International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT241061218

Architecting Enterprise-Scale Data Products: A Framework for Advanced Data Science and AI/ML Operations

Seshendranath Balla Venkata

Comcast, USA

ARTICLEINFO

Article History:

Accepted : 23 Nov 2024

Published: 15 Dec 2024

Publication Issue

Volume 10, Issue 6

November-December-2024

Page Number

1724-1734

ABSTRACT

This article presents a comprehensive framework for building enterprise-scale data products that power modern Customer & Product Analytics, Data Science, artificial intelligence, and machine learning initiatives. The article examines the foundational architecture patterns, pipeline engineering strategies, and advanced distributed computing approaches in both on-prem and cloud. These are essential for developing robust data infrastructure capable of handling complex Data Analytics, Data Science, and AI/ML workflows. The article explores critical aspects of feature engineering at scale, real-time processing capabilities, and the implementation of feature stores, while addressing the challenges of data quality, governance, legal, and security in regulated environments. The article introduces a systematic approach to integrating data products with MLOps pipelines, emphasizing the importance of automated workflows, monitoring systems, and feedback loops in production environments. The findings demonstrate that successful implementation of scalable data products requires a careful balance of

architectural decisions, technology selection, and operational practices. The article contributes to the field by providing actionable insights and architectural patterns that organizations can adopt to build resilient, scalable, and efficient data products for their Data Analytics, Data Science, and AI/ML use cases. This article establishes a foundational framework that bridges the gap between theoretical data architecture principles and practical implementation challenges in enterprise settings.

Keywords: Data Product Engineering, Data Analytics, Enterprise ML Infrastructure, Scalable Distributed Data Architecture, Data Governance Framework.

Introduction

A. Evolution of Data Products in Data Analytics, Data Science, and AI/ML Landscape

The landscape of data products has undergone a remarkable transformation in recent years, driven by the exponential growth in artificial intelligence and machine learning applications. This evolution has shifted from traditional data warehousing approaches to sophisticated, real-time data platforms that can handle petabyte-scale operations [1]. Businesses now handle a variety of Data Analytics, Data Science, and AI/ML workloads using intricate, networked data ecosystems rather than just basic ETL processes. To guarantee the reproducibility and traceability of Data Analytics, Data Science, and AI/ML experiments, strong citation and versioning procedures are essential due to the dynamic nature of contemporary data platforms.

B. Critical Role of Data Infrastructure Scalability

Data infrastructure scalability has emerged as a critical differentiator in the success of Data Analytics, Data Science, and AI/ML initiatives. Modern data platforms must simultaneously support diverse workloads, from batch processing to real-time streaming, while maintaining performance and reliability. These platforms need to handle not just the volume of data but also the velocity and variety of incoming data streams, making scalability a multidimensional challenge [2]. The ability to scale horizontally while maintaining data consistency and accessibility has become paramount for organizations deploying Data Analytics, Data Science, and AI/ML solutions in production environments. The increasing complexity of data pipelines and the need for real-time processing capabilities have pushed organizations to adopt more sophisticated infrastructure patterns.

C. Current Challenges and Pain Points

The emergence of large-scale data products has introduced significant challenges in the Data Analytics, Data Science, and AI/ML landscape. Organizations struggle with maintaining data pipeline complexity while ensuring consistent quality across environments. The need for real-time processing capabilities has created additional pressure on infrastructure design and resource allocation. Technical debt management in data infrastructure remains a persistent challenge, particularly as organizations scale their Data Analytics, Data Science, and AI/ML operations. These challenges are exacerbated by the rapid evolution of AI technologies and the increasing regulatory requirements around data governance and privacy [2]. The integration of legacy systems with modern data platforms adds

another layer of complexity to the already challenging landscape.

