



RESEARCH IN-
SIGHTS SERIES

Data Science & Analytics

Research Insights 2025–2026

A comprehensive synthesis of peer-reviewed research
on data science, analytics, and intelligent data systems

72+

Papers Reviewed

8

Research Domains

24

Pages

58.58%

LLM-BI Accuracy Gain

Atine John Bosco

Chartered Information Scientist (MBCS) · CISA · PhD Candidate

Google Cloud Digital Leader · ITIL 4 Practitioner



www.atine.info



[linkedin.com/in/atine](https://www.linkedin.com/in/atine)

1 Executive Summary

💡 Key Findings at a Glance

- **LLM-powered BI platforms** (e.g., DataLab) deliver **58.58%** accuracy improvement and **61.65%** token-cost reduction on enterprise BI tasks
- **Human-centred AutoML** is the new paradigm — moving beyond predictive optimisation toward interactive, role-aware, explainable workflows
- **Real-time ML pipelines** built on Kafka/Flink/Spark now embed anomaly detection, fraud detection, and trend analysis directly in stream processing
- **Data mesh architecture** and domain-owned pipelines are replacing centralised ETL, driven by cloud-native orchestration and event-driven patterns
- **SHAP-based interpretability** is the de facto standard for operationalising AutoML models in regulated industries
- **Responsible AI governance** remains the critical gap — no standardised quantitative benchmarks exist for fairness, bias, and explainability at scale

The data science and analytics landscape in 2025–2026 is defined by convergence: the fusion of *large language models* with traditional analytics pipelines, the maturation of *MLOps* practices, and the democratisation of machine learning through *AutoML* platforms. This report synthesises findings from **72+** **peer-reviewed papers** across eight research domains, providing ICT professionals and data leaders with actionable intelligence on where the field is heading.

The most transformative shift is the emergence of **LLM-agent architectures** as unified interfaces for business intelligence. Systems like DataLab demonstrate that multi-agent LLM frameworks — combining domain knowledge modules, inter-agent communication, and notebook-based context — can dramatically reduce the cost and complexity of enterprise analytics [10]. Simultaneously, the AutoML community is reorienting around human-centred design, recognising that predictive performance alone is insufficient for real-world deployment [4].

📄 Report Scope

Domains Covered: ML Pipelines & MLOps · AutoML & Automated Data Science · LLMs for Analytics · Real-Time Analytics · Data Engineering & Architecture · Big Data Platforms · Business Intelligence & Visualisation · Responsible AI & Governance

Sources: SciSpace, Google Scholar, ArXiv · **Papers:** 72+ · **Period:** 2024–2026

2 ML Pipelines & MLOps

2.1 The Shift from Batch ETL to Event-Driven Architectures

Modern data engineering is undergoing a fundamental architectural transition. Research by Kadakia (2025) describes the evolution from traditional batch ETL pipelines to *hybrid batch/event-driven*, cloud-native orchestration systems [1]. Key characteristics of this new paradigm include:

- **Data Mesh Architecture:** Domain-owned pipelines replacing centralised data warehouses, enabling independent scaling and governance
- **Autonomic Pipelines:** Self-healing pipeline behaviour with automated error detection, retry logic, and graceful degradation
- **Cloud-Native Orchestration:** Containerised workloads managed via Kubernetes with declarative pipeline definitions (e.g., Apache Airflow, Prefect, Dagster)
- **Event-Driven Ingestion:** Apache Kafka and Apache Pulsar as backbone for real-time data capture with guaranteed delivery semantics

2.2 MLOps: Closing the Training-Deployment Gap

The operationalisation of machine learning models remains a critical challenge. Trehan et al. (2024) demonstrate that designing robust, scalable ML pipelines requires explicit integration of *data engineering*, *model optimisation*, and *deployment best practices* [3]. Their framework identifies five pipeline stages:

1. **Ingest:** Structured/unstructured data capture with schema validation and lineage tracking
2. **Transform:** Feature engineering, normalisation, and automated quality checks
3. **Feature Store:** Centralised, versioned feature repository enabling training-serving consistency
4. **Model Training:** Experiment tracking (MLflow, Weights & Biases), hyperparameter optimisation
5. **Model Serving:** Canary deployments, A/B testing, automated rollback on performance degradation

