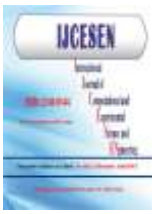
International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.4 (2025) pp. 10145-10155
http://www.ijcesen.com



ISSN: 2149-9144

Research Article

Metadata Driven Optimization of Distributed ETL Pipelines in Cloud Native Data Warehouses

Jitendra Gopaluni*

University of Houston – Clear Lake, Houston, Texas
* **Corresponding Author Email:** gopaluni1003@gmail.com **-ORCID:** 0009-0000-1446-5413

Article Info:

DOI: 10.22399/ijcesen.4742
Received : 01 March 2025
Accepted : 30 March 2025

Keywords

AI-driven optimization,
metadata-driven ETL,
cloud-native data warehouses,
distributed data pipelines,
data governance

Abstract:

The data proliferation of distributed and cloud-native systems has altered organizational approach to Extract, Transform, and Load (ETL) pipelines to support analytics and decision-making. Conventional ETL frameworks based on fixed and script-based processes become scalable, maintainable, and real-time flexible in multi-cloud environments. Our review explores the development of metadata-based ETL systems. It could separate the logic and implementation of pipelines by externalizing transformation policies, lineage, and governance policies into structured metadata. These architectures enable dynamic reconfiguration, automation, and optimization of ETL processes. It helps to promote the agility and scalability in cloud-native data warehouses such as Snowflake, Databricks, and BigQuery. The paper summarises the existing developments in distributed ETL optimization, such as metadata-aware orchestration, AI-based performance tuning, and predictive workload balancing. It discusses metadata lifecycle management, lineage tracking, and security compliance governance mechanisms through such frameworks as Apache Atlas and Azure Purview. Moreover, we mention the new trends connected with Generative AI to modernize ETL, self-healing cognitive pipelines, and sustainable metadata management with the principles of green computing. This paper will show that metadata-based design can turn ETL systems into self-optimizing data pipes. Those were adaptive and autonomous, by collating the results of scholarly and industry research. The combination of artificial intelligence and metadata governance creates the basis of the future generation of intelligent, interoperable, and sustainable cloud data ecosystems.

1. Introduction

The digital transformation of companies has triggered an unprecedented influx of data, associated with large volume, diversity, and speed. The data explosion has been driving the development of data-management architecture. The cloud-native data warehouses with the building blocks of scalable, flexible, and efficient analytics. The primary component of these contemporary data ecosystems is extract, transform, load (ETL) pipelines, which coordinate the transfer and transformation of heterogeneous origins into formatted repositories. The conventional data warehouses offer a high level of data governance and scalability needs of modern applications. The cloud-native and distributed frameworks provide the much-needed flexibility and scalability to support heterogeneous datasets. The following complexities are created by metadata management to ensure data quality, lineage, and interoperability [1].

Metadata has allows automated schema mapping, data lineage, and adaptive workflow coordination. It is highly essential to maintain the pipeline integrity and efficiency in a distributed environment. The metadata management promotes the FAIR principles of making data findable, accessible, interoperable and reusable for best practices in data-intensive fields. Despite its significance, inconsistent metadata management has been a major issue that prioritize flexibility and cost-efficient scalability. The quality of data pipelines depends on a myriad of factors such as data types, data infrastructure, the lifecycle management, and stages of processing. Mistakes like inappropriate types of data and compatibility of the issues process the data cleaning and integration. These difficulties explain the effective metadata-driven solutions are needed to detect and address

data quality issues and improve the performance of ETL processes [2]. Also, the growing sophistication of data pipelines due to the incorporation of various data streams and the implementation of sophisticated analytics makes continuous optimization of rich metadata. The emergence of Industry 4.0 in industries like agriculture, there has been increased pressure on having efficient data-management solutions of data produced by the IoT devices and sensors. The cloud-based aarchitecture have been found to be necessary for the special needs of applications such as low-latency, high-throughput, and scalability. The comparative studies emphasize the trade-offs between the user proximity, network stability, reliability which are determined by the underlying metadata management strategies [3]. The transition capabilities of the research prototype into an industrial-scale solution is dependent on proper coordination and optimization of metadata-based ETL pipelines.

The introduction of distributed data stream processing systems, such as Apache Storm, Spark streaming, Flink, and Kafka streaming of ETL. These models provide real-time analytics, which require low-latency and high-throughput networks, and fault-tolerant networks of processing data immediately upon its reception. The choice and optimization of appropriate stream processing structures depend on the presence. The quality of metadata that guides the resource allocation, fault recovery, and performance tuning [4]. The importance of metadata and optimizing distributed ETL pipes is more critical as organizations and use of real-time analytics. The scheduling of task assignments is optimized in cloud-native environment; the multi-objective algorithms can be used to trade competing objectives like makespan and resource use. The effectiveness of these optimization strategies is strictly connected with the availability of the comprehensive metadata, which informs the decision-making processes. The facilitates the adaptive reaction to the changes in workloads and the state of the system [5]. The developing cloud computing, the combination of sophisticated optimization algorithms with metadata driven ETL pipelines will be necessary in the operational excellence. This review aims to summarize existing studies on metadata-based optimization techniques, outline the most important challenges, and point out future outlooks on the evolution of the distributed ETL pipeline management.

