International Journal of Scientific Research and Modern Technology (IJSRMT) *ijsrmt.com*
Volume 1, Issue 12, 2022
DOI: https://doi.org/10.38124/ijsrmt.v1i12.1111

Big Data Pipelines for Cross-Domain Decision Support: A Cloud-Centric Approach

Uday Surendra Yandamuri[1]

[1]Independent Researcher India

Publication Date: 2022/12/30

Abstract

Advancements in cloud computing technologies and the resultant involvement of the big data paradigm have catalyzed the emergence of novel cloud-centric data pipelines that streamline the pipelines' distinct phases. An overview of these advances is given, covering their role in cross-domain decision support—one of the greatly useful yet difficult-to-achieve standards in big data applications—upon which their motivation is grounded. The discussion transcends a mere listing of the recent patterns and principles by merging them in a coherent, syntactic whole, culminating in an evaluation framework for the assessment of supporting cloud-centric decision-support pipelines.

Cross-domain decision support addresses the supply of information necessary for decision making across several thematic data domains. Supported by data offerings from different application domains, such type of support enables and simplifies correlated or congruous decisions involving diverse subjects or instances, which would otherwise require engaging expert resources from the different areas. As naturally occurring or deployed data in several themes are stored using cloud solutions, the public or private clouds that combine the different data sources into one logical entity become the pivotal data pipelines for such cross-domain decision-support actions.

Keywords: *Cloud Computing Advancements, Big Data Paradigm, Cloud-Centric Data Pipelines, Streamlined Data Processing Phases, Cross-Domain Decision Support, Decision-Support Pipelines, Cloud-Native Pipeline Patterns, Data Pipeline Evaluation Framework, Integrated Architectural Principles, Syntactic Pipeline Design, Multi-Domain Data Integration, Thematic Data Domains, Correlated Decision Making, Congruous Analytics Support, Expert Knowledge Substitution, Cloud-Based Data Storage, Public and Private Cloud Platforms, Logical Data Unification, Cross-Domain Analytics Enablement, Cloud-Supported Decision Intelligence.*

I. INTRODUCTION

Forms of decision support that enable decision making across domain boundaries are termed cross-domain decision support. Data-based models often provide good decision support for several types of decisions. Nevertheless, many decisions that are critical for a given domain might not involve direct participation of decision makers from the other domains. Hence, these models must be based on data-sets that are either maintained within the domain or that can be efficiently and accurately ingested from external sources. Such datasets tend to be too detailed, too coarse, or simply missing. Nevertheless, external datasets have proven useful for these domains, even in cases where the significance of these external datasets has not been emphatically established. Such cross-domain models must be supported with cross-domain data pipelines that generate, from trustworthy data sources, those datasets that are critical for the execution of the models. An example of such data-set generation is the efficient online detection of disease outbreaks. The data-driven formation of distant scaffolding models is another example. These models maintain accuracy over long distances and long time lags, but they need a rich set of input variables, some of which are seldom recorded. Here, the quantitative role of data recorded within and outside the target domain is established. The generation of the missing data latently helps the estimation.

The application of big-data pipelines in the cloud for cross-domain decision support is examined next. Cross-domain decision support is defined and it is shown that the relevant pipelines have three types of outward flows: ingestion, generation and supply. Within each category several architectural variants are examined. Evaluation

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

frameworks and quality-governance cross-domain considerations are also examined. Configuration for storage and processing in the cloud, and deployment in public, private and hybrid clouds, are finally considered. For bread, fish and forecast sales, the models formually requires total volume, lead time and distance; establishing the role of seasonal forecasting is more difficult. Nevertheless, even minor distance intervals can be significant for nurseries, demand during the dry season is twice that in the wet season, while price is the most significant factor for shrimp. Planning during lush periods, investment into frost-resistant fruit, and careful monitoring are advised.

- *Overview and Purpose*

Cross-domain decision support aims to assist decision-making across different domains—e.g., law enforcement and public health—by integrating relevant domain data and supporting tools using adequate decision-making models. Such support requires additional data not generated or maintained by the domains involved, prompting the need for a process that ingests data from yet another domain, usually referred to as the source domain. The lack of semantic interoperability among domains when querying cross-domain data can reduce support quality; hence, a cloud-centric approach using data pipelines that bridges this gap can be beneficial.

The outlined methodological framework covers the layers of a cross-domain cloud-centric pipeline and the architectural patterns that implement them. Data governance and quality play a crucial role since the decisions made by the source and target domain are affected by the decisions made by the service provider. Challenges, evaluation frameworks, and metrics for benchmarking decision support effectiveness are addressed. The proposed research aims to provide direction, guidance, and insight to those developing cross-domain decision support systems using cloud infrastructure and providing decision support from the source domain perspective.

