SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

A Governance-Centric Lakehouse Architecture For Ai-Ready Public Sector Decision Intelligence Using Medallion Data Engineering Patterns

Pankaj Sahani

Purdue University, West Lafayette IN sahani1987@gmail.com

Abstract

Artificial Intelligence (AI) is being used in public sector organizations to improve policymaking, service delivery, and evidence-based decision-making. Broadly, however, separate data ecosystems, poor data quality, a lack of interoperability, and inadequate governance processes remain obstacles to creating reliable, AI-enabled data analytics environments. In this paper, the researcher proposes a roadmap for the data engineering lifecycle, including governance principles, through a Governance-Centric Lakehouse Architecture for the AI-Ready Public Sector Decision Intelligence. The suggested framework builds upon bronze, silver, and gold data layers for a sequentially enhanced level of data quality, standardization, traceability, analysis, and governance aspects of data stewardship, security, privacy, compliance, and accountability for data use in AI. Additionally, the architecture includes metadata management, data lineage, and access control restrictions based on policies, thereby enabling transparent and explainable AI applications in public-sector settings. The framework integrates contemporary lakehouse features with governance needs, creating a scalable data environment that facilitates the development and utilization of trusted data assets. The framework's combination of modern lakehouse features and governance requirements enables the building and leveraging of trusted data assets to support predictive analytics, decision intelligence, and strategic planning. This study highlights a scalable and interoperable blueprint for governments seeking to rapidly digitize their operations without compromising regulatory requirements, operational efficiency, or ethical AI use in a data-informed government sphere.

Keywords: Data Lakehouse, Data Governance, Medallion Architecture, Decision Intelligence, Artificial Intelligence.

I. INTRODUCTION

To meet the growth in the volume and frequency of digital transformation initiatives, public sector organizations have changed drastically in how they manage their data. Around the globe, nations are employing data-driven approaches to tackle pressing problems, improve public-sector processes, and increasingly empower their citizens through digital services [1, 2]. With all the advancements in e-Governance, smart infrastructure, cloud computing, Internet of Things (IoT) devices, and digital public services, the volume, scale, and speed of Government data have increased by leaps and bounds [2]. The information comes from many diverse industries, from health care to tax, education to transport, environmental monitoring to social safety, that are ripe for high-level analysis and decision support with the help of Artificial Intelligence [3]. Data quality, data availability, and data governance now play a significant role in enabling public institutions to move from describing to predicting/prescribing their decisions. It is well accepted that data engineering frameworks are a fundamental layer that enables the efficient collection, analytics, integration, transformation, and robust management of data in various formats throughout the entire analytical process [4, 5]. In this context, the data lakehouse paradigm has been gaining a lot of interest because the data lake paradigm can be viewed as the "best of all worlds", in terms of scalability and flexibility of the data lakes, combined with the benefits of governance and performance of data warehouses. In addition, Medallion data engineering patterns categorize the data into three categories: bronze, silver, and gold data layers, and allow the data value to change over time, as well as the value of analysis [6]. However, the technology wasn't the only thing. Thus, data governance becomes a strategic issue, encompassing policies and standards, stewardship, control of data and metadata, security controls, privacy preservation, and compliance monitoring. To develop trustworthy audit trails and data collection suitable for AI-powered applications, it is crucial to have sound governance in place and to prepare decision intelligence that is transparent, explainable, and responsible for public-sector applications [7].

Digital government programs and initiatives entail investments, program success, and progress, but there are numerous challenges that public sector bodies will need to consider when leveraging data to operate and interact with the world digitally. The government's information systems are known to be disjointed, poorly integrated across departments, and lacking a communications architecture that allows information to flow freely across silos [8]. This results in key information being 'stovepiped' and policymakers and administrators having disparate perspectives on public services and performance. While managed by structured data warehouses, they provide structured reporting but lack the scalability and flexibility needed to support big data from semi-structured and unstructured data sources across any digital platform. Information in data lakes, however, is also marred by several challenges: poor metadata quality, lack of lineage tracking, lack of ownership information, and low trust in analysis outputs due to poorly observed governance [9]. Many of these questions are now complicated by the urgent need for systems to be based on well-governed, standardized, high-quality data sets to generate meaningful results in this rapidly evolving world of Artificial Intelligence (AI) in public administration. Alongside concerns about the transparency and explainability of algorithms, bias mitigation, privacy protection, cybersecurity, and regulatory compliance, public trust and accountability are gaining importance, especially in government contexts [10]. Each element of data governance frameworks is studied individually and in depth, as is each element of data lakehouse architectures and their readiness for an AI-first paradigm, without consolidating anything into a single data governance framework or data lakehouse under one umbrella, as it were. Sometimes, it can be difficult to define AI in the public sector, with its governance, highly sophisticated analytics, and belief in decision intelligence. This keynote will focus on the importance of a data architecture informed by governance principles and grounded in current data engineering thinking and approaches, and will feature insights on leveraging data engineering to scale AI-based analytics.

In this paper, the aim is to tackle these challenges by presenting a Governance-Centric Lakehouse Architecture for Public Sector AI-Ready Decision Making with Medallion Data Engineering Patterns. The proposed framework comprises a set of governance mechanisms aligned across the data lifecycle, becoming part of, not separate from, the data engineering process. The Bronze, Silver, and Gold architecture enables all of that - scaling up data, updating it to improve data quality and standards, understanding what data does through data lineage, thoroughly looking after data quality and consistency, and enriching the data with analysis. All layers of the architecture are systematically integrated with metadata management, data stewardship, security enforcement, privacy preservation, access management, auditability, regulatory compliance, and governance rules and controls. It also has built-in AI-readiness elements such as Trusted data assets, Explainable analytics algorithms, Policy-driven data access, and comprehensive monitoring to boost confidence in the insights and recommendations generated and applied by AI. Furthermore, the architecture will make it interoperable with various parts of the government and enable the interconnection of multiple data sources, with an integrated decision intelligence platform to enable evidence-based decision-making. When combined, modern lakehouse technologies and governance-by-design concepts can provide a scalable, secure, and trustworthy foundation for next-generation public-sector analytics. In conclusion, the frame focuses on accelerating the digital transformation of governmental institutions, enhancing the transparency of governmental processes, ensuring compliance with laws and regulations, boosting trust in government, and, finally, leveraging Artificial Intelligence to facilitate information-based decision-making and value creation.

