Sarcouncil Journal of Engineering and Computer Sciences

ISSN(Online): 2945-3585

Volume- 04| Issue- 07| 2025

Review Article **Received:** 07-06-2025| **Accepted:** 01-07-2025 | **Published:** 20-07-2025

Demystifying Modern Data Engineering: From ETL to AI-Driven Pipelines

Jimish Jitendra Kadakia
Independent Researcher, USA

Abstract: What a modern data engineer does today is radically different than a traditional Extract, Transform, and Load (ETL) developer, or even now a data engineer who applies complex machine learning (ML) models that allow organizations to rethink how to work with data. The evolution is more than just a migration from batch to event-driven, automatic, and elastic real-time data pipeline. The shift is more profound than just new hybrid batch and event-driven architectures and proof of concepts (POC) with speed and scale. In modern data architectures, new cloud-native configurations offer different advantages and scales of distributed environments/computing, containers, advanced orchestration, and intelligent decisions. New data stacks based uniquely and differentiated in the organization of analytics introduce a new complex pile of platforms and services, including cloud-native transformation platforms, orchestration platforms, a large data warehouse, stream processing technologies, and unique end-to-end processing environments. Together with machine learning and automated pipeline tuning, automated anomaly detection, and automated predictive maintenance, organizations are building more reliable,better-performing processing pipelines. Also, merging machine learning algorithms into traditional data processing started to introduce what is commonly referred to as autonomic computing - a self-healing computing model able to self-correct from faults to determine a resource provisioning plan autonomously. The decentralisation of data ownership through data mesh architecture, with an appropriate degree of governance and enhancements that enable large language model engagement with data pipelines that offer an alternative natural language interface for easier pipeline authoring and unrestricted interactions, will be the focus of future data engineering evolution. The evolution is proof of strategic maturation—an evolution toward intelligent data infrastructures that autonomously respond, manage complex analytical workloads, rely less on human interaction, and optimize operational efficiency for end-user experiences across large enterprise data environments.

Keywords: Data engineering transformation, cloud-native architectures, AI-driven pipelines, real-time processing, intelligent automation, data mesh.

INTRODUCTION

Data engineering has made the dramatic shift from being solely a technical discipline to being the essential backbone of modern business intelligence and analytics operations. This field has experienced rapid growth, with big data analytics profits expected to rise by 56 percent over the four years from $130 billion to $203 billion between 2016 and 2020 (Columbus, L. 2018).. Entrenched data professionals were previously operating based upon rigid, schedule-driven processes for transferring information through systems, while today, experienced data engineers are designing complex, intelligent pipelines that can react to business needs in real-time situations.

This evolution involves a lot more than just technological superiority; it represents a complete paradigm shift in organizational philosophy toward data as a strategic business asset. Given that businesses generate over 2.5 quintillion bytes of data per day and that 90 percent of the world's data was created in the last two years, there is an undeniable need for more advanced and automated channels (Columbus, L. 2018). The previous task-based divisions of data extraction, transforming, and loading are disappearing and being replaced with continuous workflows that are event-driven and capable of adapting, learning, and self-optimizing.

Companies that have adopted engineered methodologies realize significantly better performance metrics than traditional approaches. Organizations with the ability to process data in real-time tend to have revenue growth rates five to six times higher than organizations that only have a batch processing capability (Analytics, M. 2016). In a competitive landscape, it has become crucial to leverage analytics to obtain such an advantage, with high-performance organizations applying analytics at a significant scale and in many parts of the organization, rather than just for a single use case in a single department.

Modern data engineering can deliver observable operational improvements across multiple areas. For example, in advanced implementations, organizations delivered insights 23% faster, had 19% lower overall data processing costs, and had 67% fewer data quality issues when compared to traditional ETL implementations capability (Analytics, M. 2016). The data pipeline evolution has raised the prominence of data engineering and brought it out of the shadows as critical to enabling artificial intelligence, machine learning, and real-

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

***Corresponding Author:** Jimish Jitendra Kadakia

DOI- https://doi.org/10.5281/zenodo.16225826

time decision-making capabilities that are needed for organizational competitive positioning in the current data economy.