2. Metadata-Driven Optimization of Distributed ETL Pipelines: Key Concepts and Trends

2.1. History of ETL and Data Integration Frameworks

The development of ETL (Extract, Transform, Load) and data integration models reflects the paradigm shift in data management systems based systems to the recent cloud-native, metadata-driven systems. Originally, ETL systems designed to run on-premise were designed to be used in batch processing, with logic coupled and hard to scale with the increase in data volumes [6]. The introduction of ELT (Extract, Load, Transform) shifted the workload of transformations to the data warehouse itself taking advantage of the potential of modern databases and cloud storage to process larger. Also, it enhances the more diverse volumes of data and the flexibility and scalability of the transformation process. The growing data diversity and speed, institutions moved to the distributed ETL models, which utilize distributed computing and parallel processing to enable real-time analytics. These systems allowed the consumption and processing of structured and semi-structured data at scale, but often required manual configuration and maintenance. The latest stage of this development is the adoption of cloud-native and metadata-driven ETL systems. These systems separate pipeline logic and execution by refining transformation rules, mappings, and orchestration to structured metadata at dynamic runtime [7]. Such an approach will make it easier to adjust to changing business needs quickly, enhance automation and improve scalability. Data integration approaches based on metadata have taken the centre-stage as organisations of their modernise data integration strategies and facilitate advanced analytics of self-service access to data [8].

2.2 Data engineering metadata

Metadata is the foundation of modern data engineering that offers data assets the necessary context and organization. It is generally divided into three main types: technical metadata, which includes the structure, schema, data types, lineage and storage information. Thus allowing automated schema mapping, data validation and impact analysis modified; operational metadata, which includes process execution information, performance indicators, workflow status and error logs. Thus facilitates the lossless linkage between technical implementation and business requirements which includes business rules of usage constraints. Thus bridging the gap between technical implementation and business requirements and supporting the appropriate governance and compliance. It has automated data

discovery, cataloguing, and lineage tracking capabilities, which minimize manual work and improve the quality of data [9]. AI-based metadata management tools can automatically create, classify, and revise metadata. Further, the automating processes enable the real-time integration and governance. The intricate data ecosystem, metadata management is essential to interoperability, regulatory adherence, and effective data operations [10].

2.3 ETL Architectures that are metadata-driven

The basic characteristic of the metadata-based ETL systems is the separation of pipeline logic execution which enables on-demand adaptation and re-use. Such systems are represented using the pipeline definitions of structured metadata (comprising transformation rules, data mappings, and business logic) and executed via execution engines during data processing. This kind of division gives organisations the capability to easily partition pipelines and apply common standards without significant changes in code [11]. The metadata-based systems allow to have data quality control and regulatory compliance and data resources optimization as business rules and governance models are directly encoded within pipeline metadata. This will reduce the overhead cost of development, raise efficiency in maintenance as well as bring flexibility in the changing business environment. Metadata as a form of externalisation enables organisations to develop pipelines simplify the process of bringing new sources that create a platform to build additional data programmes, including AI-based analytics and self-service access to data [12].

2.4 Distributed and cloud-native data warehouses

The cloud-native data warehouses have changed the organizations save, process, and analyze data. It includes Snowflake, Google BigQuery, Databricks, and AWS Redshift, which provide elastic computing power, scaling storage, multi-tenant features, and the heterogenous data sources [13]. Snowflake has a multi-cluster, shared-data architecture, which decouples storage with compute. It allows extreme scalability to store the structured and semi-structured data. Its pricing strategies and a strong security posture make it applicable to both small and large organizations. The Google BigQuery is a serverless, fully administered analytics engine that delivers high-performance and emerges as an affordable choice. It copes with real-time analytics and integrates with a vast collection of Google Cloud services [14].

The databricks brings together data engineering, machine learning and analytics into a unified platform that facilitates real-time processing of data and superior metadata management. AWS Redshift provides scalable/managed data warehousing which has full integration within the AWS environment supporting both traditional and modern analytics workloads. The platforms use metadata to optimize workloads, control costs, and automate governance, becoming a key part of the current data integration strategies. It may promote higher features like time travel, sharing of data and multi-cloud interoperability, which adds to their appeal to data-driven businesses.

2.5 Metadatabased optimization research trends

Metadata-driven optimization is the focus of modern research with a focus on automation, AI-orchestrated, and durable metadata governance models. ML and AI are used to create metadata that enhance the data discovery and streamline pipeline execution. The self-healing pipelines, predictive resource management, and inbuilt governance models are the best practice which transforming the paradigm to reactive forms of data management [15].

The constant metadata acquisition and real-time integration which allows responding to changing loads with the advanced analytics. Successful metadata management provides data quality, security, and compliance, resolves the issue of interoperability and standardization. The interplay between these trends drives the evolution of smart, self-optimizing data platforms decreasing manual intervention and the operational costs. Metadata management frameworks using AI are also being developed to address data quality, ethical issues and technical limitations. Such frameworks use advanced AI technologies to either automatically generate metadata, increase resource visibility, and effectively manage large collections of digital content [16].

The future trend is the continued use of linked data, predictive analytics, and real-time ontology updates, which will further enhance the capabilities of metadata-based systems. Optimization of distributed ETL pipelines in the cloud-native data warehouse is connected with the efficient management and usage of metadata. The data ecosystems becoming more complex and dynamic, metadata-based models provide a better scalability, reliability, and value of analysis. The convergence of automation, AI, and well-developed governance systems is determining the future of data

engineering, which allows organizations to gain a better value of their information resources and reduce the manual intervention.