D. Value Proposition of Modern Data Products

Modern data products deliver substantial value through their ability to accelerate Data Analytics, Data Science, and AI/ML development cycles and reduce operational overhead. By implementing robust data citation mechanisms, organizations can ensure the reproducibility and traceability of their Data Analytics, Data Science, and AI/ML experiments [1]. These products enable faster time-to-market for ML-powered solutions while maintaining high data quality standards. The integration of automated monitoring systems ensures reliability in production environments, while streamlined feature engineering capabilities accelerate model development cycles. The value proposition extends beyond technical benefits to include improved governance, better resource utilization, and enhanced collaboration across data analytics and data science teams.

E. Scope and Organization of the Article

This article presents a comprehensive framework for building enterprise-grade data products that power Data Analytics, Data Science, and AI/ML initiatives. Beginning with foundational architecture patterns, we explore the essential components of modern data infrastructure. The discussion progresses through pipeline engineering strategies, feature management approaches, and governance frameworks. We examine practical implementation strategies with particular emphasis on scalability and operational efficiency. The integration patterns with ML workflows are discussed in detail, followed by insights into future trends in data product development. Throughout the article, we reference proven methodologies and emerging patterns that organizations can adopt to build robust, scalable data products for their Data Analytics, Data Science, and AI/ML use cases.

Foundational Architecture for Data Products

A. Core Infrastructure Components

The foundation of modern data products rests on a carefully orchestrated set of infrastructure components designed to handle the complexities of Data Analytics, Data Science, and AI/ML workloads [DivergeIT, 3]. The storage layer design implements a multi-tiered approach, combining high-performance block storage for active workloads with object storage for cost-effective data retention. This hybrid storage strategy enables organizations to optimize both performance and cost while maintaining data accessibility. Compute resource management encompasses both physical and virtual resources, with automation playing a crucial role in resource allocation and optimization. The network architecture incorporates software-defined networking principles, ensuring high-bandwidth, low-latency communication essential for distributed Data Analytics, Data Science, and AI/ML workloads. Service mesh implementations have become integral to modern data architectures, providing sophisticated service discovery, load balancing, and traffic management capabilities that enhance the reliability and observability of distributed systems.

Component Type	Key Elements	Primary Function	Implementation Considerations
Storage Layer	• Block Storage • Object Storage • File Systems	Data Persistence	• Performance Requirements • Cost Optimization • Scalability Needs
Compute Resources	• Virtual Machines	Processing Capacity	• Auto-scaling

	• Containers • Serverless		• Resource Utilization • Cost Management
Network Infrastructure	• Load Balancers • Service Mesh • API Gateways	Connectivity	• Latency Requirements • Security Needs • Traffic Management
Monitoring Systems	• Metrics Collection • Logging • Alerting	Observability	• Real-time Monitoring • Historical Analysis • Issue Detection

Table 1: Core Components of Data Product Architecture [3, 4]

B. Architectural Patterns

Modern data products leverage cloud-native design principles to achieve unprecedented levels of scalability and resilience. These principles emphasize containerization, orchestration, and automated scaling capabilities. Hybrid and multi-cloud strategies have become increasingly prevalent, allowing organizations to distribute workloads across different cloud providers while maintaining operational consistency. The microservices architecture breaks down complex data processing workflows into manageable, independently deployable services that can be scaled and maintained separately. Event-driven systems enable real-time data processing and analysis, which is crucial for Data Analytics, Data Science, and AI/ML applications that require immediate insights. This architectural approach ensures loose coupling between components while maintaining high cohesion within individual services, supporting the dynamic nature of modern data processing requirements.

C. Infrastructure as Code (IaC)

Infrastructure as Code has revolutionized the way organizations deploy and manage their data infrastructure [AWS, 4]. Configuration management systems maintain infrastructure state across environments, ensuring consistency and reproducibility of deployments. Resource provisioning follows automated workflows that create and manage infrastructure components based on predefined templates, reducing human error and increasing deployment efficiency. Environment consistency is maintained through rigorous version control of infrastructure definitions, enabling teams to track changes and roll back when necessary. Version control strategies for infrastructure code mirror software development best practices, including branch management, code review processes, and automated testing of infrastructure changes. This approach has transformed infrastructure management from a manual, error-prone process to a streamlined, automated workflow that supports rapid iteration and scaling of data products.