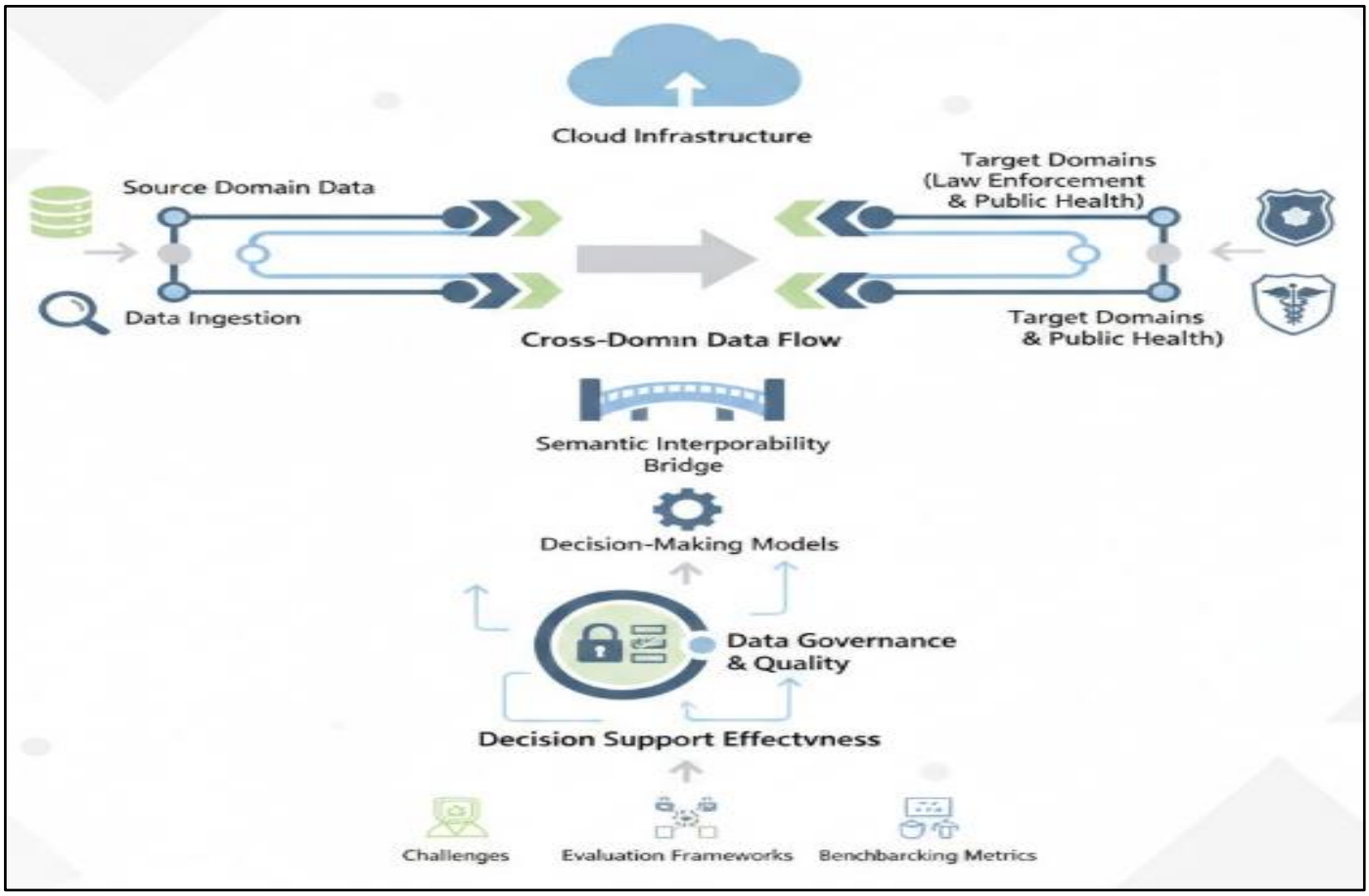

Fig 1 Bridging the Semantic Gap: A Cloud-Centric Framework for Governance-Aware Cross-Domain Decision Support Systems

II. BACKGROUND AND FOUNDATIONS

Cross-domain decision support in the context of big data pipelines is characterized by a modular and layered pipeline architecture, a cloud-centric deployment model, and Semantic Web technologies and standards for data ingestion, integration, enrichment, and quality management. The research addresses key challenges related to the ingestion and integration of heterogeneous and distributed data sources, the processing of large data sets, data governance and quality assessment, and the design of evaluation frameworks and metrics.

The decision-support requirements of heterogeneous domains are either similar or complementary, which motivates the gathering of information from multiple domains. The decision-critical data for different domains are often located in different domains, and accessing them

may require cross-domain collaboration. Identifying and accessing data from other domains depend on semantic interoperability among the cross-domain decision-support systems, which involves the design of semantic models, the development of mapping rules, and the creation of algorithms to discover knowledge across domains.

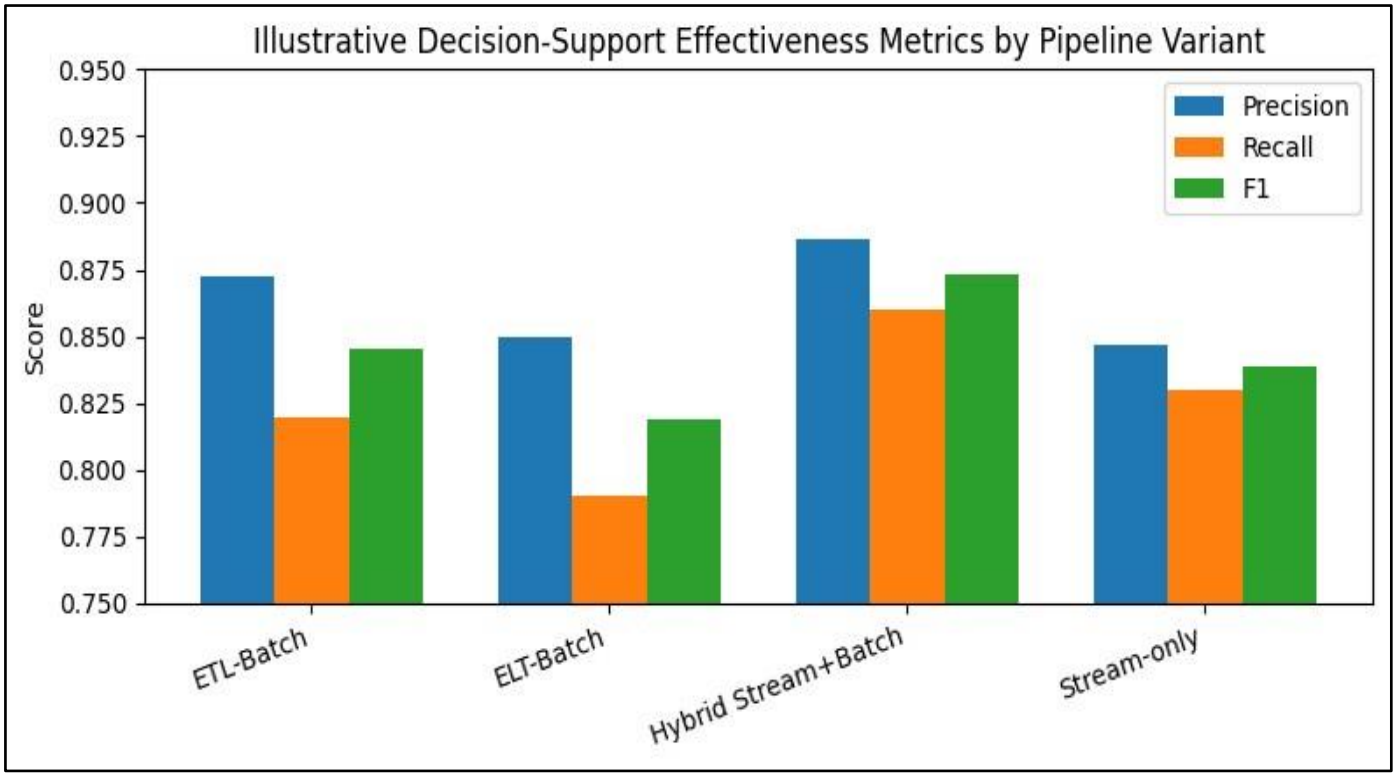

Fig 2 Derivation of Precision, Recall, and F-Measure (Step-by-step)

➢ *Precision (Positive Predictive Value)*

- Meaning: Out of everything the pipeline predicted as positive, how many were correct?

Predicted positive count:

$$\text{PredPos} = TP + FP$$

Correct predicted positives: TP

So precision is:

$$\text{Precision} = \frac{TP}{TP + FP}$$

➢ *Recall (Sensitivity / True Positive Rate)*

- Meaning: Out of all the truly positive cases, how many did the pipeline catch?

Actual positive count:

$$\text{ActualPos} = TP + FN$$

Captured positives: TP

So recall is:

$$\text{Recall} = \frac{TP}{TP + FN}$$

➢ *Cross-Domain Decision Support*

Cross-Domain Decision Support addresses the need for decision-critical data at the cross-domain level, the challenges of both data interoperability and synthesis, and the techniques employed to provide data leading to decisions that span multiple data sources across multiple domains. Data at the cross-domain level are often the hardest to come by and, for this very reason, are of great interest: lists of potential terrorists, new viruses with a possible threat of human contagion, spread of electrical blackouts, epidemic warnings, etc. Such information does not rely solely on a single domain but needs modeling inputs from different sources and domains: meteorological, socio-economical (e.g., modeling the population in attractive regions), and/or infrastructure (airports, harbors, and so on).

Decision analyses at the upper or cross-domain level require decisions that draw on data from multiple domains. The challenge here is to provide, integrate, and quality-control data from different sources in a collaborative framework that introduces knowledge and languages on the data and processes being considered (e.g., to model the arrival of potentially dangerous passengers at U.S. airports, it is important to consider the weather conditions along the arrival routes, ticket prices, and so on). The supply of reliable data at this level is thus crucial for accurate decision-making. The techniques and approaches to data supply for decision analyses involving multiple domains are not always evident and the joint construction of data of good quality and predictive power is a delicate task.