3. Metadata-Driven Architecture and Optimization Framework: Design, Execution, and Cloud Integration

Metadata-based ETL (Extract, Transform, Load) pipeline architecture represents a paradigm shift in the organisations manage, optimise, and scale their data integration processes. This methodology is based on metadata as the most important principle of organisation, which facilitates automation, flexibility and efficiency of distributed and cloud-native networks. It provide a comprehensive analysis of the architectural building blocks, metadata interpretation, optimisation plans, and cloud-native ecosystem integration (Figure 1).

This illustration gives a very broad overview of how metadata controls the ETL process - starting with a heterogeneous source of data, accessed through a metadata abstraction layer, which is executed by an engine transformation engine, and finally stored in a cloud based data warehouse, thus enabling scalable analytics.

3.1 Architectural components

The metadata-based ETL framework is derived on the basis of a layered architecture with the specific task to ensure flexibility, sustainability, and management. Its central point is the metadata repositor of truth of any pipeline settings, rules of transformation, schema definitions, and lineage data along with governance policies. It is not a passive store and updated during the lifetime of the pipeline lifecycle of the business metadata needs. The execution engine is responsible for decoding the metadata and converting it into ETL logic that can be executed. The traditional ETL systems becomes the execution engine of a metadata-based system dynamically forms and executes jobs with respect to the current metadata state. This dynamic behaviour allows the quick adaptation to the changes of the data sources or regulatory restrictions as the logic can be updated by altering metadata instead of writing code again. The layer orchestrates the ETL jobs by scheduling, monitoring and coordinating them. This layer guarantees that dependencies are observed, failures are managed with the workflow. The modern orchestration tools, including Apache Airflow or cloud-native orchestrators, are commonly combined to automate these processes, and the uses of the metadata to power schedule. The governance layer is necessary to incorporate data quality, security, and compliance requirements into the pipeline. Encoding governance rules as metadata allows organisations to the same standards into practice, track data lineage, and enable auditing and regulatory compliance. Such an approach not lowers the risk of operation but creates confidence in the data to be handled [18].

3.3 Interpretation of metadata on the runtime

The distinguishing feature of metadata-driven ETL architectures is interpreted dynamically at runtime. The execution engine reads the metadata repository to get all the information like target schema, transformation logic, business rules and the operational parameters. This metadata is interpreted and used to create the required code or execution plans. The execution engine is configured to adapt to the new structure and uses the relevant transformations and validations, but does not require any manual change of code. This feature is helpful to continuous changes in data sources and needs because it will reduce the downtime and quickening the incorporation of new information resource. It can be improved by middleware components which automate file processing, data conversion and distributed job processing. These middleware layers can be used in cloud-based ETL systems to communicate with different storage to ensure the logic metadata on the whole data landscape [19].

3.3. Optimisation strategies

Metadata based ETL framework realises optimisation in metadata-based strategies through a number of high level strategies that make use of metadata as a performance, scalability and resilience strategy. The system lowers the time to look up used metadata and speed up execution of a pipeline by caching it. The dynamic query optimisation of the most optimal execution paths, changing to suit the distribution of data, and reduces the consumption of resources using the metadata. This holds especially in large-scale distributed systems, where performance bottlenecks can have significant operational effects. Another important optimisation is the schema evolution management. Metadata repositories can trace the schema versions and variations and allow automated adaptation of ETL logic to changes in the source or target schema. The system is able to update transformation rules and mappings with data flow. The strategy minimizes the maintenance cost and facilitates the continuous integration and deployment in data engineering activities. The

adaptive allocation of resources is a significant factor in cloud-native and distributed ETL. The orchestration layer can assign the compute resources to various stages of a pipeline by tracking the operational metadata like job performance measurements and resource usage. For example, distinguishing job clusters for different ETL tasks and duplicating resources to match the workload can significantly reduce memory usage and costs. The practical implementations have shown that these kinds of strategies can enhance the efficiency of memory utilisation by 34 to 78%, with improvements and cost reductions. Another recent development is AI-based optimisation where machine learning models are used to predict resource requirements, optimise scheduling and identify anomalies before influence pipeline execution. The self-tuning systems can reduce human error and improve the performance and reliability of ETL [20].

3.4 Cloud-native-ecosystem integration

The combination of metadata-based ETL systems and cloud-native computing are a logical way forward as the modern cloud services provide scalability, flexibility, and automation. The metadata-driven architectures are supported by platforms like AWS Glue, Azure Data Factory, and Databricks, which are suited to deploy the metadata-driven architecture. In such ecosystems, the metadata repository is implemented as a managed service is available and secure. The execution engine has the ability to tap the resources of serverless compute, which is auto-scaling based on the workload requirements. Cloud-native workflow managers orchestrate tasks and are driven by metadata to perform scheduling, dependency resolution, and error handling. Governance is practiced by provide the security and compliance functionalities [21].

Middleware architectures are very instrumental in enable the integration of cloud services. They automate metadata management, file processing and job scheduling and facilitates the efficient processing of large datasets and facilitating real-time analytics. As an example, the Medallion Architecture deployed on top of AWS serverless services (Amazon S3, AWS Lambda, AWS Glue, and AWS Step Functions) can be scaled elastically, has limited operational costs, and orchestrated by events. Delta Lake adds this model by adding ACID transactions and Change Data Capture (CDC) features, which enable efficient incremental update support, and near real-time analytics.Metadata-based ETL architectures represent a significant breakthrough in data engineering, providing an architecture with layers that centralise metadata management and pipeline execution, and incorporate governance across the data lifecycle. These systems make it possible to adapt through dynamic interpretation of metadata at runtime, and to enable automated optimisation strategies with cloud-native platforms. Therefore, they offer a highly scalable, robust, and affordable data integration system that enables organisations to address the needs of modern analytics and business intelligence [22].