The implementation of these foundational elements requires careful consideration of organizational requirements, technical constraints, and future scalability needs. Successful data products build upon these foundations to create robust, maintainable, and scalable systems capable of supporting advanced Data Analytics, Data Science, and AI/ML workloads. The integration of modern infrastructure practices with traditional IT governance frameworks enables organizations to maintain control while achieving the agility needed for competitive advantage in the Data Analytics, Data Science, and AI/ML space.

Data Pipeline Engineering and Optimization

A. Pipeline Design Patterns

Modern data pipeline architectures have evolved to meet the demanding requirements of Data Analytics, Data Science, and AI/ML workloads through modular and resilient design patterns [CMU, 6]. The modular pipeline architecture enables organizations to break down complex data processing workflows into maintainable, reusable components that can be independently scaled and modified. This approach facilitates easier testing, debugging, and maintenance of pipeline components. Error handling and recovery mechanisms are implemented through systematic pattern recognition, enabling automated identification and resolution of common failure modes. Pipeline monitoring systems leverage advanced visualization techniques for real-time system health monitoring, incorporating interactive dashboards that provide deep insights into pipeline performance [IEEE, 5]. Advanced scheduling and orchestration capabilities manage complex dependencies between pipeline components while optimizing resource utilization and ensuring timely data delivery for downstream ML processes.

B. Processing Paradigms

The evolution of data processing paradigms has led to sophisticated implementations that handle both real-time and batch-processing requirements efficiently. Stream processing implementations leverage visualization-driven monitoring systems to ensure optimal performance and early detection of anomalies [IEEE, 5]. Batch processing systems utilize pattern recognition algorithms to identify optimal processing strategies and resource allocation patterns. Lambda architecture combines these insights to provide comprehensive views of data while maintaining low latency for real-time requirements. Kappa architecture simplifies the processing paradigm by treating all data as streams, with visual analytics tools supporting operational monitoring and optimization. Hybrid processing solutions adapt to varying workload characteristics, using machine learning-based pattern detection to optimize processing strategies dynamically [CMU, 6].

C. Performance Optimization

Performance optimization in data pipelines requires a multi-faceted approach focusing on resource utilization, latency management, and cost efficiency. Resource utilization strategies leverage visual analytics tools to identify bottlenecks and optimization opportunities in real time [IEEE, 5]. Latency management incorporates pattern-based optimization techniques, using historical performance data to predict and prevent processing delays. Throughput optimization techniques employ machine learning models to identify and implement optimal processing patterns across different workload types. Cost efficiency measures include pattern-based resource allocation strategies and visualization-driven decision support systems that help balance performance requirements with infrastructure costs.

These advanced pipeline engineering practices enable organizations to build robust, scalable data processing systems that support sophisticated Data Analytics, Data Science, and AI/ML workflows while maintaining operational efficiency and cost-effectiveness. The integration of visual analytics and pattern recognition technologies has transformed how organizations design, monitor, and optimize their data processing pipelines.

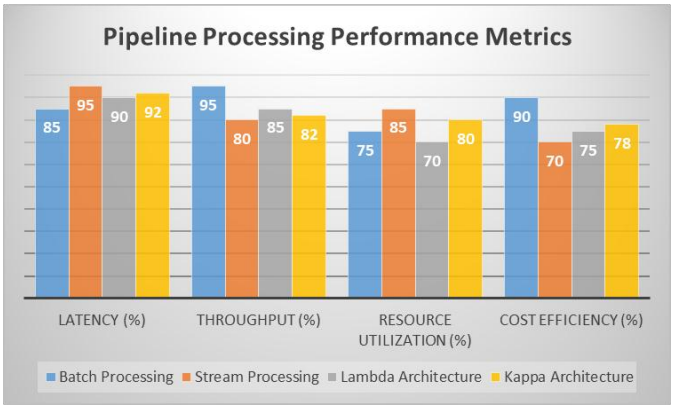

Fig. 1: Pipeline Processing Performance Metrics [5, 6]

