International Journal of Economic Practices and Theories (IJEPT)
ISSN: 2247-7225
Volume 2026, Issue 1

Cognitive Data Engineering: AI-Governed Data Quality, Lineage, and Pipeline Optimization at Scale

Velangani Divya Vardhan Kumar Bandi

Director AI/ML Engineering

vdivyavardhankb@gmail.com

ORCID ID :0009-0007-5650-4100

Abstract

Artificial intelligence (AI) is establishing itself as the next generation of technology. However, the data required to train such expandable, self-learning, powerful systems have so far been collected and pre-processed in a traditional manner. Moreover, because of the lack of governance in many such AI initiatives, these processes remain uncontrolled and often produce low-quality results. Both issues urgently need solution.

Three core principles of cognitive data engineering have been developed that are enabled by the recent expansion of AI technology. First, quality metrics and evaluation, as well as anomaly detection and correction mechanisms, have been formalized to provide a comprehensive AI-governed data-quality framework. Next, a set of metadata standards that define describe and affect Internet-scale data ecosystems is proposed. Their implementation provides a sophisticated method of capturing data lineage and provenance information and using that data for compliance and efficient data query acceleration. Third, the data pipeline architecture is designed to produce the definition and execution of complex data pipelines and orchestration in a cost-aware manner. These contributions enable an AI governance framework for complete data pipelines. Such a framework defines roles, policies, and decision rights to identify risks in the use of data pipelines, assess those risks quantitatively, and provide mitigation guidelines.

Keywords: Cognitive Data Engineering, AI-Governed Data Pipelines, Data Quality Metrics and Evaluation, Automated Anomaly Detection and Correction, Metadata Standards, Internet-Scale Data Ecosystems, Data Lineage and Provenance, Compliance-Aware Data Architecture, Cost-Aware Pipeline Orchestration, AI Governance Frameworks, End-to-End Data Pipeline Management, Risk Identification and Quantification, Policy-Driven Data Engineering, Decision Rights in AI Systems, Data Provenance Acceleration, Intelligent Metadata Management, Regulated Data Pipelines, Autonomous Data Quality Management, Enterprise AI Governance, Scalable Data Orchestration.

1. Introduction

Data engineering, the process of making data available for consumption, is a critical function in almost every industry. It encompasses the design, creation, management, and optimization of data pipelines and other systems for storing, moving, and processing large volumes of data. These capabilities have powered the development of data science, machine learning, big data, and AI, enabling rapid innovations in a broad range of application domains, including healthcare, national security, finance, and self-driving vehicles.

The advent of generative AI and, in particular, the recent development of foundation models—extremely large deep learning models that have been pretrained using self-supervision on massive datasets—has spurred a new wave of interest in using AI systems not just for analysis and prediction, but also for production workflows and applications. These new, more complex ecosystems for deploying AI systems—often referred to as AI factories—have changed the requirements and stressors on data engineering, creating challenges not only for engineering but also for the quality, lineage, and management of the underlying data pipeline support systems. The focus so far in AI has primarily been on the intelligence exhibited by the AI models and not on other key components of the wider ecosystem for deploying these models.

1.1. Overview of Cognitive Data Engineering Principles

Organizations increasingly rely on data science to extract value from data. Unlike classic software development processes, data science projects use data pipelines that vary in structure and execution over time and can even merge or split running pipelines. Data pipeline engineering is often now performed by highly skilled data scientists and analysts, who work without necessary software engineering principles and data management tools. These gaps lead to data quality issues and pipeline execution inefficiencies.

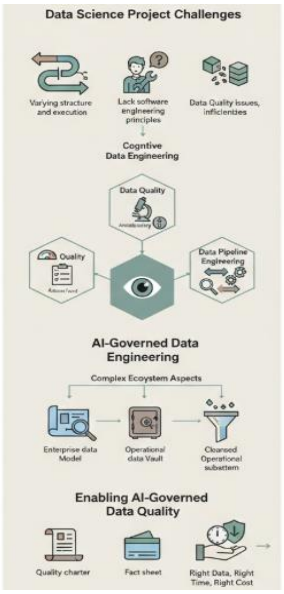

Fig 1: Cognitive Data Engineering: Leveraging AI-Governed Data Quality and Lineage Management to Bridge the Gap Between Data Science and Software Engineering

Cognitive data engineering—AI-governed data quality, data lineage management, and data pipeline engineering—allows data to be managed so that the right data are on hand at the right time and for the right cost. Essential AI-governed data quality principles include defining quality metrics and levels for various use cases; assessing data quality based on these requirements; detecting and correcting data (de)normalization, compositional, and logical anomalies; and ensuring the availability of complete operational datasets. In complex, heterogeneous ecosystems, additional aspects, such as the definition and management of an enterprise data model, an operational data vault, or a cleansed operational subsystem, become crucial. Enabling AI-governed data quality within a large enterprise also calls for elements such as a quality charter, fact sheet, and risk profile.

2. Theoretical Foundations of Cognitive Data Engineering

Data engineering focuses on data processing for analytics and learning. Though often a specialty, engineering for AI-driven services introduces unique challenges that require new support paradigms. Quality control becomes increasingly difficult as real-time data feeds from sporadic, untrusted, and rapidly evolving sources replace trusted and consistently available historical databases. Data provenance and lineage tracking gain importance due to the intricacy of AI/ML components and the growing dependence on complex data ecosystems comprising diverse and hacker-prone services. Resource-aware data scheduling mechanisms are required to continuously monitor the current and predicted load levels and control cloud resource usage across public, private, and hybrid deployments.

While the evolving data ecosystem is already highly complex and vulnerable, its management and control must also now be delegated to AI-based cognitive services with the potential for augmented self-healing that can dynamically fix detected issues with minimal human intervention. Recent work demonstrates the need for data quality control and resilience by analyzing the most common causes of ML prediction errors in enterprise systems, proposing a hierarchy of metric types and corresponding monitoring solutions and combining them into a coherent quality framework for continuous requirement specification, monitoring, and maintenance.

2.1. Data Quality in AI-Driven Systems

Data quality issues impact machine learning systems, resulting in organizations scuttling AI initiatives. The scientific literature supports these real-world experiences through normative works proposing quality metrics for data lakes, operational and exploratory repositories, and automated data pipelines. Temporal stability of quality metrics is used in change detection algorithms that identify incremental maintenance and cleaning tasks. The surface and depth of data lakes influence the quality of children observed during the actual execution of machine learning projects. Further, data provided by sensor networks is termed smart data, and a framework for defining smart data quality metrics employing time constraints is presented. Given the nature of AI systems, quality has also been viewed through the lens of AI governance, where existing data quality roles are mapped to a governance layer; the data steward role has been expanded to include data-quality-as-a-service operations and infrastructure support services.

An exploratory study using the qualities of successful companies as a basis for comparison has shown that emerging technologies promote lower quality than the ten established qualities generally accepted. However, machine learning and artificial intelligence reduce quality for education and cognitive systems. Towards this goal, machine learning housing data ask such services to inspect input datasets in an unsupervised setup. Florida, USA a set of specified features is presented, that serve the purpose of tracking dependency-based heuristic preparation in the final machine-learning service using the serviced dataset. A recommendation model is built and presented using flask and span tools that asks for other training sets along with its quality features visible in the recommendation model while preparing, thus serving the need of dependency-based heuristic allocation and preparation of data.