✓ MLOps Best Practices

- Implement **feature stores** (Feast, Tecton, Hopsworks) to eliminate training-serving skew
- Use **model registries** (MLflow, SageMaker Model Registry) for versioning and audit trails
- Automate **data quality gates** at each pipeline stage to catch drift before model

degradation

- Monitor **model performance** in production with statistical drift detection (KS-test, PSI, ADWIN)

2.3 Automated Feature Engineering

Research on automated feature engineering for ETL pipelines [2] highlights its role in improving both model performance and pipeline transparency. Key advances include:

- **Deep Feature Synthesis (DFS):** Automated generation of aggregate and transform features from relational data
- **Neural Feature Selection:** Attention-based mechanisms for identifying high-importance features in high-dimensional datasets
- **Temporal Feature Engineering:** Automated extraction of lag, rolling window, and seasonality features for time-series data
- **Explainable Features:** SHAP-integrated feature engineering pipelines that provide global and local feature attribution

3 AutoML & Automated Data Science

3.1 The Human-Centred AutoML Paradigm

The most significant conceptual shift in AutoML research is the move from *performance-first* to *human-centred* design. Lindauer et al. (2024) present a compelling call-to-action arguing that AutoML systems must prioritise user interaction, role-aware workflows, and collaborative design [4]. This paradigm shift recognises several critical limitations of current systems:



Current AutoML Limitations

- **Black-box outputs:** Most AutoML systems produce models without sufficient explanations for business stakeholders
- **Role blindness:** Systems treat all users identically, ignoring the different needs of data scientists vs. domain experts vs. executives
- **Static workflows:** Limited support for iterative, exploratory workflows that reflect real data science practice
- **Trust deficits:** Without interpretability, regulated industries cannot deploy AutoML-generated models

3.2 AutoML System Components

Karl et al. (2024) provide a comprehensive survey of AutoML system components [5], identifying the following key subsystems:

ComponentDescription & Key Tools	
Data Preprocessing	Automated imputation, encoding, normalisation (Auto-sklearn, TPOT)
Model Search	Neural Architecture Search (NAS), ensemble construction, meta-learning
Hyperparameter Optimisation	Bayesian optimisation, multi-fidelity (BOHB, Hyperband), population-based training
Neural Architecture Search	Hardware-aware NAS, once-for-all networks, differentiable NAS (DARTS)
Multi-Criteria Evaluation	Pareto-optimal model selection balancing accuracy, latency, fairness, and cost
Explainability	SHAP, LIME, counterfactual explanations, model cards

3.3 Leading AutoML Frameworks



AutoML Ecosystem 2025–2026

- **Auto-sklearn 2.0**: Bayesian optimisation + meta-learning, strong on tabular data, open-source
- **Google AutoML / Vertex AI**: Cloud-native, multi-modal (vision, NLP, tabular), enterprise-grade
- **H2O.ai AutoML**: Open-source leader for financial services, supports XGBoost, GBM, deep learning
- **DataRobot**: Enterprise platform with automated model documentation and governance features
- **AutoGluon**: Amazon’s framework, state-of-the-art on tabular benchmarks, minimal configuration
- **FLAML**: Microsoft’s fast, lightweight AutoML for constrained compute environments

3.4 Operationalising AutoML with SHAP

Stoica & Stoica (2024) demonstrate that integrating SHAP-based explanations into AutoML workflows is the key to gaining deployment confidence in regulated industries [8]. Their framework recommends:

1. Generate global SHAP feature importance plots to validate model logic against domain knowledge
2. Use local SHAP waterfall plots for individual prediction explanations in audit trails
3. Implement SHAP-based monitoring to detect feature drift in production deployments
4. Include model cards with SHAP summaries in all model documentation packages

4 LLMs for Data Analysis & Business Intelligence

4.1 The DataLab Breakthrough

The most significant empirical result in LLM-powered analytics is the **DataLab platform** developed by Weng et al. (2024) [10]. DataLab is a unified LLM-based business intelligence platform that combines:

- **Domain Knowledge Modules:** Pre-loaded business context and schema understanding
- **Inter-Agent Communication:** Structured message passing between specialised agents (data analyst, visualisation, code generation)
- **Cell-Based Notebook Context:** Persistent state management across analytical steps
- **Tool-Call Validation:** Automated verification of generated code before execution

🏆 DataLab Performance Results			
navylight	Metric	Baseline	DataLab
	Enterprise BI Accuracy	Baseline	+58.58% relative
	Token Cost	Baseline	-61.65% reduction
	BI Benchmark Rank	Competitive	State-of-the-art
	Task Completion Rate	Moderate	Significantly higher

4.2 LLM Agent Architectures for Data Science

Cui et al. (2025) provide a comprehensive taxonomy of LLM-based automated data science architectures [9]:

- **Single-Agent Orchestration:** One LLM manages the full data science workflow — suitable for simple, well-defined tasks; limited by context window and tool diversity
- **Function-Oriented Multi-Agent:** Rigid role assignment (data analyst, ML engineer, visualiser) with fixed handoff protocols — higher reliability, lower flexibility
- **Role-Oriented Multi-Agent:** Dynamic task transfer based on task complexity and agent capability — highest flexibility, requires robust coordination mechanisms

4.3 LLMs as Pipeline Interfaces

Barbon Junior et al. (2024) argue that LLMs represent a paradigm shift in how data pipelines are authored and operated [11]. Key capabilities include:

- **NL2SQL:** Natural language to SQL query generation with schema-aware prompting (accuracy: 70–85% on Spider benchmark with fine-tuned models)
- **NL2VIS:** Natural language to data visualisation (Vega-Lite, D3.js specification generation)
- **Pipeline Authoring:** Generating Apache Airflow DAGs, dbt models, and Spark jobs from natural language descriptions
- **Anomaly Explanation:** LLMs providing natural language explanations for statistical anomalies detected by classical methods



LLM Analytics Limitations

- **Semantic misalignment:** LLMs may misinterpret domain-specific terminology in SQL/pipeline generation
- **Hallucinated schemas:** Models can generate queries referencing non-existent columns or tables
- **Evaluation gaps:** Lack of standardised benchmarks for enterprise-specific BI tasks beyond academic datasets
- **Cost at scale:** Token costs for complex multi-agent workflows can be prohibitive without optimisation (DataLab addresses this with 61.65% reduction)

5 Real-Time Analytics & Streaming Pipelines

5.1 Stream Processing Architecture

Research by Kaushik et al. (2025) establishes the canonical architecture for real-time intelligent data pipelines [13]. The three-layer model consists of:

- Ingestion Layer:** Apache Kafka (high-throughput, fault-tolerant message streaming), Apache Pulsar (multi-tenancy), AWS Kinesis (cloud-native)
 - Processing Layer:** Apache Flink (stateful stream processing, exactly-once semantics), Apache Spark Streaming (micro-batch, ML integration), Apache Storm (low-latency, sub-millisecond)
 - Serving Layer:** Redis (sub-millisecond feature serving), Elasticsearch (real-time search/analytics), ClickHouse (OLAP at streaming speed)

5.2 ML-in-Stream Applications

Embedding machine learning directly within streaming pipelines enables immediate operational decisions:

navymidUse Case	ML Approach	Latency Target
Fraud Detection	Online learning, Isolation Forest	<100ms
Anomaly Detection	LSTM autoencoders, statistical control charts	<500ms
Recommendation	Collaborative filtering, bandit algorithms	<50ms
Predictive Maintenance	CUSUM, change-point detection	<1s
Sentiment Analysis	Lightweight transformer (DistilBERT)	<200ms

5.3 AutoML for Dynamic Pipeline Adaptation

Thirunagalingam (2024) demonstrates that AutoML can automate pipeline adaptation under *concept drift* — the phenomenon where data distributions change over time, degrading model performance [14]. Key strategies include:

- Drift Detection:** ADWIN, DDM, and Page-Hinkley tests for statistical distribution monitoring
 - Automated Retraining:** Triggered retraining pipelines when drift exceeds configurable thresholds