The transition to intelligent data channels evidences the broader acknowledgement across the industry that data haste, variety, and volume are and will continue to grow exponentially. Organizations will need to adapt their data processing architectures to increasingly complex systems, while maintaining reliability and performance measures. Modern data engineering is enabling organizations to continuously optimize the flow of data end-to-end, turn data processes into automated processes with ride detection and resolution capabilities, and to use predictive maintenance approaches to minimize system downtime and maximize operational performance across the enterprise-wide ecosystem of data.

The Evolution Beyond Traditional ETL

Traditional Extract, Transform, Load processes used to run based on methodical scheduling patterns that only ran overnight or on a weekly batch and represented enterprise data management practices for decades. Traditional data engineering teams would systematically pull data from multiple source systems, apply the necessary transformation logic in a staging area, and then load the transformed data into the central data warehouse. The traditional method has shown to be dependable and has a tried and true operating process, but it has serious time lags between the creation of the data and the availability of data for analytics, and very often has time windows of 24 to 168 hours, depending on which batch processing schedule was being implemented.

Today, modern data engineering teams have adopted streaming architectures and evolution technology, eliminating the traditional time constraints and typical temporal problems. The real-time analytics market has a tremendous growth potential and is projected to be worth USD 193.71 billion by 2032, supported by the growing use of artificial intelligence and the Internet of Things technologies in business (GlobeNewswire, 2025). Today's pipeline architecture reacts to change in data as that change occurs as an event and does not adhere to any predetermined batch schedule. Organizations can now leverage advanced streaming functionality to make strategic decisions based on real-time operations instead of historical operations based on the last time data was processed.

Real-time processing frameworks control continuous data streams in an operationally efficient manner, performing complex transformations and delivering event information to specific destinations within seconds or minutes of original production. Today's streaming platforms exhibit impressive throughput performance, processing millions of discrete events per second and sending data back to applications at or below microsecond response times that are required by high-performance applications to immediately access data (GlobeNewswire, 2025). The architecture shift eliminates traditional batch processing restrictions; organizations are now able to respond in real-time to changing market conditions and shifting operational needs.

Transformational logic approaches have also evolved from firm/rigid rule embeddedness (removing firm structures and rules) to highly programmable and changeable approaches that enable inertia-defying operational flexibility and long-term sustaining maintainability. Data engineering professionals today develop transformation logic with a version control system that provides innovative and collaborative workflows, testing processes, and immediate production deployment systems throughout the full application development lifecycle. The generally accepted approach is termed 'analytics engineering' that facilitates 'operational hand-offs' across disparate divisions between traditional data engineering approaches and conventional software engineering approaches so that workflows are integrated and productivity and reliability are assertively enhanced.

Companies with advanced data processing capabilities steadily produce better competitive outcomes. Data-driven decision-making is transforming all kinds of companies as they can respond faster to changes in the marketplace or customer behavior, become more efficient with their operations, and find new sources of revenue - all based on (more) insights gained from analytics. Companies employing advanced analytics publicly report (and we have no reason to believe they are not true) objectively higher revenue growth rates than those contingent on batch processing as a traditional enterprise information model. The analytics transformation is fundamental to the magnetic retention of market position, as both data volume increases quickly than the business is ready for it, and decision-making times shrink with equal disproportion. Many modern enterprise-level

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

implementations are also leveraging continuous integration and continuous deployment strategies that embrace rapid development and advancement of any transformation logic, while not taking production down or sacrificing the ultimate stability of the system.