Scope of the Paper

1. In the Public Sector Data Engineering processes and environments in the Lake House concept, it is important to be able to bring Data Governance into it.

2. To learn how to use Medallion data engineering patterns to ingest data, cleanse, transform, standardize, and prepare data for AI-ready datasets throughout the bronze, silver, and gold layers.
3. Embed governance into the data life cycle, metadata management, data stewardship, data lineage, data privacy, cyber security controls, and compliance monitoring, etc.
4. Talk about an AI readiness plan for public sector environments that fosters trusted analytics, trusted explanations of AI, decisions, and/or governance of models.
5. To build a scalable foundation for decision intelligence that supports both internal and external (cross-agency) interoperability, evidence-based policymaking, and digital government with responsible use of AI in public sector decision making.

II. Related Work

Data governance and big data architecture have been significantly impacted by the rise of data-driven governance, especially in data governance, big data platforms, lakehouse architectures, and decision intelligence powered by artificial intelligence (AI). A preliminary study revealed certain challenges in data environments that were not well understood, including a lack of data standards, data quality issues, and data proliferation [8]. With the many projects that government agencies have embarked on, increasingly involving large data repositories and online platforms, data lakes are no longer an option for handling the huge volume of structured and unstructured data. There was, however, very little governance, which sometimes resulted in poor data quality, limited discoverability of the data, and low trust in the results of analyses [9]. To overcome these restrictions, a new paradigm has been introduced to leverage the flexibility of Data Lakes and the Governability/Performance of Data Warehouses, enabling the implementation of a single analytical environment that can be supported by modern Data-Intensive Applications [10][11]. These new developments further highlight the importance and necessity of Medallion data engineering patterns, metadata-driven architectures, the relationship of data quality to data lineage, data traceability, and the trustworthiness of data analytics [12]. Paralleling this development, we are seeing increased adoption of AI systems in public organizations, along with the development of structured analytical ecosystems that ensure transparency, explainability, and accountability in their operations, which are fundamental requirements for evidence-based policymaking [13]. Consequently, there has been significant research into well-defined governance principles, particularly regarding fairness, transparency, privacy protection, enforcement, and risk management in trustworthy AI systems [14] [15]. But the literature on these subjects usually frames these concepts separately, such as governance, lakehouse engineering, and preparedness for Artificial Intelligence. For the public sector, it is not yet a well-known landscape for building decision intelligence readiness that aligns with AI, with building blocks and Medallion patterns at the lakehouse level, all centered on governance.

Table 1. Summary of Existing Research

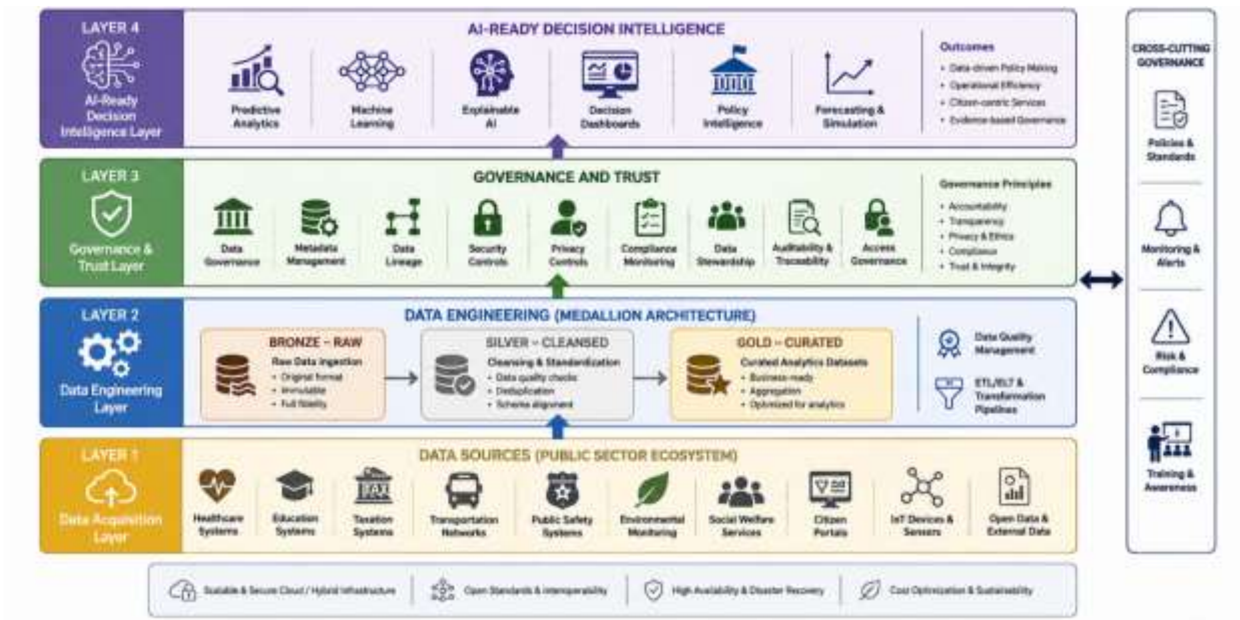
Ref.	Research Area	Research Focus	Key Contribution	Limitation
[8]	Public Sector Data Management	Data integration and information governance	Highlighted challenges associated with fragmented government data ecosystems	Limited consideration of AI-driven analytics
[9]	Data Lakes	Scalable storage and processing of heterogeneous data	Improved flexibility for handling large-scale structured and unstructured datasets	Weak governance and metadata management capabilities

[10]	Digital Government Analytics	Data-driven decision support and public-sector analytics	Demonstrated the value of analytics in policy formulation and service delivery	Limited architectural guidance for AI readiness
[11]	Lakehouse Architectures	Unified analytical platforms combining data lakes and warehouses	Enhanced scalability, reliability, and analytical performance	Governance controls are not comprehensively integrated
[12]	Medallion Data Engineering	Layered data refinement using bronze, silver, and gold zones	Improved data quality, lineage visibility, and analytical readiness	Limited focus on public-sector governance requirements
[13]	Explainable and Trustworthy AI	Transparency and accountability in AI systems	Established principles for trustworthy AI adoption	Primarily focused on model governance rather than data governance
[14]	AI Governance Frameworks	Fairness, compliance, risk management, and auditability	Strengthened governance requirements for AI deployment	Limited integration with underlying data architectures
[15]	AI-Ready Data Platforms	Metadata intelligence, interoperability, and decision intelligence	Defined requirements for AI-ready analytical ecosystems	Lack of a governance-centric lakehouse framework for public-sector environments

