are mandatory in every case; rather, they should be governed by their own conditional requirements.

This reinforces the importance of a data structure that can represent variability across incidents, vehicles, and persons, ensuring that each element is recorded only when relevant.

Entity Type	Incidence	Vehicle	Person	Hospital
P0	-	-	✓	-
VP0	-	✓	✓	-
V0	-	✓	-	-
I0	✓	-	-	-
P1	✓	-	✓	-
P2	✓	-	✓	✓
V1	✓	✓	-	-
D2	✓	✓	✓	-
D1	✓	✓	✓	✓
Pa2	✓	✓	✓	-
Pa1	✓	✓	✓	✓

Table 2. lists all case analyzed based on Figure 4 and Table 1, and specifies the corresponding data fields needed for each data entity type.
✓ indicates that the data field is required for the specified entity type in the case, and
✗ indicates that the data field is not relevant or not required for that entity type in the case.

3.3 TOWARD MORE EFFICIENT AND FLEXIBLE DATA MODELS

To handle a wide range of real-world road incident cases that identify in previous section, from the ground to the complex ones with individual involvement, vehicle participation, hospitalization status, and also events that do not involve people or vehicles. The goal is to ensure flexibility, completeness, and readiness for diverse reporting situations, both common and rare. The solution lies in rethinking the structure of road incident databases. A well-designed model should allow:

1. One incident relates to multiple vehicles, people.
2. Separate but linked storage of vehicle, and person (to avoid redundancy).
3. Accurate modeling of unusual but common real cases (e.g., vehicles with no driver, non-vehicle incidents, multi-point impacts).
4. A flexible structure enables better performance during both data collection (faster, cleaner forms) and data analysis (more precise filtering and aggregation.

Figure 5 illustrates the proposed Data Vault 2.0 model of road safety data that is capable of supporting a complex case. The analysis begins when an incident occurs:

- The tables **hub_incident** and **sat_incident** are used to store the core incident information.
- If multiple vehicles are involved, the **link_incident_vehicle** table is used to associate those vehicles with the incident.
- In a case where a vehicle collides with a pedestrian, data about the pedestrian is stored in the **hub_person** and **sat_person** tables. The absence of a link between the person and **hub_vehicle** indicates that the individual was not inside a vehicle yet was involved in the same incident.
- When multiple people are involved in a single incident, such as in public transportation crashes, the model supports this complexity by linking **hub_incident** and **sat_incident** to multiple individuals through the **link_incident_person** table.
- Each individual situation can also be associated with a different hospital that enables

detailed and tracking of medical treatment.

Moreover, if data needs to be updated or modified, it can be done without affecting the relationships between the core factors, as the detailed information is stored in the sat tables.

4. DISCUSSION

In this section, we outline how the proposed Data Vault 2.0 model facilitates a better representation of road incidents from real-world data and overcomes the drawbacks that were raised in the previous sections. The results emphasize the model's capacity to accommodate a wide range of cases, minimize redundancy and enhance flexibility for data analysis and integration.

4.1 MODEL DESIGN AND EVALUATION

The proposed model is designed using the Data Vault 2.0 approach, which emphasizes flexibility, scalability, and the ability to accommodate evolving data requirements. The model incorporates four major domains(incidents, vehicles, persons, and hospitals) and defines their interrelationships through linking tables that naturally support many-to-many associations. This structure enables the model to store detailed information.

The evaluation focused on how the model satisfies the design criteria outlined in Section 3.3, especially its ability to support complex real-world cases. Figure 6 illustrates a highly complex incident used to test the model: a bus with 4 passengers colliding with a tricycle with one passenger, together with a pedestrian and a parked vehicle. In total, the event comprises 3 vehicles and 8 persons, each with different involvement types and hospitalization outcomes.

Then, Table 3 illustrates how the proposed model utilizes the appropriate combination of Hub, Link, and Satellite tables to capture each individual's data. For example, hospitalized individuals are associated with the hospital domain, multiple passengers can be linked to the same vehicle, and people outside vehicles (e.g., pedestrians) are recorded without requiring vehicle associations. The model successfully stores all these elements without structural conflict.