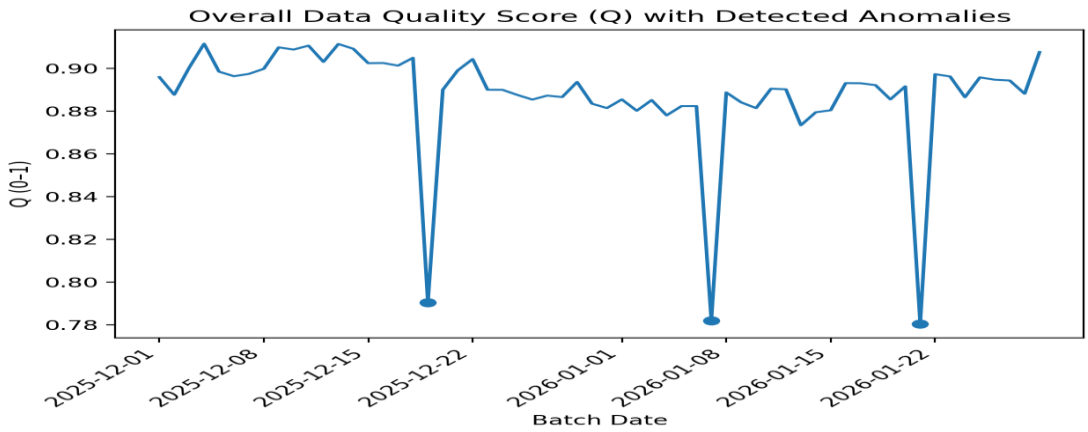

Fig 2: Multi-Facet Data Quality Scoring Model with Normalization and Weighted Aggregation

Equation 1) Data Quality Metrics → Overall Quality Score (Q)

Step 1: Define quality facets (attributes)

$$q_1, q_2, \dots, q_m$$

Each q_i is typically normalized to $[0,1]$ (0 = worst, 1 = best).

Step 2: Normalize a raw metric to a 0–1 score

Suppose a facet is measured as a raw metric x (e.g., % missing, latency, error rate). To map it to $[0,1]$, a common normalization is min–max:

- If *higher is better*:

$$q = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

- If *lower is better* (e.g., missing rate, error rate):

$$q = 1 - \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

This gives comparable scores across facets.

Step 3: Weight facets by importance (policy / use-case)

The article emphasizes that quality depends on goals/use-cases and governance policies. Assign weights $w_i \geq 0$ with:

$$\sum_{i=1}^{m} w_i = 1$$

Step 4: Compute the overall quality score

A standard governed "overall quality" KPI is the weighted sum:

$$Q = \sum_{i=1}^{m} w_i \, q_i$$

2.2. Data Lineage and Provenance

Research on data lineage has largely focused on two objectives: the management of data in large Directly Accessed Data Store (DADS) data stores and within complex data ecosystems characterized by a rich data provenance model. The latter addresses new business challenges imposed by the growing disintermediation of data consumers and producers through cloud computing and the desire for a more agile, responsive style of interaction. The proliferation of diverse data providers participating in a data-sharing ecosystem brings to the fore the old parable of "garbage in, garbage out." Consumers believe they are obligated to pay for products of higher quality than those they've been exposed to in the past. DADS stores, on the other hand, represent a vault for data writers. These stores offer online transaction processing-type cost and availability trade-offs, with more emphasis given to scalability than to quality. Management of DADS repositories is further complicated by the complexity of the last-miles providing data to these stores at scale.

Two specific objectives emerge from these considerations. The first involves supporting the full data lifecycle within complex ecosystems over timeframes that range from a few minutes to several weeks. The goal is to formalize the structure and role of data within a Cloud Service Provider (CSP) ecosystem characterized by DADS consumers, producers, and a range of data creation and repurposing services that can take different forms (ranging from highly automated provisioning to high-touch, bespoke replication). The lineage and integrity requirements are subsequently translated into a suite of new data properties and persistence mechanisms enabled by a combination of NoSQL storage and the use of provenance tracking. When viewed from this angle, the data is no longer an opaque block, but a living organism that grows, ages, matures, and dies. The second objective extends the traditional concepts of data lineage and provenance by examining in detail the source, transformation, quality, and completeness properties of the data arriving at the last-mile within a data-sharing ecosystem of multiple players.

3. AI-Governed Data Quality Framework

Quality provides foundation and merit for data science and analytics. Intrinsically, all data in data-driven systems exhibit some quality flaws. Majority of these flaws are recognized and serviced, however, there are remaining anomalies which can fall under the radar of reliability matrices. To help with discovering and servicing these flaws in data, an AI-governed quality framework has been structured which recognizes critical quality metrics, automated anomaly discovery and service methods that can be orchestrated during data pipeline execution.

Quality evaluation criteria is delineated in the quality assurance phase using a set of Quality Attributes, one for each data quality facet. The framework's next stage autogenerated condition-specific knowledge logic that can cast fault indication rules on data quality. Subsequent stage uses these indication patterns to alert an appliance

of existing issue with data sources/producers. The final stage generates machine-learning-based decision classifiers that predict whether a remediation activity is required during the service of the data in line with established quality dimensions. A knowledge system for an automatic crowd-sizing ability remains under development. An augmentation for this knowledge system defines logical decision templates that express surface conditions capable of requiring machine-assisted appraisal and servicing of data.

3.1. Quality Metrics and Evaluation

The quality of data is dependent on the goals of its analysis. Nevertheless, a range of attributes characterizes the utility of a dataset in any context. Data quality metrics collect these attributes into a formal scheme. Automating the identification of suitable metrics is a fundamental aspect of quality analysis in research, where no additional information is accessible. Conversely, production workloads typically provide extensive metadata that can be harnessed by AI to select the most relevant metrics. Data users can then assess if data quality meets the requirements of their application domain and perform exploratory testing to quantify data errors in that context.

Data quality evaluation provides quantitative oversight and alerts users to COVID-19 detection failures via anomalies in transportation networks. Quality aspects throughout a real estate data pipeline are captured. A broader AI approach combines diverse metrics to highlight even minor issues. Large-scale data flows exacerbate these problems, and machine learning addresses an arms race between noise and detection. Push-pull models in data preparation emphasize the cooperation of users and data scientists in data quality assurance. AI reveals data quality using data-specific models, scans the metadata registry for quality problems, and generates rectification code automatically. Such auto-cleaning is attractive if no further damage results.

