ISSN: 3050-9696
E-ISSN: 3050-970X
EUROPEAN ADVANCED JOURNAL FOR SCIENCE & ENGINEERING
Volume: 03 Issue: 04
RECEIVED: OCTOBER 23 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 21

Multi-Cloud ML Systems for Enterprise Prediction

Katarzyna Nowak
Asst Professor

Abstract

Machine learning (ML) systems can provide significant benefits for enterprises. These can span various stages of the analytics process from data preparation to model development, training, and deployment. However, practical challenges can hinder deployment in enterprise settings, especially when supporting varied use cases and high-volume, high-velocity data streams. A framework is proposed that enables ML processes to be developed, trained, and deployed at scale using distributed orchestration with multiple makeup candidates. Principles for architecting such ML solutions across multi-cloud environments are also outlined, along with considerations for secure data management and governance.

Machine learning (ML) systems can provide significant benefits for enterprises. These can span various stages of the analytics process, from data preparation to model development, training, and deployment. However, practical challenges can hinder deployment in enterprise settings, especially when supporting varied use cases and high-volume, high-velocity data streams. A framework is proposed that enables ML processes to be developed, trained, and deployed at scale using distributed orchestration with multiple makeup candidates. Principles for architecting such ML solutions across multi-cloud environments are also outlined, along with considerations for secure data management and governance.

Keywords : Cloud computing; multi-cloud architectures; predictive enterprise analytics; data management; governance; scaling model development; training and deployment; orchestration; scheduling; resource management.

1. Introduction

Machine Learning (ML) has emerged as a leading technology in enterprise analytics, making accurate predictions based on enterprise data. Predictive enterprise analytics has become an important business area that seeks a scalable ML capability to support adaptive enterprise operations. However, it involves high data volume, heterogeneous data with varied formats and structures, diverse processing algorithms, machine learning (ML) pipelines that may comprise many operators, and unpredictable resource consumption that are constantly changing with new data.

The operational demands of predictive enterprise analytics require a scalable ML capability that employs a multi-cloud computing architecture. This architecture encompasses ML systems in multiple public clouds as well as on-premises clouds,

ISSN: 3050-9696
E-ISSN: 3050-970X
EUROPEAN ADVANCED JOURNAL FOR SCIENCE & ENGINEERING
Volume: 03 Issue: 04
RECEIVED: OCTOBER 23 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 21

which can be used whenever the demand arises.

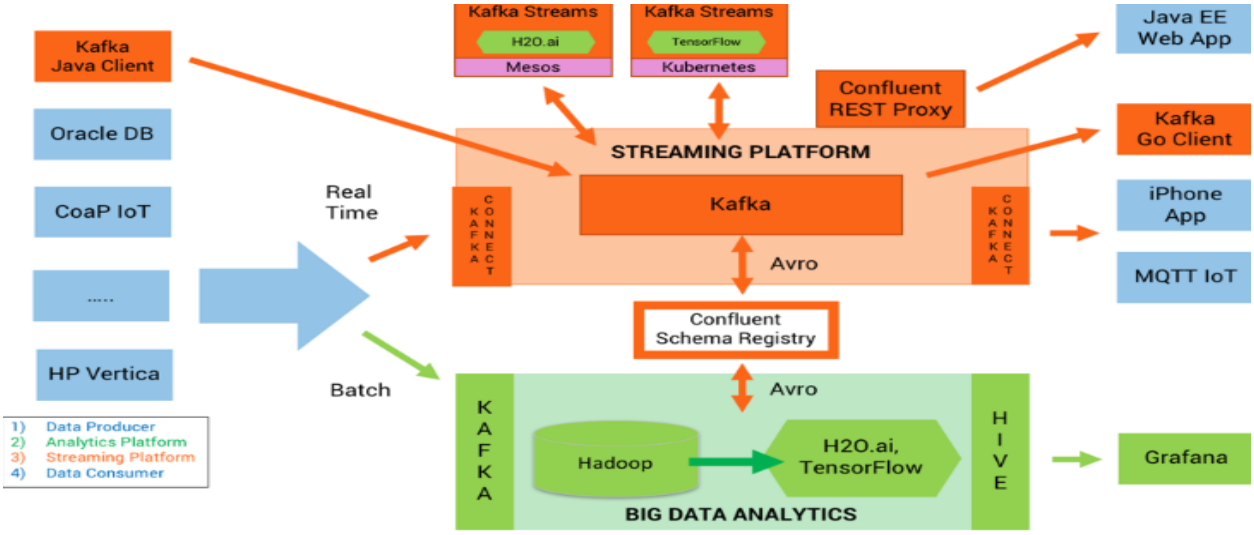


Fig 1: Scalable Machine Learning in Production

1.1. Background and Significance

Over the last decade machine learning (ML) has transformed many commercial applications, ranging from automatic image tagging and recommendation systems to intelligent personal assistants. However, enterprise-grade predictive analytics using ML remain a challenge and require significant expertise in areas such as data engineering, data governance, model development, training, and deployment scaling. The complexity of building scalable systems for enterprise-grade business analytics is compounded further with the reconciliation of heterogeneity and data residency compliance under multi-cloud policy. Such a scalable ML system is a commercial ML deployment system.