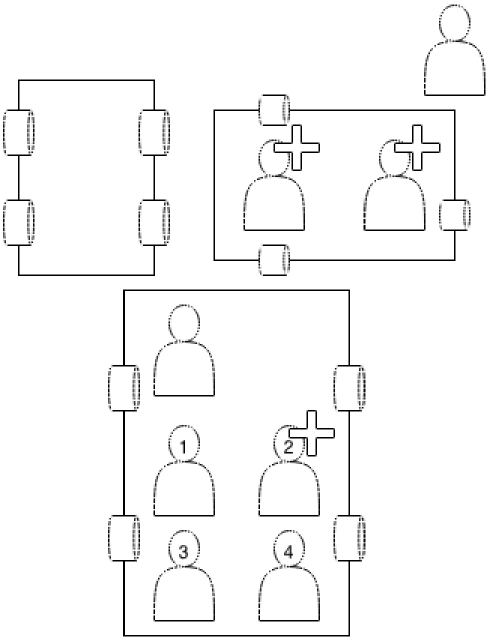


Figure 6. Example of a Complex Road Incident Scenario for Model Evaluation

Data case	Entity	**Hub**&Sat	**Hub**&Sat	**Hub**&Sat	**Link**&Sat	**Link**&Sat	Sat
	type	Incidence	Vehicle	Person	Person	Vehicle	Hospital
Bus driver	D1	✓	✓	✓	✓	✓	-
Bus Passenger no.1	P1	✓	✓	✓	✓	✓	-
Bus passenger no.2	P2	✓	✓	✓	✓	✓	✓
Bus passenger no.3	P1	✓	✓	✓	✓	✓	-
Bus passenger no.4	P1	✓	✓	✓	✓	✓	-
Tricycle driver	D2	✓	✓	✓	✓	✓	✓
Tricycle passenger	P2	✓	✓	✓	✓	✓	✓
Pedestrian	Pe1	✓	-	✓	✓	-	-
Parked vehicle	V1	✓	✓	-	-	✓	-

Table 3. Specifies the corresponding data fields needed in Scenario from Figure 6.
✓ data fields in the table is needed, while - is no need that table.

Therefore, if there is any vehicle or anyone who has been in the incident again, the personal data in **Hub_vehicle** or **Hub_person** can be reused. This methodological approach ensures that the proposed model is not only theoretically grounded but also practically tested against real-world conditions and data management requirements.

In addition, data reuse is possible in the architecture. When the vehicle and person appear in more than one incident, then their core record of the Vehicle (and Person) from **Hub_Vehicle**/**Hub_Person** can be reused with no duplication. Overall, this evaluation validates that the proposed model is theoretically sound and practically flexible for a diverse and complex set of incident scenarios encountered in actual road safety data systems.

4.2 COMPARISON OF DATA MODELS

In order to evaluate the performance and usability of a designed model, in all phases of this research, a comparison was made between traditional flat table based ones and the Data Vault 2.0 modeling techniques. Table 4 summarizes the evaluation.

The results indicate that while the flat model is easier to collect data from humans and sensors and read for end users, as it is simple, redundant free and easy to understand. However, the model becomes less efficient for analysis, integration and managing complex cases. Flat structure creates redundant data, limits modeling of many to one relationships and adds complexity when new fields are required.

Comparison Criteria for Road Safety Data Models	Flat model	Dimensional	Data Vault
		model	2.0 model
Easy to collect data from both human inputs and sensor-based sources	✓	✗	✗
Easy for humans to read and interpret	✓	✓	✗
Machine-readable and easy for systems to process	✓	✓	✓
Supports efficient data analysis	✗	✗	✓
Facilitates modification and extension of data fields	✗	✓	✓

Facilitates data integration	✗	✗	✓
Capable of handling complex data structures	✗	✗	✓
Prevents unnecessary data volume growth	✗	✗	✓

Table 4. Comparison Criteria for Road Safety Data Models
✓ means the model is effective and suitable for that specific use case,
while ✗ means the model is ineffective or unsuitable for that use case.

Before considering Data Vault 2.0, the dimensional model was also evaluated. However, this model was excluded from the analysis due to its limitations in handling complex relationships in road incident data. The dimensional model is built around a star schema. It causes problems with joins between tables at level 2 (non-central tables), making it difficult to represent many-to-many relationships. In road incidents, where multiple vehicles, persons, and other factors like pedestrians or hospital data are involved, such relationships are common. The inability of the dimensional model to naturally handle these complexities without introducing redundancy or losing key information led to its exclusion.