Significant advancements have been made in the four topic areas in the studies reviewed: Governance frameworks, Technologies, the Lakehouse, AI Governance, and the Analytics platform acceptable for action. All these aspects are addressed individually in most existing methods. Data governance research is more likely to address policy guidelines, whereas lakehouse studies are more likely to address scalability and governance-by-design less often; and AI governance frameworks, which focus on the trustworthiness of the AI model, lack some aspects of governance-by-design. Finally, an architecture that harnesses principles for governing the Lakehouse, includes built-in governance tools, addresses AI readiness, and supports public-sector decision intelligence is needed. This study seeks to fill this gap with an architecture specifically designed to support Trusted, Scalable, and AI-ready Public Sector Analytics.

III. GOVERNANCE-CENTRIC LAKEHOUSE ARCHITECTURE

In the era of government becoming increasingly data-driven and artificial intelligence (AI) influencing public decision-making, traditional public-sector issues with data architecture are making it tougher for businesses to scale with the data they are expected to process, achieve a taste of interoperability among various data types, ensure data trustworthiness, and conform to industry regulations and standards. Traditional data warehouses are great for governance but less flexible because they aren't well-suited to data from a wide range of sources. Whilst data lakes are great for scalability, it can be tricky to enforce governance, control metadata, and ensure proper data quality practices. To address these challenges, this study presents a new governance-centric lakehouse architecture that embeds governance mechanisms throughout the data engineering lifecycle and leverages the medallion model to lay the foundation for AI-driven decision intelligence. The architecture aims to reformat, ingest, clean, and structure government data into trusted, governable analytical assets that will then be fed into intelligence generation for consumption. The proposed approach will differ from the traditional approach, in which all data lifecycle operations would be governed to enable accountability, transparency, traceability, security, and compliance throughout the data lifecycle. It is capable of handling structured, semi-structured, and unstructured information across diverse public sectors, including healthcare, taxes, education, transportation, public safety, environmental monitoring, and citizen services platforms. It will enable continuous metadata management, lineage tracking, quality controls, privacy protections, and policy-based access mechanisms that create a trusted structure for using analytics and AI applications. The four layers of the architecture are: Data Acquisition Layer, Data Engineering Layer, Governance and Trust Layer, and AI-Ready Decision Intelligence Layer. Together, these layers create a unified space for public-sector agencies to convert data into action and for increased transparency, accountability,



and public trust.

Figure 1. Proposed Governance-Centric Lakehouse Architecture for AI-Ready Public Sector Decision Intelligence

3.1 Data Acquisition Layer

It is proposed that the Data Acquisition Layer is the lowest layer in the architecture, serving as the point where data is collected and integrated from various Public Sector sources. Governments generate many types of data from processes and activities across their operational systems, citizen involvement platforms, digital services, the Internet of Things (IoT), and other channels for data sharing and collection. Usually, these datasets are unstructured, inconsistent, and infrequent and low in quality, making them problematic for subsequent analytical tasks. One of the major goals of this layer is to create 1 source for structured, semi-structured, and unstructured data to be ingested into the lakehouse. Healthcare organizations, educational institutions, tax providers, transport companies, social welfare providers, and environmental and public safety providers are all data sources feeding into batch and time-series data sources. At ingestion, details of source ownership, time of collection, classification of source(s), and operational context are captured to help control the source and governance. The layer also provides interoperability, so that elements can interoperate with other elements, irrespective of the platforms they are built on and are used by governments. Once the source registration process is automated, you can easily ensure that data from new sources (in your organizational sense) is pre-tidied to comply with organizational standards and governance. This layer ensures a common, scalable way to collect data, enabling standardized data practices, discovery, and greater readiness for data analysis in the public sector.

3.2 Data Engineering Layer

The goal of the Medallion approach is to provide enriched, truthful analytical resources to better understand and analyze data in the Data Engineering Layer. This layer adds the bronze, silver, and gold models, aiming to capture the 'nubber' quality and the data's use over time. The Bronze layer contains ingested raw data, including the source location where the data was found, for audit/traceability. Data consistency is provided by the silver layer through schema cleansing, validation, pre-storing standard data, schema harmonization, and duplicate data deletion. Quality rules and transformation procedures ensure that errors, missing values, and inconsistencies are addressed systematically prior to analytical consumption. Gold is a data layer that provides a highly curated, aligned data source for reporting, analytics, machine learning, and decision-making applications. Data lineage information is maintained at all times, enabling traceability from the different source systems to the analytical outputs. The automated transformation pipelines enable scalability and reduce manual effort and complexity. Continuous data quality metrics are monitored to ensure the data is reliable, accurate, and trustworthy for analysis. The Medallion approach can ensure that an organization's data collection is easily distinguished from purely business-related data and can help the organization avoid spreading low-quality or inconsistent data, thereby supporting decision-making. It will therefore be a key element in creating AI-ready datasets that enable innovation, whether in public-sector analytics or evidence-based policy-making.