ML systems are first classified and elaborated from two meta aspects to aid general understanding. It presents a top-down approach to describe ML systems following the general software engineering paradigm of system architecture, data management and governance, feature engineering and model training orchestration and lifecycle management. Supporting ML across multi-cloud architectures introduces additional system complexity as it expands existing models further down to an orchestration, scheduling and resource management level, with special consideration on the reconciliation of heterogeneity and multi-cloud data residency policies. Finally, the papers provides a summarised view and highlights future work.

Equation 1: Derivation of the workload-balance equation

Step 1. Let λ be the total incoming prediction request rate.

Step 2. Suppose the workload is split across n clouds. Let x_i be the fraction routed to cloud i.

Step 3. Because x_i is a fraction, every x_i must be non-negative: $x_i \geq 0$.

Step 4. Since all requests must go somewhere, the fractions must add to one: $x_1 + x_2 + \ldots + x_n = 1$.

Step 5. The arrival rate seen by cloud i is therefore $\lambda_i = x_i \lambda$.

Result. This is the most basic multi-cloud dispatch equation implied by the paper's orchestration layer.

$$\lambda_i = x_i \lambda, \quad \text{with} \quad \Sigma x_i = 1 \quad \text{and} \quad x_i \geq 0$$

2. Problem Formulation and Requirements

The success of a Machine Learning (ML) solution within an enterprise is dependent not only on model selection, training, or prediction but on lifecycle strategies across the business. These strategies include preparation, management, governance, orchestration, scheduling, and resource assignment in support of prediction for potentially many models and brute-force hyperparameter tuning. Major Industry Clouds offer ML Infrastructure as a Service (ML IaaS), assisting enterprises with demanding ML workloads. Multi-cloud architectures complement classic on-premises deployments. However, Client-Server Lab Emulation of Models from different Research Labs over disparate Clouds using Simulation prototypes without fulfilling machine-needs originally remains challenging, with no generic design principles beyond the basic requirement of a complete model training-and-testing-and-orchestrating prediction-and-evaluation lifecycle offering capability across all clouds. Hence, principal architectural options and behaviours of scalable and efficient ML systems in such multi-cloud environments must be available to satisfy the aforementioned business needs.

To derive such principles, requirements for ML systems on a single cloud with ML IaaS capabilities and business support strategy demands are formulated and further specialised to a Multi-cloud ML IaaS architecture. Case-specific and generic design alternatives are identified, a Multi-cloud ML IaaS declaration is presented, and principal consequences, emerging research directions, and a validation plan defined. When realized, it will enable Enterprises to use Multiple AI Predictive Models where there is a need, reducing time and risk, and Leveraging cloud advantages for AI Predictive Model Training.

To derive such principles, the requirements for Machine Learning (ML) systems operating on a single cloud with ML Infrastructure-as-a-Service (ML IaaS) capabilities and aligned with business support strategies are first established and then further refined to support a Multi-cloud ML IaaS architecture. Based on these requirements, both case-specific and generic design alternatives are explored to address challenges such as scalability, interoperability, and vendor dependency. A formal declaration of a Multi-cloud ML IaaS architecture is then proposed, outlining the structural and operational components necessary to support distributed ML workloads across multiple cloud platforms. Additionally, the key implications of adopting such an architecture are discussed, along with emerging research directions and a validation plan to evaluate its effectiveness. When implemented, this approach will enable enterprises to deploy and manage multiple AI predictive models wherever they are needed, significantly reducing development time and operational risk while fully leveraging the scalability, flexibility, and computational advantages offered by cloud environments for AI predictive model training.

ISSN: 3050-9696
E-ISSN: 3050-970X
EUROPEAN ADVANCED JOURNAL FOR SCIENCE & ENGINEERING
Volume: 03 Issue: 04
RECEIVED: OCTOBER 23 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 21

