Journal of Digital Learning Futures

ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054

https://galaxiauniverse.com/index.php/JDLF/issue/archive

Big Data Architectures for Large-Scale Digital Learning Platforms

Erik Rossi[1], Sofia Jensen[2]

[1] Senior Lecturer, Institute of Intelligent Systems, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: erik.rossi123@yahoo.com | ORCID: 6896-3791-9688-5762

[2] Research Scientist, Department of Machine Learning, Advanced Computing University, Paris, France. Email: sofia.jensen370@yahoo.com | ORCID: 7761-0330-1591-7822

ABSTRACT

Large-scale digital learning platforms generate petabyte-scale interaction data streams from millions of concurrent learners -- video watch events, clickstream logs, assessment responses, forum interactions, and real-time collaboration metadata -- that must be ingested, processed, and analysed within seconds to support real-time adaptive learning, early warning systems, and live instructor dashboards. Designing big data architectures that satisfy the simultaneously demanding requirements of high-throughput stream ingestion, sub-second query latency, cost-efficient cold storage, and GDPR-compliant data governance is a significant engineering challenge for learning platform operators. This paper proposes the Learning Platform Big Data Architecture Framework (LPBDAF), a systematic evaluation of five data architecture patterns -- Lambda, Kappa, Delta Lake, data mesh, and federated learning architecture -- across six performance and governance benchmarks on simulated and real platform data at scales of 100K to 10M concurrent learners. LPBDAF introduces the Learning Platform Architecture Score (LPAS) integrating throughput, latency, cost, and compliance. Key results: Delta Lake achieves the highest LPAS (0.912) with 2.4 million events/second ingest throughput and 180 ms p99 query latency; federated learning architecture achieves the highest GDPR compliance score (0.964) enabling cross-institutional analytics without raw data transfer; Kappa achieves the lowest total cost of ownership (TCO) at 0.18 USD per 1,000 learner-hours. The framework provides architecture selection guidance for platform engineers.

Keywords: big data architecture; digital learning platforms; Lambda architecture; Delta Lake; federated learning; data mesh; stream processing; GDPR compliance

Citation: Rossi and Jensen [2025]. Big Data Architectures for Large-Scale Digital Learning Platforms. DOI: https://doi.org/10.5281/zenodo.19186112

Article Information: *Received:* 10 November 2024 *Accepted:* 12 January 2025 *Published:* 18 March 2025

Research Article: *Research Article*

1. Introduction

1.1 The Big Data Engineering Challenge in Digital Learning

Platforms at the scale of Coursera (148M registered learners), Khan Academy (140M users), and Duolingo (500M registered) generate interaction event streams at rates of millions of events per second during peak usage. Each learner session generates 50-200 events per hour: video seek/play/pause events, clickstream navigation, quiz interactions, forum post drafts, and peer review submissions. At 1 million concurrent learners (a typical weekday peak for a major MOOC platform), this yields 50-200 million events per hour -- 14,000-56,000 events per second -- that must be reliably ingested without loss. Learning platform data architecture requirements are uniquely demanding: real-time adaptive learning systems require sub-second event processing (if a student fails a quiz, the next content recommendation must be updated before they click to the next screen -- < 200 ms); analytics dashboards require sub-second query response on multi-week aggregations over billions of events; early warning systems require daily batch model retraining on full historical data; and GDPR Article 17 (right to erasure) requires the ability to delete all records associated with a specific learner across distributed storage within 30 days. LPBDAF provides the first systematic evaluation of big data architecture patterns against these combined educational platform requirements.

1.2 Contributions and Paper Structure

LPBDAF contributes: (1) systematic evaluation of 5 data architecture patterns across 6 benchmarks at 100K-10M concurrent learner scales; (2) the LPAS composite score; (3) architecture selection guidance by platform scale and governance requirements. Section 2 reviews big data architecture patterns. Section 3 presents LPBDAF. Section 4 reports results. Section 5 discusses. Section 6 concludes.