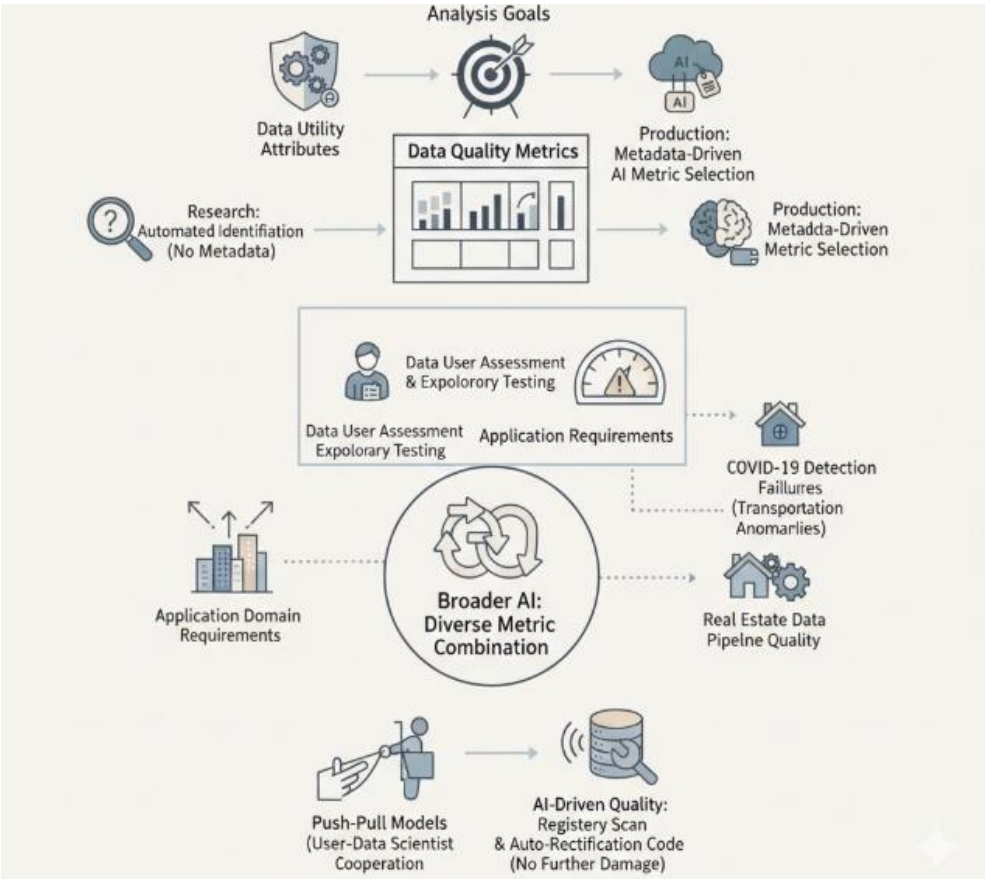

Fig 3: Adaptive Data Quality Engineering: From Metadata-Driven Metric Selection to AI-Enabled Autonomous Rectification

3.2. Anomaly Detection and Correction Mechanisms

Anomaly detection and correction mechanisms monitor the quality of the data and take remedial measures when unacceptable quality levels are detected. An additional AI model monitors the quality of the entire system, using feedback from the data quality detectors to detect anomalies in the detectors and initiate remediation actions. Remediation actions may include rerunning processes in the system to re-create bad-quality data, leaving the decision to the AI engine on which processes should be rerun. Such an engine, however, will require information about the cost of executing each process, as well as the downstream impacts because rerunning certain processes may change the input data of processes several stages downstream.

An AI engine is capable of determining if the data in an operational data ecosystem is of acceptable quality, and if not, can quickly initiate remediative actions. It receives a stream of quality scores from the data quality monitors and can use them to detect sudden shifts in the data ecosystem, determining the level of anomaly associated with the shift. Anomaly detection can be achieved through an open-source framework called HAAS (Hypothesis-Augmented Active Sampling). Based on a previous anomaly in the data ecosystem, an anomaly-detection model is built, which provides an early warning of subsequent future anomalies. The model learns a spatio-temporal representation of the ecosystem quality metrics and uses it to associate different anomalies—of similar cause—to the same category, thereby enhancing prediction accuracy. The model can be trained in a self-supervised manner.

4. Lineage Management in Complex Data Ecosystems

Increases in the volume, diversity, and velocity of data, especially unstructured content, led enterprises to establish data lakes—centralized repositories that enable massive parallel processing, provide resource agility through cloud technology, and support the storage of virtually any amount of structured and unstructured content at a fraction of the cost of traditional systems. Enterprises also adopted data fabric architectures, which facilitate access to distributed data by integrating disparate information sources and silos, thus providing seamless data sharing across disparate environments. The data fabric approach does not require data to be physically moved but provides a layer of data services that supports analytics, governance, and data sharing across a complex web of interconnected systems—including data lakes, enterprise data warehouses, relational databases, social media, and Web sources—whether on-premises or in multiple clouds. Moreover, organizations used firehoses to capture events continuously from applications and transaction systems, storing records in immutable databases.

Despite the advantages of data lakes and fabric architectures, organizations still struggled with the management of complex data ecosystems. Existing tools offered only point solutions, often specialized for specific systems, languages, or characteristics, and built with short-term needs in mind, leading to implementation silos or poorly integrated capabilities across continents and functions. Operations teams therefore lacked a complete picture of the data environment. With no definitive single source of truth for upstream and downstream data asset storage and protection, enterprises could not easily assess the impact of decisions in one part of the business on data in other parts. Quality information was out of date, incorrect, incomplete, or missing altogether. Furthermore, the risk associated with business decisions was increasing, and root-cause analysis was taking longer. As a consequence, data-driven organizations faced increased operational risk; identified anomalies were difficult to resolve, and the corrective actions often did not address the underlying issues.

Table 1. Temporal Data Quality Metrics Across Operational Batches

Date	Completeness	Accuracy	Consistency
2025-12-01	0.9369052570380036	0.8788801445304677	0.8980828284943351
2025-12-02	0.9181880608904369	0.8620314819095767	0.8954349570518084
2025-12-03	0.9259650586792301	0.885254017870187	0.9120175530620748
2025-12-04	0.9323865999298262	0.9083056783475382	0.9084797906816178

Date	Completeness	Accuracy	Consistency
2025-12-05	0.9229274539809192	0.8846744695708679	0.9076565711850199

4.1. Metadata Standards and Interoperability

Standardized metadata formats simplify the creation of lineage information and improve interoperability among tools. International standards developed by the Object Management Group and W3C substantially facilitate metadata sharing across nontrivial data ecosystems. In data lakes, data warehouse ecosystems, and other settings exposed to a broader audience, supporting metadata schemas should enable understandable browsing of the available datasets and direct linkage to corresponding metadata actualization. Examples of broadly adopted patterns that aid surviving information retrieval include the schema.org search engine metadata schema, community-contributed catalog metadata by means of the Open Data initiative, and various metadata repositories for COVID-19 datasets.

The Data Catalog Vocabulary (DCAT) is a W3C standard aiming to facilitate the discovery of datasets published on the web. DCAT allows all expected data sources to be described with a minimal, coherent, and interoperable set of properties. DCAT enables the description of catalogs of datasets and data services published on the web. It is also designed to provide an interface to data catalogs and linked datasets available on the web using common schema.org patterns, such as the Dataset and DataDownload classes. It can encode the dataset description patterns commonly used in data catalogs, including the DCAT catalog, the Data Catalog Vocabulary (DCAT), and SSN–Semantic Sensor Network Ontology, in RDF.

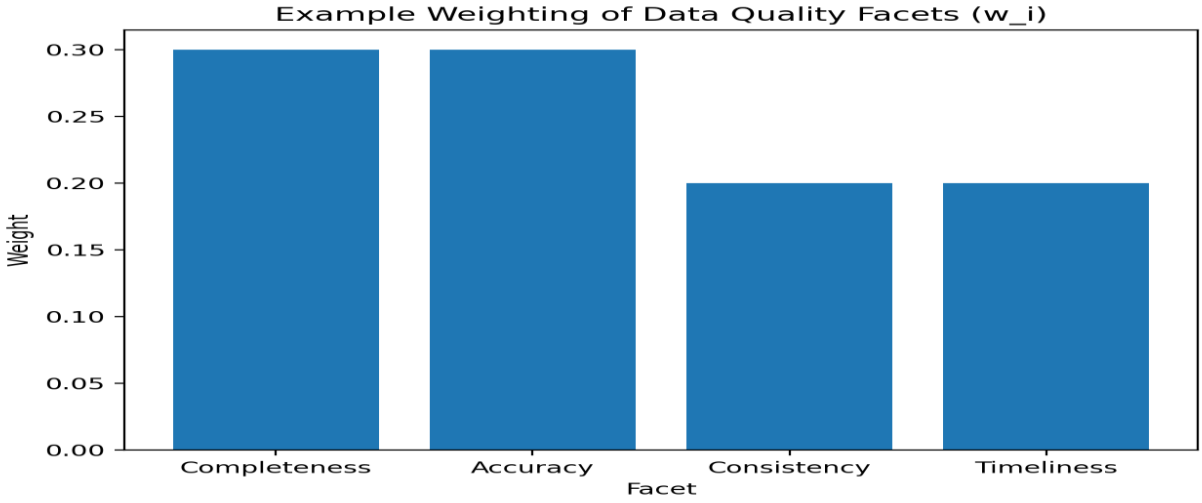

Fig 4: Metadata-Driven Lineage and Interoperability Framework for Anomaly-Aware Data Ecosystems

Equation 2) Anomaly Detection on Data Quality (Shift/Spike Detection)

Step 1: Maintain a historical baseline

Given a time series Q_t (overall quality per batch/window), compute baseline mean and standard deviation:

$$\mu = \frac{1}{T}\sum_{t=1}^{T} Q_t \; \sigma = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(Q_t - \mu)^2}$$

Step 2: Standardize the current observation (z-score)

$$z_t = \frac{Q_t - \mu}{\sigma}$$

Interpretation: how many standard deviations away from normal the current batch is.