In contrast, the Data Vault 2.0 model provides stronger support for data analysis, system integration, complex scenario representation, and schema evolution, ensuring that data volume remains manageable and logically structured. Although it creates complexity during initial setup and training, but the long-term benefits in performance and adaptability outweigh these challenges. Therefore, the Data Vault 2.0 approach demonstrates clear benefits for road safety data management, especially when the multi-source integration, large-scale analytics, and high data variability.

5. CONCLUSION

This study examined current practices in road safety data collection and management and discussed the key challenges from existing systems that are unable to capture the complexity of real-world road incidents. A significant contribution of this work is the introduction of a refined classification of data entities. By considering the relationships between the major factors of road safety areas, which are people, vehicles, incidents, and hospital-related data, this provides a more accurate and structured way to represent complex accident scenarios, minimizing redundancy and ensuring that relevant data is recorded only when applicable.

To address the limitations of traditional data models, the study also proposes a data model that offers greater flexibility, scalability, and efficiency in managing diverse road safety data. The proposed model adopts the Data Vault 2.0 approach that separates core entities from their descriptive attributes, which can ensure that complex scenarios, such as multi-vehicle collisions, pedestrian involvement, and varying injury outcomes, are addressed and also properly represented without unnecessary data duplication. This approach enhances data integrity, improves analytical capabilities, and supports seamless integration with other data sources.

The evaluation of the proposed model demonstrates its ability to handle a wide range of data entity types, reduce redundancy, and provide meaningful insights for better road safety analysis. Moreover, this flexible model allows changing of data needs, and offers a robust solution for real-time data collection and long-term road safety planning. Future work may extend this model to operational systems, integrate real-time data sources, and assess performance in national-level deployments to further validate its applicability and impact.