6 Data Engineering & Big Data Platforms

6.1 The Modern Data Stack

The *Modern Data Stack* (MDS) represents a cloud-native, composable approach to data infrastructure. Pathak (2025) documents how AI is transforming data engineering across the full stack [16]:

 Modern Data Stack Components

Layer	Function	Leading Tools
Ingestion	Batch + streaming data capture	Fivetran, Airbyte, Kafka
Storage	Cloud data warehouse/lake	Snowflake, BigQuery, Databricks
Transformation	SQL-based modelling, dbt	dbt, Spark, Trino
Orchestration	Workflow scheduling, DAGs	Airflow, Prefect, Dagster
Serving	BI, ML feature serving	Looker, Tableau, Feast
Governance	Lineage, cataloguing	DataHub, Apache Atlas, Alation

6.2 Big Data Analytics Innovations

Research on big data analytics challenges and innovations [7] identifies five key trends shaping the 2025–2026 landscape:

- Lakehouse Architecture:** Combining the flexibility of data lakes with the ACID transactions and schema enforcement of data warehouses (Delta Lake, Apache Iceberg, Apache Hudi)
- Data Fabric:** Intelligent, self-service data integration layer using AI/ML for automated metadata management and data discovery
- Federated Query Engines:** Querying data across heterogeneous sources without centralisation (Trino, Dremio, Starburst)
- Sustainable Big Data:** Green computing approaches to reduce the carbon footprint of large-scale data processing [15]
- Edge Analytics:** Processing data at the point of generation to reduce latency and bandwidth costs (TensorFlow Lite, ONNX Runtime on edge devices)

6.3 AI-Augmented Data Engineering



AI Augmentation in Data Engineering

- **Automated Schema Inference:** LLMs inferring data schemas from raw files, reducing manual cataloguing effort by up to 70%
- **Intelligent Data Quality:** ML-based anomaly detection in data pipelines identifying format errors, outliers, and referential integrity violations
- **Smart Partitioning:** Reinforcement learning optimising data partition strategies for query performance
- **Automated Documentation:** LLMs generating pipeline documentation, data dictionaries, and lineage descriptions from code

7 Business Intelligence & Data Visualisation

7.1 Conversational BI: The New Frontier

The integration of LLMs into BI platforms is creating a new category of *conversational BI* tools. Key developments include:

- **Natural Language Queries:** Business users asking questions in plain English and receiving instant SQL-backed answers (Microsoft Copilot for Power BI, Tableau Pulse)
- **Automated Narrative Generation:** LLMs converting data insights into executive-ready narratives with contextual explanations
- **Proactive Insights:** AI-driven anomaly detection surfacing unexpected patterns before users ask
- **Augmented Analytics:** Gartner’s vision of AI-assisted insight generation becoming mainstream in enterprise BI platforms

7.2 Data Visualisation Best Practices

- **Perceptual Design:** Applying pre-attentive attributes (colour, size, position) to guide attention to key insights
- **Interactive Dashboards:** Drill-down, cross-filtering, and parameter controls enabling self-service exploration
- **Accessibility:** WCAG 2.1 AA compliance, colour-blind-safe palettes (viridis, cividis), and screen reader support
- **Storytelling with Data:** Narrative structure (context → conflict → resolution) applied to analytical presentations
- **Real-Time Dashboards:** WebSocket-powered live updates replacing static reports in operational contexts

7.3 Leading BI & Visualisation Platforms

Platform	Strength	AI Features (2025–26)
Power BI	Microsoft ecosystem, enterprise	Copilot NL queries, DAX suggestions
Tableau	Visual analytics, storytelling	Tableau Pulse, Einstein Discovery
Looker	Data modelling (LookML), GCP	Gemini integration, semantic layer
Apache Superset	Open-source, SQL-first	Community AI plugins
Metabase	Self-service, simplicity	Automated insights, NL queries
Grafana	Observability, time-series	AI anomaly detection, ML plugins

8 Responsible AI & Data Governance

8.1 The Governance Gap

Despite significant advances in AI capabilities, responsible AI governance remains the most critical unresolved challenge in data science. Research consistently identifies the absence of standardised, quantitative governance benchmarks as a major barrier to enterprise AI adoption [9] [4].



