International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 6, November- December 2022 ||

DOI: 10.15680/IJCTECE.2022.0506022

Distributed Cloud Data Lakes for Intelligent Transportation Data Integration

Sathiri Dhanaraj

Independent Researcher, India

dhanrajsathiri@gmail.com

ABSTRACT: Intelligent transportation systems rely on a patchwork of independent data providers and users, hampering holistic applications that improve road safety, increase efficient travel, and reduce carbon emissions. The sheer volume, velocity, and variety of data generated by these systems call for a distributed architecture that allows geographically close data users to share data, collaborate on analytics, and leverage machine learning and statistical modelling at scale for better decision-making. Distributed data lakes, built on elastic cloud-native primitives, allow automated data ingest from multiple providers, storage in purpose-built formats, and batch and real-time analytics. Core design decisions and environmental dependencies inform a target architecture that tackles the classification, ingest, and modeling of vehicular, user-deployed infrastructure, GBFS- and event-driven source data. A proof-of-concept validation using a remote region of Ontario, Canada, proposes specific Cloudflare Workers integration and extends earlier semantic mapping of LTE data to schema evolution.

The volume, velocity, and variety of data generated by intelligent transportation systems (ITS) can support a wide range of applications, including more efficient management of road safety, reduced travel times and vehicle emissions, and increased monetization of supplier data. However, independent data providers and users, such as the City of Toronto's traffic-management system, cannot provide a complete picture. Enabling collaboration on data and analytics is key to delivering truly intelligent system features, yet geographically close data users have traditionally relied on direct links.

KEYWORDS: Intelligent transportation systems, data integration, cloud computing, big data processing, machine learning, distributed systems, cyber-physical systems, data lakes, distributed cloud storage, metadata management, interoperability, knowledge representation, crosstalk.

I. INTRODUCTION

For more than a decade, Intelligent Transportation Systems (ITS) have been generating unprecedented volumes of temporal data. ITS applications largely reside in isolated data silos, making it difficult to cross-analyze data across multiple institutes and systems for more precise insights. Deploying IT resources in data centers is not a cost-effective approach since many ITS services and applications require high elasticity with workloads that have bursty workloads. The costs associated with data storage and processing in such scenarios can be excessively high. Transferring large volumes of ITS sensor data to public cloud services can incur excessively high egress costs. Deploying a cloud architecture for data analytics, data sharing, and other purposes that combines local resources at participating sites—without transferring data to a public cloud—will address these concerns. A distributed cloud architecture comprising a data lake at each participating site, together enabling close-to-cloud elasticity for analytics and sharing ITS data, is proposed. The architecture is supporting a proposal technology with Integrated Data Lake System (IdaLS) that focuses on ITS data lakes.

To bridge the key gaps and requirements, core requirements of distributed cloud data lakes are first identified and the entire data lake ecosystem is articulated. Remaining sections present underlying aspects of the architecture, including primary data sources and ingest, data modeling, and ontologies to enhance interoperability, processing and analytics based on lake house principles for combined batch and stream analytics, aspects related to governance and security, and proposed future directions. The work closes with a discussion of these aspects and remaining gaps.

1.1. Background and Significance

The landscape of Intelligent Transportation Systems (ITS) and Connected and Autonomous Vehicles (CAV) data is vast and varied. ITS data streams include multi-modal sensor data from vehicles, smart city traffic feeds (e.g., cameras, traffic lights, and radar), GPS traces of taxis, buses, and transit services, and mobile event logs. Moreover, CAV and

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 6, November- December 2022 ||

DOI: 10.15680/IJCTECE.2022.0506022

ITS research centers are also producing volumes of practical driving scenarios and some are using crowd-sourcing to enable real-time/near-real-time collection of mixed event data (e.g., accidents, road conditions, flooding, and traffic jams). Therefore, the need for efficient management remains an urgent challenge not only for these centers but also for autonomous driving research and development companies. A distributed data lake architecture with shared data can simplify the problem.

A distributed cloud data lake is a viable proposition for their functionality, cost-effectiveness, and accessibility. Such lakes have been proven useful in various use cases in several domains. However, no research has catalogued an extensive range of ITS data together with vehicle and smart city data and made a case for applying a distributed cloud data lake. Addressing this shortcoming provides an important research contribution and offers a basis for future cloud deployment.

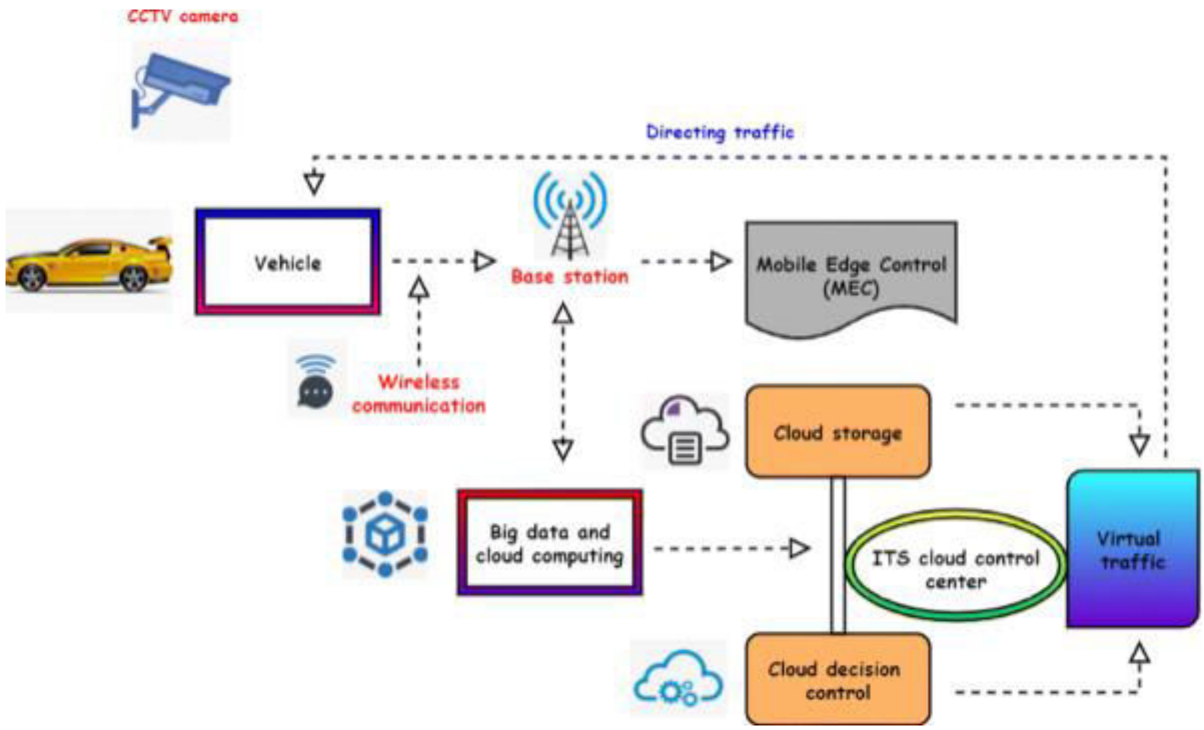


Fig 1: Big data analysis and cloud computing for smart transportation system integration

1.2. Research design

Distributing intelligent transportation system (ITS) data lakes across multiple clouds and geographic regions requires a design that considers their purpose, the data sources involved, how data is ingested and stored, the processing pipelines, the security and governance policies, and compliance with regulations. The objective is to provide a practical architecture, along with monitoring and maintenance procedures, that can be implemented or scaled to meet a specific region's analytics needs. Using a bottom-up approach, a foundation was established to model, integrate, analyze, and share the large volumes of ITS data generated by a metropolitan area.

Extensive data from LIDAR, video cameras, traffic signals, road sensors, GPS traces, and special events can be acquired, shared, and processed through an engine that collects the data at the city level and distributes it to multiple data lakes. Each of these lakes participates in a monitoring framework that ensures data accuracy and quality according to user-defined policies. Data from the lakes can be analyzed by batch and stream pipelines, machine learning model training, and real-time inference. While ITS data lakes may be small for smaller cities, they are nevertheless essential for ensuring data quality, governance, privacy, and security.