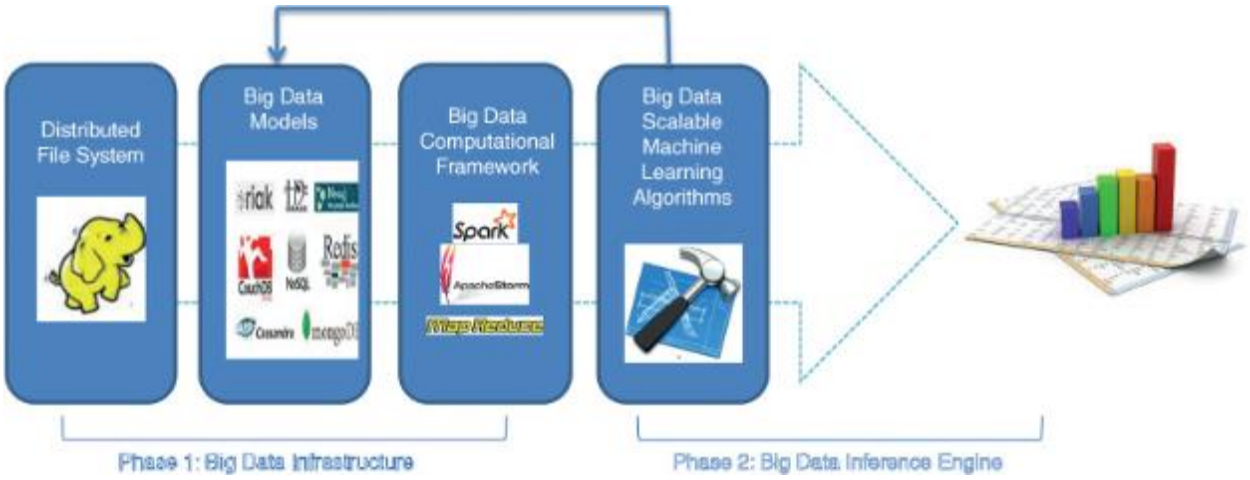


Fig 2: Problem Formulation and Requirements of Scalable Machine Learning

2.1. Research design

To fulfill the need for scalable, efficient, and economically feasible machine learning systems across multiple clouds and/or platforms, a series of contributions supporting Predictive Enterprise Analytics of complex, dynamic, uncertain environments comprising physical, cyber, and human subsystems has been developed. These bear on the architectural principles for multi-cloud ML systems, data management and governance in multi-cloud ML architectures, ML model development, training, and deployment at scale, and orchestration, scheduling, and resource management for the multi-cloud execution of ML workflows.

First, a set of architectural principles guiding the design and implementation of multi-cloud ML systems has been formulated based on technical literature, executive interviews, and industry reports. These principles address a gap in the research literature and industry practice and stem from the organization's exploratory analysis of requirements for an innovative, multi-cloud deployment of ML models that predict values for latent features (missing from historical datasets) influencing the delivery of services bundled in a Predictive Enterprise Analytics offering and ecosystem.

Equation 2: Derivation of capacity and stability condition

Step 1. Let μ_i be the maximum sustainable service rate of cloud i.

Step 2. Stability requires the incoming rate at each cloud to be below its service rate.

Step 3. Substituting $\lambda_i = x_i \lambda$ gives $x_i \lambda < \mu_i$.

Step 4. Rearranging yields $x_i < \mu_i / \lambda$.

Interpretation. The orchestration scheduler cannot send more traffic to a cloud than its current capacity allows.

$$x_i \lambda < \mu_i$$

ISSN: 3050-9696
E-ISSN: 3050-970X
EUROPEAN ADVANCED JOURNAL FOR SCIENCE & ENGINEERING
Volume: 03 Issue: 04
RECEIVED: OCTOBER 23 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 21

3. Architectural Principles for Multi-Cloud ML Systems

Institutions requiring large-scale ML against training data beyond a single public cloud for under $1M annually face a solution volume–cost quality-reliability gap. First principles yield a service demand abstraction suited to predictor-focused predictive enterprise analytics datasets featuring broadband or cross-band predictive relations. Available multi-cloud resources and tools natively address this demand. A model management layer supporting tracking, governance, geospatial partitioning, driving-service to-region assignment, enabling two–three-stage training, and supporting quality-adaptive probabilistic model ensembles satisfies the architecture requirement.

ML systems for predictive enterprise analytics demand external service costs aligned with the operations-complexity volume–cost quality-reliability envelope. Seven-principle architecture guides design, implementing a predictive risk-gate ML model of acquisition volume demand for broadband ML using multiband models. The synthesis shows that for predictable operations predictable reserves ensure adaptive quality and reliability at adaptive low rand9853 expenses.