3.3 Governance and Trust Layer

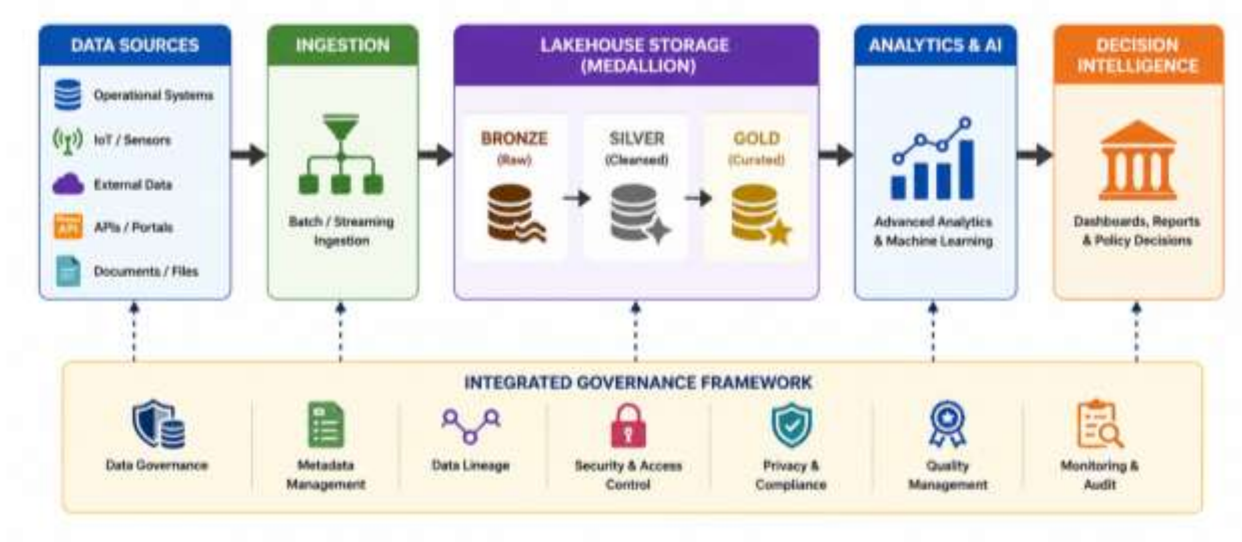
The Governance and Trust Layer is the unique characteristic of the proposed architecture, as it embeds governance controls throughout the data lifecycle. This layer of governance is an integral part of data management and analytics governance, rather than a standalone one. The major responsibilities that might be considered part of metadata management, data stewardship, security and enforcement, privacy protection, lineage tracking, compliance, and auditability. Data metadata is stored in metadata repositories, which contain information about data ownership, classification, quality, transformation history, and usage policies. There is increased visibility and accountability for data lineage through data lineage mechanisms. When protecting sensitive data, security controls include Role-Based Access Control (RBAC), authentication, authorization, encryption, and monitoring. Measures are in place to ensure privacy in compliance with all relevant regulations or ethical parameters applicable to citizen data. Data quality and data governance compliance structures are established through Stewardship Policies. All relevant data activities are audited, and organizations can analyze unusual activity to ensure regulatory compliance. Through layers that continually govern analysis outputs and ensure consistency, reliability, efficiency, fairness, and explainability, trustworthiness is achieved. As AI-based decision support tools become more prevalent across government organizations worldwide, the Governance and Trust Layer provides a platform for governments to build public trust in AI-assisted decisions.

3.4 AI-Ready Decision Intelligence Layer

The AI-Ready Decision Intelligence Layer transforms governed and curated assets into actionable insights that can be applied to support public-decision making. This layer features a variety of advanced AI/ML-driven modeling and explainable AI tools and capabilities, all consolidated in a single analytical environment. Use trusted datasets from the underlying layers to deliver better, more transparent results from analytic modeling. A few D.I. apps include resource allocation optimizer, fraud detection, public health forecasting, transportation planning, social welfare analytics, and social welfare policy impact assessment. Explainability mechanisms help users, stakeholders, and analysts understand AI systems' results and suggest ways to improve them, thereby increasing accountability and trust. Live, interactive dashboards/visualization tools can be provided to the policymaker for Operational and Strategic Intelligence. There are continuous monitoring processes to measure the performance of predictive models and to assess bias, fairness, and governance policies. Also, the layer may be used for scenario analysis and/or predictive simulation, which allows decision-makers to evaluate their options for implementing the policy before it is used in the real world. Architecture is the planned and directed access to data resources and AI-assisted analytical tools, which lead to evidence-informed governance and leading decision-making. Lastly, the layer that sits between the technical Data Management processes and the impacts experienced by the policy for citizens in the Public sector, ensuring that investments in Data are reflected in actual gains for the Government in its performance, transparency, and citizens' access to government services.

IV. GOVERNANCE FRAMEWORK INTEGRATION

Good governance is essential for trustworthy, secure, compliant, and AI-driven decision intelligence, and for better governance in public-sector data ecosystems. New lakehouse designs are scalable and flexible; however, they need to be properly governed, with governance controls in place throughout the data lifecycle. Whereas this proposed idea of governing governance through the architecture of the CCS will embed the principles of governance in every layer of operation/analysis, treating them as part of the entire idea rather than as an operation/analysis on the periphery. This staff-centric model allows businesses to be measurable, explainable/transparent, auditable to ensure regulatory compliance, and adaptable for future analysis, AI, and other use cases. The framework is centered on a single Data governance framework that integrates information, metadata, security, privacy, and AI governance, and provides a monitoring and management model for data assets from acquisition through the generation of all sorts of information, metadata, security, and privacy checks, as well as types of AI. Data governance mechanisms can be categorized in the organization as (i) responsibility for data stewardship, (ii) accountability for data stewardship, (iii) data quality management, and (iv) enforcement of data policies throughout the organization [16]. Further improvement of the discovery and transparency of the underlying data can be achieved by introducing metadata governance, consistent metadata definition, cataloging, and lineage features, which enable users to access it for related analytical workflows [17, 18]. Security governance will include deploying policies and mechanisms to prevent sensitive data from being accessed or used by third parties in inappropriate ways, such as access control, encryption, authentication, and role-based monitoring systems. Private data governance is the process of managing private data to ensure its classification, consent, anonymization, and



retention in accordance with legal, ethical, and organizational requirements. At the same time, AI governance introduces new capabilities, such as explainability, fairness evaluation, model monitoring, bias detection, and auditing, to enhance transparency in AI-driven decision-making [19]. Interoperability, accountability, and evidence-informed policy are all parts of public trust environments in public-sector organizations and could be combined within a single framework. The framework can be used to provide ongoing governance oversight through automated controls, ensuring that policy compliance, quality issues, and compliance risks are detected before they negatively impact operational performance. This ensures that governance can be integral to the entire lakehouse ecosystem, making decisions transparent, accessible, and reliable, while also being an evolving process linked to the Objectives of the public sector [20].