Equation 1: Per-source ingest rate

For a data source i:

- Sampling rate: f_i(samples/second = Hz)

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 6, November- December 2022 ||

DOI: 10.15680/IJCTECE.2022.0506022

- Payload size: s_i(bytes/sample)
- Number of producers: n_i(e.g., vehicles, sensors, cameras)

Step-by-step

1. One producer emits per second:

$$\text{bytes/sec per producer} = f_i \cdot s_i$$

2. With $n_i$producers:

$$r_i = n_i \cdot f_i \cdot s_i (\text{bytes/sec})$$

II. THEORETICAL FOUNDATIONS OF DISTRIBUTED CLOUD DATA LAKES

Distributed data lakes require a sound foundational architecture. First, the data lake layer structure, storage tiering, metadata, governance mechanisms, scalability, and interoperability are defined. Subsequently, cloud-based storage and compute components are examined, including object storage, retrieval and bulk-query primitives, elasticity and cost trade-offs, and data- and system-level security controls.

Intelligent transportation systems (ITS) must provide scalable, continuous, and timely services for planning, managing, and optimizing the road network. A distributed cloud data-lake architecture supports ITS by enabling organizations to acquire, integrate, process, analyze, and analyze vast amounts of structured and unstructured data from diverse sources running within a provincial or regional jurisdiction. The success of the data-lake paradigm within an ITS context will depend on the combination of a well-designed data-lake architecture and the effective use of distributed cloud storage and computation services.

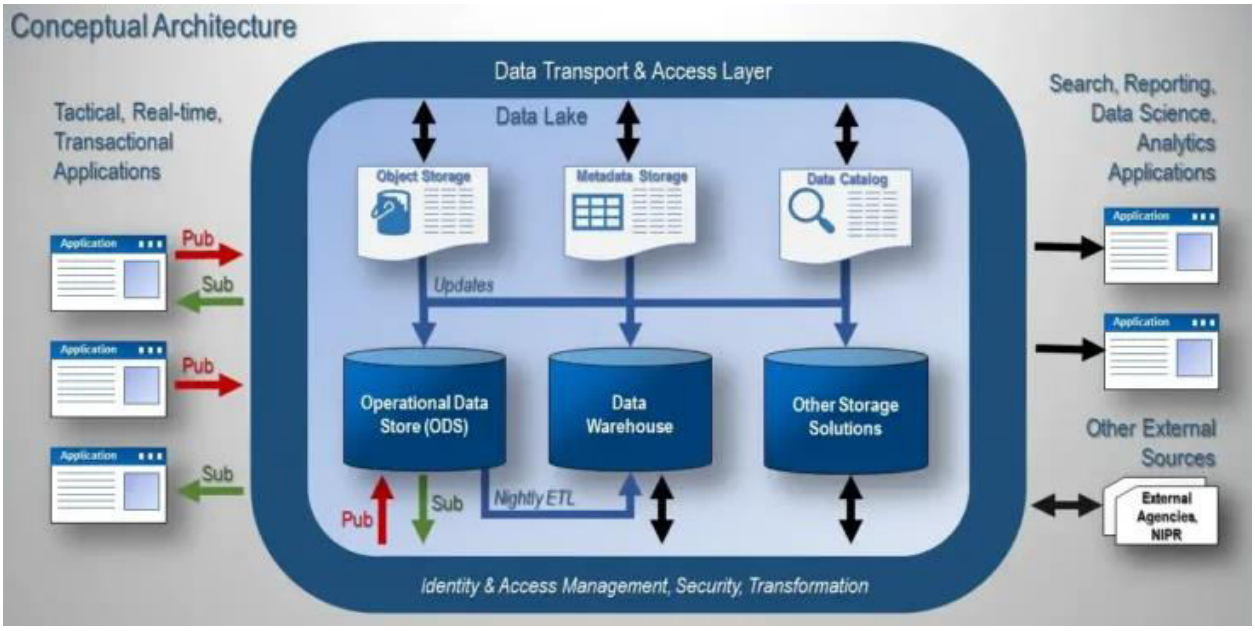


Fig 2: Theoretical Foundations of Distributed Cloud Data Lakes

2.1. Data Lake Architecture

Virtual data lakes consist of a controlled and governed collection of data lakes that are accessible in a multi-cloud environment. The layered architecture enables scaling out of workloads close to source and demand. A virtual pool integrates heterogeneous data systems into a single interoperable ecosystem, supporting analytics, machine learning, analytics, and Visualization-as-a-Service. This pool is composed of batch and streaming sensitive data lakes, a lake of identifiers and encyclopedic information, and supporting service-oriented systems. The virtual pool leverages cloud-based object storage engines and supporting compute paradigms, such as serverless computing and Dataflow. The underlying architectural building blocks support load-independent scalability, cost-efficient resource provisioning, and seamless integration of data and services from any source into a coherent ecosystem.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 6, November- December 2022 ||

DOI: 10.15680/IJCTECE.2022.0506022

Data lakes are designed to cope with the bulk of big-data collection models, in which the data is collected at a higher speed than it can be processed, and therefore, a representation of the data in the lake is often sufficient for traffic management preview and operation. The design choices do not address the operational model of a data lake that supports an integrated view of data for intelligent transport systems but rather provide support for a pool of data lakes. A data lake for batch and stream processing needs to be constructed as data arrives at source and is made available for exploration and machine learning.

Equation 2: Daily volume
Seconds per day $= 86400$.
1. Per day (bytes):

$$V_i = r_i \cdot 86400$$

2. Convert to GB (GiB-style):

$$V_i^{GB} = \frac{V_i}{1024^3}$$

$$V_{\text{total}} = \sum_i V_i$$

2.2. Cloud-Based Distributed Storage and Compute
Object storage is a natural fit for data lakes. Data is addressed via global identifiers that obscure underlying details such as physical location, thereby easing distribution and replication. Instead of file hierarchies, data is organized within a flat namespace that allows for varied collections of similar data. By representing logical metadata collections as key/value encoders, application-level challenges such cross-layer data discovery can be addressed. Many cloud providers and self-hosted solutions support S3-compatible APIs, enabling applications to be decoupled from the storage service while concurrently supporting native S3 and S3-compatible storage.
Data lakes also benefit from cloud compute primitives. Usage patterns often require either elastically provisioned concurrent processing power with minimal idle time or single-instance processes handling long-running jobs. Very high availability can be achieved at low cost via redundancy in the clients. Batch processing frameworks supporting the former case, such as Apache Spark and AWS Glue, are widely used and have established processing and enrichment patterns. Solutions for the latter case (e.g., Apache Flink, AWS Kinesis Data Streams, and Apache Beam) provide the necessary support for real-time analytics of rapidly arriving event and sensory data.
Another key quality is the role of cloud-native services in lowering maintenance and administrative effort. Utilization of these services requires accepting vendor lock-in and data-exit concerns (e.g., transfer costs, replicate-and-delete strategy) during the data-lake design process. Cloud-managed environments also help with security, providing primitives that simplify access control and encryption at rest and in transit. However, fulfilling a security and compliance policy for sensitive data remains essential; during design, these requirements must be applied to the data-lake architecture rather than to the services themselves.

III. DATA SOURCES AND INGEST FOR INTELLIGENT TRANSPORTATION SYSTEMS

A typical Intelligent Transportation System provides a variety of data sources for the distributed data lake architecture. These sources may generate data at a high frequency but with limited coverage, or they may cover the entire area of interest with a much lower sampling rate. The ingestion mechanisms that gather the data often trade off data quality for sampling frequency and ingestion cost. The aim of this section is not to delve deep into the details of the data ingested, but rather to provide a general overview of the data sources within an ITS and the quality considerations that need to be taken into account when designing the ingestion flows.