4. Distributed optimization and scalability

4.1 Distributed ETL challenges

The cloud-native architecture and ETL systems have become the foundation of the data infrastructure. These distributed pipelines allow ingestion and transformation of data dispersed sources and nonhomogeneous platforms. Also, this architectural development poses a novel performance, orchestration, and cost issues. The network latency is the challenges that facing the distributed ETL systems. The distributed pipelines may experience throughput congestion and synchronization in various locations to these drawbacks. Replication of data in distributed clusters is prone to performance degradation unless network parameters are optimised with bandwidth and packet-loss tolerance. The complexity of orchestration is another problem. Distributed ETL pipelines often integrate multiple orchestration tools (e.g., Apache Airflow, Azure Data Factory, AWS Step Functions) to orchestrate workflows across containers, nodes, and services. This multi-tool dependency adds a management load in the scheduling, dependency tracking and failure recovery. The orchestration layer should make sure that the tasks, which have dependencies, like data extraction, transformation, and validation. In the cloud-native data environment, the cost of ETL execution increases with the amount of compute utilised, network egress, and storage. The operation cost can be ruthlessly increased by inefficient orchestration or useless redundancy data shuffling between distributed nodes. Consequently, distributed ETL systems are reliant on intelligent resource provisioning and elastic scaling systems to balance performance with cost efficiency. These issues highlight the importance of metadata-based optimisation systems that can dynamically model the pattern of workloads, dynamically coordinate compute resources and eliminate unnecessary data flows of ETL system [23].

4.2 AI-based and forecasting optimisation

In the distributed ETL ecosystems, the statical optimisation methods are ineffective with the dynamic workload, schema modification and the heterogeneous infrastructure. AI and ML have played a major role in facilitating predictive and adaptive optimisation of ETL pipelines. AI-based ETL systems use metadata-driven ML models to understand the execution metrics of jobs in the past (job runtime, memory usage, I/O operations, data skew, etc.) to anticipate possible bottlenecks and modify resource-allocation policies. As an illustration, tuning Apache Spark settings using ML has led to performance improvements and cost reductions in cloud ETL workloads. The models use pipeline metadata such as data lineage, transformation dependencies and system logs to train optimisation algorithms used to predict future behaviour of workloads. As shown by Mantri [4], the integration of Spark and Snowflake can be optimised with the use of ML to decrease the time of the job by 45 % and the cost of cloud resources by up to 30 %. The smart meta-data-gathering models can check on performance and feed live information to an intelligent orchestration engine based on AI, which can change execution plans during network surge periods [24]. Moreover, reinforcement learning (RL) applied to distributed ETL scheduling has been demonstrated to work. The operational metadata can be used to teach RR agents the best task scheduling and data partitioning policies across compute nodes. These agents adjust to the workload variations and the manual tuning process is minimized. This strategy is able to not increase throughput but ensure compliance with SLA in changing data loads. Therefore, self-learning, self-healing ETL systems brought to a point of optimum performance are built on AI-driven optimisation coupled with metadata-driven automation.

4.3 Elasticity of resources and orchestration

Cloud-native data warehouses are resource-elastic, meaning that ETL workloads can automatically scale in response to changes in data volume and compute requirements. Kubernetes, Docker Swarm, and server-less computing systems are distributed orchestration systems that have transformed the scale of ETL by decoupling execution of workloads with fixed infrastructure provisioning. It was shown by Gupta and Sundararajan that the integration of Apache Spark, Kubernetes and AWS Glue with a distributed ETL reduced the query response time by 75.9 %, and the total costs reduced by 64.5 % to traditional monolithic systems. The system used metadata-aware orchestration to schedule ETL jobs across containers and the optimisation of compute usage and reduction of idle time [25].Metadata is the important in the orchestration because it gives contextual information concerning task dependencies, dataset scale, and data locality. It allows orchestrators to serve workloads that are near to sources of their data to minimize latency and network overhead. This elasticity is extended by server-less architecture which is server-less computing which uses metadata to define the conditions of triggering functions and run-time parameters. In addition, metadata registries can be used to bring standardisation to runtime configurations in distributed environments using containerised ETL micro-services. This practice aids fine-grained scaling, timely failure recovery of nodes, and smooth versioning of ETL-elements in large-scale analytics. In general, metadata-based orchestration and resource elasticity guarantee stability in performance and predictability in cost, a crucial attribute to sustainable distributed ETL optimisation.

4.4 Smart metadata indexing and parallelisation

The sheer magnitude of contemporary ETL solutions requires sophisticated metadata indexing and parallel processing plans to remain performance-efficient. Metadata indexing entails the sorting of transformation guidelines, schema differentiations, and execution conditions into orderly query-optimal metadata removes. It enables the ETL engines to fetch configuration information quickly in real-time and reducing the pipelines. Parallelisation uses metadata to identify independent transformation units which can be executed in parallel at distributed nodes. ETL frameworks distribute workload and with lower latency by splitting datasets and transformation logic to metadata-defined keys (e.g. partition columns or shard identifiers) [27]. Current metadata-based ETL systems, such as AIDEN (Artificial Intelligence-Driven ETL Networks), use adaptive partitioning algorithms to process data volume and system load in real time adjusting partition sizes [28]. The effect of this strategy is to avoid skew of data, and enhance parallel throughput in multi- node clusters. Furthermore, execution graphs, which are based upon metadata task dependencies and lineages. It may enable orchestrators to rearrange tasks to ensure that high-priority jobs enough resources and that idle nodes spend less time. ETL systems are characterised by a high level of transparency and fault tolerance as the intelligent metadata indexing and distributed parallelisation have not only high throughput and scalability with fault tolerance. These methods are the staple of cloud-native data warehouses with

distributed execution and optimisation of metadata-driven (Table 2).

