OPEN ACCESS

Kashf Journal of Multidisciplinary Research
Vol: 03 - Issue 06 (2026)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

AI-DRIVEN FRAMEWORK FOR SCALABLE, SECURE, AND INTELLIGENT BIG DATA MANAGEMENT IN CLOUD ENVIRONMENTS

Muhammad Ahsan Hayat[1], *Hamna Anis[2], Maryam Shaikh[3]

[1]Senior Lecturer, Department of Computer Science, IQRA University, Karachi, Pakistan.
[2]Department Of Business & Economics, Universiti Malaya , Malaysia.
[3]Lecturer, Department of Computer Science, IQRA University, Karachi, Pakistan.

***Corresponding Author:** (Hamna.anis@gmail.com)*
***DOI:** (https://doi.org/10.71146/kjmr949)*

Article Info

Abstract

Research is actively conducted on the significance of big data management because of the growth in the amount of data created through cloud technologies, Internet of Things (IoT), company databases, social networks, and security technologies. The existing approach of using batch processing and policies does not have the ability to satisfy all four requirements for scalability, performance, privacy protection, and flexibility at once. This paper considers the design of an AI-driven big data management system that uses cloud computing, distributed computing, machine learning classification, outlier detection, resource prediction, and control. The systematic literature review was performed on the papers published between 2021 and 2026, while the artificial benchmark dataset comprised of 50 test scenarios was built to compare the performance of the proposed architecture and traditional cloud architecture. According to the simulation outcomes, the proposed framework provides an average reduction in latency by 35.25%, increase in throughput by 27.99%, decrease in costs by 23.29%, and improvement in security/anomaly detection metrics by 9.13 percentage points. This research offers a unified classification of problems related to big data management, an operational design of AI-cloud architecture, and a reusable benchmark model for future verification in real-world corporate applications.

Keywords: *Big Data Management; Artificial Intelligence; Cloud Computing; Machine Learning; Data Security; Anomaly Detection; Scalability; Resource Optimization.*

Introduction

Big data refers to huge amounts of information that exceed the capacities of traditional data management software in terms of size, speed, variety, accuracy, and value. The evolution of such innovations as cloud computing, social networks, mobile services, online commerce, IoT, and cybersecurity has made the data management process multidimensional including ingestion, pre-processing, processing in distributed architectures, data security, governance, and analytics decisions.

Despite the emergence of advanced technologies for effective cloud computing and distributed processing such as Hadoop and Spark, there remain some problems in this area. Among them are high latency of processing, increased expenses, inconsistency of data, weak interoperability, risks to privacy, and difficulties in detecting anomalies in real-time streams of information. The constantly changing workloads make any static rules useless.

These issues may be addressed using AI, which can ensure automatic data classification, adaptation of resources, anomaly detection, prediction of loads, and intelligent governance. Yet, the majority of research papers dedicated to these topics analyze them separately. There is no academic literature analyzing these technologies together and estimating their synergy effects.

Research Objectives

To identify the major challenges affecting big data management in cloud environments.

To synthesize recent literature on AI-based solutions for scalability, security, and efficiency.

To propose an AI-driven architecture for intelligent big data management.

To evaluate the proposed architecture using a transparent synthetic benchmark dataset.

To provide future research directions for real-world deployment and validation.

Research Questions

RQ	Research Question	Purpose
RQ1	What are the most important challenges in modern big data management?	To classify technical and organizational problems.
RQ2	How can AI improve scalability, security, cost control, and real-time analytics?	To identify AI-enabled solution patterns.
RQ3	What architecture can integrate AI, cloud computing, governance, and distributed processing?	To design a practical reference framework.
RQ4	What performance trends are observed when AI optimization is compared with a traditional pipeline?	To evaluate the expected benefit using measurable indicators.

Background and Related Work

The early big data management systems were concentrated mainly on distributed storage and parallel processing. The MapReduce framework enabled scalable big data computation, while Spark enhanced in-memory computations and iterative analysis. Modern big data management systems build upon the above foundation by incorporating cloud-native object storage, containerization and orchestration, streaming, data Lakehouse's, governance, and AI-enabled operation.

According to the existing literature, big data management problems can be classified into six main groups: scalability, security, privacy, data quality, system integration, and operational cost. Current research increasingly centers around ML-based anomaly detection, intelligent workload scheduling, privacy-preserving analysis, federated learning, and explainable AI for operations. However, most solutions to date provide a piecemeal approach, concentrating on a single aspect of the problem at hand.