Vehicular Sensor Data: Vehicle tracking data captured by GPS or mobile network triangulation, sensor measurements recorded by buses and taxis, and vehicular traffic flow and speed aggregate statistics generated by roadside sensors are the most common types of vehicular sensor data. With respect to protection against privacy violation, sampling rates and aggregation strategies are often selected in such a way that the risk of revealing users' identity or their whereabouts at a certain time is mitigated. However, the investigative processing of vehicles tracking traces may require high frequency and deep data along the traces; for instance, the detection of the shortest path traveled by a certain taxi may require the corresponding 3D trace with information about the taxi's position, speed, and acceleration at certain time intervals.

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 6, November- December 2022 ||

DOI: 10.15680/IJCTECE.2022.0506022

Infrastructure and Mobility Data: Data collected by traffic cameras (multi-class object detection) and traffic signal settings, and the information about congested places inferred from multisensory event alerts are other important ITS data sources. Other vehicular mobility data provided by vehicles' users can usually be operated through the ITS without restrictions; for example, recorded historical GPS traces are not sensible to privacy violation and can be usually made available for any kind of usage. Finally, many events such as sport contests, concerts, and business fairs affect the mobility of people in and around the territory, and data articulation with external applications (social network APIs) can be modeled to process the effect of these events in a smarter way. A data integration challenge remains open with respect to the articulation of these heterogeneous data sources.

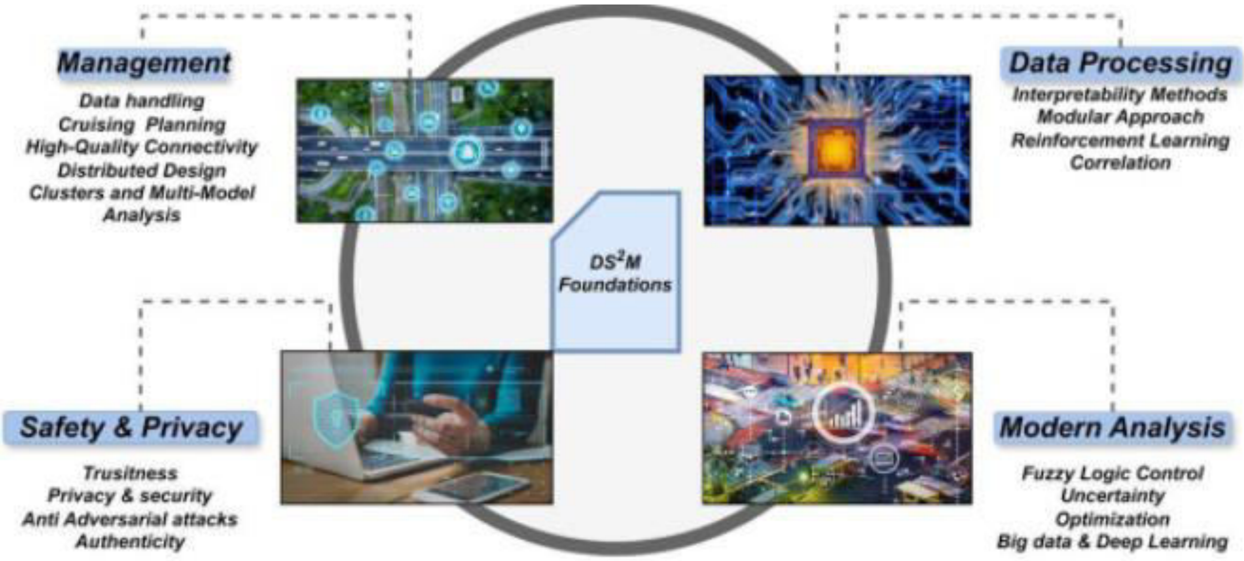


Fig 3: Data Sources and Ingest for Intelligent Transportation Systems

3.1. Vehicular Sensor Data

Vehicular sensors represent a substantial source of data for both conventional and Intelligent Transport System (ITS) applications. These sensors range from basic temperature, humidity, and light sensors to more advanced devices capable of monitoring vehicle speed and fuel consumption, collecting high-resolution GPS traces, and detecting curve radii. These data points are characterized by high sampling frequencies, ranging from several milliseconds for odometers in lightweight vehicles to a few minutes for temperature, humidity, and light data. Their overall quality also has the potential to be high, as vehicular sensors operate independently of weather, climate, and external characteristics, and are typically calibrated by the manufacturer.

The quality of the produced data is highly dependent on the service and mission conditions, being influenced on the one hand by the usage pattern of each vehicle and the number and type of users (for example, a low-time density of underground scenarios), and on the other hand by the age of the vehicle and the driving behavior of the driver-in vantage urban conditions of road infrastructure. Data produced by non-proximal driving do not pass close to the sampled site and are potentially subject to a certain number of errors. Apart from essential initial quality checks associated with pitch and roll angles, typical preprocessing procedures on these type of data are higher-level statistical filtering, aimed at replacing extreme values in light, humidity, or external temperature detected by only one faulty sensor by preprocessed statistics for the relevant site.

Equation 3: Data transfer time and throughput equations (why "ship everything to central cloud" hurts)

Let:

- Data to transfer: DGB/day
- Link throughput: BGbit/s

Step-by-step

1. Convert GB to Gbit: $D \cdot 8$
2. Time (seconds):

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 6, November- December 2022 ||

DOI: 10.15680/IJCTECE.2022.0506022

$$t = \frac{D \cdot 8}{B}$$

3. Convert to hours:

$$t_{\text{hours}} = \frac{D \cdot 8}{B \cdot 3600}$$

3.2. Infrastructure and Mobility Data

Traffic cameras capture vehicle images and help detect the vehicle count, speed, and classification both spatially and temporally. The processing can be performed on-premise, in a cloud environment, or even at the edge, depending on the latency and computing power required to meet the quality of service of the Intelligent Transportation Systems. Vehicle images can also be analyzed to detect events such as accidents or congestions, which usually need to be reported from external sources because of their short-duration nature. Traffic light data are usually aggregated on the centralized control system that allows the traffic lights to interact with one another and make decisions in real time. Road sensors (loops, inductors, etc.) create a high volume of data with a low velocity and can be integrated with data acquisition processes to include information such as debugging data. Anonymous GPS traces can also be used to identify traffic flow for large-scale movement of vehicles and buses. Event monitoring (e.g., accidents, landslides, parades, and races) is essential for both vehicular and pedestrian mobility and safety. The information can be provided externally (like accidents or road parades) or generated by the AIS or Virtual Transportation System. Several event notifications can be used jointly with smart phones and cloud infrastructure to take advantage of the real-time awareness phenomenon and provide warnings to users about the events in their proximity.

Connecting, validating, and aligning these data sources is more challenging than for vehicular sensors. Although the quality of traffic camera analysis may be refined with deep learning prediction, other information (like GPS traces) is sparse or faulty, and none of them go through a controlled quality check.

IV. DATA MODELING AND ONTOLOGIES FOR TRANSPORTATION DATA

Semantic structures enabling seamless data combination and analyses across Intelligent Transportation System and research project boundaries are identified as prerequisites for a cloud-based data lake serving disparate ITS applications and stakeholders. The resulting core semantic model, a collection of common schemas covering the essential data sources, and an extensible ontology dedicated to dynamic traffic events address these needs.

Interoperability is achieved through a combination of semantic enrichment, an ontology dedicated to ITS-AD events, and mapping strategies connecting to other external datasets. As a result, both integration and analytic reuse are facilitated. Combined with a provenance-dedicated ontology that is applicable to any ITS setting, these components provide a comprehensive interoperation foundation for the data lake and a basis for scoping future data bathing, distillation, and enrichment efforts.

Intelligent Transportation System (ITS) Data Lakes support large-scale data centers that consolidate data across heterogeneous sources and provide infrastructure support for improving situational awareness and decision making. A critical enabler within such data lakes is a semantic structure that supports seamless aggregation and analyses across project-specific clouds and research projects. Such a model simplifies data derivation and event de-identification for Privacy-Preserving Analytics and reduces the cost, time, and effort of future Data bathing, distilling, and mining.