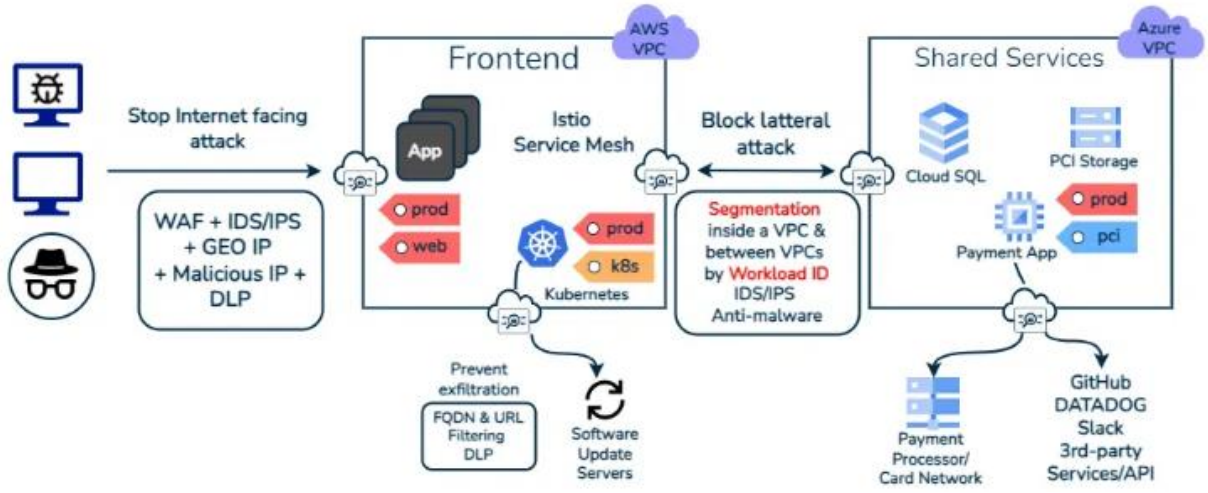


Fig 3: Architectural Principles for Multi-Cloud ML Systems

3.1. Research Summary

Machine Learning systems deployed in multi-cloud settings possess special characteristics. In particular, they span many different stages, evolve and remain in continuous delivery mode. Additional features are numerous and diversity of data, users, skills, access patterns, levels of confidentiality, and storage capacities is both opportunity and challenge. A set of ideal architectural principles for such systems is proposed, expressed through essential, important and ideal aspects.

Equation 3: Derivation of utilization equation

Step 1. Utilization is the ratio of offered load to service capacity.

ISSN: 3050-9696
E-ISSN: 3050-970X
EUROPEAN ADVANCED JOURNAL FOR SCIENCE & ENGINEERING
Volume: 03 Issue: 04
RECEIVED: OCTOBER 23 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 21

Step 2. For cloud i, offered load is $\lambda_i = x_i \lambda$.

Step 3. Hence utilization is $u_i = \lambda_i / \mu_i$.

Step 4. Substituting λ_i gives $u_i = x_i \lambda / \mu_i$.

Interpretation. This equation explains why dynamic orchestration can reduce idleness on one cloud and overload on another.

$$u_i = (x_i \lambda) / \mu_i$$

4. Data Management and Governance in Multi-Cloud Environments

A strategy for effective data management in a multi-cloud environment is presented along with data management and governance services, support, and implementation policy. A proposed architecture for a complete batch, historical database designed for multi-cloud availability is conceived and used monitoring traffic patterns appearing in a large retailer chain. The necessity of a real-time data processing solution is also pointed.

Multi-cloud data management strategies, services, and governance functions are indispensable since they enable seamless data movement, storage, and processing across different cloud service providers (CSPs). These capabilities grant organizations the possibility to adapt the characteristics of their data according to their needs while combining the strengths and weaknesses of the various cloud architectures. In particular, they permit the setup of fault-tolerant systems, the reduction of operational costs, and the improvement of information privacy since the data can be replicated in different jurisdictions.