Step 3: Flag anomalies using a threshold

$$\text{Anomaly at } t \Leftrightarrow |z_t| > k$$

Typical governed choice: $k \in [2,3]$ depending on tolerance.

Step 4 (optional): Trigger remediation decision

$$\hat{y}_t = f(\mathbf{q}_t, \mathbf{m}_t)$$

where $\mathbf{q}_t$ are facet scores at time t, and $\mathbf{m}_t$ are metadata/context features (pipeline stage, data source, lineage, etc.). Output $\hat{y}_t \in \{0,1\}$: remediate or not.

4.2. Provenance Capture Techniques

An understanding of the componential combinations in complex data ecosystems helps make intelligent design choices to enable metadata-gathering automation as referred to in Section 4.1. Provenance capture methods can therefore be applied judiciously and holistically to avoid excessive resource usage. Provenance capture techniques may therefore be classified as non-intrusive state-of-the-art mechanisms based on event sources and sinks.

Non-intrusive methods use available system-level infrastructure to capture provenance, e.g., logging facilities, network monitoring devices, or process control systems. Furthermore, logging/debugging facilities available in execution platforms may also serve provenance-gathering purposes. Such mechanisms should remain active even during normal operation. Examples include Precog, which adds provenance-gathering support to standard log generation systems, DPM, which captures provenance of display devices, or ProGraVi for BIOS-based provenance. In streaming systems, provenance can be tapped from monitoring or observability infrastructure, such as Prometheus.

Provenance may also be captured during data exchange using protocol-aware middleware, which may act as an indirect source for data transfers. Comprehensible logs can be generated by covering event sources and sinks with provenance monitoring tools. Network flow-monitoring devices may serve as indirect sources and sinks for provenance. These devices extract knowledge about the features of the traffic flowing through them, including summaries of client-server requests and responses. A causal model for the entire provenance graph is constructed by correlating the event information available in the indirect sources and sinks using the monitoring resources.

5. Scalable Data Pipeline Architecture

Sophisticated AI solutions typically rely on complex data pipelines involving intricate orchestration of various data sources, analytics components, and target systems. The structure and capabilities of the pipeline need to facilitate seamless integration into deployment environments while remaining scale-agnostic to allow deployment on-premises, on cloud platforms, or in hybrid configurations. Different paradigms for data pipeline orchestration have emerged to support such eased integration and modifiability. In particular the marriage of Workflow Management Systems (WMS) with Message-Oriented Middleware (MOM) technologies has proven suitable for deployment scenarios with significant operational variability, such as in the transport and logistics domains where pipelines process incoming sensor data streamed in chronological order, although ad-hoc processing of archived databased reports remains a key requirement.

Development approaches for heterogeneous custom-built data pipelines bring additional modifiability challenges. Cognitive Data Engineering suggests adopting a semiconductor design practice pioneered in the 1990s by Advanced Micro Devices Inc. (AMD), where assets are housed in multi-site Infrastructure as a Service (IaaS) clouds with specialised design partners engaged at a trusted services supplier level. Scheduling and resource allocation require a resource-aware approach that spans the entire pipeline lifetime and is supported by capable tools. Business Intelligence (BI) solutions, Data Integration (DI) products, and cloud-based data pipeline offerings from third-party providers present obvious cost and time advantages that have driven widespread adoption. However, naive scheduling of DI job requests can lead to excessive resource purchases.

5.1. Orchestration Paradigms and Tooling

Two principal classes of orchestration have emerged: centralized, user-driven orchestration, and distributed, automatically-driven orchestration. Centralized orchestration facilitates users in specifying data flows through predefined components. Automation tools such as Apache Airflow, Azkaban, and Oozie enable graphical interaction with controllers. However, these systems satisfy only a subset of AI govern-ance goals defined in a preceding section. The overload on human resources becomes a bottleneck in complex, multi-tenant ecosystems with extensive, frequent, and diverse data operations. Also, the underlying tools populate monitoring, control, and logging sites with many rule-compliant alerts, warnings, and errors, most of which are never investigated. These deficiencies prompted an alternative control paradigm. Dynamic, resource-aware scheduling shifts the role of the human operator from initiator to overseer. The AI-aware user-community only needs to specify data and other resources essential to complete an operation and trigger the execution. The underlying mechanism uses AI models and run-time data to select appropriate algorithms, tools, and other resources. Data integrity and the quality of the source data guide selection. AI checks and balances enable safe operation.

Tools assisting rule-compliant scheduling must pursue a well-defined set of AI governance goals. Several start-ups have introduced resource-aware ddsl tools. Data- and compute-intensive process-ensembles define look-up-datasets that constrain sources and data-quality requirements. Scheduling tools monitor availability, data topologies, and transformations and use these to assign and adapt jobs. Cost estimates address waiting times and expenses of cloud-based resources. dPipeline provides experimental demonstration. Independent Data Science and Machine Learning engineers can request direct execution. A store queries available ddsl definitions and uses these to define a multi-variable-regression task. Scheduling tools empty the adjustment-years-enumeration table enabled by upcoming Rainfall-Quality-Classification task, use the run-time data for variance-accounting, and assign a Regression model for implementation.