Figure 2. Integrated Governance Framework Across the Lakehouse Ecosystem

The cross-cutting governance solution would be all-encompassing, incorporated into the lakehouse architecture design (should it be implemented), and would govern data and analytics throughout the design. It's no longer an ad hoc activity at some point in the data life cycle, but rather an ongoing, constant partner throughout the data, all the way to AI-driven decision intelligence. At the data-source level, governance policies are also followed to ensure one's approach to onboarding data-sources (who owns them, how they are stewarded, and how metadata is registered) and establishing the governance accountability from the data-source down into the ecosystem. During health care information sharing, the following are observed: a set of data owners and stewards; a set of data classifications; a set of policies for collecting health care information; and a set of policies for analytical processing. The context of Data Governance can be leveraged to maintain data quality validation rules, schema standardization, tracking of data flow during transformation, and data lineage across the Data Engineering layer over the long term, thereby ensuring data integrity and traceability. For example, information from multiple data sources may be normalized and combined, and changes may be reverted to the complete change history. Additional measures put in place to secure sensitive information and prevent its disclosure include role-based access controls, information encryption, and active tracking and monitoring of user access and usage habits. For instance, the health data might not be made available to any entity, or access to it might be recorded and tracked but not shared. An AI governance mechanism provides an additional layer of control within the AI and decision intelligence layer, enabling detection, monitoring, review, and tracking of AI and AI-driven decisions, AI model behavior, AI bias, explainability scores, and related metrics. The AI model can be continually monitored and adjusted based on the socio-demographic, gender, and cultural composition of the targeted community, ensuring a fair and unbiased process. All of these aspects of governance constitute a full-blown Lifecycle Enterprise, in which the landscape of security, transparency, accountability, and data privacy can be realized across the enterprise's entire analytical lifecycle. Architectural design integration of governance measures to further strengthen the trust of organizations in analytics, predictive modeling, and decision support tools with the assistance of AI. Thus, the approach of Governance of digital transformation provides a solid foundation for the responsible use of AI in public sector services in an ethically sound, transparent, and legally safe manner [16] [18] [20].

V. AI READINESS ASSESSMENT MODEL

The adoption of analytical technologies is not sufficient to ensure AI implementation; the data environment must also be capable of supporting it. Despite the abundance of data in many Government agencies, much of it is fragmented, poorly governed, not interoperable, poorly managed in terms of metadata, and/or of poor quality. However, the use of AI-powered decision intelligence is much more dependent on the organization's data assets being prepared to support data-intensive workloads. Current studies emphasize the need for data quality, data governance frameworks, and organizations' readiness to make efficient and effective use of AI [21]. In addition, the NIST AI Risk Management Framework highlights governance, measurement, and monitoring as crucial elements for the trustworthy deployment of AI [22]. The proposed AI Readiness Assessment Model has been developed to measure the capability of organizations in a systematic way, over various dimensions such as data quality, data governance maturity, data metadata availability, data interoperability, security, data privacy compliance, data lineage visibility, data analytical capabilities, and incorporating a decision intelligence solution. Plus, the model for assessing the maturity of AI selection deliberately follows enterprise AI maturity models, which recommend data and technology readiness, as well as becoming AI selection capabilities and roles more mature to be ready for wider enterprise adoption of AI. Furthermore, the assessment model is supported by enterprise AI maturity models, such as the KPMG's AI Maturity Model Framework, which encourages incremental development of enterprise data, technology, governance, and organization to best prepare for enterprise AI adoption [23]. Further, stewardship, metadata management, accountability mechanisms, and policy handling emerge as important for data governance maturity [24]. Moreover, as revealed in recent work on AI readiness in digital government [25], explainability and transparency, interoperability, and trusted data assets are key to building a sound decision-intelligence system. The proposed model thus takes a maturity-based view, in which organizations develop through a series of readiness phases, from disjointed, reactive data management to governed, AI-driven ecosystems. Evaluating readiness by size and maturity levels enables public organizations to identify capability gaps,

prioritize modernization efforts, and develop strategic AI route maps for responsible implementation. Furthermore, the model can help decision-makers balance governance objectives and technological investments and provide robust, quality data ecosystems for AI systems to operate. In conclusion, the AI Readiness Assessment Model is a strategy that can help public sector organizations move towards responsible, scalable, and governance-based use of AI.

Figure 3. AI Readiness Assessment Model for Public Sector Decision Intelligence

The proposed AI Readiness Assessment Model helps public sector organizations assess their readiness to use AI and informs them of the strategic, technical, and governance decisions they need to make to become AI-ready. Results of multi-organizational data quality, governance, interoperability, metadata management, security, privacy, and data analytical readiness aspects are captured throughout the progression of the maturity stages. In the early phases, data assets are scattered across a variety of systems and are likely poorly controlled or governed, or even not controlled, with data used mostly in their more fundamental roles, such as operational reporting. For instance, government departments might keep their own databases of definitions that don't match up, limiting their capacity to learn from and collaborate with one another. Managed takes the very simplest requirements for stewardship, early-stage quality control, and enterprise-based regular data collection procedures, to ensure consistency and reliability. The Standardized level of maturing organizations is when they transition to formalized governance, a single data repository common to all organizations with metadata, data flow, and processing interoperability, and common data quality measures that enable them to share trusted information with each other and repeat standard data analytics processes. Shared metadata standards and data exchange protocols, for example, can be used in the healthcare, taxation, and social welfare sectors to service integration of citizen services. Develop them continuously in an optimized way, with automation, lineage visualization, continuous monitoring, additional cyber protection, efficiency, transparency, and trust across all activities. Typical features are real-time monitoring dashboards, automated quality checks, and policy-driven governance workflows. The AI-Ready Enterprise tier is now a fully managed, interconnected tier with information assets that are categorized, visible, protected, interconnected, and regularly monitored for Data Quality, Data Compliance, and Data Operations. As the organization gets more mature, it may be able to embed the capability for embedding Explainable AI (XAI) and predictive systems, intelligent automation, and decision intelligence systems into the confidence, transparency, accountability, and compliance aspects of it, which helps support regulatory and compliance requirements [22, 25]. The maturity progression follows current digital transformation and data maturity models to underscore the value of capability building, which must be complemented by governance to sustain the ongoing adoption of AI and the ecosystem of intelligent decisions [21],[23]. Further, the framework could be used to compare agencies and departments in terms of readiness, indicate where some need to develop capabilities, prioritize modernization investments, and outline and plan modernization roadmaps based on the agencies' and departments' readiness. This helps shape the assessment outcome and recommends robotic solutions, such as strengthening public-sector governance capacities, advancing digitalization, and laying the foundations for trust to unlock the responsible scale of AI solution assessment.