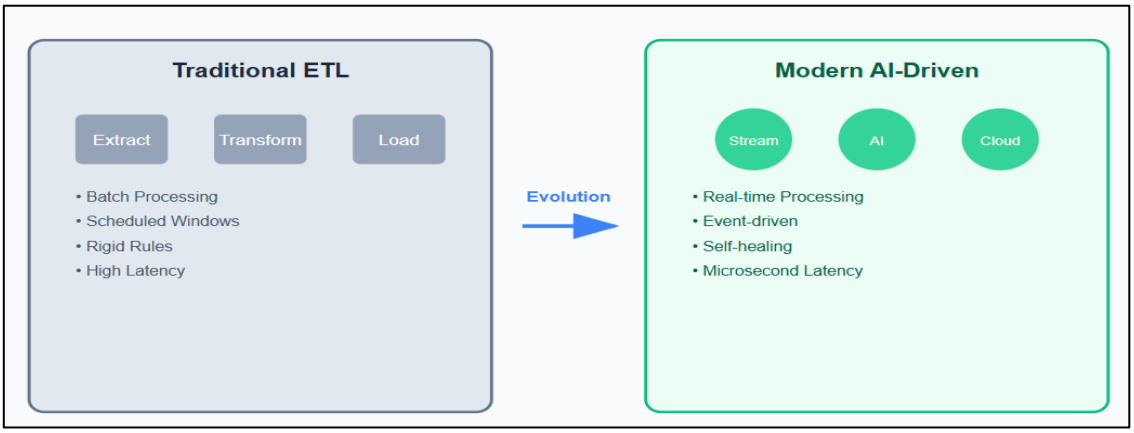


Fig 1. Evolution from Traditional ETL to Modern AI-Driven Pipelines (GlobeNewswire, 2025; Medium, 2023)

CLOUD-NATIVE ARCHITECTURE AND ORCHESTRATION

Distributed Processing Foundations

Cloud-based infrastructures have totally revolutionized data infrastructure engineering approaches by offering elastic scalability opportunities while entirely removing the traditional infrastructure management burden, which was previously a limiting operational constraint. The cloud landscape continues to show remarkable growth, with 94% of organizations running on the cloud at some capacity, and with global public cloud spending at $490 billion as of 2022 (Slingerland, C. 2025). Modern data pipeline architectures employ advanced distributed computation frameworks that can automatically scale computational resources in response to shifting workloads, real-time processing, and workloads executed without users needing to take action.

Elastic scaling capabilities allow organizations to process very large data sets when operationally performing at peak without needing to invest in the associated compute hardware that may only be required occasionally over the course of the calendar year. The impact and value of cost efficiency are real, with cloud usage reducing IT costs by roughly 15% while improving efficiency across aspects by 20-25% across industries (Slingerland, C. 2025). Containerized architectures have allowed for the level of flexibility in the pipeline process via the standardization of deploying the infrastructure in a consistent fashion to support its function across a variety of operational configurations. Data engineers systematically package all transformation logic, dependencies, and configuration parameters into portable containerized units that will execute reliably throughout the development, testing, and final production environments, thereby mitigating deployment 'friction' and enhancing pipeline reliability and operations.

Modern distributed processing frameworks allow horizontal scaling across potentially hundreds and even thousands of nodes, permitting companies to execute Petabyte-scale datasets in a reasonable time and at reasonable operating prices. Cloud-native container orchestration platforms dynamically allocate resources, managing the automatic provisioning of additional computational capacity during burst or peak periods, while optimally declining utilization during low or idle times for cost-efficient operation.

Workflow Orchestration

The latest pipeline orchestration techniques have advanced from basic scheduling tools to comprehensive workflow management systems that provide efficient frameworks with orchestrated multi-system operations and unprecedented accuracy and reliability. The global workflow orchestration market is expected to grow with future potential, with revenue projections rising from $4.53 billion in 2021 to $18.31 billion by 2030, driven by a compound annual growth rate of 16.7% in influence and dependence on data pipeline automatic management solutions (Businesswire, 2022) . Modern orchestration platforms offer complex knowledge of task interdependence and strong failure recovery, and

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

they monitor the extensive performance necessary for operational complexity in enterprise functions.

Today's orchestration systems can map workflows with multiple cloud providers, multiple processing frameworks, and heterogeneous system architecture, all while ensuring consistent operation and performance benchmarks in distributed environments. The most recent orchestration platforms can work with two triggering processes: time-based scheduling and event-based scheduling, allowing pipelines to respond in a timely manner to events such as when a file arrives, or a database entry is modified, or when an API notifies (Businesswire, 2022). Advanced orchestrators use intelligent optimization algorithms that dynamically change execution paths based on today's worker availability and evaluation of worker metrics over time, based on a lot of operational data intelligence.

Today's workflow management systems include predictive analytics features to evaluate execution patterns, to project resource allocation strategies, and to forecast potential bottleneck locations well in advance of performance degradation. Advanced orchestration platforms come complete with full dashboards that show real-time pipeline status, resource utilization, and operational information that can be consumed to manage proactively across many distributed processing specifications.