Governance Challenges 2025–2026

- **Algorithmic Bias:** No standardised cross-industry benchmark for measuring and comparing model fairness metrics
- **Explainability Standards:** Regulatory requirements (EU AI Act, GDPR Article 22) demand explanations, but implementation varies widely
- **Data Lineage:** End-to-end provenance tracking from raw data to model prediction remains technically challenging at scale
- **LLM Governance:** New challenges from LLM hallucination, prompt injection, and model memorisation of training data
- **Cross-Border Data:** Conflicting data residency requirements (GDPR, PDPA, CCPA) complicating global ML deployments

8.2 Interpretability Tools & Techniques

- **SHAP (SHapley Additive exPlanations):** Game-theoretic feature attribution providing both global and local explanations; recommended by Stoica & Stoica (2024) as the standard for AutoML operationalisation [8]
- **LIME (Local Interpretable Model-agnostic Explanations):** Perturbation-based local approximations; useful for image and text classification
- **Integrated Gradients:** Attribution method for deep learning models; required by Google’s Responsible AI practices
- **Counterfactual Explanations:** “What would need to change for a different outcome?” — increasingly required by EU AI Act
- **Model Cards:** Standardised documentation of model performance, limitations, and intended use cases (Google, Hugging Face)

8.3 Regulatory Landscape

- **EU AI Act (2024):** First comprehensive AI regulation, classifying systems by risk level; high-risk AI requires conformity assessments, human oversight, and transparency
- **GDPR Article 22:** Right to explanation for automated decisions; drives demand for

interpretable ML in EU markets

- **NIST AI RMF 1.0:** US framework for AI risk management; four functions (Govern, Map, Measure, Manage) increasingly adopted by US federal agencies
- **ISO/IEC 42001:2023:** First international standard for AI management systems; provides certification pathway for responsible AI

9 Strategic Roadmap for Data-Driven Organisations

9.1 Data Science Maturity Model

 **Five-Stage Data Science Maturity Model**

- Stage 1 — Descriptive:** Historical reporting, SQL-based dashboards, manual analysis. *Tools: Excel, basic BI*
- Stage 2 — Diagnostic:** Root cause analysis, data exploration, ad-hoc queries. *Tools: Tableau, Power BI, Python/R*
- Stage 3 — Predictive:** ML models for forecasting, classification, clustering. *Tools: scikit-learn, XGBoost, AutoML*
- Stage 4 — Prescriptive:** Optimisation, simulation, decision automation. *Tools: MLOps platforms, reinforcement learning*
- Stage 5 — Autonomous:** Self-learning systems, LLM-agent analytics, continuous adaptation. *Tools: DataLab-style platforms, LLM APIs*

9.2 Technology Convergence Map

Technology	navymid	
	Current (2025)	State Projected (2026+)
LLM-Powered BI	Early enterprise adoption	Mainstream, NL-first interfaces
AutoML	Mature for tabular data	Human-centred, multi-modal
Real-Time ML	Streaming + batch hybrid	Fully event-driven, self-adapting
Data Mesh	Pilot deployments	Enterprise-wide adoption
AI Governance	Fragmented standards	ISO 42001 compliance widespread
Quantum Data Science	Research phase	First production use cases

9.3 Recommended Action Framework



5-Step Data Science Readiness Framework

1. **Assess Data Infrastructure:** Audit current data quality, lineage, and accessibility. Implement a data catalogue (DataHub, Alation) as foundation
2. **Modernise Pipelines:** Migrate from legacy ETL to event-driven, cloud-native architectures using the Modern Data Stack
3. **Deploy AutoML:** Start with human-centred AutoML platforms (AutoGluon, H2O.ai) for high-impact use cases; integrate SHAP explanations from day one
4. **Pilot LLM Analytics:** Implement conversational BI for business users; start with NL2SQL on well-governed datasets
5. **Embed Governance:** Align with EU AI Act, NIST AI RMF, and ISO 42001; appoint dedicated AI governance roles; implement model cards and audit trails