Table 2. AI Readiness Assessment Dimensions and Evaluation Criteria

Dimension	Assessment Focus	Evaluation Criteria	Maturity Indicators	Expected Outcome
Data Quality	Accuracy, completeness, consistency, timeliness	Error rate, completeness score, validation compliance	High-quality and standardized datasets	Reliable AI training and analytics

Governance Maturity	Governance policies and stewardship	Ownership assignment, stewardship processes, policy enforcement	Formal governance framework implemented	Improved accountability and trust
Metadata Management	Data cataloging and discoverability	Metadata completeness, catalog coverage, and classification accuracy	Centralized metadata repository	Enhanced transparency and usability
Interoperability	Cross-agency data integration	Data sharing standards, API connectivity, and format compatibility	Seamless interdepartmental exchange	Improved collaboration and analytics
Security Controls	Protection of sensitive information	Access control, encryption, and monitoring effectiveness	Strong security governance	Reduced cyber and operational risks
Privacy Compliance	Regulatory and ethical compliance	Consent management, anonymization, retention policies	Compliance-driven data handling	Protection of citizen information
Data Lineage	End-to-end traceability	Transformation tracking, audit logs, source visibility	Full data traceability	Greater transparency and auditability
Analytical Capability	Support for advanced analytics	BI tools, predictive analytics, real-time processing	Mature analytical environment	Improved evidence-based decisions
AI Governance	Trustworthiness of AI systems	Explainability, bias monitoring, and model auditing	Responsible AI implementation	Increased trust in AI outcomes
Decision Intelligence	Strategic decision support	Forecasting, simulation, optimization, policy analytics	AI-enabled governance ecosystem	Enhanced policy effectiveness

VI. LIMITATIONS AND CHALLENGES

Beyond the technical aspects, there are non-technical challenges and limitations in the proposed Governance-centric Lakehouse Architecture that hinder the public sector's AI readiness. The technical environment can vary significantly across public sector organizations. There may be differing levels of legacy systems, employee mindsets, and local needs, which present challenges during integration. To create the intended shape of the AEC, significant investments are needed in infrastructure modernization, governance development, personnel skill development, and organizational change management. Moreover, problems such as inconsistencies in data models, policies, legal frameworks, and operations still hinder interoperability among various agencies. When it comes to providing the accuracy of results and recommendations from AI analysis, the quality of the data can play a crucial role. Variations in data quality due to missing or duplicated records, differences in data formats, changes in clinical history, or even aging data can affect the quality of analytical data and the results produced by the system. However, the development of a robust architecture depends on sound management and tracking of metadata, as well as the use of new tools, their features, and specialized expertise that may not be available in some government institutions. New challenges also emerge in the privacy and security aspects of data, where data in the public sector is typically loaded with highly sensitive information about citizens, which is essential to data security, while simultaneously being relevant to the analytical value that can be added. In addition, explainability, measuring fairness, and detecting bias are concepts introduced by this new era of AI governance and remain evolving areas of research, although not all have sufficient uniformity of standards across jurisdictions. Resistance by those institutions that will further inhibit implementation and impact of the proposed governance structure is also likely, since they are targets of governance improvement and data-sharing programs. The potential for enhanced transparency, accountability, and degree of AI preparedness is significant; however, key to its implementation will be thoughtful planning, steady investment, firm leadership, ownership, and continued oversight by governance of the operational, technical, and regulatory challenges.

6.1 Legacy System Integration and Interoperability Challenges

One of the key challenges in the proposed architecture is integrating legacy information systems, which are used regularly in the public sector. But a decade or more has passed since many of these applications and databases were written, and they have been deployed by different government organizations, not enough when it comes to AI or enterprise-wide analytics. They often have non-standard schemas; they don't always follow standard schemas; they seldom meet application-to-application data type-matching standards; and they are not always data-sharing-friendly. Technically, integrating all these information sources from health care, tax, education, and social welfare systems into a lakehouse can be complex and resource-intensive. Moreover, many different definitions of data and departments' classifications of metadata are found, and it is difficult to integrate. The pressure increases further if more than one jurisdiction or regional authority is involved in information exchange programs. To enable meaningful interoperability, it may be necessary to undertake substantial effort to develop common standards and transformation pipelines, and to harmonize data models. When not implemented properly, organizations may be plagued by implementation delays, increased implementation costs, and reduced analytical power.

6.2 Data Quality and Governance Maturity Constraints

The quality of the data that underpins an AI-ready architecture and the maturity of its governance are critical to its success. Organizations in the public sector often face incomplete records, duplicate data, inconsistent data formats, outdated data, and data entry inaccuracies. This can affect the trust in the analytical models and impact the outcomes of policies produced by an AI system. Ugliness was also found in governance processes and rules themselves, which vary from agency to agency: some have well-developed governance mechanisms and rules, while others have weak governance mechanisms and poor data management rules. Therefore, for governance-by-design to be implemented, significant change to the organization is needed, from defining who owns what, who needs to steward what, when, and what standards are applicable to the metadata, what checks on data quality and consistency are put in place, as well as what steps are taken for compliance. There may be a need for far-reaching measures not only in workforce training but also in cultural change programs to ensure the uptake of the governance practices. Moreover, long-term governance is required to ensure quality, and therefore, mechanisms are needed to continue monitoring, auditing, and enforcing the policy. Organizations may still have poor governance and may lack transparency, traceability, and trustworthiness in their analytical environments.