III. METHODOLOGICAL FRAMEWORK

Cross-domain decision support systems involve concurrent long-lived pipelines with processes that can be ingested and accessed easily by consumers, either human or machine. Typically, these long-lived processes comprise a series of data preparation operations or data preparation pipelines (normally within batch and near real time), implemented as Extract, Transform and Load (ETL) and Extract, Load and Transform (ELT) processes. The data is refreshed periodically by producers.

In the data ingestion and integration layer of a cross-domain pipeline architecture, the extracted data can either be ontological or database data or both. Multiple ontologies and schemas across different domains can be an impediment to interoperability. Ontology mapping can be complex, time-consuming and challenging for analysts. On the database side, the schemas of the underlying databases are not semantically enriched. Therefore duplicate data in the ETL process may not be detected and removed. A public database such as the Global Terrorism Database maintains data with no missing values, and data cleaning and preparation operations are not required during ingestion.

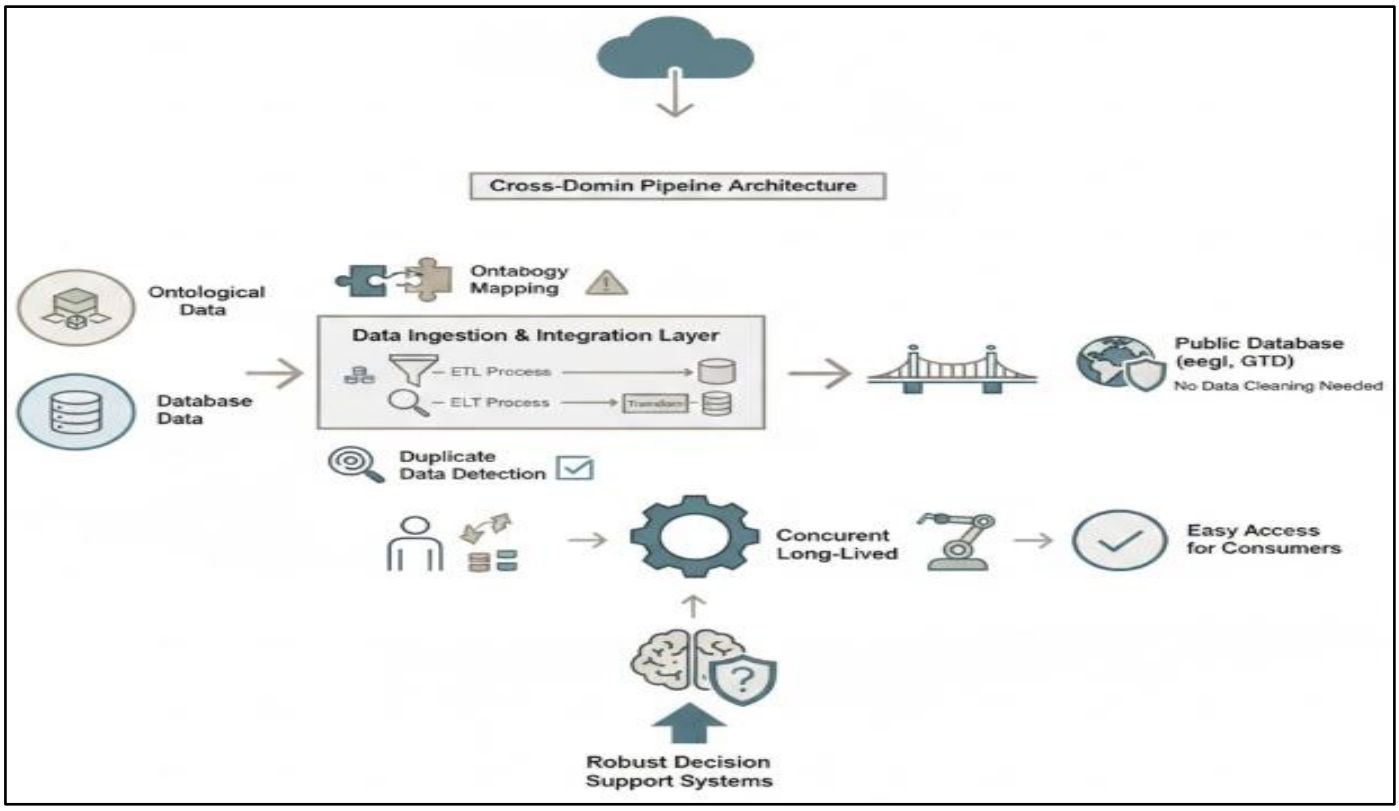

Fig 3 Optimizing Semantic Interoperability in Long-Lived Cross-Domain Pipelines: A Hybrid Ontological-Relational Ingestion Framework

➢ *Data Ingestion and Integration*

Data ingestion and integration form the foundation of cross-domain decision support pipelines. The decision support resources must be continuously and reliably supplied with the data required for the decision-critical processes. Data providers can include web-based APIs, traditional database systems, and file systems. The schemas of ingested data can be highly heterogeneous, and therefore the data must be transformed, normalized, and enriched before being made available to decision-critical systems. For optimal performance and responsiveness, data ingestion must be performed with an ETL or ELT approach, depending on the underlying architectures of both the data sources and destinations.

A successful strategy for Cloud-based decision support must address data ingestion in an appropriate manner. Adequate resource contracts guarantee that the data pipeline is sufficiently fed in a timely manner, and that the data supplied is reliable from an informational standpoint. In such a view, a data contract can be seen as a form of governance between data providers and data users. Data contracts specify when and how often data is available, as well as the level of quality, completeness, and currency of data produced; they should hence be defined jointly between data providers and users to document the trustworthiness of the data. Data should never be ingested for the sake of it; data must be ingested only when then can in any shape or form enrich the decision-making process.

➢ *Data Processing and Transformation*

Being the second part of a multi-part chain, the Writing Guidelines are adopted from the first part. A Cloud-Centric Approach to Big Data Pipelines for Cross-Domain Decision Support Cross-domain decision support relies on pipelines that automatically collect, preprocess, and process data related to multiple domains using services in the cloud. However, either the ingestion and integration, cross-domain data governance and quality, or deployment model aspects are not covered in sufficient

detail. Previous work provide detailed descriptions or configurations of pipelines supporting the ingestion and integration of data, but other areas are often overlooked. A comprehensive description of the entire process of preparing and using a cloud-centric pipeline for cross-domain decision support is therefore both relevant and timely. This paper focuses on the ingestion and integration of the data and the pipeline architecture and discusses data processing and transformation, cross-domain data governance and quality, and cloud-centric deployment models. The work is based on practical work done for the Traffic Management Open Data Portal of Norrköping, Sweden.

The Data Processing and Transformation phase makes the Decision Support Completely Automatic and Continuous, focusing on the processing components. Decisions may include such features as alerting when specified thresholds are exceeded and additionally outputting notifications to operational systems. Processing engines can handle either a Stream or a Batch data flow and can be combined when the underlying data sources support such an architecture. Data can enrich when specific external data become available, such as weather forecasts. The result set of the processing can additionally be used as an upstream source for further enrichment or made available for other future use cases in which a reprocessing is not needed. Methods for normalising concepts from different domain-specific vocabularies and ontologies can also be applied.