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 6, November- December 2022 ||

DOI: 10.15680/IJCTECE.2022.0506022

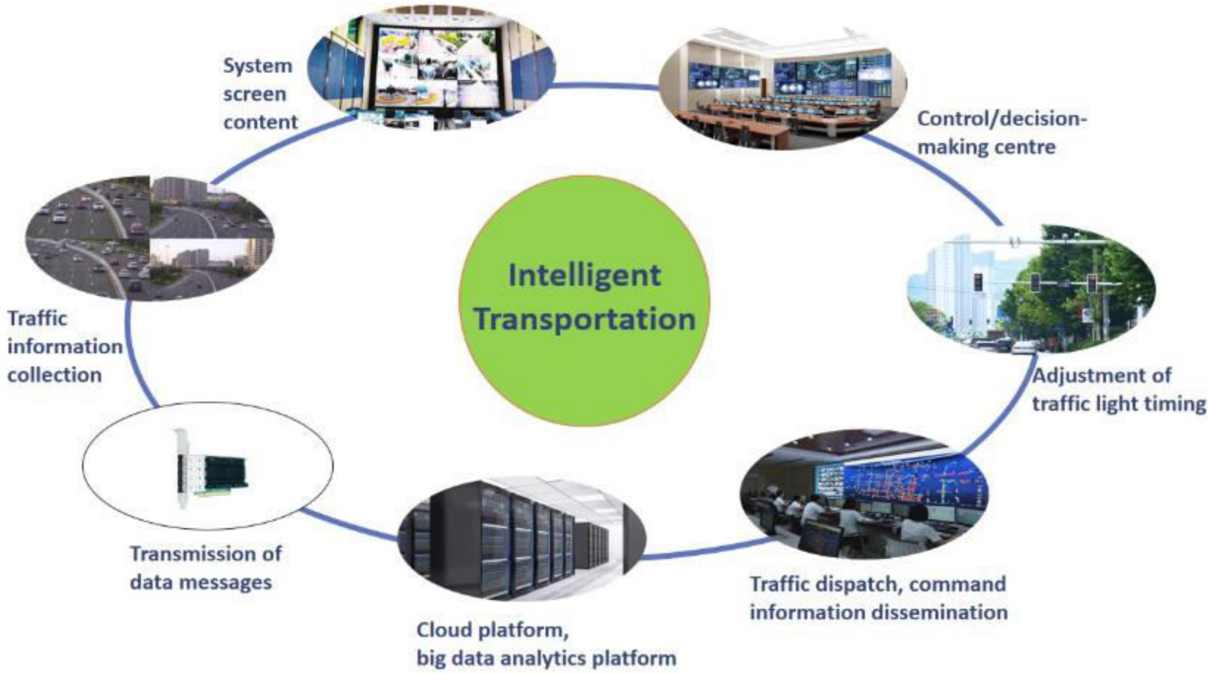


Fig 4: Intelligent transportation based on big data

4.1. Common Schemas and Interoperability

Intelligent Transportation Systems relies on a multitude of heterogeneous data sources but the majority of these sources are not standardized. Such lack of interoperability complicates data fusion and Analytics. To solve this limitation, a common schema has been proposed for the two primary data sources: vehicular sensors and infrastructure.

Supporting an Intelligent Transportation System based on the integration of heterogeneous data sources involves addressing four challenges: (1) Identifying and mapping the schemas of the various data sources, (2) Enabling semantic enrichment and interoperability using ontologies, (3) Ensuring compliance with aspects related to privacy, security, and regulatory requirements, and (4) Supporting the processing and Machine Learning Analytics. The first challenge deals with the majority of the data schemas provided by the sensors included in the vehicles and by infrastructure sensors, as described. These sensors produce different types of data streams that report the state of the tracked aspect at different time frequency, with different features, and with diverse degrees of quality. Some of these characteristics should be tackled at the data-integration level. An overall schema has been defined that covers the majority of the current and expected data sources, along with mappings for the most relevant medium- and high-level quality attributes.

In the absence of a common schema, the persistence and integration costs rise immensely as data-integration transformations must be recreated each time an analysis is to be performed. In addition to the basic integration schema, the integrations are mapped to the Common Data Model (CDM) standard proposed by the Open Data and Connected Vehicles Working Group and designed to support Different Use Cases in Data Managed Data Lakes. The mappings follow the general recommendations and best practices proposed for enabling the inclusion of third-party data in a Data Lake that do not follow the CDM schemas. Such mappings are updated as necessary depending on Changes in Data Lakes.

Equation 4: Centralized: ship all data out

1. Daily egress cost:

$$C_{\text{day}} = p \cdot D$$

2. Monthly (≈30 days):

$$C_{\text{month}} = 30 \cdot p \cdot D$$

If only a fraction α(e.g., 0.1 = 10%) must be shared out-of-region:

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 6, November- December 2022 ||

DOI: 10.15680/IJCTECE.2022.0506022

$$C_{\text{month,dist}} = 30 \cdot p \cdot (\alpha D)$$

4.2. Semantic Enrichment and Linkage

Ontologies provide a formal definition of the terms used in a domain. A vocabulary is reused from multiple standards and made available for external use. Data is linked to external datasets, e.g. the source, age and gender statistics from the New Zealand census, so that information is accessible for operational tasks. Provenance information about the generation of each dataset is added whenever possible, describing, for example, the geolocation from georeferencing or the de-identification of people through facial recognition software.

In order to organize and enable semantic interlinking of the available transportation data through semantic data lakes, a formal ontology model is being designed and implemented in a modular way. The Transportation Domain Ontology (TDO) serves as the core ontology for the domain. The TDO reuses concepts from existing vocabularies, particularly the W3C's Semantic Sensor Network ontology, and includes specific sensor types and models that are not included in it. A Sensor Data Ontology (SDO) incorporates domain-relevant statistical functions like mean and mode and is linked to a geography ontology, thus allowing a formal description of the set of sensors that generate data used in Intelligent Transportation System-based applications.

V. PROCESSING AND ANALYTICS IN DISTRIBUTED DATA LAKES

Processing and analytic pipelines define data transformation and storage processes before being made accessible for analysis. Batch and stream processing pipelines use frameworks and tools tuned to distinct requirements for data latency, fault tolerance, and quality. Analytic needs for machine learning model training and real-time inference then drive other design decisions.

Two processing frameworks coexist in the distributed cloud data lake. Data flow into Apache Spark through batch jobs or as streams via Apache Flink. Latency, scalability, throughput, and fault-tolerance requirements dictate which approach to use. For data that supports internal or external machine-learning services, preprocessing or monitoring pipelines are specific to the service needs. Key parts of the two types of ML pipelines share features, and monitoring pipelines can also be repurposed for batch training and enrichment.