5. Metadata management, governance, and security

5.1 Metadata lifecycle management

Metadata in cloud-native data ecosystems serves as the basic building block to distributed ETL optimization, automation, traceability, and governance across the entire data lifecycle. Metadata management involves four key processes, including creation, storage, update, and retirement, which need a systematic approach to ensure the multi-cloud environments. The creation stage begins with the initial definition of the data, schema, and transformation rules within the ETL. The definitions are extracted into configuration files, application programming interfaces and templates of transformation logic. The systems like AWS Glue and Apache Atlas. Metadata is identified dynamically as schema discovery that includes both structural and semantic attributes [28]. The metadata retirement which is inevitable in compliance and performance optimization. The unnecessary metadata records are pruned or moved to low-cost storage levels of an object store of the active catalog (Figure 2).

This diagram depicts the governance process of a metadata-based ETL system, which describes the relationships between metadata repository, compliance engine, and audit that apply policies, maintain data quality, and regulatory compliance in the cloud-native environments.

5.2 Data lineage and provenance

The metadata governance which implies the ability to track the origin, changes, and flow of a data element through systems. The lineage metadata in distributed ETL systems records the entire transformation of the incoming raw data to analytical consumption. Thus offering the insight into data dependencies, transformation logic and operational states. The integration of lineage metadata into the ETL processes ensure that the data integrity is verified by the technical and compliance teams. The metadata-based AIDEN platform, the creation of lineage graphs through the analysis of transformation scripts results in a visual display of the way datasets change within distributed settings. Metadata helpful in impact analysis, which can be used to explain the schema rules at the top of the lineage propagate to analytical models at the bottom. It is useful in a multi-cloud warehouse environment where schema drift and reports. Other tools like Apache Atlas and Azure Purview offer an application programming interface of dynamic lineage tracking and incorporates metadata into larger governance processes. Auditability and regulatory compliance is further facilitated by provenance metadata which captures temporal, spatial and agential contexts of applied transformations. Gandha has suggested that provenance tracking within ETL pipelines can help decrease the time spent on debugging and compliance audit by 40 %. Because it provides verifiable provenance of every transformation and data movement activity [29].

5.3 Security, compliance and auditing

The metadata is playing an increasing role in the process of automation and orchestration. It has become a premium asset of security threats, such as unauthorized access, the exfiltration of data and its manipulation. The metadata repositories may hold business logic that is sensitive, schema definitions, and reference credential. Metadata encryption at rest and in transit the basic security precautions. The distributed ETL systems usually use symmetric encryption (AES -256) to store catalogs and Transport Layer Security (TLS) to communicate between services. The access to metadata repositories should be controlled using role-based access control (RBAC) and attribute-based access control (ABAC) to promote fine-gained permission control that complies with the corporate policies[31]. The process of compliance auditing requires unchangeable documentation of every metadata movement. Apache Atlas, Collibra, and Informatica Enterprise Data Catalog are the platforms to maintain an audit trail of all metadata modifications, access to provide traceability on such standards as the General Data Protection Regulation (GDPR), the Health Insurance Portability and Accountability Act (HIPAA), and ISO 27001. Gupta and Sundararajan were show that allowing integrated governance layers in distributed ETL systems on Databricks and Azure Data Factory with security concerns, minimizes manual scaling. The dynamism controls through the use of metadata tags, organizations are able to automate data classification and encryption during pipeline execution [25].

5.4 Data mesh and fabric paradigms integration

The new data mesh and data fabric architectures have redefined organizations deal with data ownership and data governance. Metadata is the semantic weaving and fabric of distributed data

domains, which is used to guarantee interoperability and discoverability in federated environments. A data mesh decentralizes data ownership by delegating datasets responsibility to domain teams with global interoperability by using standardised metadata contracts. Metadata-based ETL models are well suited to this paradigm of quality constraints, and transformation logic in machine-readable metadata representations that can be cross-domain-accessible. The data fabric architectures integrate data management by providing a virtualized, metadata-driven layer that abstracts physical data store the integrated metadata-based ETL into these new architectures, organizations can achieve real decentralization and centralized governance each domain is responsible for its own pipelines and adheres to enterprise-wide metadata standards. Finally, the metadata incorporation in the data mesh which regulates the flow of data, imposes policies, and balances the distributed analytics platform scalability [31].

6. Future directions and research challenges

The growing use of multi-cloud and hybrid data ecosystems by organizations and it has become a fundamental requirement of metadata-based ETL frameworks. Metadata schema,, are largely vendor-specific leading to poor interoperability across systems like AWS Glue, Azure Data Factory and Google Cloud Data Catalog. Such fragmentation hinders smooth coordination and management of distributed ETL activities between heterogeneous cloud providers. Future studies focus on the design of standardised metadata-exchange models that would support the sharing of metadata across platforms of transformational logic (Figure 3). The possible methods are the integration of open metadata standards like Governance (OMAG) framework within the ODPi Egeria program. These standards might help reduce the cost and complexity of integration by making metadata flow through ETL systems underlying systems by using common ontologies and schema registries. Besides, lineage and governance metadata interoperability is critical to consistent auditing and compliance monitoring in distributed ETL systems. However, the necessary logical involves significant progress in semantic harmonisation, metadata version control and policy synchronisation across the different governance structures. The problems represent dynamic research areas that are necessary to the new generation of metadata-based distributed ETL systems. The high rate of growth data and the metadata has compounded the energy consumption of the cloud-based ETL systems. As a result, sustainable data engineering has become an essential research agenda, which predicts efficiency and metadata management. The metadata-driven systems are expected to reflect the concept of the computing and storage can be optimised to reduce energy use. Similarly, active metadata analytics may be used to guide the judicious scheduling of ETL jobs in times of low carbon intensity to datacentres that are fed by renewable energy sources. Based on these observations, future studies can explore carbon-conscious metadata tagging and allow organisations to measure and reduce the environmental footprint of ETL tasks. The metadata catalogues that balance performance with sustainability of lineage information and shrinking metadata repositories without affecting governance integrity. These will be necessary as business organizations of metadata objects in a multi-cloud environment.