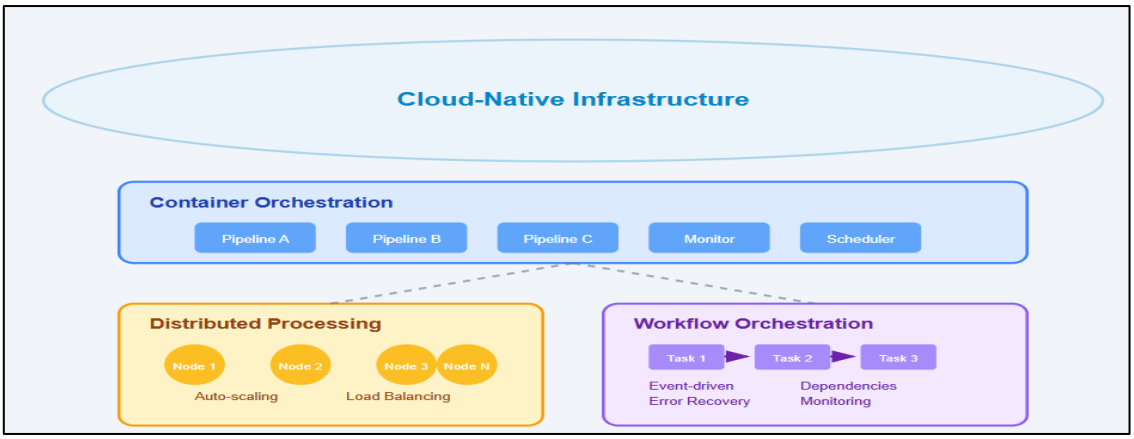


Fig 2. Cloud-Native Architecture Components and Orchestration Framework (Slingerland, C. 2025; Businesswire, 2022)

Key Technologies in the Modern Stack

Current data engineering environments are designed from multiple categories of specialized technologies, and they collectively make up overall data processing "ecosystems" to be able to complete enterprise-sized workloads with as low an overhead level as could be achieved before. Data transformation platforms represent the new foundation in this area as they provide a pathway for analysts to create transformation logic in an easy-to-use and familiar syntactic model. These new platforms take the complexity of distributed processing frameworks away from the analysts. It is worth noting that the global data integration market is anticipated to grow rapidly annually, with valuations reaching $19.6 billion by 2026, showcased with a compound annual growth rate of 13.6% from a 2021 baseline, continuously fueled by the significant demand for integrated connectivity across enterprise data systems (MarketsandMarkets, 2021). Some advanced platforms can automatically create optimal code and manage widespread complex system dependencies without the need for human intervention, while significantly reducing operational costs and average time to production.

In recent years, a variety of workflow management or orchestration tools have emerged that offer advanced visual interfaces with advanced dependency management systems, strong error handling, and other monitoring functions that are required in production environments. Today, orchestration systems integrate with different data sources and destination systems with a seamless, unified control plane that centralizes the management of data operations across heterogeneous technology ecosystems. Organizations can integrate workflows across multiple cloud providers, on-premise systems, and hybrid approaches, meaning organizations can maintain operational consistency no matter how complex their underlying infrastructure is, as they support real-time data processing needs.

Cloud data warehouses represent an entirely new level of storage and compute architectures that

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

offer virtually unlimited scaling for storage capacity by taking advantage of a pay-per-use model that aligns costs with variable consumption characteristics. Today's platforms apply architectural separation of storage and compute, allowing organizations to scale either one in whichever direction they need, depending on operational needs and budget. When it comes to adopting collaborative and self-service analytics, organizations leveraging advanced data capabilities experience 23% higher profitability than their competitors who lack a multi-dimensional analytics capability (SarahLee, 2025). Having a loosely coupled architecture enables organizations to resource dynamically, providing the ability to automatically scale on double the compute capacity during periods of high processing, and minimize cost during periods of low utilization.

Stream processing frameworks provide capabilities for the real-time ingestion and transformation of data, allowing organizations to build reactive applications that respond to events as they occur in the real world. Next-generation streaming frameworks may abstract away the complexity of distributed stream processing while having simplified programming interfaces that enable organizations to develop solutions at reduced cost and timeframes compared to the disruptive systems that are in conflict with conventional data engineering strategies. Modern streaming frameworks process millions of unique events per second, with latencies that are measured in microseconds, and can support high-throughput applications that must have immediate data availability and enable real-time analytical observations.