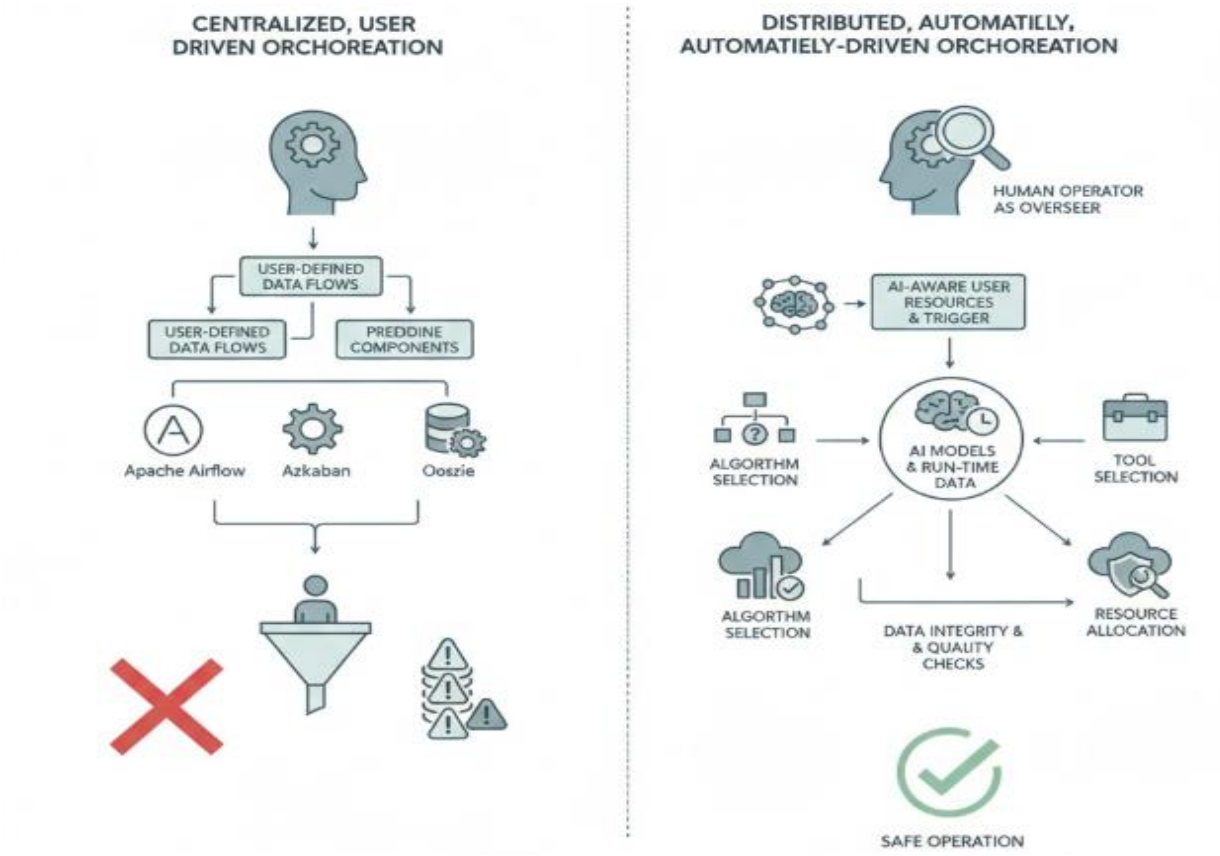

Fig 5: From User-Initiated to AI-Driven Orchestration: Autonomous Resource-Aware Scheduling for Governed Data-Flow Ensembles

5.2. Resource-Aware Scheduling and Cost Optimization

Formal orchestration of distributed data pipelines facilitates controlled processing of complex data ecosystem resources. Effective data pipeline management minimizes execution time and external service costs while balancing the presence of expense-intense and cumulative queries across the execution timeline. Resource-aware scheduling, designed using advanced queue management techniques, prioritizes the execution of active queries with temporary resources over those relying on expensive services.

The orchestration of complex data ecosystems often relies on established techniques, tools, and systems. These resources primarily target construction, validation, schedule management, and health status monitoring, while sophisticated scheduling remains a distinct challenge. Managing many queries simultaneously, especially when data ecosystems provide services whose execution incurs additional costs, poses a major issue. Resource-aware scheduling mitigates this challenge by focusing on the efficient management of temporary resources, such as those used for intermediate data materialization. The objective is to minimize costs incurred from costly underlying services, e.g. text analysis or machine translation, while parallelizing the execution of active queries requiring these services. Advanced queue management techniques support the prioritization of active queries reliant on temporary resources, without overexpensing the underlying costly services.

By defining different application queue classes, the architecture supports multiple queue management strategies. As all queries are now partitioned, it becomes possible to regulate the maximum number of parallel queries by class.

Table 2. Cost Profile of Data Pipeline Jobs Under Naïve Scheduling

Job	Service $/hr	Runtime (hr)	Naive cost ($)
J5	7.38	2.29	20.72
J6	11.65	1.55	19.68
J7	8.74	0.58	5.05
J8	12.75	2.09	29.86
J9	6.96	0.6	4.2
J10	9.77	2.15	24.5

6. AI Governance for Data Pipelines

Cognitive Data Engineering represents the next logical step of Data Engineering for the cognitive era. Data pipelines that propel AI and analytics workloads must thus be able to guarantee the impact of their outputs on the value generation of the respective organizations. Data pipelines must become industry-grade Artificial Intelligence and Machine Learning products that are kept, governed, and maintained across the complete lifecycle by solution development, IT service provisioning, business users, data governance or data protection functions.

AI governance for the data pipeline layer enables the definition of AI governance roles, policies, and decision rights in an AI governance model. It also enables risk assessment for AI and analytics solutions in a risk management framework, covering both data used by the AI components and outputs consumed by AI consumers. Provided that the above-mentioned dimensions are in place, business users of the data pipeline layer can generate high-quality data products and services while delivery teams leverage a collaborative approach for monitoring the operational state and possible issues with the solutions in production. AI governance for data pipelines is guided by the GPAI principles around a risk-oriented approach for securing the AI life cycle. A structured approach is proposed for establishing those components so that industry-grade products and services are delivered for increasing the business value derived from horizontal data ecosystems.

Equation 3) Risk Scoring Using Thresholds vs Normal Ranges

Step 1: Define risk indicators (metrics)

Let risk be represented by indicators $r_1, \ldots, r_p$ (e.g., availability state, SLA violations, cost spikes, quality drops).

Step 2: Normalize each indicator relative to its historical normal range

If a metric has historical mean μ_i and std σ_i:

$$\tilde{r}_i = \frac{r_i - \mu_i}{\sigma_i}$$

Then convert it to a bounded "risk contribution," e.g. via sigmoid:

$$g(\tilde{r}_i) = \frac{1}{1 + e^{-\tilde{r}_i}}$$

Step 3: Aggregate into a pipeline risk score

$$R = \sum_{i=1}^{p} \alpha_i \, g(\tilde{r}_i), \quad \sum \alpha_i = 1$$

Step 4: Governance rule for alerting/closure

$$\text{Trigger mitigation} \Leftrightarrow R > \tau$$

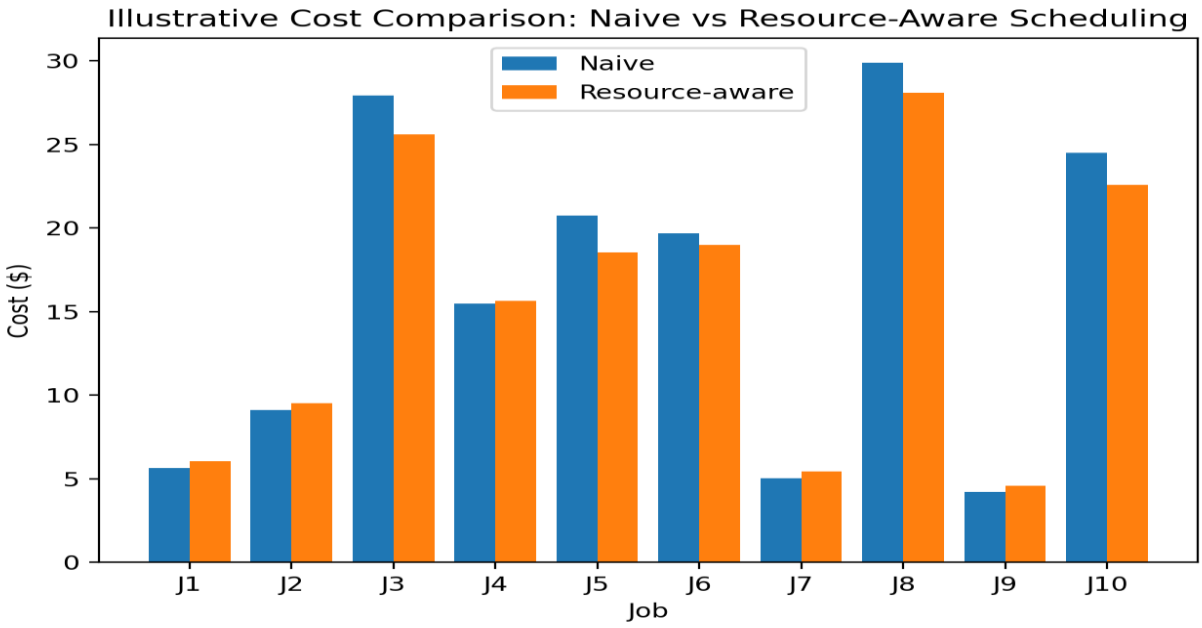

Fig 6: AI-Governed Pipeline Risk Scoring and Threshold-Based Alert Framework

6.1. Roles, Policies, and Decision Rights

Using AI to govern data pipelines means determining roles within the data ecosystem, defining the policies that govern those roles, and establishing decision rights over the AI-assisted data pipelines. These three elements of governance work together to determine which human actors can override decisions made by AI components within the data pipeline, and the broader impact of those decisions on the universe of data that the ecosystem supports.