This value shows the endless feedback loop as metadata is analyzed by AI agents to accelerate the optimization, the improved results update the metadata warehouse, and through this, the autonomous and intelligent improvements are converted to iterative progression of ETL systems.

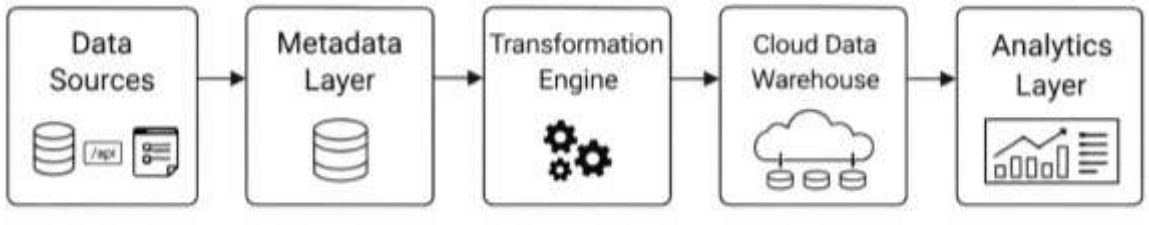


Figure 1: *Metadata-Driven ETL Pipeline Conceptual Architecture.*

Table 2. *Performance Metrics of Distributed ETL Optimization Approaches*

Metric	Traditional ETL	Metadata-Driven ETL	AI-Optimized Metadata-Driven ETL
Execution Latency	High (Batch-oriented processing with static	Medium (Dynamic orchestration based on	Low (Predictive workload balancing using AI/ML

	scheduling; prone to bottlenecks in large data loads)	metadata configurations; improved parallelism)	models; near real-time optimization)
Scalability	Limited (Tightly coupled scripts and hardware dependencies)	High (Containerized and microservice-based deployment; metadata enables modular scaling)	Very High (Elastic scaling with predictive orchestration and AI-driven provisioning)
Fault Tolerance	Moderate (Manual recovery and limited logging)	High (Centralized metadata catalog supports recovery and lineage tracking)	Very High (Self-healing pipelines using AI-driven anomaly detection and auto-correction)
Maintenance Overhead	High (Manual scripting; poor reusability)	Low (Reusability through metadata templates and dynamic configuration)	Very Low (Automated optimization via reinforcement learning and AI-driven tuning)
Cost Efficiency	Moderate (Static resource allocation leads to over-provisioning)	High (Elastic scaling reduces idle compute; metadata reduces redundancy)	Very High (AI-assisted resource allocation; continuous cost optimization based on metadata feedback)
Governance and Compliance	Low (Hard-coded transformations hinder lineage tracking)	High (Metadata provides full lineage and governance mapping)	Very High (AI enhances metadata integrity, anomaly detection, and compliance enforcement)
Automation Level	Minimal (Manual scheduling and error handling)	Medium (Automated orchestration with metadata-defined workflows)	Extensive (End-to-end automation via predictive orchestration and cognitive agents)