Feature Engineering and Data Processing at Scale

A. Automated Feature Engineering

The evolution of automated feature engineering has revolutionized the way organizations approach data preparation for ML models [Databricks, 7]. Feature discovery methods employ sophisticated algorithms to identify relevant data transformations and interactions, particularly on temporal and spatial relationships in complex datasets. Automated selection techniques leverage statistical analysis to evaluate feature importance across different data distributions and scales. Feature validation systems implement comprehensive quality checks, ensuring that generated features maintain consistency across both training and serving environments. Version control for features has become increasingly critical, especially when dealing with time-series data and evolving feature distributions, requiring sophisticated tracking mechanisms for feature lineage and dependencies [Datenbank-Spektrum, 8].

Stage	Activities	Tools/Methods	Quality Checks
Discovery	• Data Analysis • Pattern Recognition	• Statistical Analysis • ML Algorithms	• Statistical Significance • Business Relevance
Engineering	• Feature Creation • Transformation	• Automated Tools • Custom Scripts	• Data Quality • Performance Impact
Validation	• Testing • Performance Evaluation	• Validation Framework • Monitoring Tools	• Accuracy Metrics • Resource Usage
Deployment	• Production Release • Monitoring	• Feature Store • Monitoring Systems	• Production Performance • Business Impact

Table 2: Feature Engineering Lifecycle Components [7, 8]

B. Distributed Processing

Distributed processing architectures form the backbone of scalable feature engineering systems, enabling organizations to handle massive datasets efficiently [Databricks, 7]. Parallel processing frameworks distribute computation across clusters of machines, implementing sophisticated workload management and fault tolerance mechanisms. Data partitioning strategies optimize data distribution based on spatio-temporal characteristics, ensuring efficient time series and location-based features processing. Resource allocation methods employ adaptive algorithms that dynamically adjust computing resources based on feature computation complexity and data volume. Load balancing techniques ensure even distribution of processing tasks while maintaining system stability under varying workload conditions, which is particularly important for real-time feature generation scenarios.

C. Feature Store Implementation

Feature stores have emerged as a critical component in modern ML infrastructure, serving as a centralized repository for managing and serving features at scale [Datenbank-Spektrum, 8]. Online and offline storage mechanisms are designed to handle different types of features, with special consideration for spatio-temporal data structures and their unique access patterns. The feature-serving architecture implements sophisticated caching strategies to optimize feature retrieval, particularly for frequently accessed

temporal and spatial features. Caching strategies are tailored to different feature types, with specific optimizations for time-series data and spatial relationships. Update mechanisms ensure consistency between online and offline feature values while maintaining system performance, with particular attention to time-sensitive feature updates and their propagation across the system.

These advanced feature engineering practices enable organizations to build robust, scalable systems that can handle complex data types while maintaining performance and reliability. The integration of temporal and spatial aspects in feature engineering has become increasingly important for modern ML applications, requiring sophisticated processing and storage solutions.

Data Quality, Governance, and Security Framework

A. Quality Management

The foundation of reliable Data Analytics, Data Science, and AI/ML systems rests on robust data quality management frameworks that ensure data accuracy, completeness, and consistency [IBM, 9]. Data validation rules establish standardized controls across the organization, implementing comprehensive checks that span format validation, semantic consistency, and business rule compliance. Quality metrics and monitoring systems leverage automated dashboards and alerting mechanisms to provide real-time visibility into data health. Organizations employ sophisticated anomaly detection algorithms to identify data quality issues proactively, preventing downstream impacts on ML models. Data lineage tracking has become increasingly critical, maintaining detailed records of data transformations and dependencies that support both compliance requirements and quality management initiatives.

B. Governance Implementation

Modern data governance frameworks have evolved to address the complex requirements of Data Analytics, Data Science, and AI/ML systems while ensuring regulatory compliance and ethical use of data [Axamit, 10]. Policy management systems implement standardized governance models that adapt to changing regulatory requirements while maintaining organizational efficiency. Compliance controls are embedded throughout the data lifecycle, with specific attention to regulatory requirements such as GDPR, CCPA, and industry-specific standards. Documentation standards maintain comprehensive metadata about data assets, supporting both operational efficiency and compliance requirements. Change management processes ensure the controlled evolution of data systems through structured workflows that balance innovation with risk management [IBM, 9].