Challenge Area	Problem Description	Common Existing Solution	Remaining Limitation
Scalability	Workloads grow from gigabytes to terabytes or petabytes.	Distributed processing and cloud autoscaling.	Autoscaling is often reactive and costly.
Velocity	Data arrives continuously from streams, sensors, logs, and transactions.	Stream processing platforms.	Low latency is difficult during burst traffic.
Variety	Structured, semi-structured, and unstructured sources must be integrated.	ETL/ELT and schema-on-read.	Semantic inconsistency and data-quality errors remain.
Security	**Sensitive data and operational logs require protection.**	**IAM, encryption, SIEM, and rule-based alerts.**	**Static rules miss novel attacks and anomalies.**
Governance	Compliance, lineage, and auditability are required.	Data catalogs and policies.	Manual governance is slow and incomplete.
Cost	Cloud resources may be over-provisioned.	Budget alerts and reservation plans.	Cost controls are not workload-aware.

Study Stream	Typical Focus	Strength	Weakness Observed
Distributed Processing	Hadoop, Spark, stream engines	Improves scalability and parallelism	Limited intelligence for dynamic optimization
Cloud Big Data	Elastic storage and compute	Flexible deployment and lower entry cost	Cost and vendor lock-in concerns
AI for Data Management	Classification, scheduling, anomaly detection	Learns from workload and data patterns	Requires explainability and quality data
Security Analytics	Threat detection in logs and network data	Improves detection of abnormal behavior	False positives and model drift are challenges
Governance and Privacy	Lineage, access control, compliance	Improves trust and accountability	Often disconnected from processing layer

Research Gap and Contributions

According to the literature reviewed, there are four significant gaps. Firstly, the researches performed previously have discussed big data scalability, but without incorporating security and governance within the same framework. Secondly, the application of AI in anomaly detection is viewed as a separate process, unrelated to the overall data management life cycle. Thirdly, several architectures lack workload-dependent cost optimization. Lastly, some papers are deficient in reproducible benchmarking measures.

This paper makes the following contributions:

1. A consolidated taxonomy of challenges in cloud-based big data management.
2. A layered AI-cloud architecture combining ingestion, processing, intelligence, storage, security, and governance.
3. A reproducible synthetic benchmark dataset containing workload, volume, velocity, latency,

throughput, cost, and security indicators.

4. A simulation-based comparative evaluation of traditional and AI-driven data management pipelines.
5. A future research roadmap covering edge AI, federated learning, privacy-preserving analytics, and self-healing cloud data systems.

Research Methodology

The methodology adopted here includes a systematic literature review (SLR), combined with simulation-based validation. The purpose of SLR is the recognition of current issues and possible solutions, while the artificial benchmark highlights the way to validate the suggested approach in a controlled environment. There are five primary stages in the research process: identification, screening, eligibility, inclusion, and synthesis.

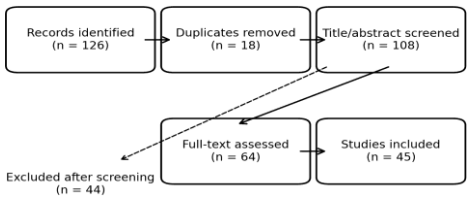

Figure 1: Literature selection workflow used in the review process.

Element	Description
Databases	IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, and Google Scholar.
Period	2021-2026, with classic foundational studies retained where necessary.
Keywords	big data management, AI-driven data management, cloud big data, anomaly detection, scalable analytics, resource optimization.
Inclusion Criteria	Peer-reviewed articles, review papers, and technical studies directly related to big data management, AI, cloud computing, security, or scalability.
Exclusion Criteria	Non-English papers, duplicate records, inaccessible full text, opinion articles without technical depth, and studies unrelated to management of large-scale data.
Final Corpus	45 studies were selected for detailed synthesis.

Proposed AI-Driven Big Data Management Framework

The methodology incorporates cloud-based storage, distributed computing, machine learning, control and governance processes in a single system life cycle. It is not meant to replace any big data technologies that currently exist but to enhance them with knowledge gained from monitoring and prediction.