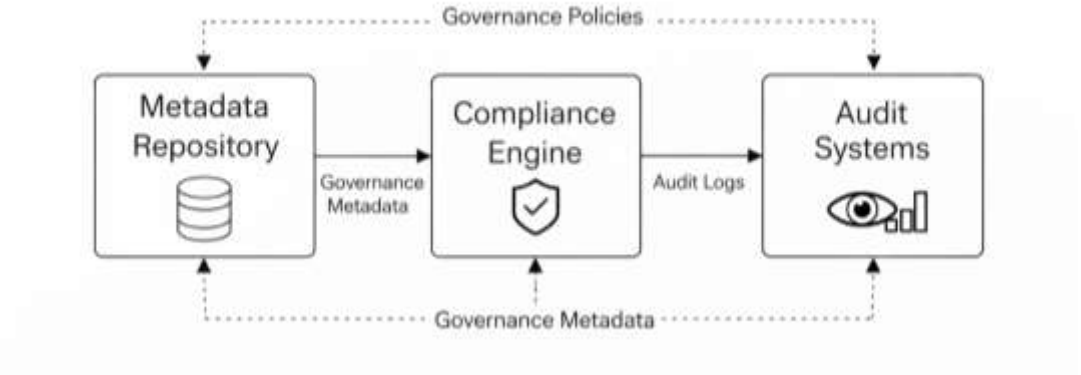


Figure 2: *Cloud-native ETL Metadata Governance Model.*

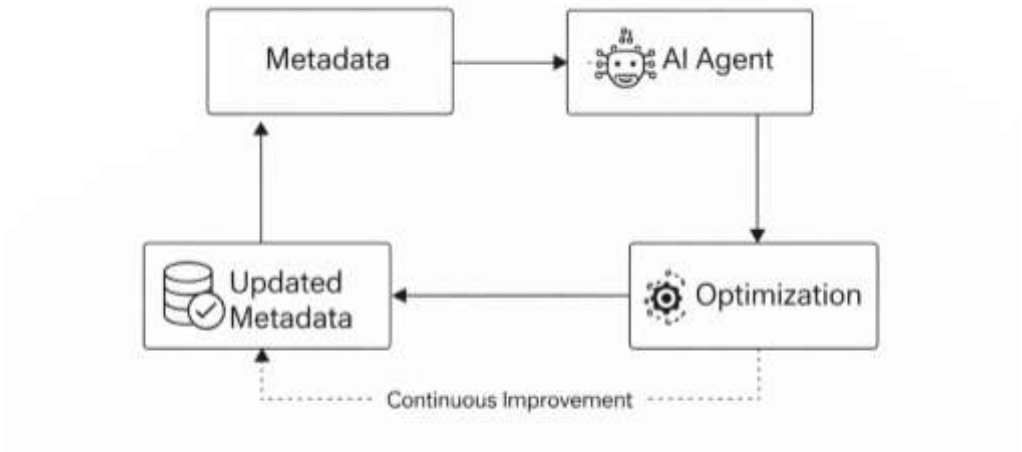


Figure 3: *AI-Optimized Metadata-based Optimization Loop.*

7. Conclusions

Structured metadata artefacts are transforming ETL processes in cloud-native data ecosystems using metadata-driven architectures, which decouple business logic and execution. Unlike the non-computational, script-based paradigms, these models support flexible, automated and scalable data pipelines. Metadata has become an operational layer, which coordinates real-time execution, provides compliance and provenance of lineage. ETL pipelines are provisioned with predictive workload balancing and self-tuning with the implementation of AI-based optimisation. AIDEN and GenAI-ETL are examples of emerging prototypes that will allow self-healing autonomous pipelines, and their conceptualisation through data-mesh and data-fabric concepts will ensure decentralised but controlled data management, which will mark the emergence of intelligent, sustainable and self-developing cloud data ecosystems.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Kondylakis, Haridimos, Varvara Kalokyri, Stelios Sfakianakis, Kostas Marias, Manolis Tsiknakis, Ana Jimenez-Pastor, Eduardo Camacho-Ramos, et al. 2023. “Data Infrastructures for AI in Medical Imaging: A Report on the Experiences of Five EU Projects.” *European Radiology Experimental* 7 (1): 20.

[2] Foidl, Harald, Valentina Golendukhina, Rudolf Ramler, and Michael Felderer. 2024. “Data Pipeline Quality: Influencing Factors, Root Causes of Data-Related Issues, and Processing Problem Areas for Developers.” *The Journal of Systems and Software* 207 (111855): 111855.

[3] Debauche, Olivier, Saïd Mahmoudi, Pierre Manneback, and Frédéric Lebeau. 2022. “Cloud and Distributed Architectures for Data Management in Agriculture 4.0 : Review and Future Trends.” *Journal of King Saud University - Computer and Information Sciences* 34 (9): 7494–7514.

[4] Isah, Haruna, Tariq Abughofa, Sazia Mahfuz, Dharmitha Ajerla, Farhana Zulkernine, and Shahzad Khan. 2019. “A Survey of Distributed Data Stream Processing Frameworks.” *IEEE Access: Practical Innovations, Open Solutions* 7: 154300–316.

[5] Malti, Arslan Nedhir, Mourad Hakem, and Badr Benmammar. 2024. “A New Hybrid Multi-Objective Optimization Algorithm for Task Scheduling in Cloud Systems.” *Cluster Computing* 27 (3): 2525–48.

[6] Wen, Lei, Hengshun Qian, and Wenpan Liu. 2022. “Research on Intelligent Cloud Native Architecture and Key Technologies Based on DevOps Concept.” *Procedia Computer Science* 208: 590–97.

[7] Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM, 51*(1), 107–113.

[8] Abadi, D. J. (2018). Data management in the cloud: Limitations and opportunities. *IEEE Data Engineering Bulletin, 41*(1), 3–9.

[9] Amorim, Ricardo Carvalho, João Aguiar Castro, João Rocha da Silva, and Cristina Ribeiro. 2017. “A Comparison of Research Data Management Platforms: Architecture, Flexible Metadata and Interoperability.” *Universal Access in the Information Society* 16 (4): 851–62.

[10] Oyighan, Diseiye, Ejiro Sandra Ukubeyinje, Boma T. David -West, and Bolaji David Oladokun. 2024. “The Role of AI in Transforming Metadata Management: Insights on Challenges, Opportunities, and Emerging Trends.” *Asian Journal of Information Science and Technology* 14 (2): 20–26.

[11] Rozony, F. Z., Aktar M. N. A., M. Ashrafuzzaman, and A. Islam. 2024. “A Systematic Review of Big Data Integration Challenges and Solutions for Heterogeneous Data Sources.” *Academic Journal on Business Administration, Innovation & Sustainability* 4 (04): 1–18.

[12] Alonso, Juncal, Leire Orue-Echevarria, Valentina Casola, Ana Isabel Torre, Maider Huarte, Eneko Osaba, and Jesus L. Lobo. 2023. “Understanding the Challenges and Novel Architectural Models of Multi-Cloud Native Applications – a Systematic Literature Review.” *Journal of Cloud Computing Advances Systems and Applications* 12 (1).