6.3 Privacy, Security, and Regulatory Compliance Risks

Citizen information in a public-sector data ecosystem may contain highly sensitive data, such as health care, financial, education, social welfare, and public safety data. A challenge for the proposed architecture is developing higher-level analytics for these datasets and ensuring their security. If an organization is the target of an insider attack, cyberattack, or data breach, it can be easy to fool people, leading to legal and regulatory consequences. Security and privacy (including encryption, access control, audit windows, and audit monitoring) are included in the architecture, but maintaining security in large-scale, distributed environments can be complicated. Depending on jurisdictional or regional requirements, compliance can be challenging for those who share information across jurisdictions or regions. Some privacy-protecting approaches to analytics, anonymization procedures, and consent processes may also be complicated and involve calculations. Moreover, the limits of suitable analyses with respect to privacy protections can be difficult to determine, as overly invasive laws will limit what data-driven programs can achieve. With the growing use of AI, there are heightened concerns about data ethics, algorithmic transparency, and accountability, which add to the dimensions of governance.

6.4 AI Governance, Organizational Adoption, and Resource Limitations

What is important to note is that AI governance is becoming an important part of responsible use of AI. Currently, very few organizations have established AI governance frameworks that operationalize their ethical AI model behavior, identify bias, assess fairness, and ensure explainability. However, introducing AI governance components into AI frameworks may likewise lead to concrete, waterproofed concerns regarding a lack of expertise in waterproofed governance, changing laws, and stakeholders' technological capacities. In addition, complex issues such as competing governance agendas and limited staff access to and financial resources in public sector organizations may impede the incorporation of better governance practices. Furthermore, good governance plans might face obstacles to implementation related to internal coherence, and data- and AI-sharing procedures could pose major resistance. But if they don't know the new technology and are “asked” to do new work and take on new responsibilities, challenges to the new governance-led decision-making processes can be difficult for employees accustomed to the SOC. The distance between the experts and the lack of people with practical experience in the topics themselves may also result in slow progress in implementing them. The proposed architecture requires investments in technology, but it is even more essential to establish a robust commitment at the organizational level, at the case and market leadership levels, and among dedicated staff, and to continually reinforce improvement commitments, so as to keep Governance goals aligned with society's needs and new AI technologies.

VII. FUTURE RESEARCH DIRECTIONS

The proposed Governance-Centric Lakehouse Architecture provides a strong foundation for making AI a ‘first-class citizen' for decision-making in the public sector, as the future is likely to bring many more developments in the AI, digital government, and data governance space. Going forward, it is important that the research further develops limited governance practices for data management outside of AI-based tools and be used to support governments' governance of service delivery and operational planning. With the introduction of new technologies such as Generative AI/Large Language Models (LLMs), Federated learning, Digital Twins, and Autonomous intelligent agents, new Governance questions are emerging across the fields of Accountability, Transparency, Explainability, and Risk Management. Meanwhile, the public sector is likely to be part of increasingly complex, multi-jurisdictional, and multi-agency environments populated by other external entities, requiring new engagements in interoperability and governance. Moving forward, additional research should focus on governance-oriented architectures for real-time decision intelligence, privacy-friendly analytics, and sovereign AI initiatives, as well as on continuous compliance monitoring, while ensuring compliance and maintaining citizens' and reputation's trust. Moreover, intelligent automation in the governance processes unlocks opportunities for greater efficiency, scalability, and responsiveness, even in the realm of large data. Although this will address current governance and AI readiness challenges, the performance and scalability must be demonstrated. This could be done by applying the architecture to 'real-world' scenarios, following it in practice over time, and studying its application across different agencies.

7.1 Federated and Cross-Agency Governance Ecosystems

The future direction requires exploring a new 'federated governance' approach that enables secure data sharing and joint decision-making across Government agencies and jurisdictions. A greater degree of inter-sectoral public service collaboration is needed, especially when working across sectors such as public safety, healthcare, education, transportation, taxation, and social welfare, which are usually autonomous.

Organizations have the option of putting in place federated governance architectures for managing (and maintaining) trust-based information sharing among federated entities while still retaining control over data ownership and custodianship. Other distributed metadata research areas include policies and processes for interoperable support of autonomy, distributed metadata management, federated lineage tracking, and cross-agency stewardship frameworks. In addition, further evaluations of the governance arrangements for international coordination and cross-border data exchange programs are required. More recently developed technologies like decentralized IDM, blockchain-based audit trails, and distributed trust may also offer value-added functionality in federated governance settings. Good solutions to these problems will enable the building of a scalable and interoperable public sector ecosystem that supports collaborative analytics and evidence-based decision-making in public policy, with transparency, accountability, and compliance.

7.2 Autonomous Governance and AI-Assisted Data Stewardship

It can be more difficult to perform governance processes for data ecosystems manually in larger, more complex ones. The autonomous governance mechanisms that entail the use of Artificial Intelligence, the automatic generation of metadata, the enforcement of governance policies, the monitoring of data quality, the analysis of data lineage, the verification of conformance, and recommendations for data stewardship should thus be studied in the context of the future. Then, at any time, an intelligent governance agent could review data pipelines, suggest corrective actions as needed, check whether anyone had violated any policies, and identify any anomalies uncovered during the review of the data, with minimal human effort. AI stewardship systems can, for instance, easily recognize sensitive data, find missing data metadata, or suspicious access behaviors that could signal security breaches. Thus, further research into developing trust in the governance mechanism set up by the AI will be required, as it must take into account the AI's decisions and have a goal within the organization itself. With the right solutions, these innovations can help to scale the governance system, minimize administration burden, and increase the responsiveness of government information systems.

7.3 Real-Time Decision Intelligence and Digital Twins

Real-time decision intelligence and digital twin technologies will likely prove more impactful for the future of public-sector analytics. Future studies could focus on governance-driven lakehouse approaches to facilitate the emerging digital paradigm for city systems, including transportation, health, the environment, and other key public utilities. Digital twins can be used to model policy interventions, predict operational consequences, and optimize resource allocation prior to real-world decisions. But to ensure accurate and reliable digital twins, real-time data flows need to be managed, quality assurance systems need to be automated, monitoring systems need to be continuous and robust, and interoperability needs to be ensured. This may involve the research of streaming analytics architectures and governance models, dynamic policy co-enforcement, real-time lineage management, and explainable simulation environments. Furthermore, more research is required on how AI forecasting and optimization can be implemented in the digital twin system in a manner that is non-discriminatory, transparent, and accountable.