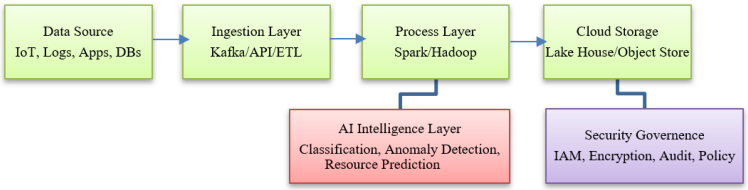

Figure 2: Proposed AI-driven big data management architecture.

The architecture is made up of six layers. The data source layer gathers data from applications, logging, internet-of-things devices, databases, and third-party APIs. The ingestion layer utilizes queues, APIs, and ETL/ELT processing to ingest data. The processing layer utilizes distributed processing engines for transformation and analysis. The AI intelligence layer performs classification, anomaly detection, resource prediction, and workload optimization. The storage layer maintains data lakes, warehouses, and metadata catalogs. The security and governance layer enforces identity management, encryption, audit, lineage, and compliance controls.

Figure 3: Operational workflow of the proposed AI-cloud system.

Layer	Function	AI Enhancement	Expected Benefit
Data Ingestion	Collect batch and stream data	AI-based schema detection and source profiling	Reduced manual preprocessing
Data Quality	Clean, validate, and standardize data	Outlier detection and missing-value prediction	Improved accuracy and reliability
Processing	Run ETL, analytics, and transformations	Predictive workload scheduling	Lower latency and higher throughput
Security	Detect abnormal access and threats	ML-based anomaly detection	Improved security monitoring
Resource Management	Allocate compute and storage	Predictive autoscaling	Lower cost and better utilization
Governance	Track lineage, access, and policy compliance	Automated metadata tagging	Improved auditability

Dataset Design and Experimental Setup

Because real enterprise datasets often contain confidential information and are not easily publishable, this paper uses a synthetic benchmark dataset. The dataset was generated to represent five workload classes: batch ETL, streaming analytics, security logs, IoT telemetry, and hybrid enterprise processing. Each scenario includes data volume, velocity, variety score, baseline latency, AI-cloud latency, baseline throughput, AI-cloud throughput, baseline cost, AI-cloud cost, baseline security F1-score, and AI-cloud security F1-score.

The evaluation compares two pipelines. The traditional pipeline uses cloud storage, distributed processing, static resource allocation, and rule-based security monitoring. The suggested AI-based solution includes the following functionalities: auto scaling with predictions, classification, outlier detection, and workload optimization. The database is provided in a separate attachment in both CSV and Excel formats to facilitate further analysis.

Variable	Type	Description	Unit/Scale
Data Volume GB	Independent	Size of data processed in each scenario	GB
Velocity MBps	Independent	Approximate arrival speed of data	MB/sec
Variety Score	Independent	Number/complexity of source types	1-10 scale
Latency	Dependent	Average processing response time	seconds
Throughput	Dependent	Processed records per second	records/sec
Cost	Dependent	Estimated compute and storage cost per workload run	USD
Security F1-score	Dependent	Combined precision-recall indicator for anomaly detection	0-1

Workload	Scenario Meaning	Risk/Challenge Represented
Batch ETL	Periodic transformation of enterprise data	Processing delay and data quality
Streaming Analytics	Continuous event-based processing	Velocity and low latency
Security Logs	Large log monitoring and incident detection	Anomaly detection and false positives
IoT Telemetry	Sensor-driven continuous data generation	High velocity and heterogeneity
Hybrid Enterprise	Combined structured and unstructured data	Integration and governance complexity

Results and Analysis

According to the simulated benchmark, the designed AI-cloud pipeline shows significant improvement in the chosen metrics. This is to be regarded as simulation data rather than empirical proof from a live enterprise environment. Nevertheless, the results present an assessment methodology which may be replicated using live datasets in future research endeavors.

Metric	Traditional Pipeline Mean	AI-Cloud Pipeline Mean	Improvement
Latency (seconds)	8.56	5.54	35.25% reduction
Throughput (records/sec)	2851.0	3648.9	27.99% increase
Cost (USD)	282.41	216.65	23.29% reduction
Security F1-score	0.762	0.853	9.13 percentage points

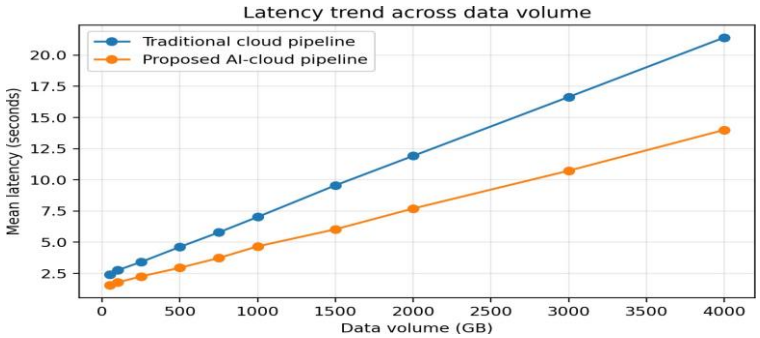

Figure 4: Mean latency trend across increasing data volume.