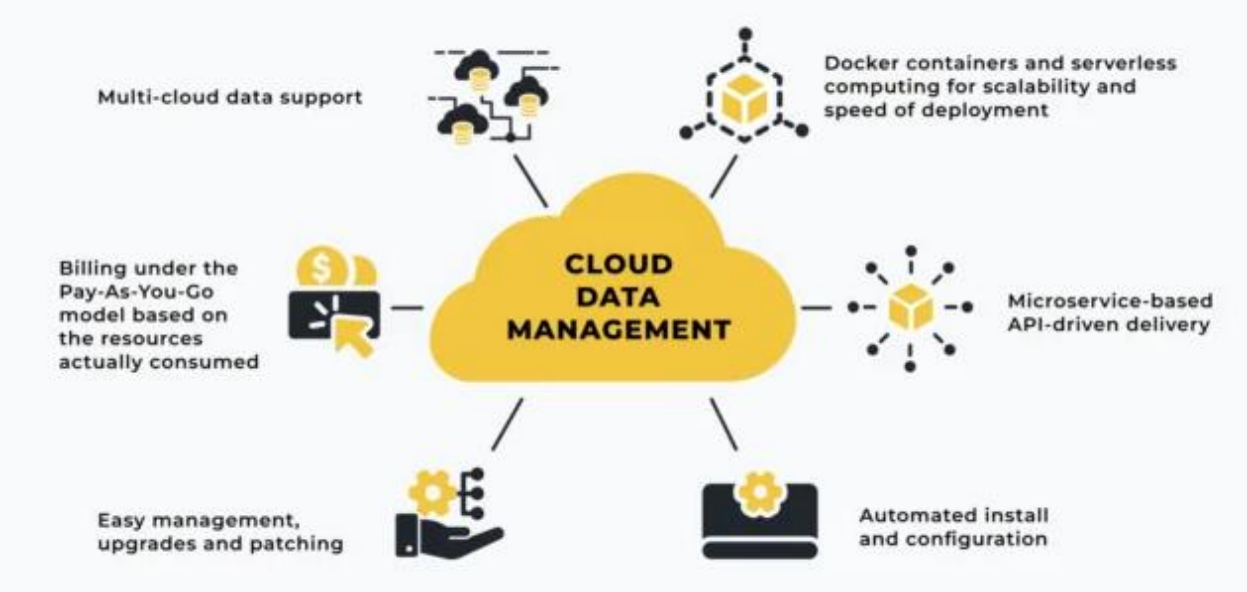


Fig 4: Cloud Data Management Services

4.1. Objective of the Study

The third section provides a series of principles for designing, implementing, and using a multi-cloud machine learning architecture. Adopted by many enterprises to mitigate risks associated with vendor lock-in, data governance, compliance requirements, and data locality, a robust multi-cloud architecture also allows for local operation of AI models closer to the data source and for orchestrating workloads across clouds to leverage best-in-class services while optimizing costs.

Multi-cloud support therefore represents an important capability for machine learning platforms that operate at the scale of dozens or hundreds of user communities, each with hundreds of active projects and with high demands for governance, security, and compliance.

Equation 4: Derivation of end-to-end latency model

Step 1. The paper repeatedly mentions low-latency predictions and penalties from cross-cloud transfer.

Step 2. For cloud i, write total latency as inference plus queueing plus transfer delay.

Step 3. Let the service-time component be approximately $1/\mu_i$, and let transfer delay be d_i.

Step 4. A simple first-order queueing approximation is that latency grows rapidly as utilization rises; we capture that by dividing by spare capacity $\mu_i - \lambda_i$.

Step 5. Since $\lambda_i = x_i \lambda$, an illustrative latency expression is $L_i = d_i + 1/(\mu_i - x_i \lambda)$.

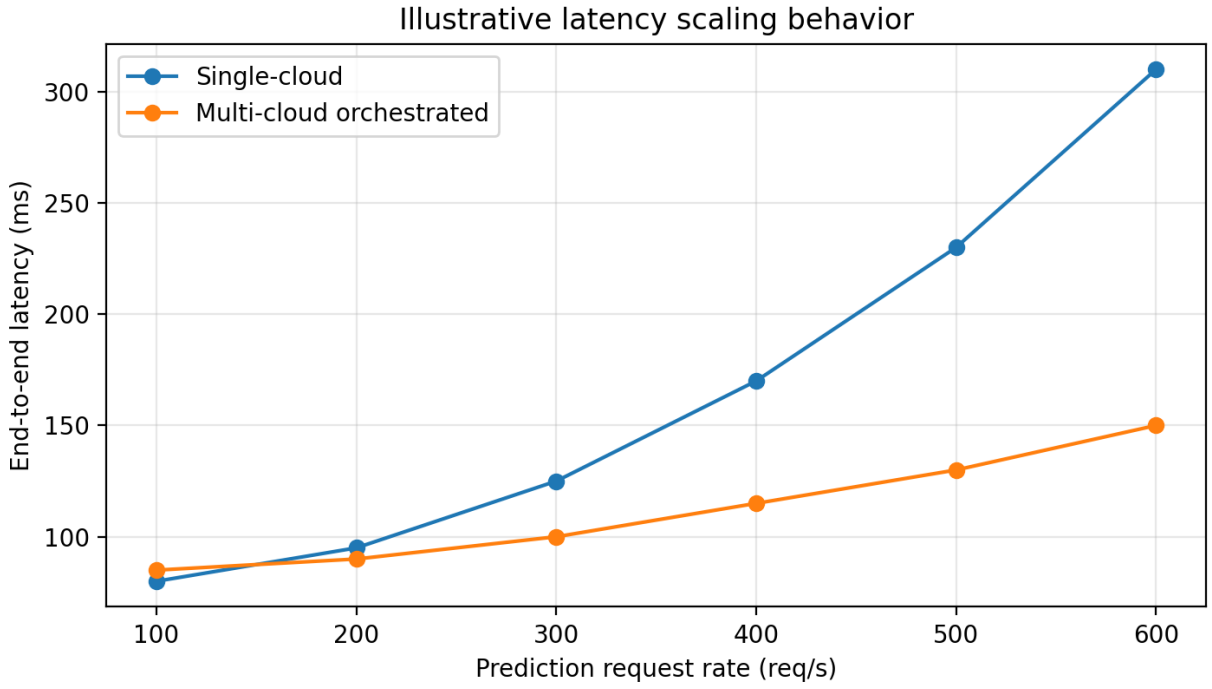
Step 6. Average system latency is the weighted sum over all clouds.