C. Security Architecture

The security architecture for data products implements multiple layers of protection while enabling necessary access for Data Analytics, Data Science, and AI/ML workflows [Axamit, 10]. Authentication systems leverage industry-standard identity management practices, incorporating multi-factor authentication and single sign-on capabilities that align with enterprise security frameworks. Authorization controls implement role-based access control (RBAC) and attribute-based access control (ABAC) mechanisms, ensuring precise control over data access and manipulation. Encryption standards maintain data confidentiality through standardized protocols and key management practices, with specific considerations for protecting sensitive ML model features and training data. Audit logging systems maintain comprehensive records of all data access and modifications, supporting both security monitoring and compliance reporting requirements.

These frameworks collectively ensure that organizations can maintain high data quality standards while meeting regulatory requirements and protecting sensitive information. The integration of quality, governance, and security considerations throughout the data lifecycle enables organizations to

build trustworthy Data Analytics, Data Science, and AI/ML systems that can operate effectively in regulated environments while maintaining operational efficiency.

Integration with ML Workflows

A. MLOps Integration

Modern ML systems require sophisticated integration patterns that seamlessly connect data products with operational workflows [Google Cloud, 11]. CI/CD pipeline integration has evolved to encompass the entire ML lifecycle, implementing automated testing and deployment processes that handle both code and model artifacts. Model registry systems maintain versioned repositories of models, their configurations, and associated metadata, enabling comprehensive governance and traceability. Experiment tracking capabilities have become increasingly sophisticated, capturing detailed information about model training runs, including feature selection decisions, hyperparameter configurations, and performance metrics. Deployment automation leverages containerization and orchestration platforms to enable reproducible model deployments across different environments.

B. Training Data Management

Effective management of training data has become a critical success factor in ML operations [Google Cloud, 11]. Dataset version control systems implement immutable snapshots of training data, maintaining clear lineage between models and their training datasets. Training data preparation pipelines automate the process of data cleaning, transformation, and feature engineering, ensuring the reproducibility of model training processes. Data augmentation techniques are integrated into automated pipelines, systematically expanding training datasets to improve model robustness. Cross-validation set management employs automated splitting strategies that maintain statistical consistency while preventing data leakage between training and evaluation sets.

C. Production Systems

The deployment of ML models in production requires a robust infrastructure that ensures reliable model serving and monitoring capabilities [Google Cloud, 11]. Model serving infrastructure leverages container orchestration platforms to enable scalable, reliable model inference services. Performance monitoring systems implement comprehensive observability frameworks that track technical metrics, business KPIs, and data quality indicators in real time. A/B testing frameworks are integrated into the deployment pipeline, enabling controlled experiments with automated metric collection and analysis. Feedback loops are systematically implemented to capture production data and performance metrics, facilitating continuous improvement of both models and their underlying data products.

These integration patterns ensure that data products effectively support the entire ML lifecycle, from experimentation to production deployment, while maintaining quality and reliability throughout the process. The emphasis on automation and reproducibility enables organizations to scale their ML operations efficiently while maintaining governance and control.

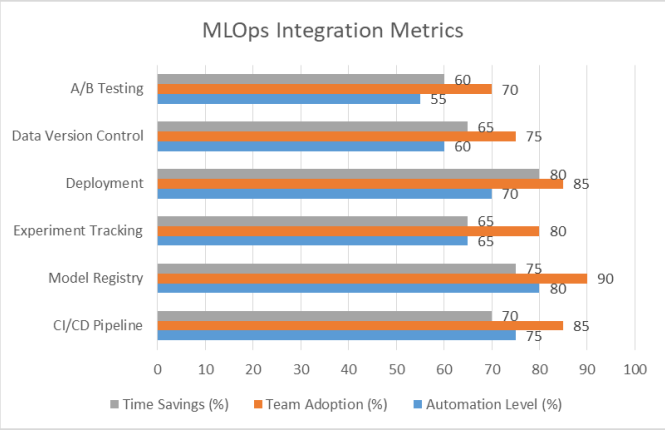

Fig. 2: MLOps Integration Metrics [11]

