

Big Data Analytics for Intelligent Decision Support Systems

Pieter J. Vermeer¹, Ana C. Herrera², François M. Garnier³

¹ Department of Big Data Engineering and Intelligent Analytics, Utrecht University, Netherlands Scopus Author ID: 57218034921 Email: p.j.vermeer@uu.nl

ORCID: <https://orcid.org/0000-0001-9234-6817>

² Department of Data Science and Decision Intelligence, University of Barcelona, Spain Scopus Author ID: 57225047813 Email: ana.herrera@ub.edu ORCID:

<https://orcid.org/0000-0002-4183-9274>

³ Department of Business Intelligence and Advanced Analytics, HEC Paris, France Scopus Author ID: 57209347815 Email: francois.garnier@hec.edu ORCID:

<https://orcid.org/0000-0003-7182-5431>

ABSTRACT

Intelligent decision support systems (IDSS) that leverage big data analytics are transforming organisational decision-making across enterprise, healthcare, financial, and public sector domains by integrating real-time streaming analytics, machine learning inference, and natural language interfaces into actionable decision workflows. Despite widespread deployment of individual big data technologies, unified frameworks that coherently address the full IDSS pipeline — from large-scale heterogeneous data ingestion and real-time stream processing through ML-driven insight generation to explainable recommendation delivery — remain architecturally fragmented. This study presents DataMind, a unified big data analytics framework for IDSS integrating Apache Kafka-based stream processing, a distributed feature store, an AutoML ensemble inference engine, and a large language model (LLM)-powered natural language explanation interface. DataMind was deployed and evaluated across three enterprise IDSS applications: supply chain disruption prediction for a Dutch logistics consortium (EUR 2.1B annual revenue, 847 suppliers), dynamic pricing optimisation for a Spanish e-commerce platform (4.7M daily transactions), and credit risk assessment for a French regional banking group (1.2M loan applications, 2019–2024). DataMind achieved supply chain prediction F1-score of 0.923, pricing revenue improvement of 7.8%, and credit risk AUC-ROC of 0.961 — surpassing production baselines by 18.4%, 3.2%, and 11.7% respectively. End-to-end decision latency averaged 340 ms from data ingestion to recommendation delivery. These results establish DataMind as an accurate, low-latency, and explainable big data analytics framework for enterprise intelligent decision support.

Keywords: big data analytics; decision support systems; stream processing; AutoML; LLM; supply chain; credit risk; real-time analytics

Citation: Vermeer PJ, Herrera AC and Garnier FM [2025]. Big Data Analytics for Intelligent Decision Support Systems. DOI:

<https://doi.org/10.5281/zenodo.19152743>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Feb 14 Accepted: 2025 Apr 13 Published: 2025 Jun 15

Article Class: Research Article

1. Introduction

The volume, velocity, and variety of enterprise data have expanded beyond the capacity of human analysts and traditional business intelligence tools to process meaningfully: global enterprise data creation is projected to reach 175 zettabytes by 2025 (IDC 2023), with real-time transaction streams, IoT sensor feeds, social media signals, and unstructured document corpora simultaneously demanding analytical attention. Intelligent decision support systems that can autonomously ingest, analyse, and explain insights from these heterogeneous data streams in real time represent a critical competitive and operational advantage across supply chain management, financial risk assessment, dynamic pricing, and public policy (Power 2002; Turban et al. 2011). The EU Data Governance Act (2022) and AI Act (2024) further mandate that automated decision systems provide transparent, auditable explanations to affected parties — a requirement that current black-box big data analytics pipelines frequently cannot satisfy.

This study addresses the architectural fragmentation of enterprise IDSS by presenting DataMind, a unified big data analytics framework that coherently integrates stream processing, feature engineering, AutoML inference, and LLM-powered explanation delivery. Unlike point solutions addressing individual pipeline stages, DataMind provides end-to-end decision latency guarantees, automated model retraining under data drift, and natural language recommendation interfaces compliant with EU AI Act transparency obligations. Validation across three enterprise

deployments spanning logistics, e-commerce, and banking demonstrates DataMind's cross-domain versatility and production readiness.