References

References

- [1] Kadakia, J. J. (2025). *Demystifying Modern Data Engineering: From ETL to AI-Driven Pipelines*. Zenodo. <https://doi.org/10.5281/zenodo.16225825>
- [2] Anonymous. (2024). *Automated Feature Engineering for ETL in Machine Learning Pipelines*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5728983
- [3] Trehan, A. et al. (2024). *Data Engineering for Scalable Machine Learning: Designing Robust Pipelines*. International Journal of Computer Engineering and Technology, 15(6). https://doi.org/10.34218/ijcet_15_06_157
- [4] Lindauer, M., Karl, F., Klier, A. et al. (2024). *Position: A Call to Action for a Human-Centered AutoML Paradigm*. arXiv:2406.03348. <https://doi.org/10.48550/arxiv.2406.03348>
- [5] Karl, F., Thomas, J., Elstner, J. et al. (2024). *Automated Machine Learning*. In: AutoML Book. Springer. https://doi.org/10.1007/978-3-031-64832-8_1
- [6] Patel, K. (2024). *AutoML and Automated Data Science by Democratizing AI through End-to-End Automation*. International Journal for Science Technology and Engineering. <https://doi.org/10.22214/ijraset.2024.64555>
- [7] Anonymous. (2024). *Big Data Analytics with Machine Learning: Challenges, Innovations, and Applications*. Journal of Engineering and Computer Innovations Research. <https://jecir.com/index.php/jecir/article/view/20>
- [8] Stoica, F., & Stoica, L. F. (2024). *AutoML Insights: Gaining Confidence to Operationalize Predictive Models*. IntechOpen. <https://doi.org/10.5772/intechopen.1004861>
- [9] Cui, W., Zhang, G., Wang, B. et al. (2025). *A Review of Automated Data Science Based on Large Language Models*. IEEE ICNC-FSKD. <https://doi.org/10.1109/icnc-fskd67701.2025.11198037>
- [10] Weng, L., Tang, Y., Feng, Y. et al. (2024). *DataLab: A Unified Platform for LLM-Powered Business Intelligence*. arXiv:2412.02205. <https://doi.org/10.48550/arxiv.2412.02205>
- [11] Barbon Junior, S., Ceravolo, P., Groppe, S. et al. (2024). *Are Large Language Models the New Interface for Data Pipelines?* arXiv:2406.06596. <https://doi.org/10.48550/arxiv.2406.06596>
- [12] Wang, J. Y., Sukiennik, N., Li, T. et al. (2024). *A Survey on Human-Centric LLMs*. arXiv:2411.14491. <https://doi.org/10.48550/arxiv.2411.14491>
- [13] Kaushik, S., Dokka, K., Patel, D. et al. (2025). *Real-Time Analytics with Intelligent Data Pipelines and ML Detection*. IEEE ICSIT. <https://doi.org/10.1109/icsit65336.2025.11295098>

-
- [14] Thirunagalingam, A. (2024). *Transforming Real-Time Data Processing: The Impact of AutoML on Dynamic Data Pipelines*. <https://doi.org/10.69888/ftsint.2024.000213>
 - [15] Kazanskaia, A. N. (2025). *Future Trends in Big Data Analytics: Emerging Technologies, AI, and Sustainability*. <https://doi.org/10.64357/neya-gjnps-hr-bg-dt-10>
 - [16] Pathak, A. D. (2025). *The Transformative Impact of AI on Data Engineering, Data Science, and Business Intelligence*. World Journal of Advanced Engineering Technology and Sciences. <https://doi.org/10.30574/wjaets.2025.15.1.0365>

About the Author

Atine John Bosco is a Chartered Information Scientist (MBCS), Certified Information Systems Auditor (CISA), Google Cloud Digital Leader, ITIL 4 Practitioner, and PhD Candidate in Computer Science. With over 15 years in ICT leadership across Africa and international organisations, Atine specialises in digital transformation, data strategy, cybersecurity governance, and emerging technologies.



www.atine.info



[linkedin.com/in/atine](https://www.linkedin.com/in/atine)

Research Insights Series · Data Science & Analytics 2025–2026

72+ papers · 8 domains · Published May 2026

This document is freely available at www.atine.info