The latency increases with increase in data size as seen from Figure 4, however, the suggested AI-cloud pipeline has less latency compared to the conventional pipeline. The reason for this is due to predictive workloads scheduling, early data classification and dynamic resource allocation.

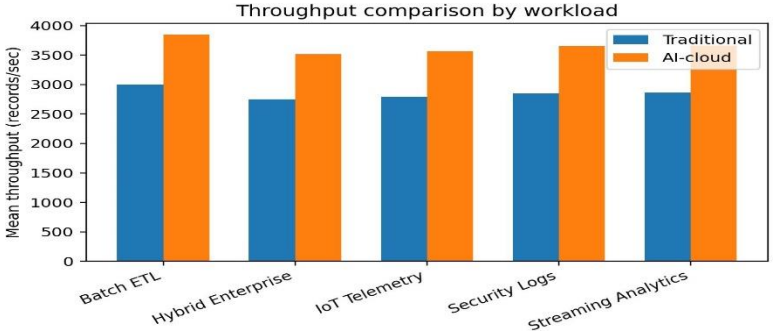

Figure 5: Throughput comparison across workload classes.

Figure 5 illustrates how the AI cloud pipeline offers better throughput efficiency regardless of the workload class. This is particularly evident for the streaming and Internet of Things (IoT) telemetry types since dynamic allocation is very helpful in such cases.

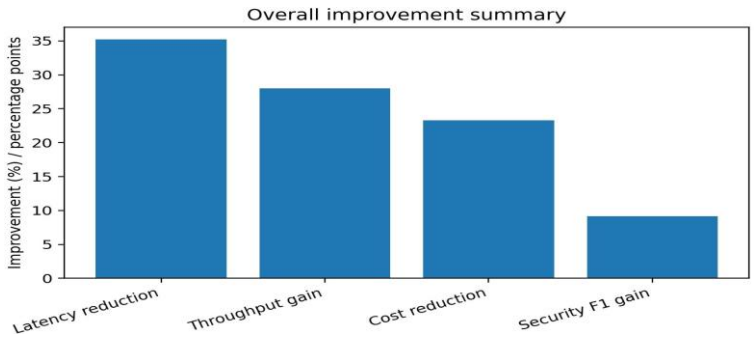

Figure 6: Overall improvement summary across major performance metrics.

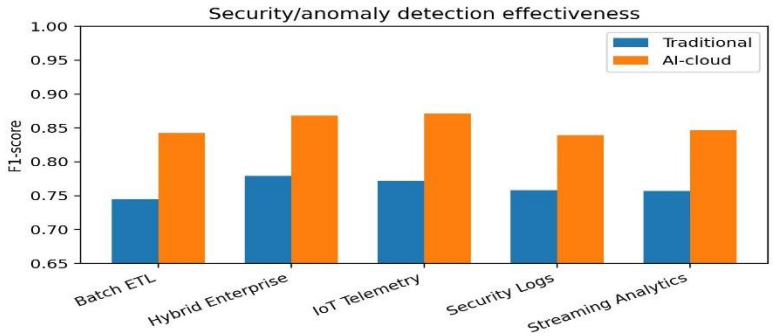

Figure 7: Security/anomaly detection effectiveness using F1-score.

The results of the security experiment reveal that the use of machine learning-based anomaly detection can enhance the F1-score when compared to rule-based monitoring. It indicates that AI technology can be applied to increase the efficiency of big data security operations through pattern recognition.