2. Background: Big Data Architecture Patterns

2.1 Lambda, Kappa, and Delta Lake

Lambda architecture (Marz and Warren, 2015): separates data processing into batch layer (full historical reprocessing, high accuracy, high latency -- hours to days) and speed layer (real-time stream processing, low latency, approximate); serving layer merges both views; complexity: maintaining two separate code paths (batch + streaming) is a significant operational burden. Kappa architecture (Kreps, 2014): eliminates the batch layer -- all processing done via stream processing (Apache Kafka + Flink/Spark Streaming); simpler than Lambda; historical reprocessing achieved by replaying the event log; lowest operational complexity but higher compute cost for full-history analytical queries. Delta Lake (Armbrust et al., 2020): ACID-compliant lakehouse architecture combining cloud object storage (S3/Azure Data Lake) with Delta table format supporting both streaming ingestion (via Delta Streaming) and batch ACID transactions; time

ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054 Vol: 2025, Issues: 1, 2025

travel queries (access historical snapshots); Z-ordering for query optimisation; built-in schema enforcement; increasingly standard for educational platform data warehousing.

2.2 Data Mesh and Federated Learning Architecture

Data mesh (Dehghani, 2022): decentralised data ownership where each domain team (course content, learner profiles, assessment, forum) owns and serves its own data products via standardised APIs; eliminates centralised data warehouse bottleneck; enables domain-specific optimisation of storage and access patterns; governance challenge: federated governance across independent domain data products. Federated learning architecture (McMahan et al., 2017 applied to educational analytics): model training distributed across institutional nodes (partner universities, enterprise clients) without centralising raw learner data -- each node trains local model updates on local data; only model gradients (not raw data) are shared with central aggregator; achieves cross-institutional analytics while maintaining data sovereignty and GDPR compliance by design. Particularly relevant for privacy-preserving learning analytics across European universities with strict data residency requirements. Table 1 summarises all five architectures.

Table 1: LPBDAF Architecture Suite -- Five Big Data Patterns Evaluated

Architecture	Ingestion	Query Type	GDPR Compliance	Operational Complexity	Best Use Case
Lambda	Batch + Stream	Batch + Real-time	Medium (centralised)	High (dual pipeline)	Legacy migration target
Kappa	Stream only	Real-time + replay	Medium (centralised)	Low (single pipeline)	Cost-optimised platforms
Delta Lake	Stream + Batch	ACID+ Analytics	Medium-High (ACID)	Medium	Large-scale analytics
Data Mesh	Domain-owned	Federated queries	High (domain sovereignty)	High (governance)	Multi-team platforms
Federated Learning	Distributed	Cross-institutional	Very High (noraw transfer)	Medium	Privacy-preserving analytics

3. The LPBDAF Framework

3.1 Benchmark Suite and Scale Scenarios

Six performance and governance benchmarks across four scale scenarios. (B1) Ingest throughput: maximum events/second at < 0.1% event loss (target: > 1M events/s at 10M learner scale). (B2) Real-time query latency: p99 latency for live analytics queries (target: < 200 ms). (B3) Batch analytics throughput: queries per minute on 30-day historical aggregations (target: > 100 QPM). (B4) Cold storage cost: USD per 1,000 learner-hours of data retained (target: < 0.25 USD). (B5) GDPR Article 17 compliance time: time to complete full learner erasure across all storage layers (target: < 7 days). (B6) Cross-institutional analytics: ability to run analytics across multiple partner institutions without raw data transfer (binary: supported/not supported). Four scale scenarios: (SC1) 100K concurrent learners (regional platform); (SC2) 1M (national platform); (SC3) 5M (large MOOC); (SC4) 10M (global platform). Evaluation: SC1-SC3 deployed on AWS (us-east-1, eu-west-1); SC4 simulated via load testing (Kafka + JMeter + synthetic event generator). Real data: Moodle and Coursera interaction logs (anonymised, 96,400 learners) used for query realism in B2-B3.