5.1. Batch and Stream Processing

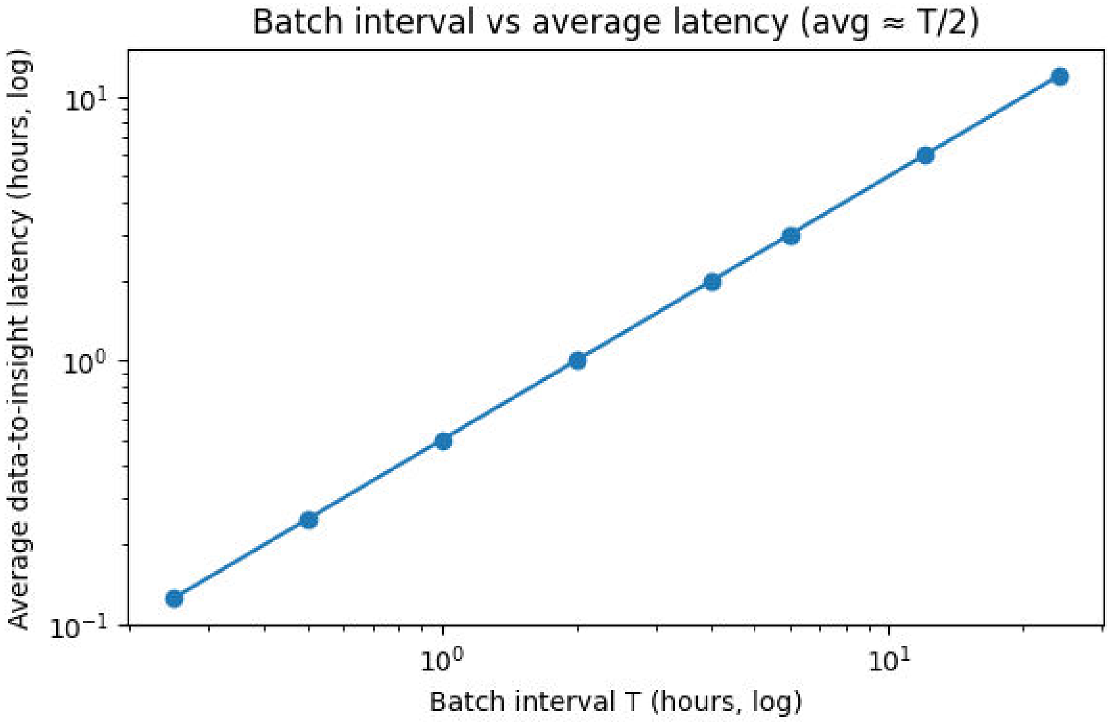
Intelligent Transportation Systems data are primarily processed via batch analytics, since most of the generated data can tolerate latencies measured in hours or days, and transactional support and fault-tolerance are not required. Such processing occurs via a mix of native technologies provided by the data lake (e.g., Amazon Athena) and external frameworks (e.g., Apache Spark or Apache Hadoop), and supports analytics models that generate actionable information, provide feedback for the data production process, or implicitly enhance data by addressing quality issues. However, a subset of sensors is also able to provide data feeds with lower latencies, where events must be captured and ruled instead. Those streams are ingested and processed in the lake as they arrive to provide action-oriented reports, such as alerts or control rules. The latency requirements of those applications is on the scale of seconds, but fault tolerance is also important. Hence, receptive streams are processed using appropriate frameworks (e.g., Apache Kafka with Kafka Streams, Apache Flink, Apache Spark Structured Streaming, etc.), that allow the definition of resilient end-to-end data flows.

Triggering is performed outside the lake, using external monitors. Whenever one or more incidents occur, the related data streams are processed through a dedicated pipeline that performs the required data enrichment — possibly with the help of external data sources — and outputs the enriched data to the lake, either in a totally different quality layer (e.g., alerts) or merged with the non-sensitive background flows already present in the lake. These background flows with lower security requirements are used to periodically retrain/update diffusion models and detect anomalous behaviours or trends, whose existence are reported to operators for further inspection. Furthermore, when classifying vehicle behaviours, a change-detection mechanism is employed for triggering the retraining of the model, based on a much lower latency.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 6, November- December 2022 ||

DOI: 10.15680/IJCTECE.2022.0506022



5.2. Machine Learning and Real-Time Inference

Machine learning is critical to developing value-added services for intelligent transportation systems. In the context of a cloud-based distributed cloud data lake, models can be trained on large, historical datasets and deployed directly into the lake for batch inference on incoming data streams. The processing framework can allocate resources as needed for inference and routing in accordance with the predicted outcomes. The incoming traffic can be routed to the most suitable machine learning feature computation according to the state of the analytics and the current traffic conditions, e.g. to a model predicting traffic congestion, the availability of alternative routes, or an incident detection model monitoring the traffic camera feeds.

Machine learning can be used at various levels in the data lakes of ITS applications. At the operational layer for real-time inference in low-latency use cases, at the operational enrichment layer for low-latency online inference of features of the incoming streams to be consumed at the analytical layer, and at the analytical layer for ready batch training of models on the historical and batch-computed streams. The features of the stream eligible for online machine learning should be stored in a feature store using a pattern similar to an operational data mart in order to alleviate the burden of recreating the features from raw or moderately-processed data directly in the model. The monitoring of the traffic state or the detection of incidents can also be additionally supported through video analytics using the video streams from the traffic cameras.

Equation 5: Stream: end-to-end latency (pipeline sum)

For a streaming pipeline:

- ingest/queue latency L_q
- processing latency L_p
- serialization/network latency L_n
- sink/index latency L_s

Total:

$$L_{\text{stream}} = L_q + L_p + L_n + L_s$$

VI. GOVERNANCE, SECURITY, AND COMPLIANCE

Adopting a distributed cloud data lake architecture does not eliminate the need for governance, policy enforcement, and transparency. Future-proofing these aspects of ITS support-multiplicity equals greater exposure and attack surface for

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 6, November- December 2022 ||

DOI: 10.15680/IJCTECE.2022.0506022

policy violations, effective and cost-efficient governance becomes ever more variable based on the characteristics of the operator's deployment of heterogeneous, geographically-dispersed clouds. Ensuring these opinionated best practices remain an enabler and not an inhibitor is critical.

Incorporating best-of-breed governance technologies, artistic effect and utility can be maximized however. Neither ITS operator infrastructure nor an ITS managed service typically, is sufficiently capitalised to develop a proprietary platform for cloud policy enforcement. Yet sovereign policing of third-party clouds can usefully be enabled by the emergence of standalone specialised offerings packaged as both security and compliance primitives. Governance becomes ever more a-technology and a-management business risk function than an operation centre for operational hygiene. Two such critical risks for ITS become data privacy and security made legally unable to be violated with isolation of sensitive data into high-cost storage anding without valid technical reasons operating in a shared cloud without encryption of sensitive data at rest and as it traverses the WAN. Deploying these governance policies maximally-effectively therefore means adoption of least-privilege principles across any technology that includes a notion of access control and of supporting active demand for tracing and deidentification. These risk policies require supporting native governance features in the underlying storage and compute platforms however.crypto-keys key management processes must themselves be obtained by correlation into a single enterprise domain key management service. However, while crash-resilience or assurance against spoiler-failure/hypervisor-apocalypse forbids shorting non-transferrable assets, the crash- or even spoiler-failure-resilience or assurance against hypervisor-apocalypse of cloud-laen virtual assets proxies or services for dynamically-crowding security-related governance policies naturally warrants deployment predominantly in pure-play cloud assets smeared as thinning of null-cover insurance against machine or security halo failures.

6.1. Access Control and Encryption

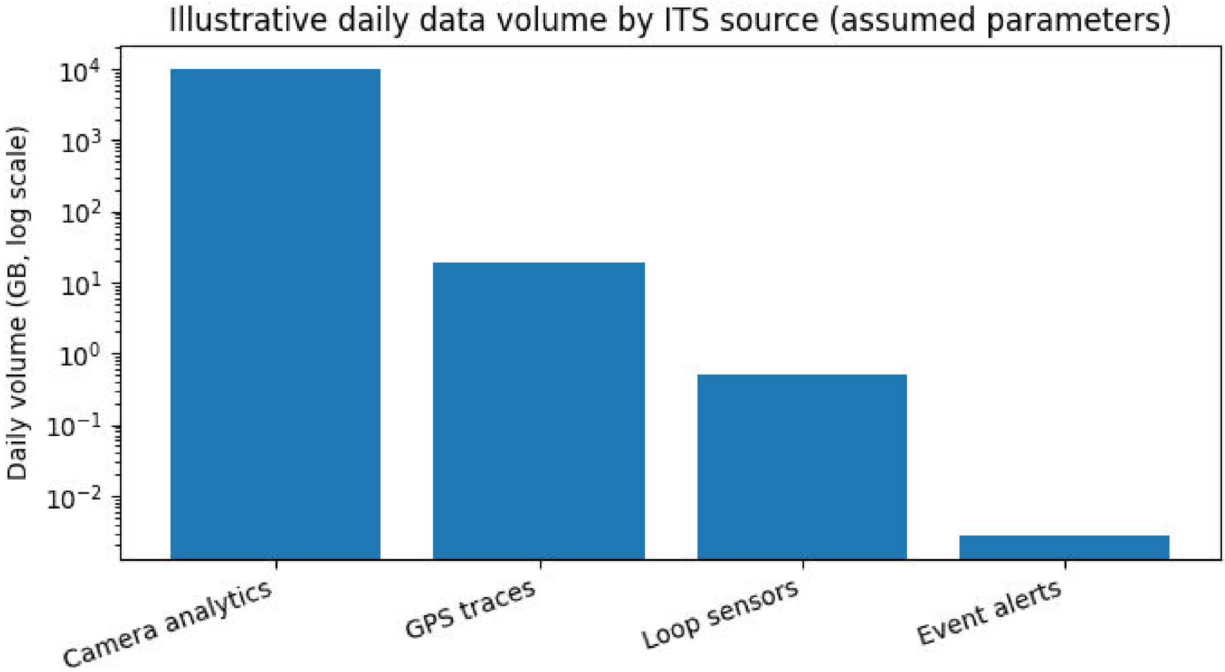
IT governance encompasses processes, decision-making structures, and accountability. In distributed cloud data lakes, governance covers security policy definition, risk assessment, and compliance with laws and regulations. The most relevant security policy areas concern access control and data encryption. The policy aimed at regulating access control to the data lake should define roles that grant access for specific data operations on specific data resources. Roles definition should follow the principle of least privilege: subjects should hold only those privileges that are essential for performing their jobs. Risks associated with the absence of these policies include data leakage and massive costs due to unauthorized usage.