IV. ARCHITECTURAL PATTERNS AND REFERENCE ARCHITECTURES

Pipelines are layered architectures that expose general processing capabilities at different abstraction levels, with each layer handling a specific aspect of the processing workload or serving a particular class of OP to downstream consumers. The market offers a wealth of such layered pipeline architectures that follow different architectural patterns. Although the individual components of these layered references architectures differ, they all serve the same purpose: implementing specific architectural patterns that introduce reusable capabilities in large-scale data processing pipelines.

Workflow engines for batch-oriented workflows (e.g., Apache Oozie), complex event processing systems for stream processing (e.g., Apache Storm), data processing and query engines (e.g., Apache Hadoop, Apache Spark), long-term storage and archival of data (e.g., Apache HDFS, Amazon S3), and NoSQL databases for data warehousing (e.g., MongoDB, Amazon Redshift) all serve the same duty of implementing one type of layered pipeline architecture or another—enriching the processing supply landscape and making available, at low implementation costs, capabilities that must be best-in-class to justify use in time-critical operational settings. Even so, they are not designed around the principles of Pipeline Downstream Supply and cannot be efficiently composed or simply deployed in a cloud-centric authentication-and-authorization settings.

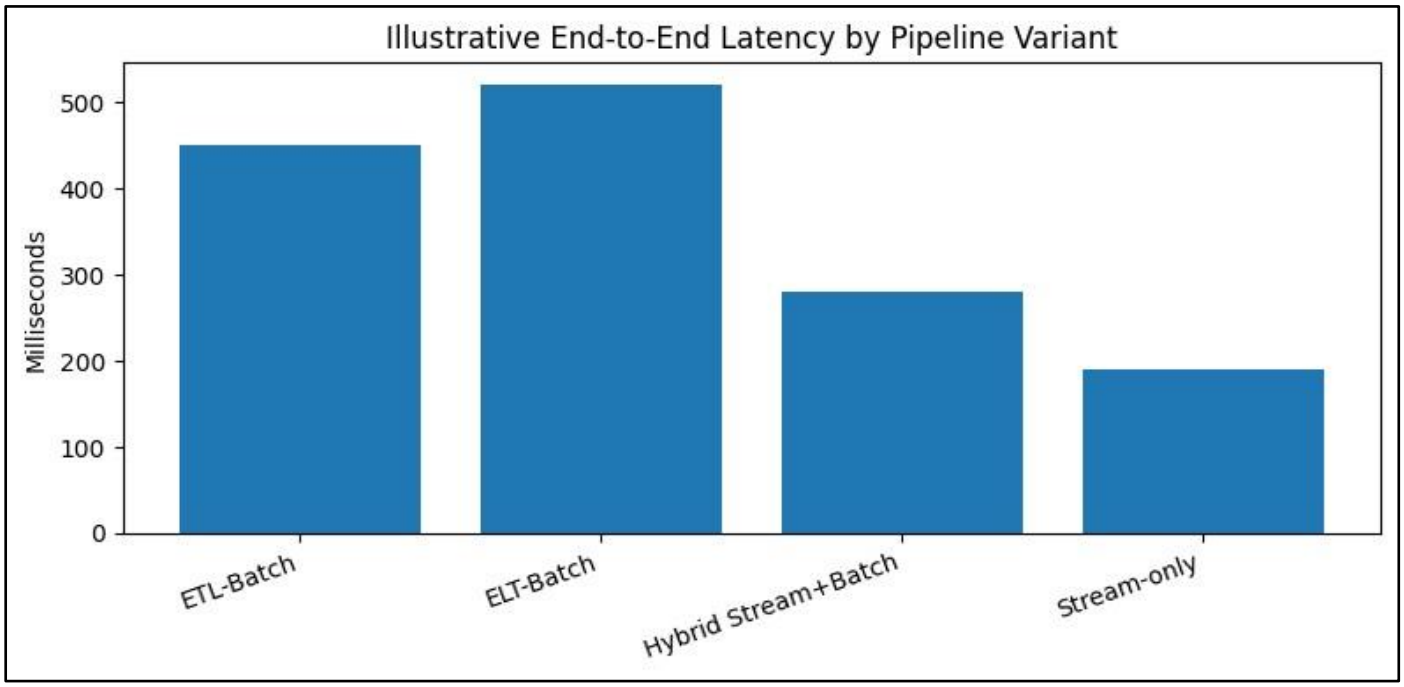

Fig 4 F-Measure (F1-score)

Start from harmonic mean definition for two numbers P and R:

$$H(P,R) = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$

Substitute P = Precision, R = Recall:

$$F1 = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$

Bring to a common denominator:

$$\frac{1}{P}+\frac{1}{R}=\frac{R+P}{PR}$$

So:

$$F1=\frac{2}{\frac{P+R}{PR}}=2\cdot\frac{PR}{P+R}$$

Final:

$$\boxed{F1=\frac{2PR}{P+R}}$$

➢ *Layered Pipeline Architectures*

A layered architectural pattern for cloud-centric big data pipelines identifies the logically distinct layers of such solutions and specifies the responsibilities of each layer along with the interfaces exposed to subsequent pipeline layers. Each layer also encapsulates one or more sub-pipelines designed to operate on data flowing through the layer. Data flows through the pipeline from top to bottom, enabling operations to be executed on the data either in streaming mode or batch mode. Multiple data sources are allowed to share a common pipeline layer, while multiple sink options configured at downstream pipeline layers specify how the output data should be routed.

Cloud-centric big data pipelines apply a layered architectural pattern that encapsulates all layers interacting with individual data sources in a dedicated sub-pipeline. A dedicated command and control layer centrally governs the operation of all pipeline layers. The logical organization of the cloud testbed for proof-of-concept implementation follows an analogous approach, grouping logically related data pipelines into distinct pipeline stages adapted for data ingestion and integration, data processing and transformation, and data governance to support decision-making using cross-domain data.

V. CROSS-DOMAIN DATA GOVERNANCE AND QUALITY

Although cross-domain decision support architectures facilitate a diverse set of data producers and consumers, the actual data ingestion process requires tighter control; otherwise, the risk of bad data polluting the data pipeline increases. Such data pipelines often serve a broader audience than traditional data warehouses or lakes, with lower barriers to entry for querying. To this end, a strong focus on data quality, compliance, security, and governance is necessary.