Data Product Tier Implementation and Value Realization

A. Data Product Tiers and Classification

The implementation of data products follows a structured tier-based approach that ensures progressive data refinement and value creation throughout the organization[12]. The Bronze tier serves as the foundational layer, focusing on raw data ingestion through standardized collection mechanisms. This tier maintains original data integrity through immutable storage while establishing initial data lineage tracking, creating essential building blocks for higher-level data products. The system ensures reliable data capture while preserving the source information in its original form.

The Silver tier transforms raw data into standardized formats through comprehensive data quality rules and validations. This intermediate layer establishes common data models across various sources, enabling cross-functional data understanding throughout the organization. Through cleaned and normalized data structures, the Silver tier bridges the gap between raw data and business-ready information, ensuring consistency and reliability in data processing.

At the highest level, the Gold tier delivers business-ready, purpose-specific datasets that support direct integration with analytical tools and enable automated decision-making processes. This tier provides optimized data for ML model training while maintaining high-level data quality guarantees. The Gold tier represents the pinnacle of data refinement, where information is fully prepared for business consumption and advanced analytics applications.

B. Data Product Features and Governance

A robust data product framework requires comprehensive product definition elements, including clear use case documentation, specifications, and defined business and technical requirements[12]. Organizations must establish ownership and responsibility matrices while implementing comprehensive metadata tagging systems that align with business context and value propositions. This foundation ensures clear accountability and understanding of data product purposes and capabilities.

The discoverability framework encompasses searchable data catalogs enhanced with detailed metadata, supporting cross-platform data lineage tracking and granular access control mechanisms. Published data dictionaries and schemas, combined with self-service discovery portals, enable users throughout the organization to efficiently locate and utilize relevant data products.

Quality management systems form the backbone of reliable data products, implementing automated quality validation pipelines and business rule compliance checking. Technical quality metrics monitoring, combined with regular quality reporting mechanisms, ensures continuous improvement in data product delivery and reliability.

C. Impact on Data Analytics, Data Science, and AI/ML Operations

Data products significantly enhance Data Analytics, Data Science, and AI/ML capabilities by providing validated, reliable data inputs for analysis and hypothesis testing. The systematic approach to data management ensures consistent analytical foundations and enables real-time decision support capabilities. This structured environment supports more accurate and reliable machine learning implementations.

Model performance enhancement is achieved through comprehensive training datasets that enable sophisticated pattern recognition and improved prediction accuracy. The quality and diversity of data products contribute to better model generalization capabilities while reducing potential biases through carefully curated and validated data sources.

The continuous optimization framework establishes automated feedback loops and robust performance monitoring systems. These mechanisms enable iterative refinement processes and support version

control for both models and data, facilitating continuous learning and improvement in AI/ML systems.

D. Service Level Agreements and Documentation

Effective data product operations rely on clearly defined service level objectives and comprehensive performance and availability metrics[12]. Quality standards and guarantees, combined with support and maintenance procedures, ensure reliable and consistent data product delivery. The establishment of clear escalation pathways maintains operational efficiency and user satisfaction.

Documentation requirements encompass complete metadata records, detailed data lineage tracking, and comprehensive usage guidelines. Quality metrics documentation and compliance requirements ensure that data products meet both technical and regulatory standards while maintaining usability and reliability.

This comprehensive framework ensures that organizations can effectively implement and manage data products while maximizing their value for Data Analytics, Data Science, and AI/ML initiatives. The structured approach to data product management, from raw data ingestion to business-ready insights, provides a scalable foundation for advanced analytics and machine learning applications. Through careful attention to quality, governance, and usability standards, organizations can build and maintain sophisticated data products that drive business value and enable advanced analytical capabilities.