The data composing a cloud-based distributed data lake are stored in S3 buckets, and Amazon enforces an access control mechanism based on access control lists. Data encryption, in turn, is a fundamental data protection technology that should always be activated. Sensitive data must be encrypted when written to disk and decrypted only in memory during computation. Data minimization is a privacy-by-design principle that requires sensitivity parameters to be defined by the data owner and the key management to be properly configured.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 6, November- December 2022 ||

DOI: 10.15680/IJCTECE.2022.0506022



6.2. Privacy, Anonymization, and Compliance

Intelligent Transportation Systems (ITS) collect and use large amounts of data capable of identifying human activities and movement. Such data is often sensitive, containing personal identification information (PII) and attribution. Collecting and using PII raises concerns concerning data privacy, warranting near-legislative compliance by data holders. Specifically, ITS data holders should adhere to every aspect of the Organisation for Economic Co-operation and Development (OECD) Privacy Guidelines by allowing individuals to remain anonymous and ensuring their privacy is not violated at all levels of information processing. ITS data holders should ensure a minimal amount of necessary data is collected and data is retained only for the necessary period of time. Moreover, audit trails should be available to verify compliance with any such rule or principle. Privacy-preserving technologies can ensure privacy protection and compliance with the privacy aspects of the OECD Privacy Guidelines.

Measures such as data minimization and prolonged anonymization of sensitive data help avoid privacy concerns. Data can be anonymized by removing PID while still allowing detection of trends without a specific attribution to individuals. Nevertheless, special attention must be paid to the data processing flow and possible combinations of various data sources. De-identification techniques can be applied to data streams affecting vehicle trajectory recognized as sensitive in conjunction to avoid the identification of a specific individual. When such techniques are used, relying expressly on data minimization, ITS data are no longer considered under GDPR and similar PI protection legislation. Data holders nevertheless remain responsible for establishing adequate data management policies to protect non-personal and proprietary data shared en-masse with the scientific community.

VII. CONCLUSION

Distributed Cloud Data Lakes for Intelligent Transportation Data Integration contribute a theoretical foundation for a distributed cloud-based data lake architecture enabling end-to-end data analytics and machine learning for Intelligent Transportation Systems. It explores distributed cloud data lakes fundamentally, offering a new braid of cloud computing, data lakes, and data-oriented applications. Intelligent Transportation Systems data serve as a reference application to illustrate the design of data sources, ingestion, storage, processing, analytics, and governance, validating the architecture with a selection of representative data characteristics and integration, processing, and compliance requirements.

Future directions include designing a complete data lake for Intelligent Transportation Systems and other applications and addressing the distributed nature of the data generation and consumption processes. Expanded support for Spark SQL-enabled batch and Spark Streaming-enabled short-processing-time operations (shorter than those supported by

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 6, November- December 2022 ||

DOI: 10.15680/IJCTECE.2022.0506022

Spark Streaming) is a candidate. Machine Learning and real-time inference models could be defined, monitored, and displayed directly from the lake. Within the broader cloud-development paradigm, integration with other cloud-based services—e.g., integrating analysis such as ML models into dashboards, or serverless-event-based architectures with Data Lake-as-a-service providence—would allow end users to upload, share, manage, and interact with their data in a single operation. Last, the architecture could serve as the foundation for a distributed intelligent-seminar environment, endorsed and published by the University Design Institute Consortium.

Architecture	Daily transfer time (h)	Monthly egress cost ($)
Centralized (ship all data to public cloud)	11.008361106117567	23778.059989213943
Distributed lakes (process locally; share only derived features)	0.11008361106117566	2377.8059989213943

Table : Illustrative centralized vs distributed cost/latency comparison

7.1. Future Directions

Future research should continue to develop Intelligent Transportation Systems data lakes based on an evolving set of widely accepted best practices. The following areas are especially ripe for exploration. First, there is significant scope for improvement in semantic enrichment, enabling data lakes to better fulfill their role in data integration. Although parsers are being developed for vehicular sensor data, new sources and types of data will always demand more tailored enrichment pipelines that capture all relevant features and relationships. More work is also needed to align existing vehicular data with common schemas, since most lakes contain vehicular traces yet offer no semantic mapping or best-practice guidelines. Such guidance should also extend to other emerging data modalities, such as metaverse, drone, or spaceborne data.

As these and other lakes mature, stakeholders in regional deployments will be called upon to address cross-lake interoperability challenges. Also critical is the definition of data governance policies for additional sensitive data sources, such as traffic cameras or mobility traces. Privacy-preserving techniques will need to be put on a more solid theoretical footing, supporting auditable compliance with laws and regulation. Further development of best-practice guidelines will also be key, outlining vulnerability and risk assessments, quality and provenance tracking, and distribution of responsibilities among stakeholders.

REFERENCES

[1] Avinash Reddy Segireddy. (2022). Terraform and Ansible in Building Resilient Cloud-Native Payment Architectures. International Journal of Intelligent Systems and Applications in Engineering, 10(3s), 444–455. Retrieved from https://www.ijisae.org/index.php/IJISAE/article/view/7905

[2] Goutham Kumar Sheelam, "Semiconductor Innovation for Edge AI: Enabling Ultra-Low Latency in Next-Gen Wireless Networks," International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), DOI: 10.17148/IJARCCE.2022.111258

[3] Mrabet, M., & Sliti, M. Integrating machine learning for the sustainable development of smart cities. Frontiers in Sustainable Cities, 6.

[4] Davuluri, P. N. (2020). Improving Data Quality and Lineage in Regulated Financial Data Platforms. Finance and Economics, 1(1), 1-14.

[5] Li, H., Wei, H., Zhao, W., & Zheng, X. Research on geographic information data circulation supports the construction of digital China. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLVIII-1/W2-, 97–104.

[6] Rongali, S. K. (2020). Predictive Modeling and Machine Learning Frameworks for Early Disease Detection in Healthcare Data Systems. Current Research in Public Health, 1(1), 1-15.

[7] Armbrust, M., Das, T., Davidson, A., Ghodsi, A., Or, A., Rosen, J., Stoica, I., Wendell, P., Xin, R., & Zaharia, M. (2021). Delta Lake: High-performance ACID table storage over cloud object stores. Proceedings of the VLDB Endowment, 13(12), 3411–3424.

[8] Inala, R. Advancing Group Insurance Solutions Through Ai-Enhanced Technology Architectures And Big Data Insights.