Roles relate to the distinct set of responsibilities that a particular actor assumes relative to the data ecosystem of interest. The highest-level role in an organization that owns or manages a data ecosystem is that of

Chief Data Officer (CDO). One of the CDO's main responsibilities is determining who has the right to make decisions about larger investments. Behind the CDO are a variety of other specialized roles, including Data Architect, Data Engineer, Data Scientist, Data Custodian, and Data Steward. These and other roles are determined on a per-ecosystem basis.

Common types of policies include service-level agreements (SLAs), which define how parameters such as data freshness and data lineage should be maintained for a given dataset; data decision policies, which define the factors or metrics that should be used to make decisions over a specific dataset; and approval and communications policies, which define decision authorities and communication lines for various audiences. Together, these policies automate — to the point of being fully AI-governed — user-defined actions over datasets that require little human intervention. Such policies transfer the risk from any single user to the multitude of users of a data asset.

6.2. Risk Assessment and Mitigation

Data pipeline risk management requires a comprehensive understanding of key pipeline components, the interdependencies and systems integrated by the pipeline, the values of qualitative and quantitative operational metrics, and the implications of deviation from ranges for these metrics. Formal risk specification must define metrics considered to indicate risk and thresholds that, when exceeded, generate alerts or trigger closure actions. These thresholds are best expressed relative to normal ranges calculated from historical data. The set of components observable within the data pipeline and the established operational status of each component must also be considered. Categorization of pipeline components according to their operational status provides an effective basis for risk assessment. Quantifying qualitative operational states enables automatic evaluation for the presence of abnormal conditions based on value ranges. Availability of service components, together with a union of qualitative and quantitative metadata parameters defining the pipeline, provide the foundation for risk assessment.

Risk mitigation heuristically employs decentralized anomaly-detection/response modules directly attached or co-located with pipeline tasks. These actively monitor the aspects of pipeline resource consumption directly attributable to each task component, and detect anomalies either by supervised learning—modeling normal operating conditions—and signaling when the pipeline deviates from those conditions, or by unsupervised learning, clustering usage patterns and signaling outliers. In practice, both approaches are often used concurrently and unified into a unified model-checking controller. Process state information, quality metrics on both the source and generated datastores, and operating state of integrated tooling are also integrated, to prevent resource consumption outside normal operating conditions for both the pipeline and its components.

7. Conclusion

The principles of Cognitive Data Engineering proposed in this work enable the judicious use of AI models to improve data quality, manage data provenance and lineage, and build scalable Data as a Service infrastructure for real-time data pipelines. AI models can help identify and assess important quality metrics of datasets produced in automated data ecosystems, detect anomalies in their values and distributions, and suggest correction actions. Quality degradation can lead to unreliable and sometimes harmful insights when the end models are based on supervised, generative or reinforcement learning systems. Quality assessment is far more challenging in this case and needs to be automatically embedded in the production cycle of the datasets (e.g., prediction of model performance with respect to a set of quality metrics). Integrating the quality assessment and correction loops in a holistic manner is an additional aspect that needs to be researched.

In complex data ecosystems, trustworthy data provenance and lineage are essential for the responsible and ethical use of data as well as for enabling interoperability among datasets. Provenance and lineage information needs to be systematically captured, published and curated across the ecosystem to build trust, and promote interoperability and reuse of datasets. Solution proposals need to bring down the effort required for provenance capture to zero for the producers of data. This can be achieved by using specialised tools that capture provenance using observational, inferential, or user-annotation techniques. Cognitive Data Engineering proposes a novel automated Data Pipeline architecture to support Data as a Service infrastructure for data-rich real-time

applications and AI-enabled autonomic mechanisms for reducing Total Cost of Ownership (TCO). The technology enables the periodic execution of data pipelines as an alternative to hard real-time execution to reduce the TCO without compromising support for real-time data analytics.

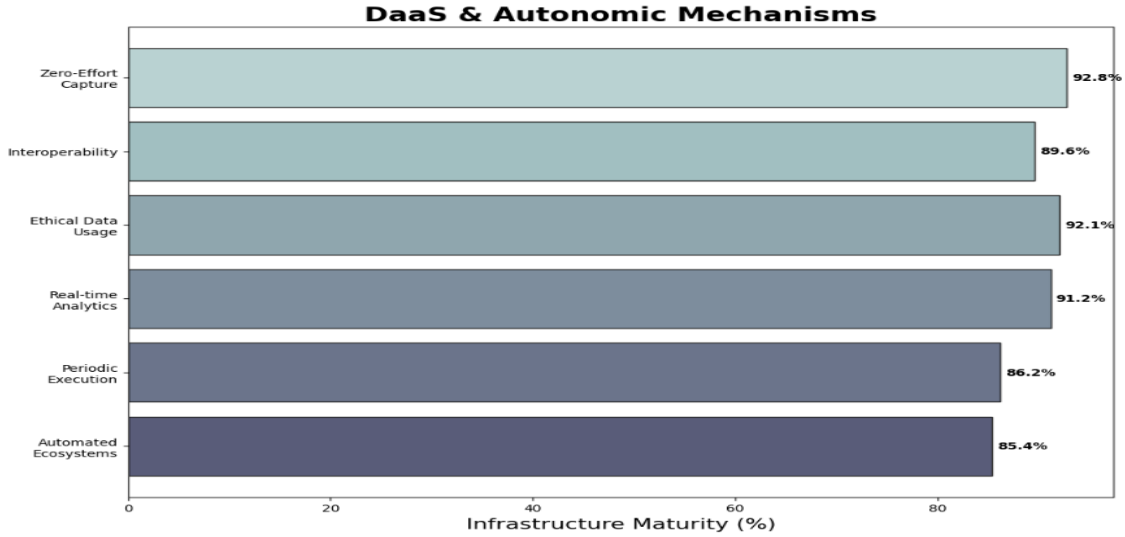

Fig 7: DaaS & Autonomic Mechanisms

7.1. Summary and Future Directions

Cognitive Data Engineering represents the design of large-scale data ecosystems aided by AI governing the quality, lineage, and cost of data pipelines. Data quality assurance receives particular emphasis, with provisioning activities for detecting and correcting quality anomalies integrated into the cognitive feedback loop. Quality metrics tailored for a diverse set of data stakeholders capture a wide variety of aspects, ranging from correctness to concept drift. AI governance provides the necessary role definition and policies to constitute and maintain an organization's data assets as a business service.

Despite much ground covered, a number of avenues remain to be pursued. First, all these aspects should be integrated into a cohesive end-to-end framework for cognitive data engineering, including supporting tooling such as a knowledge graph, workflow automation, and orchestration engine. Second, machine-learning approaches capable of addressing every aspect of AI-Powered Data Quality should be researched and developed. Third, similar treatment should be applied to Data Pipeline Cost Engineering. Special attention would need to be given to risk detection, assessment, and mitigation, as that aspect of the governance function is currently underexplored. Finally, further insights to be derived from conceptualizing Data Pipeline Quality as a supervised learning task should be pursued.