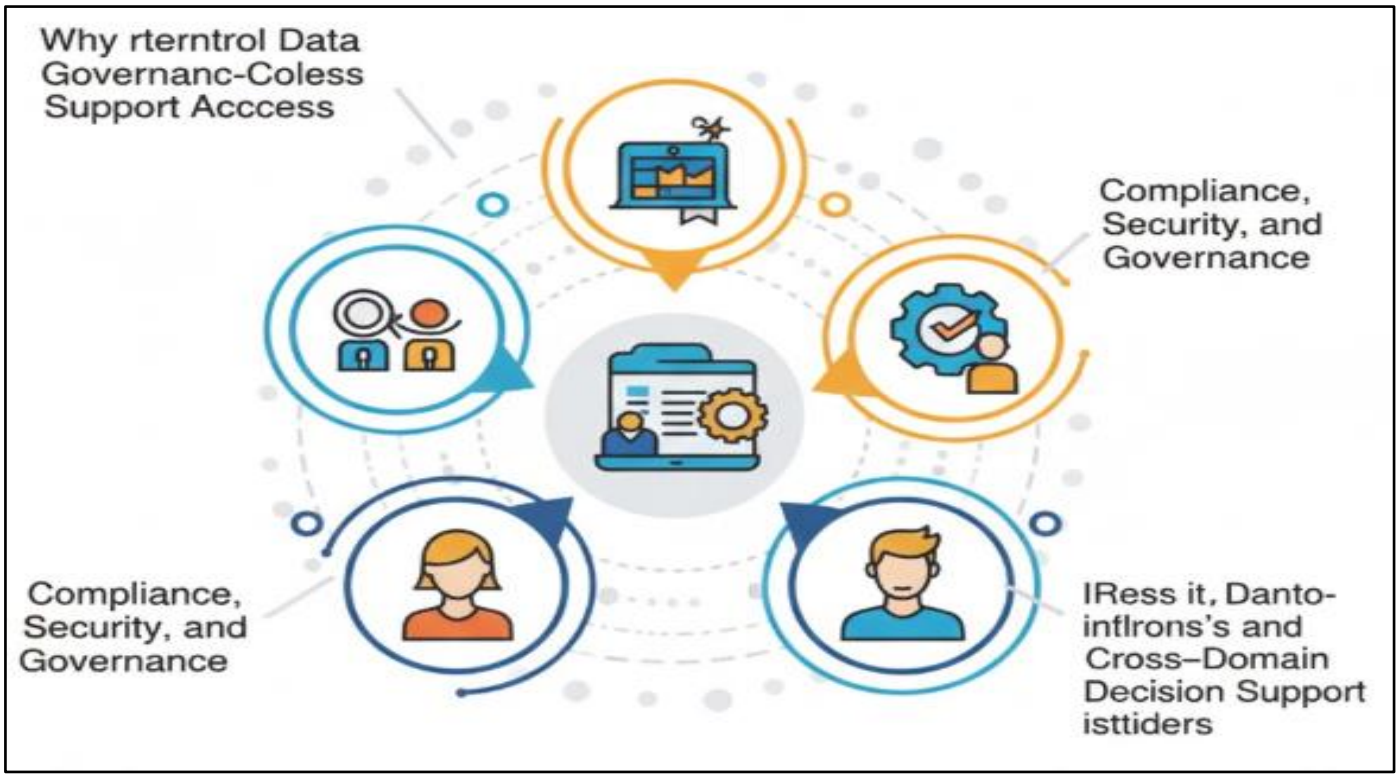

Fig 5 Securing the Pipeline: A Robust Governance Framework for End-to-End Data Quality and Compliance in Cross-Domain Decision Support

Maintaining data quality and legal compliance in cross-domain environments where data producers and consumers do not belong to the same organization is especially challenging, as organizations are rarely motivated to invest in data quality, security, or compliance of data consumed by external parties. A malicious consumer of data can just as easily query the data as an approved consumer, bypassing any data access restrictions put in place. Governance of the entire end-to-end pipeline—including data ingestion, processing, and consumption—helps to establish control over these challenges. In particular, metadata capture, lineage, governance policies, compliance checks, and external

feedback are all relevant for maintaining data quality and compliance across domains.

➢ *Metadata and Lineage*

To fulfill the data quality assurance aspect of cross-domain value-adding applications, the proposed pipelines need to incorporate metadata management and lineage monitoring capabilities. The absence of such capabilities can jeopardize several aspects of applying the data to decision support, notably addressing legal requirements on disclosure of metadata (e.g., GDPR), the correct attribution of data used to conduct the decision-making analysis, proof that the data satisfies quality requirements for the specific usage (e.g., completeness, timeliness), and proof that the data can be used for the purpose considering its origin and potential limitations.

The first two aspects of fulfilling these requirements rely on the design of a metadata management strategy, while the remaining two can be addressed through the adoption of adequate monitoring of the data lineage preserved across the different operations performed in the pipeline—both for the purpose of steering data through the data quality assessments defined by the monitoring strategy and of providing decision makers with proofs of compliance and correct usage. The definitions of the metadata management strategy and the data lineage capture plan can rely on the same governance policy and compliance framework previously defined for the ingestion process, since similar aspects are driven by the presence or absence of a data quality assurance process and by the presence or absence of a compliance-checking process.

VI. CLOUD-CENTRIC DEPLOYMENT MODELS

Public, private, and hybrid cloud deployment models can be used for Cloud Computing-based Data Pipelines, both for Data Lakes and for traditional Data Warehouses. The choice of model determines how the full Data Pipeline, Data Lake or Data Warehouse is deployed in the Cloud. Organizations often use a model based on deployment decisions (private, public, or hybrid) closer to the periphery of the Data Pipeline and a model closer to the core of the Data Pipeline based on other criteria (for example, cost, security, and latency). For instance, the core of a Data Lake may be on a public Cloud (e.g. Amazon S3) while Data Sources are on Private Clouds and some Data Pipelines reside on Private Clouds closer to these Data Sources. Such configuration reduces latency and incurs minor cost while still benefiting from the scale of public Cloud resources.

Public Cloud Service Providers, such as Amazon, provide massive resources on a pay-per-use basis. However, organizations are often reluctant to transfer sensitive Data, such as that in Banking and Financial Services, to public Clouds because of privacy and regulatory requirements. Private Clouds provide the control and security desirable for Business Intelligence decision Making, but at high cost. A Hybrid Cloud model combines the best from both models, often at lower cost and higher security, but planning and governing a Hybrid Cloud model is significantly more complex.