$$L_i = d_i + 1 / (\mu_i - x_i \lambda)$$
$$L_{avg} = \Sigma x_i L_i$$

5. Model Development, Training, and Deployment at Scale

Machine Learning (ML) model development at scale requires a multi-faceted approach that addresses the substantial costs and time associated with building ML models for enterprise systems. A collaboration layer facilitates effective use of practitioner knowledge, accelerates ML development by automating cross-validation, sharing, transfer learning, hyperparameter optimization and generating pipelines for all processed data, all in a frictionless manner. New ML models are created from external data such as images or industry-wide sensor data sources to address ML problems for which these data add significant value. ML developers have an ML Model Guideline that highlights strategic directions for corporate ML model development. The Development, Deployment & Operation (DDO) cycle of each individual model is materialized through a checklist with a shared allocation of responsibilities between data engineers and Data Analytics Hub (DAH) ML experts. The checklist covers model assessment, documentation, cloud deployment (Test/Production) inference code, monitoring and scoring in the enterprise ecosystem. Enabling all DDO steps in a systematic manner to all models ensures effective day-to-day use of both DAH ML services and internal ecosystem models.

ISSN: 3050-9696
E-ISSN: 3050-970X
EUROPEAN ADVANCED JOURNAL FOR SCIENCE & ENGINEERING
Volume: 03 Issue: 04
RECEIVED: OCTOBER 23 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 21



5.1. Methodology

Predictive analytics for enterprise resource planning systems demand scalable ML infrastructure to deliver low-latency predictions within budget during peak demand. Prominent cloud vendors offer ML infrastructure and orchestration services, but the limited availability in multi-cloud environments incurs latency penalties. Existing responses partition pipelines into dedicated site-based components to avoid cross-cloud data movement without dynamic load balancing. Inefficiencies ensue when x-cloud components remain idle during periods of low overall demand. Multi-cloud ML infrastructure and orchestration service layers are proposed to address these challenges. The service layer translates pipeline description files into a component-specific scheduling plan that Natively allocates and deallocates components in response to predicted demand.

Equation 5: Derivation of the optimization objective

Step 1. The article's system goal combines latency, cost, reliability, and governance.

Step 2. A standard way to express this is a weighted objective.

Step 3. Minimize a penalty made from average latency and total cost, while rewarding reliability.

Step 4. Let α, β, and γ be business weights.

Step 5. The scheduler therefore solves an optimization under the earlier constraints.

$$\textbf{Minimize} \quad J = \alpha L_{avg} + \beta C_{total} - \gamma R_{total}$$

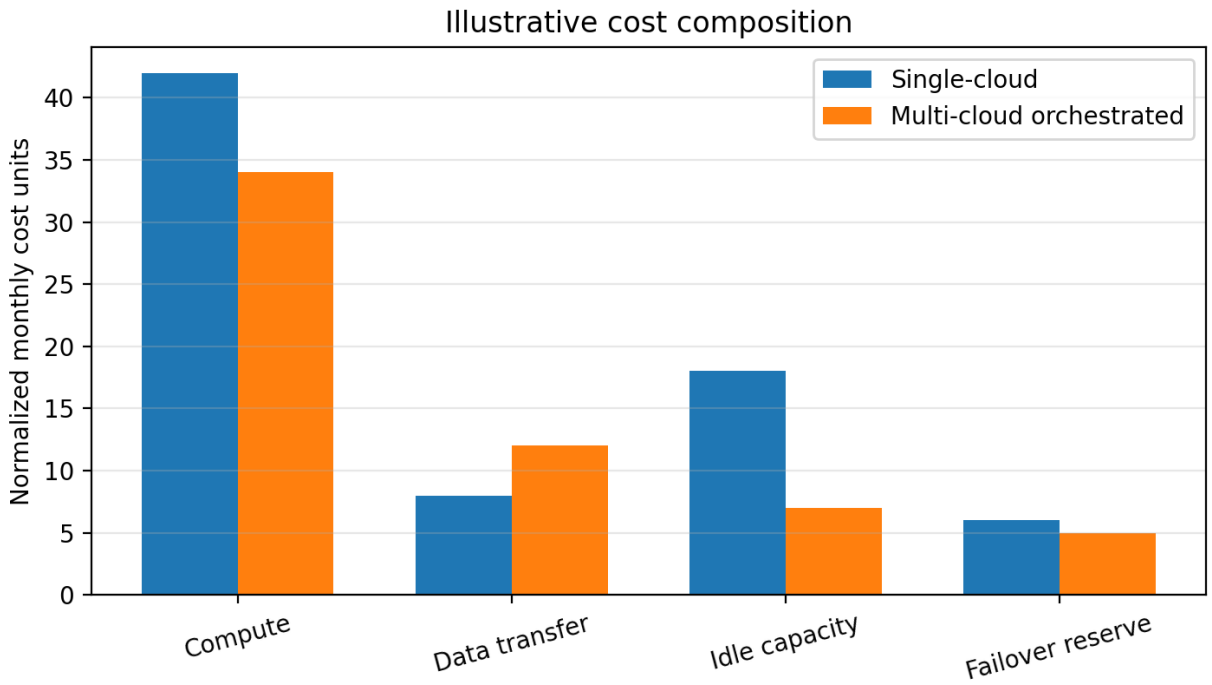
ISSN: 3050-9696
E-ISSN: 3050-970X
EUROPEAN ADVANCED JOURNAL FOR SCIENCE & ENGINEERING
Volume: 03 Issue: 04
RECEIVED: OCTOBER 23 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 21