Conclusion

The development of world-class data products for Data Analytics, Data Science, and AI/ML applications represents a critical capability for modern organizations, requiring careful consideration of architecture, infrastructure, and operational practices. Through our comprehensive examination of foundational architectures, pipeline engineering, feature management, governance frameworks, and MLOps integration patterns, we have established a robust framework for building scalable, reliable data products. The integration of cloud-native technologies, automated pipeline engineering, and sophisticated feature management capabilities enables organizations to handle increasingly complex Data Analytics, Data Science, and Data Analytics, Data Science, and AI/ML workloads while maintaining operational efficiency. The tiered approach to data products - progressing from Bronze through Silver to Gold - provides a structured framework for data refinement and value creation, ensuring organizations can effectively support both operational needs and advanced analytics requirements. The emphasis on data quality, governance, and security ensures that these systems can operate effectively in regulated environments while meeting stringent compliance requirements. As organizations continue to scale their Data Analytics, Data Science, and AI/ML initiatives, the adoption of these practices and patterns, particularly the structured approach to data product development and classification, will become increasingly critical for success. Future developments in this space will likely focus on enhanced automation, improved observability, and more sophisticated integration patterns between data products and ML systems. Organizations that successfully implement these principles while maintaining flexibility to adapt to emerging technologies will be well-positioned to leverage Data Analytics, Data Science, and AI/ML capabilities for competitive advantage, building and maintaining sophisticated systems that deliver sustainable business value while ensuring operational excellence and regulatory compliance.

References

[1]. S. Pröll and A. Rauber, "Scalable data citation in dynamic, large databases: Model and reference implementation," 2013 IEEE International Conference on Big Data, 2013, pp. 1-8.

Available: https://ieeexplore.ieee.org/document/6691588

[2]. A. Badshah, A. Daud, R. Alharbey, A. Banjar, and B. Alshemaimri, "Big data applications: overview, challenges and future," Artificial Intelligence Review, vol. 57, no. 290, 2024. Available: https://link.springer.com/article/10.1007/s10462-024-10938-5

[3]. DivergeIT, "7 Components of IT Infrastructure: Definitions & Functions," DivergeIT, 2023. Available: https://www.divergeit.com/blog/components-of-it-infrastructure

[4]. AWS, "What is Infrastructure as Code?," Amazon Web Services, 2023. Available: https://aws.amazon.com/what-is/iac/

[5]. T. von Landesberger, D. W. Fellner, and R. A. Ruddle, "Visualization system requirements for data processing pipeline design and optimization," IEEE Transactions on Visualization and Computer Graphics, vol. 23, no. 8, pp. 2028-2041, 2017. Available: https://doi.org/10.1109/TVCG.2016.2603178

[6]. CMU Software Engineering Institute, "Using Machine Learning to Detect Design Patterns," Carnegie Mellon University, Software Engineering Institute's Insights (blog), 2020. Available: https://insights.sei.cmu.edu/blog/using-machine-learning-to-detect-design-patterns/

[7]. Databricks, "Feature Engineering at Scale," Databricks Blog, 2021. Available: https://www.databricks.com/blog/2021/07/16/feature-engineering-at-scale.html

[8]. C.-M. Forke and M. Tropmann-Frick, "Feature Engineering Techniques and Spatio-Temporal Data Processing," Datenbank-Spektrum, vol. 21, pp. 237-244, 2021. Available: https://link.springer.com/article/10.1007/s13222-021-00391-x

[9]. IBM, "What is Data Governance?," IBM, 2023. Available: https://www.ibm.com/topics/data-governance

[10]. Axamit, "Data Governance Framework: Models, Examples, and Key Requirements," Axamit Blog, 2023. Available: https://axamit.com/blog/data-governance/data-governance-framework/

[11]. Google Cloud, "MLOps: Continuous delivery and automation pipelines in machine learning," Google Cloud Architecture Center, 2023. Available: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning

[12]. Dehghani, Z., "Data Mesh," O'Reilly Media, 2022. Available: https://www.oreilly.com/library/view/data-mesh/9781492092384/