[9] Kotlinski, M., & Calkowska, J. K. (2022). U-space and UTM deployment as an opportunity for more complex UAV operations including UAV medical transport. Journal of Intelligent & Robotic Systems, 106, 12. https://doi.org/10.1007/s10846-022-01681-6

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 6, November- December 2022 ||

DOI: 10.15680/IJCTECE.2022.0506022

[10] Chava, K., Chakilam, C., & Recharla, M. (2021). Machine Learning Models for Early Disease Detection: A Big Data Approach to Personalized Healthcare. International Journal of Engineering and Computer Science, 10(12), 25709–25730. https://doi.org/10.18535/ijecs.v10i12.4678
[11] Kalisetty, S., Vankayalapati, R. K., Reddy, L., Sondinti, K., & Valiki, S. (2022). AI-Native Cloud Platforms: Redefining Scalability and Flexibility in Artificial Intelligence Workflows. Linguistic and Philosophical Investigations, 21(1), 1-15.
[12] Sriram, H. K. (2022). Advancements in Credit Score Analytics using Deep Learning and Predictive Modeling Techniques. Available at SSRN 5255128.
[13] Bifet, A., & Gavaldà, R. (2007). Learning from time-changing data with adaptive windowing. Proceedings of the 2007 SIAM International Conference on Data Mining, 443–448.
[14] Muthusamy, S., Kannan, S., Lee, M., Sanjairaj, V., Lu, W. F., Fuh, J. Y., ... & Cao, T. (2021). Cover Image, Volume 118, Number 8, August 2021. Biotechnology and Bioengineering, 118(8), i-i.
[15] Gondhi, P. K. FinTech cloud-based data lakes: Performance, governance, and scalability. Journal of Computer Science and Technology Studies, 7(2), 1–12.
[16] Vadisetty, R., Polamarasetti, A., Guntupalli, R., Raghunath, V., Jyothi, V. K., & Kudithipudi, K. (2021). Privacy-Preserving Gen AI in Multi-Tenant Cloud Environments. Sateesh kumar and Raghunath, Vedaprada and Jyothi, Vinaya Kumar and Kudithipudi, Karthik, Privacy-Preserving Gen AI in Multi-Tenant Cloud Environments (January 20, 2021).
[17] Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile Networks and Applications, 19(2), 171–209.
[18] Dwaraka Nath Kummari. (2022). Fiscal Policy Simulation Using AI And Big Data: Improving Government Financial Planning. Kurdish Studies, 10(2), 934–945. https://doi.org/10.53555/ks.v10i2.3855
[19] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
[20] Gadi, A. L. The Role of Digital Twins in Automotive R&D for Rapid Prototyping and System Integration.
[21] Das, T., Zhu, A., Li, S., Narayanamurthy, S., & Bhat, P. (2013). Distributed and fault-tolerant streaming computation in Spark. Proceedings of the ACM Symposium on Cloud Computing, 1–12.
[22] Siva Hemanth Kolla. (2022). Knowledge Retrieval Systems for Enterprise Service Environments. International Journal of Intelligent Systems and Applications in Engineering, 10(3s), 495–506. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/8037
[23] Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. Communications of the ACM, 51(1), 107–113.
[24] Paleti, S. (2022). Financial Innovation through AI and Data Engineering: Rethinking Risk and Compliance in the Banking Industry. Available at SSRN 5250726.
[25] DeCandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Vosshall, P., & Vogels, W. (2007). Dynamo: Amazon's highly available key-value store. Proceedings of the 21st ACM Symposium on Operating Systems Principles, 205–220.
[26] Sriram, H. K., ADUSUPALLI, B., & Malempati, M. (2021). Revolutionizing Risk Assessment and Financial Ecosystems with Smart Automation, Secure Digital Solutions, and Advanced Analytical Frameworks.
[27] Dwork, C. (2008). Differential privacy: A survey of results. Proceedings of the 5th International Conference on Theory and Applications of Models of Computation, 1–19.
[28] Varri, D. B. S. (2021). Cloud-Native Security Architecture for Hybrid Healthcare Infrastructure. Available at SSRN 5785982.
[29] Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering, 19(1), 1–16.
[30] Dwaraka Nath Kummari,. (2022). Machine Learning Approaches to Real-Time Quality Control in Automotive Assembly Lines. Mathematical Statistician and Engineering Applications, 71(4), 16801–16820. Retrieved from https://philstat.org/index.php/MSEA/article/view/2972
[31] Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). "Counting your customers" the easy way: An alternative to the Pareto/NBD model. Marketing Science, 24(2), 275–284.
[32] Inala, R. (2022). Engineering Data Products for Investment Analytics: The Role of Product Master Data and Scalable Big Data Solutions. International Journal of Scientific Research and Modern Technology, 155-171.
[33] Davuluri, P. N. (2020). Improving Data Quality and Lineage in Regulated Financial Data Platforms. Finance and Economics, 1(1), 1-14.
[34] Kalisetty, S., & Ganti, V. K. A. T. (2019). Transforming the Retail Landscape: Srinivas's Vision for Integrating Advanced Technologies in Supply Chain Efficiency and Customer Experience. Online Journal of Materials Science, 1, 1254.
[35] Ghemawat, S., Gobioff, H., & Leung, S. T. (2003). The Google file system. Proceedings of the 19th ACM Symposium on Operating Systems Principles, 29–43.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 6, November- December 2022 ||

DOI: 10.15680/IJCTECE.2022.0506022

[36] Varri, D. B. S. (2022). A Framework for Cloud-Integrated Database Hardening in Hybrid AWS-Azure Environments: Security Posture Automation Through Wiz-Driven Insights. International Journal of Scientific Research and Modern Technology, 1(12), 216-226.
[37] Yandamuri, U. S. (2021). A Comparative Study of Traditional Reporting Systems versus Real-Time Analytics Dashboards in Enterprise Operations. Universal Journal of Business and Management, 1(1), 1–13. Retrieved from https://www.scipublications.com/journal/index.php/ujbm/article/view/1357
[38] Gottimukkala, V. R. R. (2022). Licensing Innovation in the Financial Messaging Ecosystem: Business Models and Global Compliance Impact. International Journal of Scientific Research and Modern Technology, 1(12), 177-186.
[39] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.
[40] Vadisetty, R., Polamarasetti, A., Guntupalli, R., Raghunath, V., Jyothi, V. K., & Kudithipudi, K. (2022). AI-Driven Cybersecurity: Enhancing Cloud Security with Machine Learning and AI Agents. Sateesh kumar and Raghunath, Vedaprada and Jyothi, Vinaya Kumar and Kudithipudi, Karthik, AI-Driven Cybersecurity: Enhancing Cloud Security with Machine Learning and AI Agents (February 07, 2022).
[41] Hellerstein, J. M., Haas, P. J., & Wang, H. J. (1997). Online aggregation. Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data, 171–182.
[42] Garapati, R. S. (2022). Web-Centric Cloud Framework for Real-Time Monitoring and Risk Prediction in Clinical Trials Using Machine Learning. Current Research in Public Health, 2, 1346.
[43] Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. Proceedings of the 2008 IEEE International Conference on Data Mining, 263–272.
[44] Amistapuram, K. (2022). Fraud Detection and Risk Modeling in Insurance: Early Adoption of Machine Learning in Claims Processing. Available at SSRN 5741982.
[45] Davuluri, P. S. L. N. (2021). Event-Driven Compliance Systems: Modernizing Financial Crime Detection Without Machine Intelligence. Journal of International Crisis and Risk Communication Research , 339–354. https://doi.org/10.63278/jicrcr.vi.3636
[46] Meda, R. (2022). Integrating Edge AI in Smart Factories: A Case Study from the Paint Manufacturing Industry. International Journal of Science and Research (IJSR), 1473-1489.
[47] Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. Communications of the ACM, 57(7), 86–94.
[48] Segireddy, A. R. (2020). Cloud Migration Strategies for High-Volume Financial Messaging Systems.
[49] Meda, R. Enabling Sustainable Manufacturing Through AI-Optimized Supply Chains.
[50] Amistapuram, K. (2021). Digital Transformation in Insurance: Migrating Enterprise Policy Systems to .NET Core. Universal Journal of Computer Sciences and Communications, 1(1), 1–17.
[51] Kleppmann, M. (2017). Designing data-intensive applications. O'Reilly Media.
[52] Nagabhyru, K. C. (2022). Bridging Traditional ETL Pipelines with AI Enhanced Data Workflows: Foundations of Intelligent Automation in Data Engineering. Available at SSRN 5505199.
[53] Lahiri, M., & Venkatasubramanian, S. (2013). Robust record linkage. Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, 101–112.
[54] Rongali, S. K. (2021). Cloud-Native API-Led Integration Using MuleSoft and .NET for Scalable Healthcare Interoperability. Available at SSRN 5814563.
[55] Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). Mining of massive datasets (2nd ed.). Cambridge University Press.
[56] Rongali, S. K. (2022). AI-Driven Automation in Healthcare Claims and EHR Processing Using MuleSoft and Machine Learning Pipelines. Available at SSRN 5763022.
[57] Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing, 7(1), 76–80.
[58] Choudhary, V., Kartik, & Bala, N. Cloud-based data lake. International Conference on Artificial Intelligence and Quantum Computation-Based Sensor Application (ICAIQSA), 1–5.
[59] Lin, J., Kolcz, A., & Szymanski, B. K. (2012). Large-scale machine learning at Twitter. Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, 793–804.
[60] Sheelam, G. K. Power-Efficient Semiconductors for AI at the Edge: Enabling Scalable Intelligence in Wireless Systems. International Journal of Innovative Research in Electrical, Elec-tronics, Instrumentation and Control Engineering (IJIREEICE), DOI, 10.
[61] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 6, November- December 2022 ||