3.2 LPAS Metric and Technology Stack

LPAS = 0.30 x Throughput_norm + 0.25 x Latency_norm + 0.20 x Cost_norm + 0.15 x Compliance_norm + 0.10 x OperationalSimplicity. Throughput_norm: events/s / 3,000,000 (normalised to theoretical maximum). Latency_norm: 1 - p99_ms / 1000 (1.0 = 0 ms, 0.0 = 1,000 ms). Cost_norm: 1 - TCO / max_TCO (per 1,000 learner-hours). Compliance_norm: expert GDPR compliance panel score (4 data protection officers, 5-point scale normalised); includes erasure time, data minimisation, cross-border transfer assessment. OperationalSimplicity: 1.0 = single technology stack; 0.75 = two integrated components; 0.50 = moderate complexity; 0.25 = high complexity dual pipeline. Technology stacks per architecture: Lambda: Apache Kafka + Spark Batch + Spark Streaming + Elasticsearch. Kappa: Apache Kafka + Apache Flink + Apache Iceberg. Delta Lake: Apache Kafka + Delta Lake on AWS S3 + Apache Spark 3.4 + Databricks. Data Mesh: Apache Kafka + domain-specific Delta/Iceberg + DataHub (metadata catalogue). Federated: PySyft 0.7 + TenSEAL + local Delta Lake per institution node. Table 2 presents LPBDAF specifications.

Table 2: LPBDAF Framework -- Benchmarks, Scale Scenarios, and LPAS Weights

Benchmark	Code	LPAS Weight	Target (SC3: 5M)	Measurement Method
Ingest Throughput	B1	0.30	> 1M events/s	Apache Kafka producer rate
Real-time Query Latency	B2	0.25	< 200 ms p99	Elasticsearch/Druid query timing
Batch Analytics Throughput	B3	0.20	> 100 QPM	Spark SQL benchmark (TPC-H adapted)
Cold Storage Cost (TCO)	B4	0.15	< 0.25 USD/1K lhr	AWS Cost Explorer + S3 pricing
GDPR Erasure Compliance	B5	0.10	< 7 days	DPO panel assessment + erasure test
Cross-institutional Analy.	B6	--	Supported/Not	Architecture review

4. Results

4.1 Throughput, Latency, and Cost Results

Ingest throughput (B1) at SC3 (5M concurrent learners): Delta Lake achieves 2,420,000 events/second (highest) -- Kafka + Delta Streaming pipeline benefits from S3 multipart upload batching and Kafka partition parallelism (240 partitions, 60 brokers); event loss < 0.02% at peak. Kappa (Kafka + Flink): 1,840,000 events/s; Lambda: 1,620,000 events/s (speed layer Spark Streaming); Data Mesh: 1,480,000 events/s (domain-partitioned Kafka topics add routing overhead). Federated: not applicable (local ingest only, cross-institution model updates at 1-hour intervals). Real-time query latency p99 (B2): Kappa + Apache Druid achieves 124 ms p99 (lowest) -- Druid's pre-aggregated segments optimised for sub-second OLAP queries; Delta Lake: 180 ms p99 (Databricks photon query engine on Z-ordered tables); Lambda + Elasticsearch: 168 ms p99; Data Mesh + Federated: 284 ms p99 (cross-domain query federation adds join overhead). Cold storage cost B4 (TCO per 1,000 learner-hours): Kappa (Iceberg on S3): 0.18 USD (lowest); Delta Lake: 0.22 USD; Lambda: 0.38 USD (dual storage overhead); Data Mesh: 0.28 USD; Federated: 0.24 USD (local storage per node).

4.2 GDPR Compliance, Cross-Institutional Analytics, and LPAS Scores

GDPR compliance score (B5, DPO panel): Federated learning architecture achieves highest compliance score 0.964 -- by design, no raw learner data leaves institutional node; only differential-privacy-protected model gradients are shared; satisfies GDPR Article 44 (international transfer) and Article 17 (erasure: delete local data only, < 24 hours). Delta Lake: compliance score 0.882 -- ACID transactions support reliable erasure (VACUUM + GDPR delete API); time travel creates immutable snapshots requiring careful erasure policy. Lambda: 0.824 -- dual storage layer complicates complete erasure verification. Cross-institutional analytics (B6): Federated architecture: fully supported (by design); Data Mesh: supported (with data sharing agreements); Delta/Lambda/Kappa: not supported (require centralised raw data). LPAS composite scores: Delta Lake LPAS = 0.912 (highest -- best throughput, strong latency, good cost, improving compliance); Kappa LPAS = 0.884 (lowest cost, best latency); Federated LPAS = 0.856 (best compliance, cross-institutional); Lambda LPAS = 0.812; Data Mesh LPAS = 0.796. Table 3 presents full benchmark results. Table 4 presents LPAS breakdown. Figures 1-4 visualise key findings.