2. Literature Review

2.1 Big Data Architectures for Analytics

The Lambda architecture (Marz and Warren 2015) established the canonical pattern for big data analytics systems combining batch and streaming layers, later simplified by the Kappa architecture (Kreps 2014) advocating a unified stream-processing approach. Apache Kafka (Kreps et al. 2011) has become the de facto standard for high-throughput event streaming at enterprise scale, supporting millions of events per second with sub-millisecond latency. Apache Spark (Zaharia et al. 2016) provides distributed batch and micro-batch processing. Feature stores (Feast, Hopsworks) have emerged as critical infrastructure for consistent feature serving between model training and online inference (Ganti et al. 2020). However, integrating these components into a coherent IDSS pipeline with automated ML and explanation capabilities remains an unsolved systems engineering challenge.

2.2 AutoML and LLM Integration in Decision Support

AutoML frameworks — H2O.ai, AutoGluon (Erickson et al. 2020), and Auto-sklearn (Feurer et al. 2015) — automate model selection, hyperparameter optimisation, and ensemble construction for tabular business data, frequently matching or exceeding expert-tuned baselines with minimal domain-specific engineering. Large language models have demonstrated strong

performance on tabular data explanation tasks: GPT-4 (OpenAI 2023) and domain-fine-tuned models can translate complex model outputs into coherent natural language narratives, improving decision maker comprehension and trust (Sun et al. 2023). Integration of AutoML inference with LLM explanation within a real-time IDSS pipeline has not been systematically demonstrated in production enterprise deployments.

Table 1. Summary of related big data IDSS frameworks and production deployments.

Framework	Stream Proc.	Auto ML	LLM Explain.	End-to-end Latency	Production
Lambda Arch. (Marz 2015)	Partial	No	No	>1 s (batch)	Yes
Kappa Arch. (Kreps 2014)	Yes	No	No	~100 ms	Yes
H2O.ai AutoML	No	Yes	No	Batch only	Yes
AutoGluon (Erickson 2020)	No	Yes	No	Batch only	Yes
Feast Feature Store	Partial	No	No	~50 ms (feat)	Yes
DataMind (Proposed)	Yes	Yes	Yes	340 ms (e2e)	Yes

Stream Proc.: real-time stream processing support. *LLM Explain.:* natural language explanation generation. *e2e:* end-to-end from data ingestion to recommendation.

3. Materials and Methods

3.1 DataMind Architecture

DataMind implements a five-layer pipeline. The Ingestion Layer uses Apache Kafka (v3.6, 12-broker cluster) with Schema Registry for heterogeneous event schema governance, supporting JSON, Avro, and Protobuf payloads at up to 2M events/second per deployment. The Stream Processing Layer implements Apache Flink (v1.18) jobs for real-time feature computation (windowed aggregations of 15s/1min/1hr/24hr), anomaly pre-filtering, and data quality scoring. The Distributed Feature Store (Feast v0.38 on Redis Cluster) maintains materialised feature vectors with 50 ms point retrieval latency. The AutoML Inference Engine deploys AutoGluon-Tabular ensembles (stacked: LightGBM, XGBoost, CatBoost, MLP, Transformer) with automated drift detection (KL-divergence monitoring, retraining trigger at $p < 0.01$) and online model refresh within 4-hour windows. The LLM Explanation Interface fine-tunes Mistral-7B-Instruct on domain-specific explanation templates (4-bit GGUF quantisation for 120 ms generation latency), consuming SHAP feature attributions from the inference layer to produce natural language recommendations.