DOI: 10.15680/IJCTECE.2022.0506022

[62] Vadisetty, R., Polamarasetti, A., Guntupalli, R., Rongali, S. K., Raghunath, V., Jyothi, V. K., & Kudithipudi, K. (2021). Legal and Ethical Considerations for Hosting GenAI on the Cloud. International Journal of AI, BigData, Computational and Management Studies, 2(2), 28-34.
[63] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Proceedings of the International Conference on Learning Representations, 1–12.
[64] Ramesh Inala. (2022). Cross-Domain MDM Integration Using AI-Driven Data Governance: A Case Study In Financial Technology Architecture. Migration Letters, 19(2), 280–304. Retrieved from https://migrationletters.com/index.php/ml/article/view/11982
[65] Montoya, D. Y., Neto, A. M., & da Silva, A. S. (2016). A survey of entity resolution in big data. Journal of Big Data, 3(1), 1–22.
[66] Aitha, A. R. (2021). Optimizing Data Warehousing for Large Scale Policy Management Using Advanced ETL Frameworks.
[67] Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster computing with working sets. Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing, 1–7.
[68] Varri, D. B. S. (2022). AI-Driven Risk Assessment and Compliance Automation in Multi-Cloud Environments. Available at SSRN 5774924.
[69] Meda, R. (2021). Digital Infrastructure for Predictive Inventory Management in Retail Using Machine Learning. International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), DOI, 10.
[70] Segireddy, A. R. (2021). Containerization and Microservices in Payment Systems: A Study of Kubernetes and Docker in Financial Applications. Universal Journal of Business and Management, 1(1), 1–17.
[71] Zhai, C., & Massung, S. (2016). Text data management and analysis: A practical introduction to information retrieval and text mining. ACM & Morgan Claypool.
[72] Davuluri, P. N. (2020). Event-Driven Architectures for Real-Time Regulatory Monitoring in Global Banking.
[73] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135–146.
[74] Keerthi Amistapuram , "Energy-Efficient System Design for High-Volume Insurance Applications in Cloud-Native Environments," International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering (IJIREEICE), DOI 10.17148/IJIREEICE.2020.81209
[75] Goutham Kumar Sheelam. (2022). Reconfigurable Semiconductor Architectures For AI-Enhanced Wireless Communication Networks. Kurdish Studies, 10(2), 1027–1040. https://doi.org/10.53555/ks.v10i2.3867
[76] Kothapalli Sondinti, L. R., & Syed, S. (2022). The Impact of Instant Credit Card Issuance and Personalized Financial Solutions on Enhancing Customer Experience in the Digital Banking Era. Universal Journal of Finance and Economics, 1(1), 1223. Retrieved from https://www.scipublications.com/journal/index.php/ujfe/article/view/1223
[77] Gottimukkala, V. R. R. (2021). Digital Signal Processing Challenges in Financial Messaging Systems: Case Studies in High-Volume SWIFT Flows.
[78] Bhasin, H., & Bhatia, P. (2020). Clickstream data mining for web analytics and customer behavior modeling: A review. ACM Computing Surveys, 53(6), 1–34.
[79] Kolla, S. H. (2021). Rule-Based Automation for IT Service Management Workflows. Online Journal of Engineering Sciences, 1(1), 1–14. Retrieved from https://www.scipublications.com/journal/index.php/ojes/article/view/1360
[80] Uday Surendra Yandamuri. (2022). Cloud-Based Data Integration Architectures for Scalable Enterprise Analytics. International Journal of Intelligent Systems and Applications in Engineering, 10(3s), 472–483. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/8005
[81] Abedjan, Z., Golab, L., & Naumann, F. (2016). Profiling relational data: A survey. The VLDB Journal, 24(4), 557–581.
[82] Yandamuri, U. S. (2022). Big Data Pipelines for Cross-Domain Decision Support: A Cloud-Centric Approach. International Journal of Scientific Research and Modern Technology, 1(12), 227–237. https://doi.org/10.38124/ijsrmt.v1i12.1111
[83] Dwaraka Nath Kummari. (2022). AI-Driven Audit Frameworks For Enhancing Compliance In Modern Manufacturing Systems. Migration Letters, 19(S8), 2150–2177. Retrieved from https://migrationletters.com/index.php/ml/article/view/11912
[84] Davuluri, P. N. Event-Driven Compliance Systems: Modernizing Financial Crime Detection Without Machine Intelligence.
[85] Baesens, B., Van Vlasselaer, V., & Verbeke, W. (2021). Fraud analytics using descriptive, predictive, and social network techniques: A guide to data science for fraud detection (2nd ed.). Wiley.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 5, Issue 6, November- December 2022 ||

DOI: 10.15680/IJCTECE.2022.0506022

[86] Avinash Reddy Aitha. (2022). Deep Neural Networks for Property Risk Prediction Leveraging Aerial and Satellite Imaging. International Journal of Communication Networks and Information Security (IJCNIS), 14(3), 1308–1318. Retrieved from https://www.ijcnis.org/index.php/ijcnis/article/view/8609

[87] Buccella, A., Cechich, A., Saurin, F., Montenegro, A., Rodríguez, A., & Muñoz, A. A context-based perspective on frost analysis in reuse-oriented big data-system developments. Information, 15(11), 661. https://doi.org/10.3390/info15110661

[89] Garapati, R. S. (2022). AI-Augmented Virtual Health Assistant: A Web-Based Solution for Personalized Medication Management and Patient Engagement. Available at SSRN 5639650.

[90] Gottimukkala, V. R. R. (2020). Energy-Efficient Design Patterns for Large-Scale Banking Applications Deployed on AWS Cloud. power, 9(12).

[91] Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable machine learning in healthcare. Proceedings of the ACM Conference on Health, Informatics, and Data Science, 1–10.

[92] Aitha, A. R. (2022). Cloud Native ETL Pipelines for Real Time Claims Processing in Large Scale Insurers. Available at SSRN 5532601.

[93] Bores, J., Meyer, H., Underwood, E., Sirychenko, M., Langhout, W., von Döhren, P., et al. Review and synthesis of best practices in governance and land-use policies to implement TEN-N. ARPHA Preprints. https://doi.org/10.3897/arphapreprints.e139236

[94] Kolla, S. K. (2021). Architectural Frameworks for Large-Scale Electronic Health Record Data Platforms. Current Research in Public Health, 1(1), 1–19. Retrieved from https://www.scipublications.com/journal/index.php/crph/article/view/1372

[94] Akanfe, O. A. (2022). Advancing digital financial inclusion: Data privacy, regulatory compliance, and cross-country cultural values in digital payment systems use (Doctoral dissertation, The University of Texas at San Antonio).

[95] Maguluri, K. K., Pandugula, C., Kalisetty, S., & Mallesham, G. (2022). Advancing Pain Medicine with AI and Neural Networks: Predictive Analytics and Personalized Treatment Plans for Chronic and Acute Pain Managements. Journal of Artificial Intelligence and Big Data, 2(1), 112-126.

[96] Ahsan, M. S., & Pathan, A. K. A comprehensive survey on the requirements, applications, and future challenges for access control models in IoT: The state of the art. IoT, 6(1), 9. https://doi.org/10.3390/iot6010009