[13] Alonso, Juncal, Leire Orue-Echevarria, Valentina Casola, Ana Isabel Torre, Maider Huarte, Eneko Osaba, and Jesus L. Lobo. 2023. “Understanding the Challenges and Novel Architectural Models of Multi-Cloud Native Applications – a Systematic

Literature Review." *Journal of Cloud Computing Advances Systems and Applications* 12 (1).

[14] Munson, Jacob, Thomas Cuezze, Siddat Nesar, and Dominique Zosso. 2025. "A Review of Large Language Models and the Recommendation Task." *Discover Artificial Intelligence* 5 (1).

[15] Singh, Gaurav, and Adarsh Maurya. 2025. "The Role of Metadata in Data Curation for Enhancing Discoverability in Large Datasets." *International Journal of Web of Multidisciplinary Studies* 2 (1): 31–37.

[16] Balabanov, O. S., and Institute of Software Systems NAS of Ukraine. 2019. "Big Data Analytics: Principles, Trends and Tasks (a Survey)." *Problemy Programmirovaniya. Problems in Programming*, no. 2: 047–068.

[17] Xu, Xi, Jianqiang Li, Zhichao Zhu, Linna Zhao, Huina Wang, Changwei Song, Yining Chen, Qing Zhao, Jijiang Yang, and Yan Pei. 2024. "A Comprehensive Review on Synergy of Multi-Modal Data and AI Technologies in Medical Diagnosis." *Bioengineering (Basel, Switzerland)* 11 (3): 219.

[18] Eweje, Adeoluwa, and Francis Ohaegbu. 2021. "Advances in Modern Data Stack Architectures for Scalable Data Integration and Business Intelligence." *International Journal of Multidisciplinary Research and Growth Evaluation* 2 (5): 538–50.

[19] Jahanshad, Neda, Petra Lenzini, and Janine Bijsterbosch. 2024. "Current Best Practices and Future Opportunities for Reproducible Findings Using Large-Scale Neuroimaging in Psychiatry." *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology* 50 (1): 37–51.

[20] Chanda, D. (2024). Automated ETL Pipelines for Modern Data Warehousing: Architectures, Challenges, and Emerging Solutions. *The Eastasouth Journal of Information System and Computer Science*, *1*(03), 209–212.

[21] Eweje, Adeoluwa, and Francis Ohaegbu. 2021. "Advances in Modern Data Stack Architectures for Scalable Data Integration and Business Intelligence." *International Journal of Multidisciplinary Research and Growth Evaluation* 2 (5): 538–50.

[22] Munappy, Aiswarya Raj, Jan Bosch, and Helena Homström Olsson. 2020. "Data Pipeline Management in Practice: Challenges and Opportunities." In *Product-Focused Software Process Improvement*, 168–84.

[23] Ragazou, Konstantina, Ioannis Passas, Alexandros Garefalakis, and Constantin Zopounidis. 2023. "Business Intelligence Model Empowering SMEs to Make Better Decisions and Enhance Their Competitive Advantage." *Discover Analytics* 1 (1).

[24] Mantri, A. (2023). *Advanced ML Techniques for Optimizing ETL Workflows with Apache Spark and Snowflake. Journal of Artificial Intelligence & Cloud Computing,* **2(3)**, 339– 347.

[25] National University Bangladesh, Gazipur, Bangladesh, Hosne Ara Mohna, Tonmoy Barua, Manager, Facilities and Administration, MetLife, Bangladesh, Mohammad Mohiuddin, Data Engineer, NCC Bank PLC, Dhaka, Bangladesh, Md Mostafizur Rahman, and Assistant Manager, Teletalk Bangladesh Ltd, Dhaka, Bangladesh. 2022. "Ai-Ready Data Engineering Pipelines: A Review of Medallion Architecture and Cloud-Based Integration Models." *American Journal of Scholarly Research and Innovation* 01 (01): 319–50.

[26] Vishwanadham Mandala. 2018. "Meta-Orchestrated Data Engineering: A Cloud-Native Framework for Cross-Platform Semantic Integration." *Global Research and Development Journals* 3 (12).

[27] Seenivasan, D. 2024. "AI Driven Enhancement of ETL Workflows for Scalable and Efficient Cloud Data Engineering." *International Journal of Engineering and Computer Science* 13 (06): 10–18535.

[28] Bhatlawande, S., Rajandekar, R., & Shilaskar, S. (2024). *Implementing Middleware Architecture for Automated Data Pipeline over Cloud Technologies. IEEE 13th International Conference on Communication Systems and Network Technologies (CSNT)*, **13(1)**, 506–513.

[29] Machireddy, Jeshwanth Reddy. 2023. "Data Quality Management and Performance Optimization for Enterprise-Scale ETL Pipelines in Modern Analytical Ecosystems." *Journal of Data Science, Predictive Analytics, and Big Data Applications* 8 (7): 1–26.

[30] Shivaramakrishna, D., and M. Nagaratna. 2023. "A Novel Hybrid Cryptographic Framework for Secure Data Storage in Cloud Computing: Integrating AES-OTP and RSA with Adaptive Key Management and Time-Limited Access Control." *Alexandria Engineering Journal* 84 (December): 275–84.

[31] Vattumilli, P. K. (2024). *Metadata-Driven ETL Pipelines: A Framework for Scalable Data Integration Architecture. International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, **6(6)**, Article 61224.

[32] Ananthakrishnan Kunju, A. (2024). *Autonomous GenAI Agents for Legacy-to-Cloud ETL Modernization. Journal of Artificial Intelligence General Science (JAIGS)*, **1(1)**, 55–72.