References

[1] Prashanth, B. S. (2026). Prediction of bank transaction fraud using TabNet—An adaptive deep learning architecture. Decision Support Systems.

[2] Naik, A. V., Sheelam, G. K., Panchakatla, N., Muthukumaran, K., & Saranya, K. (2025). Comprehensive Analysis on Depression Detection From Social Media Using Deep Learning and Transformer Architectures. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1–8). IEEE. 2025 International Conference on Communication, Computer, and Information Technology (IC3IT). https://doi.org/10.1109/ic3it66137.2025.11341160

[3] Wu, C., Xu, J., Wang, K., Han, W., & Chai, H. (2026). ETTracker: A fund tracking framework for anti-money laundering on Ethereum. Expert Systems with Applications, 296, 128900.

[4] Tieu, T. H. T., (2026). Integrating the fraud triangle with machine learning for financial misstatement detection. Cogent Business & Management.

[5] Pallapu, S. R., Aitha, A. R., K, Sudhakar., Vandhana, K., & Chelladurai, S. (2025). GAN-Augmented Transformer Framework for Cross-Domain Video Style Transfer. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1–6). IEEE. https://doi.org/10.1109/ic3it66137.2025.11341104.

[6] Rodríguez Valencia, L., (2025). A systematic review of artificial intelligence applied to financial fraud detection and anti-money laundering. Journal of Risk and Financial Management, 18, 612.

[7] Gadimov, E., & Mustafayev, E. (2025). Real-time suspicious detection framework for financial data and fraud prevention. Discover Internet of Things, 5, 1–22.

[8] Chary, D. V., Meda, R., C, J. S. Mary., Narasimhachari, J. P., & A S, Y. (2025). TriFusionFormer: Tri-Modal Fusion Transformer Using Gated Modality Control and Multi-Scale Attention for Emotion Recognition. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1–8). IEEE. 2025 International Conference on Communication, Computer, and Information Technology (IC3IT). https://doi.org/10.1109/ic3it66137.2025.11341646.

[9] Alexandre, C. R., (2023). Incorporating machine learning and a risk-based strategy for anti-money laundering decision support. Expert Systems with Applications, 211, 118500.

[10] Jensen, R. I. T., & Iosifidis, A. (2022). Qualifying and raising anti-money laundering alarms with deep learning. Expert Systems with Applications, 201, 117105.

[11] Pamisetty, A., Paleti, S., Adusupalli, B., Singireddy, J., Inala, R., & Nagabhyru, K. C. (2025). Explainable AI Systems for Credit Scoring and Loan Risk Assessment in Digital Banking Platforms. In 2025 IEEE 13th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS) (pp. 1478–1483). IEEE. https://doi.org/10.1109/idaacs68557.2025.11322144

[12] Brummer, C., & Yadav, Y. (2019). Fintech and the innovation trilemma. Georgetown Law Journal, 107(2), 235–307.

[13] Barberis, J., & Chishti, S. (2020). The RegTech book: The financial technology handbook for investors, entrepreneurs and visionaries. Wiley.

[14] Radha, S., Gottimukkala, V. R. R., Thottara, S., Vandhana, K., & J, Gokulraj. (2025). Adaptive Video Streaming Over 5G Networks Using Deep Reinforcement Learning with Closed-Loop Feedback Mechanism for Bitrate Control. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1–6). IEEE. 2025 International Conference on Communication, Computer, and Information Technology (IC3IT). https://doi.org/10.1109/ic3it66137.2025.11341184

[15] European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

[16] European Parliament and Council of the European Union. (2022). Regulation (EU) 2022/2554 on digital operational resilience for the financial sector (DORA). Official Journal of the European Union.

[17] Bhasgi, S. S., Garapati, R. S., B, Ayshwarya., Sasikala, M., & J, Srinivasan. (2025). Medical Image Fusion of Magnetic Resonance Imaging and Computed Tomography Using Learned Wavelet Complex Adapter. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1–6). IEEE. https://doi.org/10.1109/ic3it66137.2025.11340892

[18] Financial Action Task Force. (2020). Opportunities and challenges of new technologies for AML/CFT. FATF.

[19] Financial Action Task Force. (2021). Guidance on digital identity. FATF.

[20] P, R., Nagabhyru, K. C., C, M., Srinu, M., Kaur, H., & N, N. (2025). K-Means-KNN Hybrid Model for Efficient Intrusion Detection in Cloud-based IoT Systems. In 2025 10th International Conference on Communication and Electronics Systems (ICCES) (pp. 1583–1588). IEEE. 2025 10th International Conference on Communication and Electronics Systems (ICCES). https://doi.org/10.1109/icces67310.2025.11336840

[21] Basel Committee on Banking Supervision. (2013). Principles for effective risk data aggregation and risk reporting (BCBS 239). Bank for International Settlements.

[22] National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0). U.S. Department of Commerce.

[23] Bargavi, N., Athawale, S. G., Amistapuram, K., & Aitha, A. R. (2026). Safeguarding Consumer Data in Digital Insurance: Legal Frameworks and Ethical Imperatives. International Insurance Law Review, 34(S1), 272-284.

[24] International Organization for Standardization. (2018). ISO/IEC 27001:2018 Information security management systems—Requirements. ISO.

[25] International Organization for Standardization. (2019). ISO/IEC 27002:2019 Information security controls. ISO.

[26] Jagtap, S., Inala, R., Venu, M., & Divya, T. V. (2025, October). Large-Scale Crowd Flow Prediction Using Temporal Convolutional Network with Spatio-Temporal Attention. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1-6). IEEE.

[27] Committee of Sponsoring Organizations of the Treadway Commission. (2013). Internal control—Integrated framework. COSO.

[28] U.S. Federal Reserve. (2011). Supervisory guidance on model risk management (SR 11-7). Board of Governors of the Federal Reserve System.

[29] Ramana, B., Sheelam, G. K., Pandya, T., Rai, A. K., Kumar, V. A., & Kukreti, A. (2025). Exploring the Potential of NOMA in 6G Through Comparative Analysis with OMA Techniques. In 2025 IEEE 5th International Conference on ICT in Business Industry & amp; Government (ICTBIG) (pp. 1–6). IEEE. 2025 IEEE 5th International Conference on ICT in Business Industry & Government (ICTBIG). https://doi.org/10.1109/ictbig68706.2025.11323270

[30] European Banking Authority. (2019). Guidelines on ICT and security risk management. EBA.

[31] Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Now Publishers.

[32] Gupta, D. K., Purushotham, K., Dheer, G., P, S., Gottimukkala, V. R. R., & Kapoor, S. (2025). Semantic Feature Learning Using Transformer-Based Deep Neural Networks. In 2025 IEEE 5th International Conference on ICT in Business Industry & amp; Government (ICTBIG) (pp. 1–6). IEEE. https://doi.org/10.1109/ictbig68706.2025.11323734

[33] Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021). Extracting training data from large language models. USENIX Security Symposium, 2633–2650.

[34] Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation does not exist in the GDPR. International Data Privacy Law, 7(2), 76–99.

[35] Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. Statistical Science, 17(3), 235–255.

[36] R, Lathakumari. K., Varri, D. B. S., Atreya, M., B, Madhumala. R., & Khemka, S. (2025). Pearson Correlation Coefficient and Agglomerative Clustering with Gated Recurrent Unit Integrated with Linear Attention for Cyber-Physical Control and Monitoring System in Next-Generation Industrial Systems. In 2025

2nd International Conference on Software, Systems and Information Technology (SSITCON) (pp. 1–6). IEEE. https://doi.org/10.1109/ssitcon66133.2025.11342101

[37] Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. IEEE Symposium Series on Computational Intelligence, 159–166.