Conclusion

As AI and data-informed government emerge, and with new regulations shifting more and more towards accountability and scalable governance, the need for a new generation data architecture arises. Even though the amount of data created by Governments is important, there remain a few challenges related to Information Islands, information silos, and information fragmentation; the lack of Governance and Data Management; and limited Interoperability and readiness for AI. This is often challenging for public institutions when they have to travel from raw data to relevant derived information that can support effective policy-making and on-the-ground service provision, as imposed by the high barriers illustrated. To mitigate these concerns, this paper introduces a Governance-Centric Lakehouse Architecture for AI-Ready Public Sector Decision Intelligence that directly incorporates governance into the data engineering lifecycle, using Medallion data engineering patterns. The four pillars of the proposed Architecture are: Data Acquisition Layer, Data Engineering Layer, Governance and Trust Layer, and AI-Ready Decision Intelligence Layer, which together govern, manage, and utilize data assets in the public sector. The framework takes a more comprehensive approach to governance than the traditional one that treats governance as a separate management task; it embeds governance tasks across the data lifecycle to enable ongoing surveillance of data quality, metadata management, data lineage, security and privacy, compliance, and AI accountability. Plus, embedding the AI Readiness Assessment Model provides an organization with a roadmap for assessing readiness, identifying gaps, and mapping out a path to responsible AI adoption. It shows the promise and potential of Governance by design in enhancing transparency, trustworthiness, interoperability, and analytical reliability in the public-sector data ecosystem. The proposed architecture lays the foundation for scaling up to address some of the challenges in heritage applications (as noted above), such as governance maturity, privacy protection, and organizational readiness. To sum up, the Architecture offers a promising pathway for public sector organizations to advance their decision intelligence initiatives, foster high-quality public services, continue digital transformation, and ensure the trustworthy, ethical, and legal use of Artificial Intelligence in modern governance.

References

[1] Department of Economics. (2024). United Nations E-Government Survey 2024: Accelerating Digital Transformation for Sustainable Development-With the addendum on Artificial Intelligence. Stylus Publishing, LLC.

[2] Sysoyeva, L., & Usarralde, M. J. M. (2025). From security to digital governance in OECD: Evolution and some current considerations since the digital transformation. Salud, Ciencia y Tecnología-Serie de Conferencias, (4), 1401.

[3] ONUNGA, J. (2025). Digital transformation in public administration and data-driven decision-making: a review of Turkana County Government.

[4] Burk, S., Sweenor, D., & Miner, G. (2021). It's all analytics-part II: Designing an integrated AI, analytics, and data science architecture for your organization. Productivity Press.

[5] Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021, January). Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In Proceedings of CIDR (Vol. 8, No. 1, p. 28). sn.

[6] Strengholt, P. (2025). Building Medallion Architectures: Designing with Delta Lake and Spark. " O'Reilly Media, Inc."

[7] Wiselka, M. (2024). Development of Modern Data Platform using Medallion Architecture.

[8] Day, A., & Hart, V. (2007). The architectural guidebook: from Palladio to pod. arq: Architectural Research Quarterly, 11(2), 151-158.

[9] Boukraa, D., Bala, M., & Rizzi, S. (2024). Metadata management in data lake environments: a survey. Journal of Library Metadata, 24(4), 215-274.

[10] Aljumaili, M., Karim, R., & Tretten, P. (2016). Metadata-based data quality assessment. VINE Journal of Information and Knowledge Management Systems, 46(2), 232-250.

[11] Khatri, V., & Brown, C. V. (2010). Designing data governance. Communications of the ACM, 53(1), 148-152.

[12] International, D. (2017). DAMA-DMBOK: Data management body of knowledge. Technics Publications, LLC.

[13] Armbrust, M., Das, T., Sun, L., Yavuz, B., Zhu, S., Murthy, M., ... & Zaharia, M. (2020). Delta Lake: high-performance ACID table storage over cloud object stores. Proceedings of the VLDB Endowment, 13(12), 3411-3424.

[14] RISK, C. (2021). Risk management framework. CORPORATE REPORT.

[15] Ercan, S. I. A., & Akin, O. (2024, September). International Insights on AI Risk: Comparative Analysis of International Organizations. In 2024, the 8th International Artificial Intelligence and Data Processing Symposium (IDAP) (pp. 1-8). IEEE.

[16] Sun, Y., Parker, T., Gierschmann, L., Shams, S., Canmetin, T., Duteil, M., ... & Chin, Z. S. (2026). Defining AI Models and AI Systems: A Framework to Resolve the Boundary Problem. arXiv preprint arXiv:2603.10023.

[17] Bhat, J. (2024). Designing Enterprise Data Architecture for AI-First Government and Higher Education Institutions. International Journal of Emerging Research in Engineering and Technology, 5(3), 106-117.

[18] HOSSAIN, M. S., RAHAMAN, M. M., & BISWAS, B. AI-DRIVEN THREAT DETECTION IN PUBLIC SECTOR CYBERSECURITY, INTEGRATING MACHINE LEARNING INTO NATIONAL SECURITY SYSTEMS.

[19] BELLO, A. D., ELEBE, O., HAMMED, N. I., OMOEGUN, G. O., & FADAYOMI, O. (2024). A Cybersecurity Risk Management and Regulatory Compliance Framework for Financial Institutions.

[20] Valiki, S. (2026). Beyond Medallion: Next-Generation Lakehouse Architectures for Real-Time AI-Driven Enterprise Decision Systems. International Journal of Research Publications in Engineering, Technology and Management (IJRPETM), 9(2), 614-631.
[21] Pamula, V. K. (2026, February). The Medallion Architecture in Practice: A Framework for Building Scalable and Governed Data Lakehouses on Microsoft Fabric. In 2026, IEEE 5th International Conference on AI in Cybersecurity (ICAIC) (pp. 1-5). IEEE.
[22] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020, January). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 33-44).
[23] Kumar, N., Parekha, C., & Sheth, R. (2025). Exploring 6G Wireless Networks: A Comprehensive Analysis. Virtual Reality and Augmented Reality with 6G Communication, 51-88.
[24] Alhassan, I., Sammon, D., & Daly, M. (2018). Data governance activities: A comparison between scientific and practice-oriented literature. Journal of enterprise information management, 31(2), 300-316.
[25] Şeren, G. Y., & Esen, Ö. (2025). Can Artificial Intelligence Investments Drive Environmental Sustainability? Insights from OECD Countries.