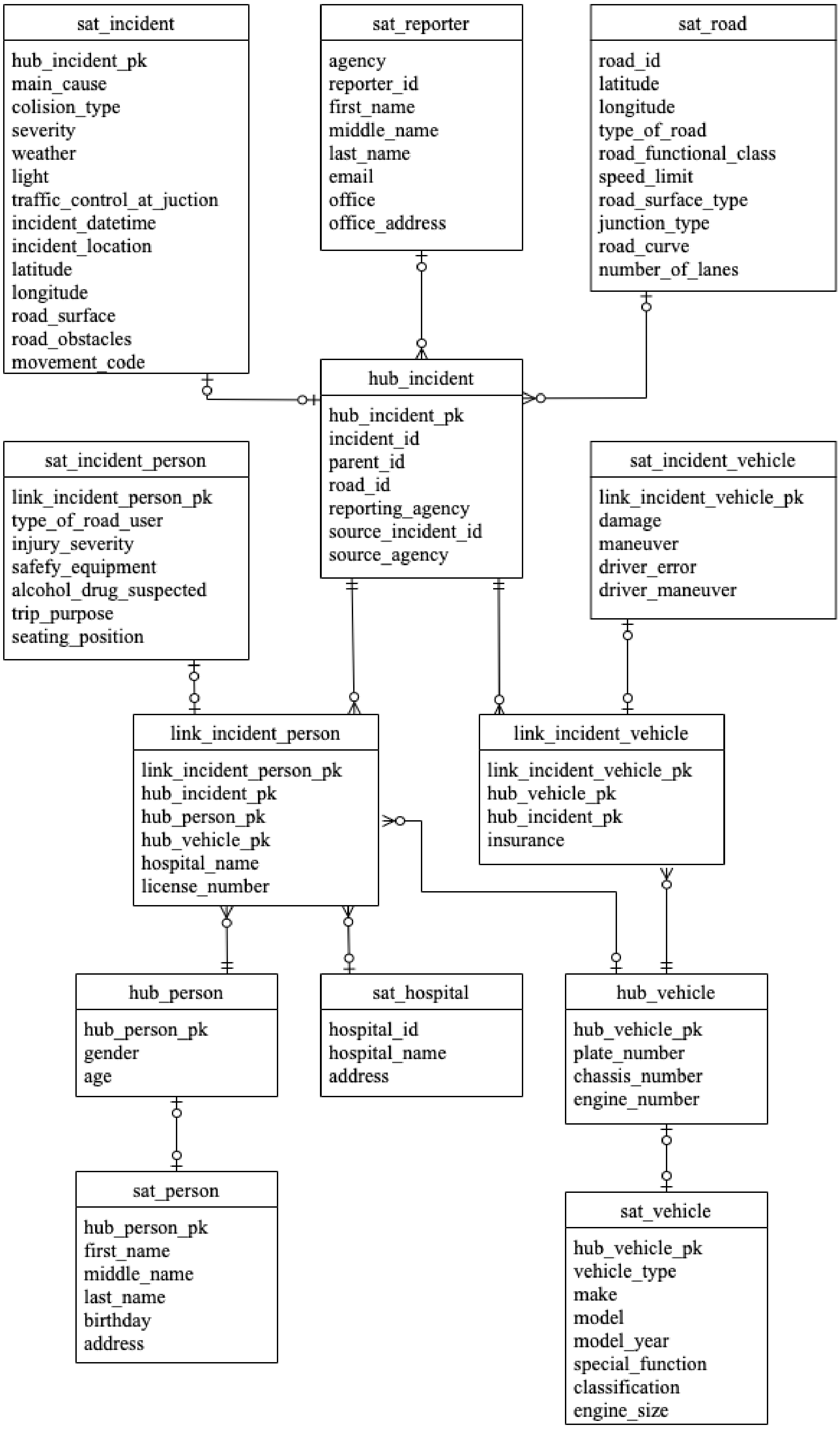


Figure 5. The proposed data model

REFERENCES

Al Aamri, A., Padmadas, S. S., Zhang, L.-C., & Al-Maniri, A. (2025). Improving the Efficiency and Quality of Road Crashes Data in the Sultanate of Oman: Evidence-Based Recommendations. Journal of Road Safety, 36(2), 63–74. https://doi.org/10.33492/JRS-D-25-1-2563701

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

World Bank. (2011) India: Improving Road Safety – Road Accident Data Management System (RADMS) in Tamil Nadu.

Peden, M., Scurfield, R., Sleet, D., Mohan, D., Hyder, A.A., Jarawan, E., Mathers, C. (2004) World report on road traffic injury prevention. World Health Organization, Geneva.

World Health Organization and The World Bank (2010) Data systems: a road safety manual for decision-makers and practitioners. World Health Organization, Geneva.

Asia-Pacific Telecommunity (APT) (2021) Final draft of APT report on traffic accident record and its analysis method's guidelines in Asia. Asia-Pacific Telecommunity.

Asian Development Bank (2025) CAREC Road Crash Investigation Manual. Asian Development Bank, Manila.

World Bank and Global Road Safety Facility (2018) DRIVER: The World Bank's sustainable solution for road crash data management. Global Road Safety Facility, Washington, D.C.

World Bank and Global Road Safety Facility (2019) Data for Road Incident Visualization, Evaluation and Reporting: Lowering the Barriers to Evidence-Based Road Safety Management in Resource-Constrained Countries. World Bank, Washington, D.C.

World Health Organization. (2021). Global plan for the Decade of Action for road safety 2021–2030. World Health Organization.

Kuppusamy, P. (n.d.) Data lake model to modern educational organizations.

Seah, B.K. and Selan, N.E. (2014) Design and implementation of data warehouse with data model using survey-based services data. In Proceedings of the 4th International Conference on Innovative Computing Technology (INTECH), pp. 58–64.

Nagm Aldeen, Y. (2020) Data warehouse – dimensional model vs normalized model, July 2020.

Komninos, G., Singh, D., and Dimaline, S. (2021) Design and build a data vault model in Amazon Redshift from a transactional database.

Yessad, L. and Labiod, A. (2016) Comparative study of data warehouses modeling approaches: Inmon, Kimball and Data Vault.

Linstedt, D., and Olschimke, M. (2015). Building a Scalable Data Warehouse with Data Vault 2.0. Morgan Kaufmann.

Chaturvedi, Bharat. (2025). Scalable Data Warehousing Using Data Vault 2.0 Design Pattern. International Journal of Computer Engineering & Technology, 16. 79-88. 10.34218/IJCET_16_03_007.

Rob, P., & Coronel, C. (2014). *Database Systems: Design, Implementation, & Management*. Cengage Learning.

Connolly, T., & Begg, C. (2015). *Database Systems: A Practical Approach to Design, Implementation, and Management*. Pearson.