The confluence of the point solutions creates synergies that augment the capabilities of each platform into a complete data engineering fingerprint that can enable multi-faceted analytical workloads across diverse industry verticals. Evidence illustrates the substantial impacts encountered by organizations transitioning to data-driven decision-making. Companies that have clearly defined analytics strategies achieved higher productivity levels of 5% and 6% higher profits than organizations that continued to rely on traditional decision-making strategies (SarahLee, 2025). Contemporary Platforms are now embedding machine-learning algorithms that automatically optimize performance metrics, can successfully predict resource-consumptive behaviors based on historical data usage, and implement predictive maintenance outcomes that minimize system downtime in order to maximize operational throughput within distributed processing environments.

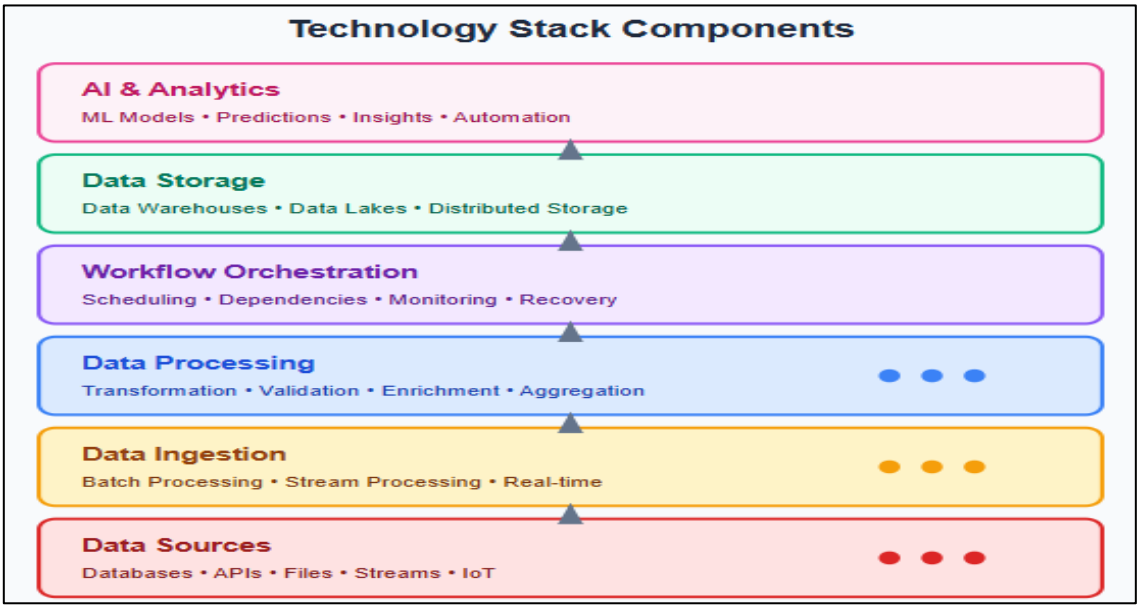


Fig 3. Modern Data Engineering Technology Stack Layers (MarketsandMarkets, 2021; SarahLee, 2025)

AI INTEGRATION AND INTELLIGENT AUTOMATION

Automated Pipeline Optimization

The arrival of artificial intelligence has drastically altered how businesses can operate data pipelines and autonomously manage them across enterprise environments. Machine learning algorithms analyze large volumes of pipeline performance data to find areas of possible improvement that humans or operators may not recognize while

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

conducting standard performance monitoring activities. There is sufficient evidence demonstrating the use of artificial intelligence technologies in various industries that have shown potential to enhance the efficiency and decision-making processes of firms. In particular, AI applications seem to work best in complex processing environments where improved decision making and efficiency arise out of complex structures such as pattern recognition and predictive analysis (Almotairi, K. H. *et al.,* 2023). Modern AI systems can autonomously change what resources should be allocated and what alternative execution patterns should be considered. They can also examine historical performance patterns and real-time system load data to identify optimal scheduling times to query access data in a timely manner.

Intelligent query optimization systems analyze SQL pattern structures and third-party historical access patterns to formulate recommendations for indexing, partitioning, and materialized views. In conjunction with automation, performance management becomes easier and less stressful in the face of increasing data volumes and more complex operations, caused by the demand for enterprise decision-making. New-age optimization systems are smarter and use deep learning algorithms to adapt to constantly changing workload patterns, allowing access or performance improvements to be implemented by the system autonomously, without human involvement or requiring system downtime. Reliability during workload variability, maintaining expectations across critical business operations, is a desired characteristic of an optimized pipeline or data processing workflow.