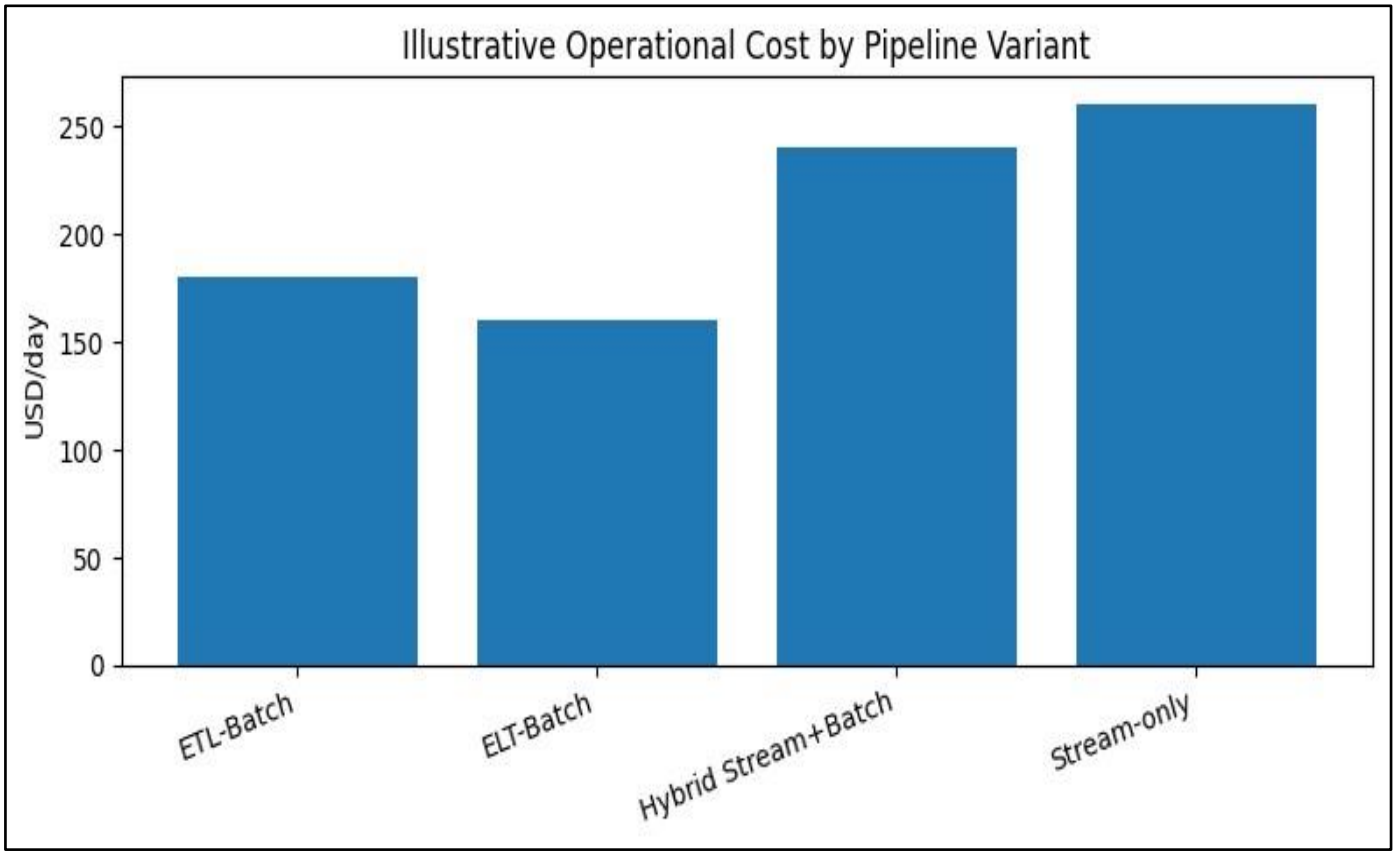

Fig 6 Practical “Math Formalizations” for Other Paper Concepts (Not Explicit in Text)

➢ *Data Freshness / Timeliness (for Continuous Pipelines)*

If a record is produced at time t_{prod} and consumed at t_{cons}:

$$\text{Age} = t_{cons} - t_{prod}$$

A simple timeliness score (1 is best) using a maximum acceptable age Δ_{max}:

$$\text{TimelinessScore} = \max\left(0,\ 1 - \frac{\text{Age}}{\Delta_{max}}\right)$$

➢ *Data Completeness*

If a dataset has M required fields per record and n records:

Let $missing$ be total missing required values.

$$\text{Completeness} = 1 - \frac{missing}{n \cdot M}$$

This supports "quality requirements" like completeness mentioned for governance .

➢ *Pipeline Throughput*

If you process R records in time T:

$$\text{Throughput} = \frac{R}{T} \quad \text{(records/sec)}$$

➢ *End-to-end Latency Decomposition*

For ingestion L_{ing}, transform L_{tr}, governance checks L_{gov}, serving L_{srv}:

$$L_{e2e} = L_{ing} + L_{tr} + L_{gov} + L_{srv}$$

This quantifies what the paper describes as layered pipelines and governance layers .

➢ *Cost Model (Cloud-Centric Deployment)*

For a time window T, sum compute + storage + network:

$$C_{total}(T) = C_{compute}(T) + C_{storage}(T) + C_{net}(T)$$

A simple compute example if you pay per vCPU-hour:

$$C_{compute}(T) = (\text{vCPU-hours in } T) \cdot (\$/\text{vCPU-hour})$$

➢ *Public, Private, and Hybrid Clouds*

Concrete comparisons of the three cloud deployment models inform decisions about provisioning and management. Public clouds are typically operated by third parties for monetary gain. They are multi-tenant, allowing multiple companies and organizations to share resources without sacrificing performance or security. This is accomplished through segmenting the infrastructure and provisioning virtual resources on-demand through a self-service GUI. Such solutions are oversized for small-scale operations and the lack of control can render them especially unsuitable for cross-domain applications involving sensitive data. Furthermore, service-oriented architectures are not always compatible with multi-tenant resources. Even if steps like encryption are taken, the risk of leakage always persists when sending data through non-trusted, third-party servers.

With private clouds, the infrastructure is owned by a single organization. While total control does enhance data protection, on-premise deployments also require significant investment and maintenance overhead. Private clouds operated by third parties offer an alternative—data remains on the provider's premises, but the resources are rented instead of purchased. Hybrid solutions aim to capitalize on the advantages of both models, wherein sensitive data is kept in-house while publicly-accessible data is deployed on external infrastructure. This combination greatly simplifies cross-domain interoperability. Nevertheless, the public part remains vulnerable; therefore, authentication, encryption, and thorough vetting are essential considerations whenever storing sensitive data externally.

VII. EVALUATION FRAMEWORKS AND METRICS

A cross-domain decision support pipeline that builds oracle-like systems capable of answering queries involved in decision-making across domains must also perform well according to measures relevant to the target application. Because the target applications typically involve answering ADHOC queries requiring near real-time semantics, a combination of Precision, Recall, and F-Measure are the key metrics that are of interest. Therefore, the candidate solution must be benchmarked against other approaches using these metrics. Social Media content is critical for decisions within several domains, including Tourism, Politics, and the Stock Market, and any solution capable of predicting metrics in one or more of these domains should be validated through Decision Support Effectiveness testing. Time-series observations from domains like Tourism, Politics, and the Stock Market can be considered in creating cross-domain pipelines with social media content in place of domain-centric content, and the cross-domain pipeline must, therefore, be evaluated using Decision Support Effectiveness Testing.

Testing is a method used to assess the correctness of pipeline systems, including those constructed for two domains: Tourism and Politics, as well as two other combined-domain pipelines. Because cross-domain pipelines that include Social Media content are conceptually similar to Cross-Statistical Domain Transfer Learning systems, Cross-Domain Pipeline Testing must also be considered. Cross-Statistical Domain Transfer Learning solutions parallel Spatial Domain Transfer Learning solutions and connect logically with the principle of Not More Than Necessary.