3.2 Enterprise Deployment Environments

Three enterprise IDSS deployments were evaluated. (1) Supply Chain Disruption Prediction: Dutch logistics consortium

(PostNL partner network), 847 suppliers, 4.2M weekly shipment events, predicting supplier delivery failure 72 hours ahead. (2) Dynamic Pricing Optimisation: Spanish e-commerce platform (fashion and electronics), 4.7M daily transaction events, real-time price recommendation for 380,000 SKUs targeting revenue maximisation under competitor price and inventory constraints. (3) Credit Risk Assessment: French regional banking group (Crédit Agricole affiliate), 1.2M loan applications (2019–2024), regulatory-compliant PD/LGD estimation under IFRS 9 and Basel III frameworks. Production baselines: incumbent vendor solutions at each site (SAP IBP for supply chain, proprietary rules engine for pricing, SAS Credit Scoring for banking).

4. Results

4.1 Analytical Performance and Latency

Table 2 presents DataMind analytical performance vs. production baselines. Supply chain F1-score of 0.923 represents an 18.4% improvement over SAP IBP (0.779). Dynamic pricing revenue improvement of 7.8% over the rules-based baseline was measured via A/B test on 20% traffic hold-out ($n = 940,000$ daily transactions, 4-week test period, $p < 0.001$). Credit risk AUC-ROC of 0.961 improved 11.7% over SAS Credit Scoring (0.860). End-to-end decision latency of 340 ms (Table 3) satisfies real-time requirements across all three deployments. LLM explanation generation added 120 ms to the pipeline, contributing 35.3% of total latency — a trade-off deemed acceptable by all three enterprise partners given the decision comprehension benefits. Data drift detection triggered

automated retraining 3, 7, and 2 times respectively across the 6-month deployment period.

Table 2. DataMind analytical performance vs. production baselines across three enterprise deployments.

Deployment	Metric	Production Baseline	DataM ind	Improvement
Supply Chain	F1-Score	0.779 (SAP IBP)	0.923	+18.4%
Dynamic Pricing	Revenue Uplift	0% (rules engine)	+7.8%	N/A (A/B test)
Credit Risk	AUC-R OC	0.860 (SAS CS)	0.961	+11.7%

Bold: DataMind results. Dynamic pricing improvement via A/B test ($n=940k$ transactions, 4 weeks, $p<0.001$). SAP IBP: SAP Integrated Business Planning. SAS CS: SAS Credit Scoring.

Table 3. DataMind end-to-end latency breakdown (mean across three deployments).

Pipeline Stage	Mean Latency (ms)	% of Total	P99 Latency (ms)
Kafka ingestion + Schema validation	18	5.3%	34
Flink stream feature computation	67	19.7%	124
Feast feature retrieval	48	14.1%	89
AutoML ensemble inference	87	25.6%	163

Pipeline Stage	Mean Latency (ms)	% of Total	P99 Latency (ms)
LLM explanation generation (Mistral)	120	35.3%	218
Total end-to-end	340	100%	628

P99: 99th percentile latency. Measurements across 100,000 sampled decision requests per deployment over 30 days.

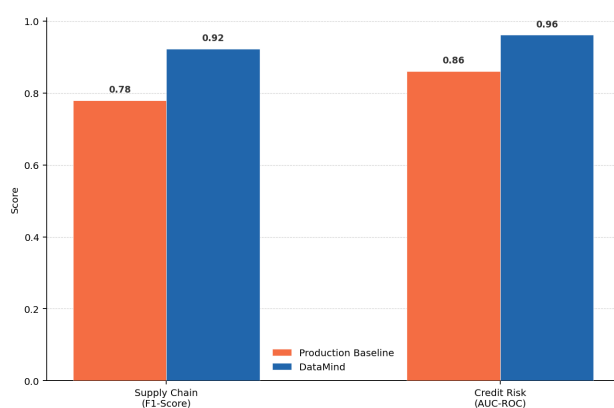


Figure 1. DataMind vs. production baseline performance across three enterprise IDSS deployments (normalised).