Anomaly Detection and Quality Monitoring

AI-powered monitoring systems carry out continuous assessment of data quality metrics, pipeline efficiency metrics, and business metrics to identify unusual patterns that deviate from established operational baselines. Sophisticated monitoring platforms are quite good at accurately detecting anomalies using complex implementation of machine learning models, which learn normal operational patterns through continuous, ongoing observation and generate an automated alert when deviant events take place - often identifying problems before business users have noticed service has been disrupted or they have experienced a data quality issue (Almotairi, K. H. *et al.,* 2023). Intelligent systems can rapidly assess and process a vast amount of operational data while still maintaining a high degree of accuracy in assessing differences between regular variances from identified baselines compared to true anomalies, which require action.

Sophisticated monitoring platforms also have parameters to correlate anomalies across the different stages of the pipeline, allowing for rapid identification of root causes, which would require extensive manual research otherwise. Advanced correlation algorithms catch complex correlations, e.g., engaging with sudden increases in processing time, related to changes to some characteristic related to the source data - eliminating more time wasted reaching resolution of a performance issue and reducing ongoing operational impacts across downstream systems with no service disruption to the business while allowing ongoing availability of continuous service.

Predictive Maintenance

Machine learning models perform holistic examinations of pipeline health metrics to assess potential system failure before a real critical problem occurs in production environments. The use of AI-enabled maintenance evolves operational capabilities, as artificial intelligence technologies allow organizations to shift from reactive to proactive maintenance approaches that foster resource efficiencies and mitigate unplanned system failures (Choudhary, S. 2024). More advanced models take in multiple factors in proactive maintenance related to impending system failure, while managing resource-use patterns and controlling for historical error rate behaviors, along with mapping external dependency relations for recommendations for maintaining better system reliability while minimizing operating costs.

Predictive maintenance capabilities encompass more than performance metrics concerning technology; they include total business impact assessments to quantify the business effects of various failure scenarios and failures as measured by potential financial consequences. More advanced models estimate business costs in relation to failure types, allowing technical teams to weight maintenance tasks according to potential operational impact rather than relying solely on technical severity metrics. AI technologies drive continuing changes to traditional approaches to maintenance across enterprise contexts (Choudhary, S. 2024).

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

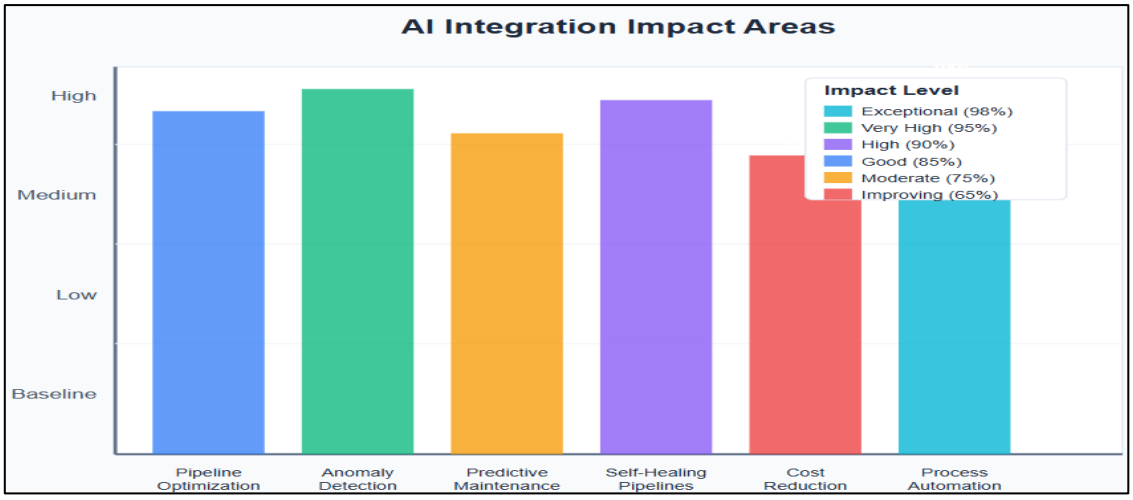


Fig 4. AI Integration Impact Areas in Data Engineering (Almotairi, K. H. *et al.*, 2023; Choudhary, S. 2024)