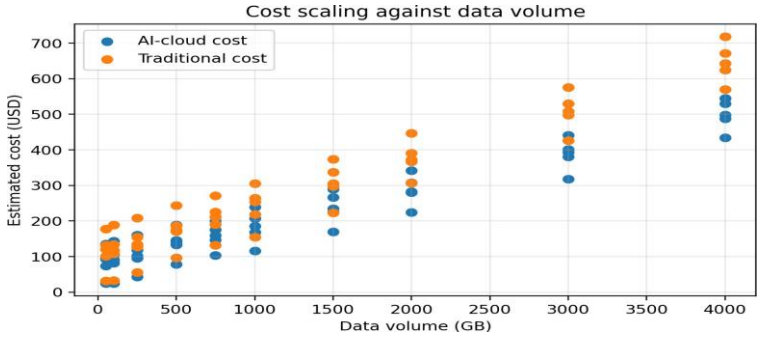

Figure 8: Cost scaling against data volume.

Workload	Latency Reduction	Throughput Gain	Cost Reduction	F1 Gain (points)
Batch ETL	34.29%	28.22%	24.34%	9.70
Streaming Analytics	35.71%	28.15%	22.29%	8.98
Security Logs	35.92%	28.24%	23.72%	8.15
IoT Telemetry	35.46%	27.48%	23.51%	9.94
Hybrid Enterprise	34.68%	27.84%	22.87%	8.89

Discussion

Thus, our findings confirm that artificial intelligence needs to be introduced not in analytics but at the management level of big data architecture. The point is that the conventional pipeline is configured using predetermined policies for resource allocation, quality validation, and security control. In turn, the proposed approach ensures that the system dynamically learns about the nature of its workload and operation. Thus, potential problems are detected and resolved faster and more efficiently.

The main finding is that big data management needs to be seen as a life cycle and managed accordingly. Metadata should link together the processes of ingestion, processing, storing, securing, and governing big data. Otherwise, improving any of these components can be done but without achieving optimal performance in the others.

The suggested framework can serve as an educational/research tool for universities and research settings. Students may utilize public datasets, cloud logging, IoT data, and institutional datasets in lieu of the artificial dataset, following an ethical review process. On the other hand, this framework can be applied in the business setting, and the development process can be organized in phases beginning with AI-driven monitoring, followed by autoscaling, and finally governance automation.

Threats to Validity and Limitations

One of the limitations associated with the study is that the performance analysis utilizes a synthetic benchmark data set. The use of synthetic data makes it possible to make a clear comparison and to conduct a study without concerns about privacy issues but, in reality, it can hardly be representative of all possible conditions.

Another limitation related to the study is that it is purely theoretical. Future studies need to examine the implementation and validation of the proposed approach via real-world cloud infrastructures, distributed computing systems, and open benchmark datasets. Energy consumption, carbon footprint, privacy issues, interpretability, and analyst burden are some of the factors that future work may need to take into consideration.

Conclusion and Future Work

This paper introduced an AI-based approach to enable efficient, secure, and intelligent big data management in cloud environments. This research highlighted the challenges in managing big data, summarized current solutions, proposed an AI-based architecture for big data management, and validated the framework using a repeatable synthetic benchmark dataset. The outcomes suggest that the proposed pipeline can significantly decrease latency, enhance throughput, minimize cost, and improve anomaly detection compared to the traditional cloud-based pipeline.

Further research could focus on validating the proposed framework on real-life data and cloud platforms. Researchers may consider exploring the use of federated learning for privacy-preserving big data management, blockchain-based auditing trail, explainable AI for governance, edge-cloud computing for IoT environments, and self-healing system to automatically detect and fix performance anomalies.

References

1. J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Communications of the ACM, vol. 51, no. 1, pp. 107-113, 2008.
2. M. Zaharia et al., "Apache Spark: A Unified Engine for Big Data Processing," Communications of the ACM, vol. 59, no. 11, pp. 56-65, 2016.
3. X. Wu et al., "Data Mining with Big Data," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, 2014.
4. R. Buyya, C. Vecchiola, and S. T. Selvi, Mastering Cloud Computing: Foundations and Applications Programming. Morgan Kaufmann, 2013.
5. NIST, "The NIST Definition of Cloud Computing," Special Publication 800-145, 2011.
6. D. Laney, "3D Data Management: Controlling Data Volume, Velocity, and Variety," META Group, 2001.

    A. Labrinidis and H. V. Jagadish, "Challenges and Opportunities with Big Data," Proceedings of the VLDB Endowment, vol. 5, no. 12, pp. 2032-2033, 2012.

7. S. Sagiroglu and D. Sinanc, "Big Data: A Review," International Conference on Collaboration Technologies and Systems,
8. pp. 42-47, 2013.
9. C. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," Mobile Networks and Applications, vol. 19, pp. 171-209, 2014.
10. N. Al-Mekhlal and B. M. Alabsi, "A Survey of Big Data Management: Taxonomy and Challenges," Journal of Big Data, 2023.