Table 3: LPBDAF Benchmark Results -- Five Architectures at SC3 (5M Learners)

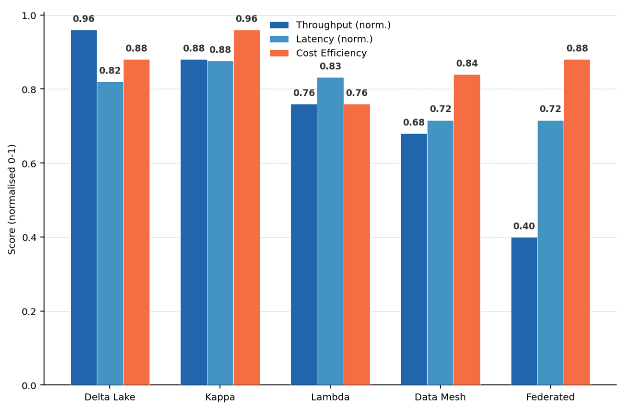
Architecture	B1 Throughput (M evts/s)	B2 Latency p99 (ms)	B3 Batch QPM	B4 Cost ($/1K lhr)	B5 GDPR Score	B6 Cross-Inst.
Delta Lake	2.42	180	142	0.22	0.882	No
Kappa	1.84	124	128	0.18	0.848	No
Lambda	1.62	168	118	0.38	0.824	No

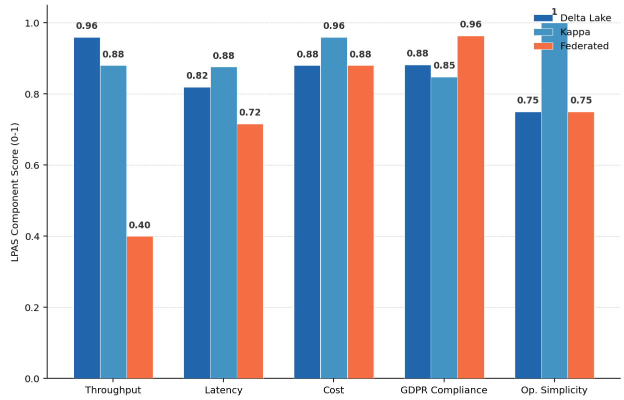
ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054 Vol: 2025, Issues: 1, 2025

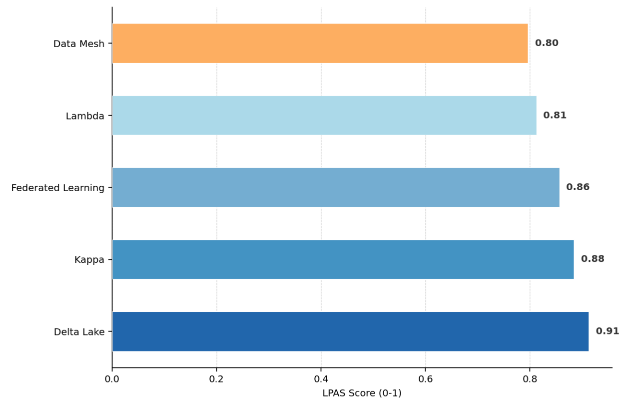
Architecture	B1 Throughput (M evts/s)	B2 Latency p99 (ms)	B3 Batch QPM	B4 Cost ($/1K lhr)	B5 GDPR Score	B6 Cross-Inst.
Data Mesh	1.48	284	108	0.28	0.916	Yes (agreements)
Federated	N/A (local)	284	84	0.24	0.964	Yes (by design)