Future Trends and Implications

The data engineering landscape continues evolving toward greater automation and intelligence, with self-healing pipeline architectures becoming increasingly sophisticated in autonomous failure recovery capabilities. Contemporary systems extend far beyond simple task restart mechanisms, implementing advanced capabilities to adjust processing parameters dynamically, switch to alternative data sources seamlessly, or modify execution strategies based on comprehensive failure pattern analysis. The global intelligent process automation market demonstrates remarkable expansion potential, with valuations projected to grow from $15.80 billion in 2023 to $51.35 billion by 2032, representing a compound annual growth rate of 13.3% driven by increasing enterprise demand for autonomous operational capabilities across diverse industry sectors (Precedence Research, 2024). Advanced self-healing systems leverage machine learning algorithms that analyze historical failure patterns to predict optimal recovery strategies, significantly reducing system downtime and minimizing human intervention requirements across complex data processing environments.

Data mesh architectures have gained considerable traction as organizations increasingly recognize the inherent limitations of traditional centralized data platform approaches. The data mesh paradigm treats data as a strategic product, with domain-specific teams assuming ownership of respective data pipelines while maintaining adherence to common standards and interface specifications. The shift toward decentralized data architectures represents a fundamental transformation in organizational data management philosophies, requiring a comprehensive rethinking of data governance frameworks, quality assurance protocols, and discoverability mechanisms across distributed enterprise environments. Intelligent process automation technologies enable organizations to streamline operations, reduce costs, and improve efficiency by automating repetitive tasks and complex workflows that previously required extensive manual oversight (Precedence Research, 2024).

The integration of large language models (LLMs) into data engineering workflows demonstrates revolutionary potential for democratizing pipeline development processes across technical skill levels. Advanced AI systems possess capabilities to generate sophisticated transformation logic from natural language descriptions, automatically document complex pipeline architectures, and provide intelligent recommendations based on best practices derived from analysis of thousands of similar implementations. The Asia Pacific generative AI market reflects an exceptional growth trajectory, with regional adoption accelerating rapidly as organizations recognize the transformative potential of AI-powered development tools for enhancing productivity and innovation capabilities (The AI Journal, 2024). Large language models enable non-technical domain experts to participate directly in data pipeline creation, reducing development bottlenecks and accelerating time-to-value for analytical initiatives.

Contemporary AI-powered development environments provide intelligent code completion, automated testing generation, and comprehensive documentation creation that significantly reduce manual development overhead while improving code quality and maintainability. The

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

democratization of data engineering through natural language interfaces enables broader organizational participation in data pipeline development, fostering innovation and reducing dependency on specialized technical resources. Advanced language models incorporate domain-specific knowledge and industry best practices, automatically suggesting optimizations and identifying potential issues during pipeline design phases, thereby enhancing overall system reliability and performance across diverse operational environments. The Asia Pacific region continues to lead generative AI adoption, with organizations leveraging advanced language models to transform traditional development workflows and accelerate digital transformation initiatives (The AI Journal, 2024).

CONCLUSION

The transition from classical, inflexible ETL processes as part of broader data engineering practices to AI-enabled data pipelining ecosystems involves a massive change beyond technology. Modern data infrastructures have become smart, self-managing to create possibilities for autonomous optimization, predictive maintenance, and self-adjustment to time- and operational context change, and can act on proposed data flows and pipelines. Cloud-native architecture enhances scalability and flexibility using distributed computing frameworks, containerized deployments, and scaling options to achieve performance and cost efficiencies. The inclusion of artificial intelligence across the data engineering lifecycle has enabled levels of capability that were not possible just a few years ago, from automated pipeline optimization, intelligent anomaly detection, and even predictive failure prevention or optimization are some examples of increasing reliability and reducing operational overhead. Today's advanced orchestration tools provide more seamless synchronization of workflows that span complex connections to multiple systems and environments of cloud providers in a unified way to manage diverse data ecosystems in one omnichannel control plane. The emergence of data mesh architectures is a testament to the further evolution of data infrastructures, which have become decentralized data ownership systems that balance scale with governance requirements. The inclusion of large language models represents even further democratized pipeline developments that will allow organizations' personnel to contribute to pipeline development while still following enterprise-level quality processes that keep data secure. Organizations that adopt these innovative data engineering principles are setting themselves up to be innovation leaders and to stay ahead of the competition in a world that is becoming more and more data-driven. The future will more than likely continue to show a convergence of artificial intelligence, cloud computing, and data processing tools leading to yet an even smarter and more autonomous data infrastructure that will, with less and less human interaction, be capable of satisfying complex analytical needs while ensuring that organizations receive the most value through better decision-making and operational efficiency.