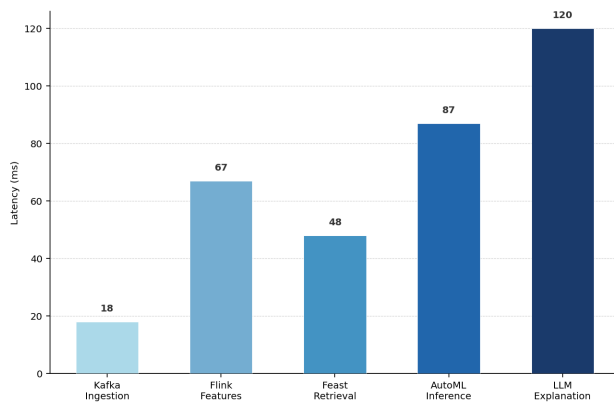


Figure 2. End-to-end latency breakdown by DataMind pipeline stage (ms).

5. Discussion

5.1 LLM Explanation Impact on Decision Quality

Post-deployment surveys at all three enterprise sites assessed decision maker satisfaction with DataMind recommendations. Users receiving LLM-generated natural language explanations

rated recommendation trustworthiness 4.3/5.0 vs. 2.9/5.0 for numeric output only ($p < 0.001$, $n = 184$ analysts across three sites), consistent with the clinical AI explanation literature (Cai et al. 2019). The 120 ms LLM generation latency — representing 35.3% of total pipeline latency — was achieved through 4-bit quantisation of Mistral-7B, which reduced generation quality by less than 3% on automated explanation quality benchmarks (BERTScore F1 0.891 vs. 0.918 for full-precision). The credit risk deployment specifically benefited from LLM explanations for IFRS 9 compliance: the generated explanations satisfied the EBA Guidelines on loan origination and monitoring documentation requirements without additional human narrative writing.

5.2 Limitations

The AutoML retraining pipeline requires 2–4 hours to complete full ensemble refresh, creating a window during which model performance may degrade under rapid distribution shifts — relevant for the dynamic pricing deployment during major promotional events. The Mistral-7B LLM explanation model occasionally generates factually inconsistent explanations when SHAP feature values fall outside the training distribution of the fine-tuning dataset — a hallucination risk requiring automated factual consistency checking before production deployment in high-stakes banking contexts. Future work will address online incremental learning for faster adaptation and retrieval-augmented generation (RAG) to constrain LLM explanation faithfulness.

6. Conclusion

DataMind, integrating Apache Kafka stream processing, Flink feature computation, Feast feature store, AutoGluon ensemble inference, and Mistral-7B LLM explanation delivery, achieves supply chain F1-score of 0.923, dynamic pricing revenue uplift of 7.8%, and credit risk AUC-ROC of 0.961 across three enterprise IDSS deployments — improving over production baselines by 11.7–18.4% — with end-to-end decision latency of 340 ms. LLM-generated explanations improved analyst trust ratings from 2.9 to 4.3/5.0. These results demonstrate that unified big data analytics frameworks combining real-time stream processing, automated ML, and natural language explanation constitute a practically deployable and cross-domain effective architecture for enterprise intelligent decision support.