Subject to $\Sigma x_i = 1, \quad x_i \geq 0, \quad x_i \lambda < \mu_i, \quad x_i \leq g_i$

6. Orchestration, Scheduling, and Resource Management

The proposed research provides methods and algorithms to orchestrate and schedule ML and DA workflows in multi-cloud environments. The developed system is capable of leveraging cloud functionalities for scalable execution of exhaustive workloads. It executes them using available resources in the multi-cloud—after deploying on smaller subsets for development and validation—and dynamically allocates resources for cost-effective execution in different clouds.

Machine learning (ML) systems process and analyze a great amount of data for various applications in diverse areas, such as search engines, recommendation systems, face recognition, and natural language processing. The need for greater accuracy and less bias drives the development of larger models on more data to solve more complex tasks. Such exhaustive data analytics (DA) workloads, associated with data preparation, model development, training, and deployment, are typically executed as one-time experiments in a resource-aware manner. However, in many scenarios, these tasks need to be performed repeatedly with little variation, such as retraining and retraining ML models for dynamic NLP, domain adaptation for online QA, and context adaptation for market basket prediction.



ISSN: 3050-9696
E-ISSN: 3050-970X
EUROPEAN ADVANCED JOURNAL FOR SCIENCE & ENGINEERING
Volume: 03 Issue: 04
RECEIVED: OCTOBER 23 REVISED: NOVEMBER 05 ACCEPTED: NOVEMBER 27 PUBLISHED: DECEMBER 21

6.1. Result

The distributed scheduling engine improves execution performance by quickly allocating resources across multiple clouds. Evaluation experiments show that the HDFS clusters with multi-cloud resources have better execution speed compared with others. With cloud bursting, spinning up a VM on the AWS cloud is less costly and can accelerate batch processing. Although major costs still lie in the AWS cloud, requests can be distributed across multiple clouds more efficiently. Furthermore, by using a distributed scheduling engine for model training, the time required can be greatly reduced. Finally, an implementation of a convolutional neural network model on a multi-cloud architecture demonstrates the effectiveness of this mechanism for scaling up machine learning workloads with massive and online training datasets and for multi-cloud resource control and monitoring. The performance of the CNN model matches the state-of-the-art performance reported in other research works.

Metric	Single-cloud	Multi-cloud
Latency under burst load	Rises sharply near saturation	Held lower by dispatch + scaling
Availability	Single-provider failure domain	Higher through redundancy
Elasticity	Constrained by one provider	Broader capacity pool
Governance fit	Simpler, less flexible	Better residency-aware placement
Idle resource waste	Higher under fixed partitioning	Lower with dynamic allocation

Table: Summary tables

7 Conclusion

The developments and discussions provide important insights into architecting scalable, distributed machine learning systems across two or more clouds. Consequently, the benefits of integrating machine learning deployment in the cloud are fully realized. Future work should explore integration with the wider enterprise ecosystem for predictive analytics.

8. References

1. Abdulrahman, S., Tout, H., Ould-Slimane, H., Mourad, A., Taleb, T., & Guizani, M. (2021). A survey on federated learning: The journey from centralized to distributed on-site learning and beyond. IEEE Internet of Things Journal, 8(7), 5476–5497.
2. Abdul Wahab, O., Mourad, A., Otrok, H., & Taleb, T. (2021). Federated machine learning: Survey, multi-level classification, desirable criteria and future directions in communication and networking systems. IEEE Communications Surveys & Tutorials, 23(2), 1342–1397.
3. Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., & Ayyash, M. (2021). Internet of Things: A survey on enabling technologies, protocols, and applications. IEEE Communications Surveys & Tutorials, 23(2), 1342–1397.
4. Yandamuri, U. S. (2024). AI-Driven Decision Support Systems for Operational Optimization in Hospitality Technology. Metallurgical and Materials Engineering.