[38] Thutari, R. T., Garapati, R. S., B M, Manjula., R K, Supriya., & M, Senbagan. (2025). Adaptive Access Control and Authentication Management for IoT Using Attention-GRU and Reinforcement Learning. In 2025 2nd International Conference on Software, Systems and Information Technology (SSITCON) (pp. 1–6). IEEE. https://doi.org/10.1109/ssitcon66133.2025.11342003.

[39] Quah, J. T. S., & Sriganesh, M. (2008). Real-time credit card fraud detection using computational intelligence. Expert Systems with Applications, 35(4), 1721–1732.

[40] Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review. Decision Support Systems, 50(3), 559–569.

[41] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[42] Kumar, I., Nagabhyru, K. C., G, Naveen. I., V, Prabhakaran. M., & V, Sruthy. K. (2025). Adaptive Meta-Knowledge Transfer Network with Feature Hallucination and Attention for Low-Shot Object Detection in Aerial Images. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1–6). IEEE. 2025 International Conference on Communication, Computer, and Information Technology (IC3IT). https://doi.org/10.1109/ic3it66137.2025.11341447

[43] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

[44] Russell, S., & Norvig, P. (2021). Artificial intelligence: A modern approach (4th ed.). Pearson.

[45] Babaiah, Ch., Dobriyal, N., Shamila, M., Aitha, A. R., Patel, S. P., & Upodhyay, D. (2025). Intelligent Fault Detection and Recovery in Wireless Sensor Networks Using AI. In 2025 IEEE 5th International Conference on ICT in Business Industry & amp; Government (ICTBIG) (pp. 1–6). IEEE. https://doi.org/10.1109/ictbig68706.2025.11323980.

[46] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD Conference, 1135–1144.

[47] Rongali, S. K. (2025, August). AI-Powered Threat Detection in Healthcare Data. In 2025 International Conference on Artificial Intelligence and Machine Vision (AIMV) (pp. 1-7). IEEE.

[48] Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS Quarterly, 36(4), 1165–1188.

[49] Van der Aalst, W. (2016). Process mining: Data science in action (2nd ed.). Springer.

[50] Ehrmann, T. S., Bull, D. L., Phipps, E. T., Brown, G. H., & Kolla, H. N. (2025). Identifying Increased MJO Dimensionality through Canonical Polyadic Decomposition. Authorea Preprints.

[51] Zaharia, M., Das, T., Li, H., Shenker, S., & Stoica, I. (2016). Discretized streams: Fault-tolerant streaming computation at scale. Communications of the ACM, 59(6), 80–87.

[52] Carbone, P., Katsifodimos, A., Ewen, S., Markl, V., Haridi, S., & Tzoumas, K. (2015). Apache Flink: Stream and batch processing in a single engine. IEEE Data Engineering Bulletin, 38(4), 28–38.

[53] Ashokkumar, S., Amistapuram, K., C, Bharathi., M, Dhanamalar., & J, Gokulraj. (2025). Attention-Guided Spatial Temporal Framework for Deepfake Detection on Social Video Platforms. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1–6). IEEE. https://doi.org/10.1109/ic3it66137.2025.11341690

[54] Ongaro, D., & Ousterhout, J. K. (2014). In search of an understandable consensus algorithm (Raft). USENIX Annual Technical Conference, 305–319.

[55] Hunt, P., Konar, M., Junqueira, F. P., & Reed, B. (2010). ZooKeeper: Wait-free coordination for internet-scale systems. USENIX Annual Technical Conference.

[56] Srikanth, T., Segireddy, A. R., Elavarasi, S. A., K, S. M. Reddy., & K, M. Krishnan. (2025). STaSFormer-SGAD: Semantic Triplet-Aware Spatial Flow-Guided Spatio-Temporal Graph for Anomaly Detection in Surveillance Videos. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1–7). IEEE. https://doi.org/10.1109/ic3it66137.2025.11341322

[57] Gilbert, S., & Lynch, N. (2002). Brewer's conjecture and the feasibility of consistent, available, partition-tolerant systems. ACM SIGACT News, 33(2), 51–59.

[58] GUNTUPALLI, R. (2025). EXPLAINABLE AI IN CLINICAL DECISION SUPPORT: INTERPRETABLE NEURAL MODELS FOR TRUSTWORTHY HEALTHCARE AUTOMATIONEXPLAINABLE AI IN CLINICAL DECISION SUPPORT: INTERPRETABLE NEURAL MODELS FOR TRUSTWORTHY HEALTHCARE AUTOMATION. TPM–Testing, Psychometrics, Methodology in Applied Psychology, 32(S9 (2025): Posted 15 December), 462-471.

[59] Newman, S. (2021). Building microservices (2nd ed.). O'Reilly Media.

[60] Pareyani, S., Goswami, S., Geetha, Y., Dimri, S. K., Niharika, D. S., & Amistapuram, K. (2025). Smart Resource Allocation in Wireless Sensor Networks Through AI Techniques. In 2025 IEEE 5th International Conference on ICT in Business Industry & amp; Government (ICTBIG) (pp. 1–6). IEEE. https://doi.org/10.1109/ictbig68706.2025.11323661

[61] Hohpe, G., & Woolf, B. (2004). Enterprise integration patterns. Addison-Wesley.

[62] Richter, P., & Dinh, T. (2020). Event-driven architectures: Concepts and practices. IEEE Software, 37(5), 12–20.

[63] PIONEERING SELF-ADAPTIVE AI ORCHESTRATION ENGINES FOR REAL-TIME END-TO-END MULTI-COUNTERPARTY DERIVATIVES, COLLATERAL, AND ACCOUNTING AUTOMATION: INTELLIGENCE-DRIVEN WORKFLOW COORDINATION AT ENTERPRISE SCALE. (2025). Lex Localis - Journal of Local Self-Government, 23(S6), 8598-8610. https://doi.org/10.52152/a5hkbh02

[64] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., & Zaharia, M. (2010). A view of cloud computing. Communications of the ACM, 53(4), 50–58.

[65] Beyer, B., Jones, C., Petoff, J., & Murphy, N. R. (2016). Site reliability engineering: How Google runs production systems. O'Reilly Media.

[66] Yandamuri, U. S. (2026). AI-Enabled Workflow Automation and Predictive Analytics for Enterprise Operations Management. Management, 3(1), 15-24.

[67] CNCF. (2023). Cloud native security whitepaper (2nd ed.). Cloud Native Computing Foundation.

[68] FinOps Foundation. (2024). FinOps framework: Principles, capabilities, and practices for cloud financial management. FinOps Foundation.

[69] Guntupalli, R. (2025). Federated Deep Learning for Predictive Healthcare: A Privacy-Preserving AI Framework on Cloud-Native Infrastructure. Vascular and Endovascular Review, 8(16s), 200-210.

[70] Google Cloud. (2021). Cloud FinOps: Managing cloud costs at scale. Google.

[71] Dutta, P., Mondal, A., Vadisetty, R., Polamarasetti, A., Guntupalli, R., & Rongali, S. K. (2025). A novel deep learning rule-based spike neural network (SNN) classification approach for diagnosis of intracranial tumors. International Journal of Information Technology, 1-8.

[72] Microsoft. (2023). Cloud adoption framework: Cost management and governance. Microsoft.

[73] Ehrmann, T., Bull, D. L., Phipps, E., & Kolla, H. (2025). Identifying Increased Dimensionality in the Madden-Julian Oscillation through Canonical Polyadic Decomposition. AGU25.

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