REFERENCES

1. Columbus, L. "10 Charts That Will Change Your Perspective Of Big Data's Growth," *Forbes*, (2018) https://www.forbes.com/sites/louiscolumbus/2018/05/23/10-charts-that-will-change-your-perspective-of-big-datas-growth/
2. Analytics, M. "The age of analytics: competing in a data-driven world." *McKinsey Global Institute Research* (2016).
3. GlobeNewswire, "Real-Time Analytics Market Size to Surpass USD 193.71 Billion by 2032 Owing to Rising Adoption of AI and IoT Technologies," (2025). https://www.globenewswire.com/news-release/2025/03/24/3047968/0/en/Real-Time-Analytics-Market-Size-to-Surpass-USD-193-71-Billion-by-2032-Owing-to-Rising-Adoption-of-AI-and-IoT-Technologies-Research-by-SNS-Insider.html
4. Medium, "Competing in the Data Age: How Data-Driven Decision-Making Transforms Companies," (2023): https://medium.com/@empa.consulting/competing-in-the-data-age-how-data-driven-decision-making-transforms-companies-2ab2176bdafc
5. Slingerland, C. "90+ Cloud Computing Statistics: A 2025 Market Snapshot," *CloudZero*, (2025). https://www.cloudzero.com/blog/cloud-computing-statistics/
6. Businesswire, "Global Workflow Orchestration Market Report 2022: Featuring Key Players Microsoft, IBM, Cisco, Oracle & More - ResearchAndMarkets.com," (2022). https://www.businesswire.com/news/home/20220908005894/en/Global-Workflow-Orchestration-Market-Report-2022-Featuring-Key-Players-Microsoft-IBM-Cisco-Oracle-More---ResearchAndMarkets.com

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher

7. MarketsandMarkets, "Data Integration Market worth $19.6 billion by 2026," Markets and Markets, 2021. https://www.marketsandmarkets.com/PressReleases/data-integration.asp
8. SarahLee, "8 Surprising Analytics Stats for Boosting Data-Driven Decision Making," *NumberAnalytics*, (2025). https://www.numberanalytics.com/blog/8-surprising-analytics-stats-boosting-data-driven-decision-making
9. Almotairi, K. H., Hussein, A. M., Abualigah, L., Abujayyab, S. K., Mahmoud, E. H., Ghanem, B. O., & Gandomi, A. H. "Impact of artificial intelligence on COVID-19 pandemic: a survey of image processing, tracking of disease, prediction of outcomes, and computational medicine." *Big Data and Cognitive Computing* 7.1 (2023): 11.
10. Choudhary, S. "AI And The Future Of Work," *Forbes*, (2024). https://www.forbes.com/councils/forbestechcouncil/2024/07/12/ai-and-the-future-of-work/
11. Precedence Research, "Intelligent Process Automation Market Size, Share, and Trends 2024 to 2034," (2024). https://www.precedenceresearch.com/intelligent-process-automation-market
12. The AI Journal, "Asia Pacific Generative AI Market Size, Share & Trends Analysis Report 2024-2030 Featuring Huawei, Alibaba, Baidu, Google, IBM, Microsoft, Genie AI, MOSTLY AI, and D-ID – ResearchAndMarkets.com," (2024). https://aijourn.com/asia-pacific-generative-ai-market-size-share-trends-analysis-report-2024-2030-featuring-huawei-alibaba-baidu-google-ibm-microsoft-genie-ai-mostly-ai-and-d-id-researchandmarkets-com/

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:

Kadakia, J. J. " Demystifying Modern Data Engineering: From ETL to AI-Driven Pipelines." *Sarcouncil Journal of Engineering and Computer Sciences* 4.7 (2025): pp 948-956.

Copyright © 2022 The Author(s): This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 (CC BY-NC-ND 4.0) International License

Publisher: SARC Publisher