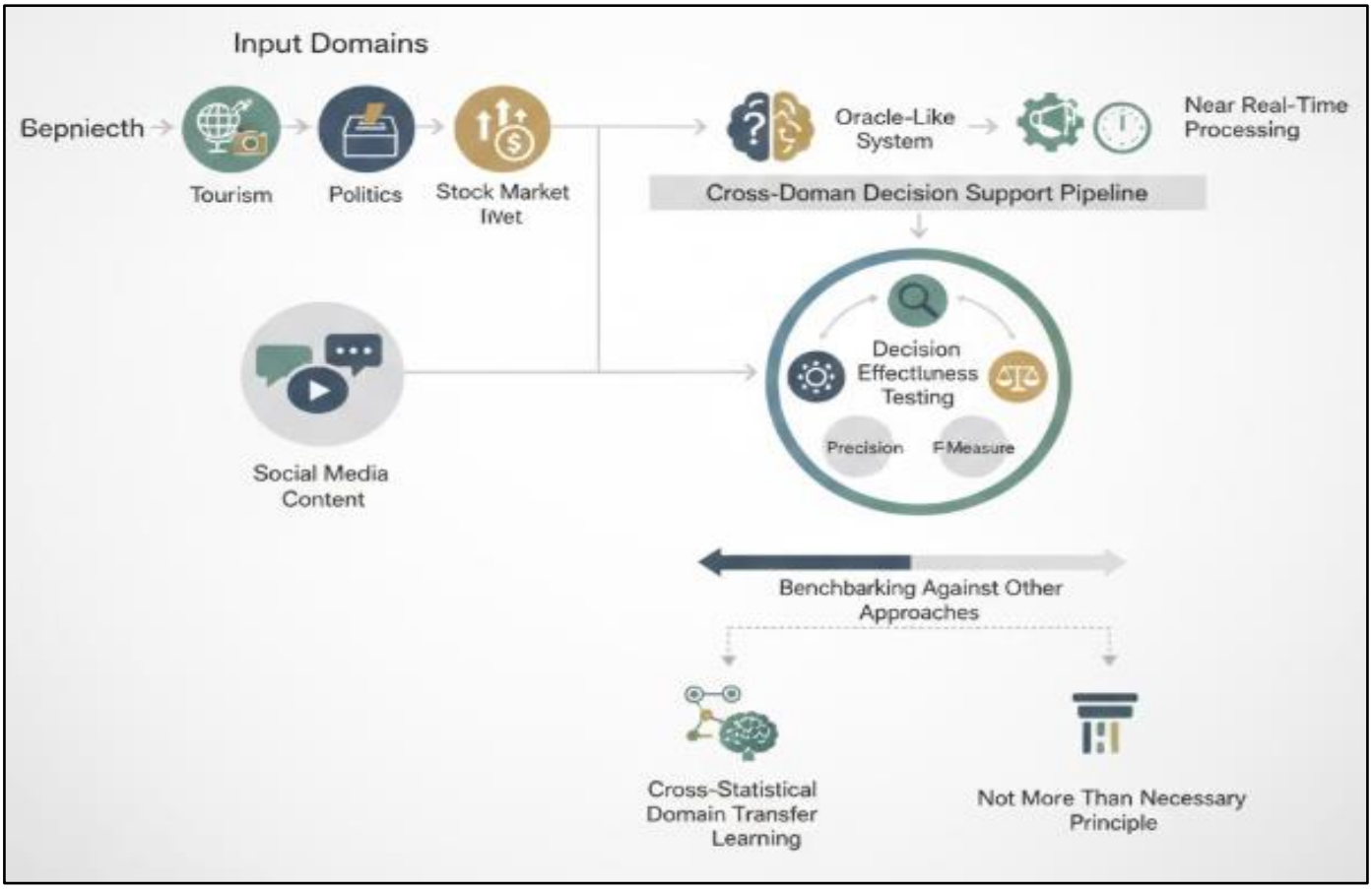

Fig 7 Benchmarking the Oracle: A Transfer Learning Framework for Evaluating Decision Support Effectiveness in Cross-Domain Social Media Pipelines

➢ *Decision Support Effectiveness*

The effectiveness of automated decision support for cross-domain and multi-stakeholder scenarios can be evaluated with generic benchmarking frameworks and tailored effectiveness metrics. Cross-domain decision support aims at the integration of data from a diverse number of origins with heterogeneous semantics for multiple domains in order to transform, validate, and deliver data for data-hungry decision-making processes. Accordingly, the adequacy-oriented evaluation of the conducted approach is twofold: both the individual data quality of each of the constituent domains and the support these data sources provide for generic automation of complex queries are of primary consideration.

Several data quality metrics are used to assess the information. For the purpose of determining the adequacy of the candidate solution for supporting time-series intelligent queries, the information in its integration is benchmarked against five common intelligent queries centered around time-series closeness, seasonality, and periodicity. The evaluation is performed over real-world datasets, and the results confirm the feasibility and appropriateness of the approach for integrating high-quality information and automatically supporting intelligent queries addressing the decision needs of multiple distinct domains. Real-life cloud-based infrastructures and topologies are leveraged to support the proposed solutions and demonstrate the applicability and practical relevance of the examinations.

VIII. CONCLUSION

Cross-domain decision-support requirements necessitate the implementation of components and services in several business domains using fundamentally different technologies and environments. The cloud-centric approach allows organizations to focus budget, time, and expertise on their own domain while delegating the implementation of the cross-domain interactions to others. The provided discussion identifies the key technical considerations for cloud-centric development and deployment of big data pipelines that support cross-domain decision making. Evaluation issues that address the effectiveness of such a pipeline demonstrate the challenges involved, how these challenges have been overcome in benchmarks, and the implications for the future development of such systems. While many cloud services facilitate the ease of big data development and operations, important quality-monitoring concerns remain essential in any design.

Increased demands for cross-domain decision support, together with greater interdependence among business domains and often limited budgets, have caused organizations to look outside their own walls for such capabilities. The high usage and increase of cloud services have pointed an affordable route for the development and implementation of pipelines capable of generating decision-critical data across diverse domains. Such services are key enablers for the implementation of the requirements of organizations that are core suppliers, core

buyers, or regulators within the general guidelines of the Business Model Management Perspective. The quality of these support pipelines remains essential to ensure that the resources used for such services are not wasted.