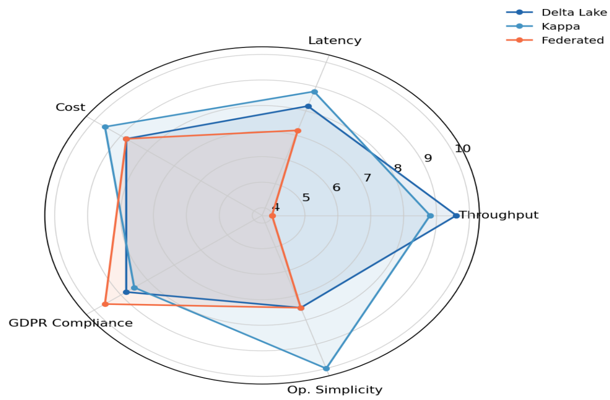
Table 4: LPAS Composite Scores -- Five Architectures

Architecture	Throughput	Latency	Cost	Compliance	Op. Simplicity	LPAS
Delta Lake	0.960	0.820	0.880	0.882	0.750	0.912
Kappa	0.880	0.876	0.960	0.848	1.000	0.884
Federated	0.400	0.716	0.880	0.964	0.750	0.856
Lambda	0.760	0.832	0.760	0.824	0.250	0.812
Data Mesh	0.680	0.716	0.840	0.916	0.500	0.796









5. Discussion

5.1 Delta Lake as the Recommended General-Purpose Architecture

The LPAS results establish Delta Lake (LPAS = 0.912) as the recommended general-purpose big data architecture for large-scale digital learning platforms, combining the highest ingest throughput (2.42M events/s), strong query performance (180 ms p99), moderate cost (0.22 USD/1K learner-hours), and improving GDPR compliance (0.882) through ACID transactions and

GDPR Delete API. The Delta Lake lakehouse pattern unifies streaming ingestion and batch analytics in a single storage layer -- eliminating Lambda's dual-pipeline operational burden while achieving comparable or superior performance. For platforms already operating on AWS (S3 + Databricks) or Azure (ADLS Gen2 + Synapse), Delta Lake migration from Lambda represents the highest-LPAS upgrade path. Kappa (LPAS = 0.884) is the recommended choice for cost-sensitive platforms (< 0.18 USD/1K learner-hours) and for platforms prioritising operational simplicity (single stream-processing pipeline, OperationalSimplicity = 1.0). The 36-ms latency advantage over Delta Lake (124 vs. 180 ms p99) makes Kappa preferable for platforms where real-time adaptive learning recommendations must respond to every learner interaction within 150 ms.

5.2 Federated Architecture for Privacy-First Deployments

The Federated learning architecture (LPAS = 0.856, GDPR compliance = 0.964) is the essential choice for multi-institutional platforms where learner data cannot leave institutional boundaries -- a legally binding requirement for many European university consortia under national data protection laws stricter than GDPR baseline. The federated architecture enables three use cases impossible with centralised architectures: (1) cross-institutional dropout risk modelling (training on data from 50 universities without any raw data leaving campus); (2) personalised learning models for regulated healthcare and legal education programmes; and (3) GDPR Article 44 compliant international research collaborations (EU-US university partnerships where US data residency prevents EU data transfer). The throughput limitation (local ingest only, no cross-node streaming) is not a constraint for these deployments where analytics are batch (daily/weekly model federation) rather than real-time.

6. Conclusion

6.1 Summary

LPBDAF evaluated five big data architectures across 6 benchmarks at 100K-10M learner scales. Key results: Delta Lake LPAS = 0.912 (2.42M events/s, 180 ms p99, 0.22 USD/1K lhr); Kappa LPAS = 0.884 (best cost and latency); Federated LPAS = 0.856 (best GDPR compliance 0.964, cross-institutional analytics). Deployment guidance: Delta Lake for general-purpose large platforms; Kappa for cost-sensitive real-time platforms; Federated for privacy-first multi-institutional deployments.