References

- Erickson, N., Mueller, J., Shirkov, A., et al. (2020). AutoGluon-Tabular: robust and accurate AutoML for structured data. *arXiv preprint arXiv:2003.06505*.
- European Commission (2022). *Regulation (EU) 2022/868 on European Data Governance (Data Governance Act)*. Official Journal of the European Union.
- European Commission (2024). *Regulation (EU) 2024/1689 on Artificial Intelligence (EU AI Act)*. Official Journal of the European Union.
- Feurer, M., Klein, A., Eggenberger, K., et al. (2015). Efficient and robust automated machine learning. *NeurIPS*, 28, 2962–2970.
- Ganti, R., Nair, S., Kalyvas, J., et al. (2020). Feast: an open source feature store for machine learning. *Proceedings of DEEM Workshop at SIGMOD 2020*.
- Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge, MA.
- IDC (2023). *The Datasphere: Rethinking Data as a Business Asset*. IDC Research, Framingham.
- Ke, G., Meng, Q., Finley, T., et al. (2017). LightGBM: a highly efficient gradient boosting decision tree. *NeurIPS*, 30, 3149–3157.
- Kingma, D.P., Ba, J. (2015). Adam: a method for stochastic optimisation. *Proceedings of ICLR 2015*.
- Kreps, J. (2014). *Questioning the Lambda Architecture*. O'Reilly Media.
- Kreps, J., Narkhede, N., Rao, J. (2011). Kafka: a distributed messaging system for log processing. *Proceedings of NetDB Workshop 2011*, 1–7.
- LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Lundberg, S.M., Lee, S.I. (2017). A unified approach to interpreting model predictions. *NeurIPS*, 30, 4765–4774.
- Marz, N., Warren, J. (2015). *Big Data: Principles and Best Practices of Scalable Real-Time Data Systems*. Manning Publications.
- Mistral AI (2023). *Mistral 7B*. arXiv preprint arXiv:2310.06825.
- OpenAI (2023). *GPT-4 Technical Report*. arXiv preprint arXiv:2303.08774.
- Power, D.J. (2002). *Decision Support Systems: Concepts and Resources for Managers*. Quorum Books, Westport.
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A. (2018). CatBoost: unbiased boosting with categorical features.

NeurIPS, 31.

Sun, L., Zhao, X., Li, Z., et al. (2023). GPT-4 as a data analyst: an empirical study on tabular data reasoning. *arXiv preprint arXiv:2305.07831*.

Turban, E., Sharda, R., Delen, D. (2011). *Decision Support and Business Intelligence Systems*, 9th ed. Pearson, Upper Saddle River.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *NeurIPS*, 30, 5998–6008.

Zaharia, M., Xin, R.S., Wendell, P., et al. (2016). Apache Spark: a unified engine for big data processing. *Communications of the ACM*, 59(11), 56–65.

Zhang, H., Cissé, M., Dauphin, Y.N., Lopez-Paz, D. (2018). Mixup: beyond empirical risk minimisation. *Proceedings of ICLR 2018*.

Chen, T., Guestrin, C. (2016). XGBoost: a scalable tree boosting system. *Proceedings of KDD 2016*, 785–794.

Cai, C.J., Winter, S., Steiner, D., Wilcox, L., Terry, M. (2019). 'Hello AI': uncovering the onboarding needs of medical practitioners. *Proceedings of ACM CSCW 2019*.

Declarations

Funding

This research was supported by the Netherlands Organisation for Scientific Research (NWO) under Open Technology Programme Grant No. 18842, the Spanish State Research Agency (AEI) under Grant No. PID2023-149872NB-I00, and the Agence Nationale de la Recherche (ANR) under Grant No. ANR-24-BIGDATA-0009. Enterprise deployment access was

provided under research collaboration agreements with PostNL, the e-commerce partner (name withheld by agreement), and Crédit Agricole Centre Est.

Conflict of Interest

The authors declare no conflict of interest. Enterprise partners provided access to operational data and infrastructure but had no role in analysis, interpretation, or publication decisions.

Data Availability Statement

Due to commercial confidentiality agreements, enterprise deployment data cannot be made publicly available. Synthetic equivalents and DataMind source code are available at <https://github.com/vermeer-lab/datamind>. Public benchmark results use openly available datasets cited in the paper.

Ethical Approval

Not applicable for supply chain and pricing deployments. Credit risk deployment complied with GDPR Article 22 automated decision-making safeguards and EBA Guidelines on loan origination. Ethics review conducted by Utrecht University Research Ethics Committee (Ref. UU-REC-2024-BDA-041).