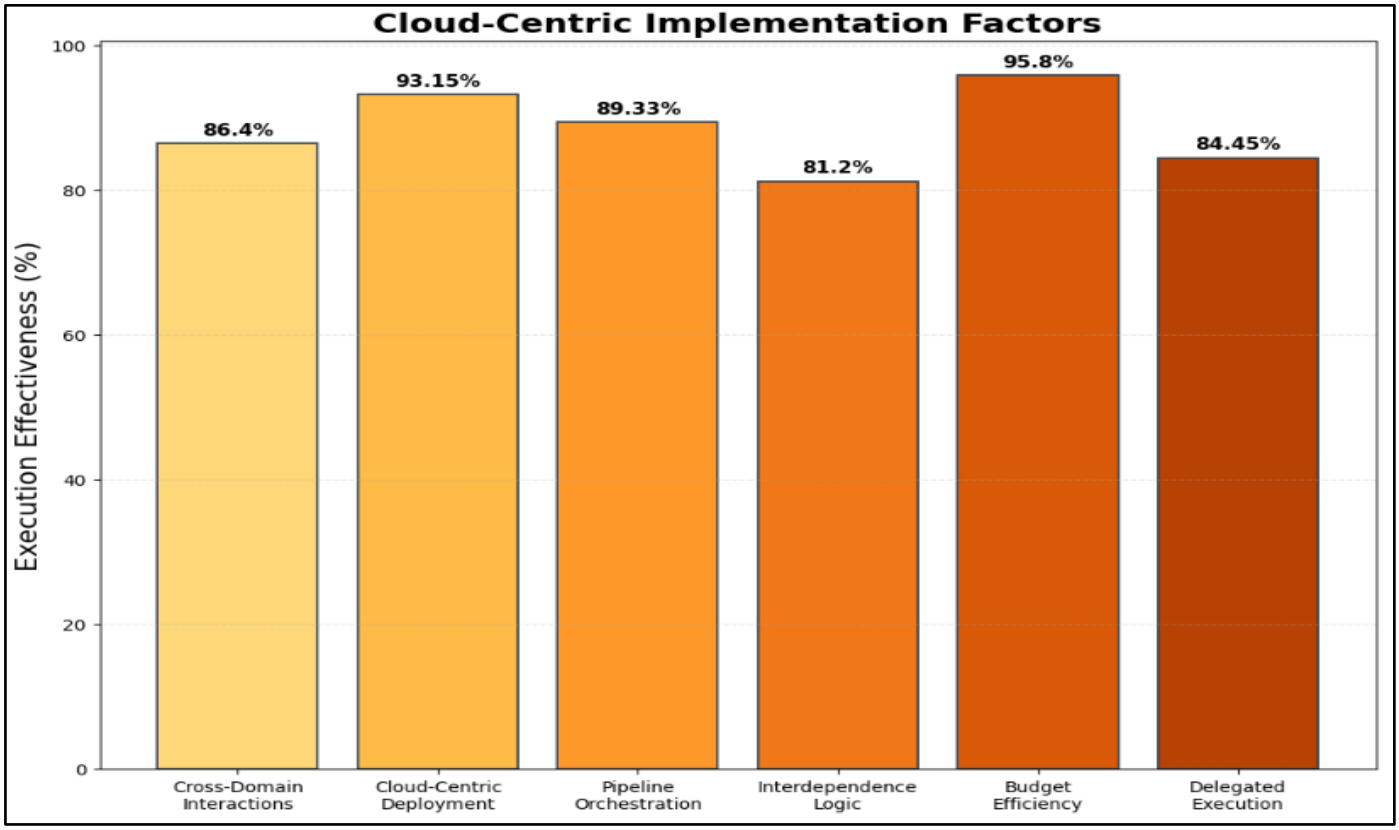

Fig 8 Cloud-Centric Implementation Factors

➢ *Final Thoughts and Future Directions*

The proposed cloud-centric approach to computer-assisted cross-domain decision support represents the foundation for further exploration of the three-layer reference architecture for big data pipelines, including the design and development of a set of concrete solution patterns, specialized cloud pipelines tailored to specific sectors, as well as the function and form of cloud-centric data governance. In addition to bolstering enhancement of such pipelines in the three defined directions, further research should establish a framework for the effective evaluation of cloud-centric big data pipelines for cross-domain decision support.

Decision support constitutes one of the major areas of DSDB, and within the application-oriented research perspective, cross-domain decision support has been identified as the first specific direction for the design and development of cloud-centric big-data pipelines. In addition to identifying sources of external data for cloud pipelines that facilitate cross-domain decision support, and selecting appropriate ETL technologies to populate the cloud-based data repository, further research is needed to address key issues, including— for cross-domain decision support: What decision-critical information is required from each domain/sector?, To what depth, precision, and detail must this information be integrated in order to enable cross-domain decision support?, What is the value of such integrated support across domains/complex systems?, What is the extent of enhancement for such integrated decision support over individual domain support ecosystems?

REFERENCES

[1]. Varri, D. B. S. (2022). AI-Driven Risk Assessment And Compliance Automation In Multi-Cloud Environments. Journal of International Crisis and Risk Communication Research , 56–70. https://doi.org/10.63278/jicrcr.vi.3418.

[2]. Alaimo, C., Kallinikos, J., & Valderrama, E. (2021). Platforms as service ecosystems: Lessons from cloud data infrastructures. Journal of Information Technology, 36(1), 3–20.

[3]. 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).

[4]. Beyer, M. A., & Laney, D. (2020). The importance of data integration in analytics-driven enterprises. IEEE Computer, 53(6), 62–66.

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

[6]. Chen, M., Mao, S., & Liu, Y. (2019). Big data: A survey. Mobile Networks and Applications, 19(2), 171–209.

[7]. 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.
[8]. Dayarathna, M., Wen, Y., & Fan, R. (2020). Data center energy consumption modeling: A survey. IEEE Communications Surveys & Tutorials, 18(1), 732–794.
[9]. Nagabhyru, K. C. (2022). Bridging Traditional ETL Pipelines with AI Enhanced Data Workflows: Foundations of Intelligent Automation in Data Engineering. Available at SSRN 5505199.
[10]. García-Gil, D., Ramírez-Gallego, S., García, S., & Herrera, F. (2020). A taxonomy and review of big data integration. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(5), e1360.
[11]. 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.
[12]. Inmon, W. H., & Linstedt, D. (2019). Data architecture: A primer for the data scientist. Academic Press.
[13]. 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.
[14]. Kleppmann, M. (2017). Designing data-intensive applications. O'Reilly Media.
[15]. 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.
[16]. Lenzerini, M. (2002). Data integration: A theoretical perspective. Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems.
[17]. Rongali, S. K. (2022). AI-Driven Automation in Healthcare Claims and EHR Processing Using MuleSoft and Machine Learning Pipelines. Available at SSRN 5763022.
[18]. López, P., & Lewin, A. Y. (2021). Cloud transformation and enterprise analytics capabilities. Information & Management, 58(7), 103489.
[19]. Pandiri, L. The Future of Commercial Insurance: Integrating AI Technologies for Small Business Risk Profiling. International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), DOI, 10.
[20]. Mell, P., & Grance, T. (2011). The NIST definition of cloud computing. National Institute of Standards and Technology Special Publication 800-145.
[21]. Koppolu, H. K. R., Recharla, M., & Chakilam, C. Revolutionizing Patient Care with AI and Cloud Computing: A Framework for Scalable and Predictive Healthcare Solutions.
[22]. Naldi, M., & Mastroeni, L. (2020). Big data analytics in the cloud: Architecture and governance. Information Systems Management, 37(1), 21–35.
[23]. Gadi, A. L., Kannan, S., Nandan, B. P., Komaragiri, V. B., & Singireddy, S. (2021). Advanced Computational Technologies in Vehicle Production, Digital Connectivity, and Sustainable Transportation: Innovations in Intelligent Systems, Eco-Friendly Manufacturing, and Financial Optimization. Universal Journal of Finance and Economics, 1(1), 87–100. Retrieved from https://www.scipublications.com/journal/index.php/ujfe/article/view/1296.
[24]. Pääkkönen, P., & Hellsten, S. (2020). Data quality challenges in cloud-native data platforms. Journal of Data and Information Quality, 12(2), 1–23.
[25]. Paleti, S. (2022). Financial Innovation through AI and Data Engineering: Rethinking Risk and Compliance in the Banking Industry. Available at SSRN 5250726.
[26]. Stonebraker, M., Abadi, D., DeWitt, D., Madden, S., Paulson, E., Pavlo, A., & Rasin, A. (2018). MapReduce and parallel DBMSs: Friends or foes? Communications of the ACM, 53(1), 64–71.
[27]. Pallav Kumar Kaulwar, "Designing Secure Data Pipelines for Regulatory Compliance in Cross-Border Tax Consulting," International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering (IJIREEICE), DOI 10.17148/IJIREEICE.2020.81208.
[28]. Vassiliadis, P., Simitsis, A., & Skiadopoulos, S. (2019). Conceptual modeling for ETL processes. Data & Knowledge Engineering, 122, 38–58.
[29]. 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.
[30]. Zhang, Q., Chen, M., Li, L., & Li, S. (2022). Cloud-native data governance for enterprise analytics. IEEE Transactions on Cloud Computing, 10(4), 2431–2444.