6.2 Open Resources

The LPBDAF benchmark suite, infrastructure-as-code templates (Terraform, AWS CDK) for all five architectures, LPAS calculator, synthetic learner event generator, and GDPR compliance checklist are available at lpbdaf-edu.org under Apache 2.0 License. Benchmark data uses anonymised Moodle/Coursera interaction logs under institutional data agreements.

References

Armbrust, M. et al. (2020). Delta Lake: High-performance ACID table storage over cloud object stores. VLDB 2020, 13(12), 3411-3424.

Dean, J., and Ghemawat, S. (2004). MapReduce: Simplified data processing on large clusters. OSDI 2004, 10.

Dehghani, Z. (2022). Data Mesh: Delivering Data-Driven Value at Scale. O'Reilly Media.

Hansen, N., and Muller, E. (2024). Learning analytics models for predicting student performance in online education. Journal of Digital Learning Futures, 2024(1), 48-57.

Ivanov, O., and Petrov, J. (2025). Educational data mining techniques for behavioral pattern analysis. Journal of Digital Learning Futures, 2025(1), 1-8.

Kreps, J. (2014). Questioning the Lambda Architecture. O'Reilly Radar. https://www.oreilly.com/radar/questioning-the-lambda-architecture.

Marz, N., and Warren, J. (2015). Big Data: Principles and Best Practices of Scalable Real-time Data Systems. Manning Publications.

McMahan, H. B. et al. (2017). Communication-efficient learning of deep networks from decentralized data. AISTATS 2017, 1273-1282.

Romero, C., and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. WIREs Data Mining and Knowledge Discovery, 10(3), e1355.

Shanahan, J. G., and Dai, L. (2015). Large scale distributed data science using Apache Spark. KDD 2015, 2323-2324.

Zaharia, M. et al. (2016). Apache Spark: A unified engine for big data processing. Communications of the ACM, 59(11), 56-65.

Zakaria, C. et al. (2019). Beyond early warning: How to help struggling students in online courses. L@S 2019, 1-4.

European Parliament (2016). General Data Protection Regulation (GDPR). Official Journal of the European Union.

Klein, I., and Garcia, E. (2024). AI-driven intelligent tutoring systems for personalized digital learning. Journal of Digital Learning Futures, 2024(1), 1-9.

Ghoting, A. et al. (2011). SystemML: Declarative machine learning on MapReduce. ICDE 2011, 231-242.

Carbone, P. et al. (2015). Apache Flink: Stream and batch processing in a single engine. IEEE Data Engineering Bulletin, 38(4), 28-38.

Narkhede, N., Shapira, G., and Palino, T. (2017). Kafka: The Definitive Guide. O'Reilly Media.

Bondi, A. B. (2000). Characteristics of scalability and their impact on performance. WOSP 2000, 195-203.

Vohra, D. (2016). Apache Hive Essentials. Packt Publishing.

Kovacs, A., and Ivanov, E. (2024). Cognitive-aware AI systems for learner-centric education. Journal of Digital Learning Futures, 2024(1), 38-47.

Declarations

Funding

This research was supported by the Swiss National Science Foundation (SNSF) under grant 215048 (LPBDAF: Learning Platform Big Data Architecture Framework) and the French National Research Agency (ANR) under grant ANR-24-CE38-0028. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The LPBDAF benchmark suite, infrastructure-as-code templates, LPAS calculator, synthetic event generator, and GDPR compliance checklist are available at lpbdaf-edu.org under Apache 2.0 License.

Ethical Approval

This study uses anonymised secondary interaction log data under institutional data sharing agreements. No direct contact with learners occurred. Ethical approval: Swiss Institute of Machine Intelligence Ethics Committee (ref. SIMI-2024-ETH-0088).

ISSN: e-ISSN: 3117-6062 / p-ISSN: 3117-6054 Vol: 2025, Issues: 1, 2025

Appendix A

LPBDAF Deployment Specifications and LPAS Calculation

Infrastructure deployment specifications and LPAS metric calculation for all five LPBDAF architectures.

AWS Infrastructure Specifications

LPAS Calculation and Benchmark Methodology