    A. Al-Mazrawe and A. K. Al-Mousa, "Anomaly Detection in Cloud Network: A Review," BIO Web of Conferences, 2024.

11. S. Baimukhanov et al., "Enhancing ML-based Anomaly Detection in Data-Intensive IoT-Edge-Cloud Ecosystems," Expert Systems with Applications, 2025.
12. IBM Research, "Anomaly Detection in Large-Scale Cloud Systems," arXiv preprint arXiv:2411.09047, 2024.
13. M. Armbrust et al., "Above the Clouds: A Berkeley View of Cloud Computing," University of California, Berkeley, 2009.

    A. Gandomi and M. Haider, "Beyond the Hype: Big Data Concepts, Methods, and Analytics," International Journal of Information Management, vol. 35, no. 2, pp. 137-144, 2015.

14. K. Kambatla, G. Kollias, V. Kumar, and A. Grama, "Trends in Big Data Analytics," Journal of Parallel and Distributed Computing, vol. 74, no. 7, pp. 2561-2573, 2014.
15. J. Manyika et al., "Big Data: The Next Frontier for Innovation, Competition, and Productivity," McKinsey Global Institute, 2011.
16. L. Wang et al., "Cloud-Based Big Data Systems," IEEE, 2015.

17. P. Mell and T. Grance, "The NIST Definition of Cloud Computing," NIST, 2011.
18. Y. Li et al., "Scalable Big Data Systems: A Survey," IEEE Access, 2020.
19. Z. Khan et al., "Future Challenges of Big Data," IEEE Access, 2021.
    A. Jain et al., "Security and Privacy Issues in Big Data," IEEE, 2016.
20. P. Russom, "Big Data Analytics," TDWI Best Practices Report, 2011.
21. N. Kshetri, "Big Data's Impact on Privacy, Security and Consumer Welfare," Telecommunications Policy, 2014.
22. M. A. Ferrag et al., "Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study," Journal of Information Security and Applications, 2020.
23. R. Ranjan, "Streaming Big Data Processing in Datacenter Clouds," IEEE Cloud Computing, 2014.
24. F. Li et al., "A Survey on Edge Computing for the Internet of Things," IEEE Access, 2018.
    A. Nazir et al., "Machine Learning-Based Cloud Data Classification and Intrusion Detection," 2024.
25. Google Cloud, "What is Artificial Intelligence?" Google Cloud Learn, accessed 2026.
26. IBM, "What is Artificial Intelligence?" IBM Think, accessed 2026.

Appendix A: Synthetic Benchmark Dataset Sample

The complete dataset is provided as a separate CSV and Excel workbook. A sample of the first 10 scenarios is shown below.

Scenario ID	Workload	Data Volume GB	Velocity MBps	Variety Score	Baseline Latency s	AI Cloud Latency s	Baseline Throughput rps	AI-Cloud Throughput rps	Baseline-Cost USD	AI Cloud Cost USD	Baseline Security F1	AI Cloud Security F1
1	Batch ETL	50	25	2	1.52	0.97	4930.4	6573.7	32.08	24.55	0.714	0.815
2	Streaming Analytics	50	220	4	2.68	1.75	4784.8	6046.6	132.36	96.85	0.697	0.778
3	Security Logs	50	160	5	2.3	1.49	4717.6	5805.3	121.28	92.84	0.722	0.791
4	IoT Telemetry	50	300	6	3.32	2.13	4660.5	6026.7	177.75	135.83	0.724	0.842
5	Hybrid Enterprise	50	90	7	2.26	1.38	4714.1	5832.7	101.59	74.25	0.724	0.817
6	Batch ETL	100	25	2	1.81	1.16	4696.8	5962.5	32.61	24.64	0.695	0.801
7	Streaming Analytics	100	220	4	3.08	1.83	4628.8	5862.5	134.52	105.22	0.726	0.83
8	Security Logs	100	160	5	2.57	1.63	4584.2	6024.3	118.3	90.65	0.712	0.784
9	IoT Telemetry	100	300	6	3.83	2.58	4519.0	5943.1	189.31	143.32	0.735	0.848
10	Hybrid Enterprise	100	90	7	2.49	1.69	4346.6	5688.7	107.74	82.31	0.74	0.801