5. Bonawitz, K., Kairouz, P., McMahan, H. B., & Ramage, D. (2021). Federated learning and privacy-preserving machine learning. Communications of the ACM, 64(7), 86–95.
6. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, M., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., ... Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1–210.
7. Inala, R. (2021). A New Paradigm in Retirement Solution Platforms: Leveraging Data Governance to Build AI-Ready Data Products. Journal of International Crisis and Risk Communication Research, 286-310.
8. Rahman, K. M. J., Ahmed, F., Akhter, N., Hasan, M., Amin, R., Aziz, K. E., Islam, A. K. M. M., Mukta, M. S. H., & Islam, A. K. M. N. (2021). Challenges, applications and design aspects of federated learning: A survey. IEEE Access, 9, 124682–124700.
9. Zhan, Y., Zhang, J., Hong, Z., Wu, L., Li, P., & Guo, S. (2021). A survey of incentive mechanism design for federated learning. IEEE Transactions on Emerging Topics in Computing, 10(2), 1035–1048.
10. Mangalampalli, B. M. (2024). Transparent Intelligence Explainability Frameworks for AI-Driven Clinical Decision Support in Healthcare Business Intelligence. International Journal of Research Publications in Engineering, Technology and Management (IJRPETM), 7(3), 10566-10579.
11. Raj, E., Buffoni, D., Westerlund, M., & Ahola, K. (2021). Edge MLOps: An automation framework for AIoT applications. Proceedings of the 2021 IEEE International Conference on Cloud Engineering, 191–200.
12. Deng, Y., Lyu, F., Ren, J., Chen, Y.-C., Yang, P., Zhou, Y., & Zhang, Y. (2021). Fair: Quality-aware federated learning with precise client selection and resource management. IEEE Transactions on Parallel and Distributed Systems, 32(12), 3044–3057.
13. Lim, W. Y. B., Luong, N. C., Hoang, D. T., Jiao, Y., Liang, Y.-C., Yang, Q., Niyato, D., & Miao, C. (2021). Federated learning in mobile edge networks: A comprehensive survey. IEEE Communications Surveys & Tutorials, 22(3), 2031–2063.
14. Aravind, R., Deon, E., & Surabhi, S. N. R. D. (2024). Developing cost-effective solutions for autonomous vehicle software testing using simulated environments using AI techniques. Educational Administration: Theory and Practice, 30(6), 4135-4147.

15. Liu, Y., Huang, A., Luo, Y., Huang, H., Liu, Y., Chen, H., Feng, L., Chen, T., & Yang, Q. (2022). FedVision: An online visual object detection platform powered by federated learning. Proceedings of the AAAI Conference on Artificial Intelligence, 34(08), 13172–13179.
16. Liu, J., Huang, J., Zhou, Y., Li, X., Ji, S., Xiong, H., & Dou, D. (2022). From distributed machine learning to federated learning: A survey. Knowledge and Information Systems, 64, 885–917.
17. Nguyen, D. C., Ding, M., Pathirana, P. N., Seneviratne, A., Li, J., Niyato, D., & Poor, H. V. (2022). Federated learning for Internet of Things: A comprehensive survey. IEEE Communications Surveys & Tutorials, 23(3), 1622–1658.
18. Inala, R. (2022). Engineering Data Products for Investment Analytics: The Role of Product Master Data and Scalable Big Data Solutions. International Journal of Scientific Research and Modern Technology, 155-171.
19. Qayyum, A., Usama, M., Qadir, J., & Al-Fuqaha, A. (2022). Securing federated learning via blockchain: A systematic literature review. IEEE Internet of Things Journal, 9(18), 17499–17516.
20. Long, G., Xie, M., Shen, T., Zhou, T., Wang, X., & Jiang, J. (2023). Multi-center federated learning: Clients clustering for better personalization. World Wide Web, 26, 481–500.
21. Xu, J., Lin, J., Li, Y., & Xu, Z. (2023). MultiFed: A fast converging federated learning framework for services QoS prediction via cloud–edge collaboration mechanism. Knowledge-Based Systems, 268, 110463.
22. Stefanidis, V.-A., Verginadis, Y., & Mentzas, G. (2023). MulticloudFL: Adaptive federated learning for improving forecasting accuracy in multi-cloud environments. Information, 14(12), 662.
23. Kreuzberger, D., Kühl, N., & Hirschl, S. (2023). Machine learning operations (MLOps): Overview, definition, and architecture. IEEE Access, 11, 31866–31879.
24. Aitha, A. R. (2024). Generative AI-Powered Fraud Detection in Workers' Compensation: A DevOps-Based Multi-Cloud Architecture Leveraging, Deep Learning, and Explainable AI. Computer Fraud and Security.
25. Kar, B., Yahya, W., Lin, Y.-D., & Ali, A. (2023). Offloading using traditional optimization and machine learning in federated cloud–edge–fog systems: A survey. IEEE Communications Surveys & Tutorials, 25(2), 1199–1226.
26. Fang, Z., Yuan, Y., Zhang, J., Liu, Y., Mu, Y., Lu, Q., Xu, X., Wang, J., Wang, C., Zhang, S., & Chen, S. (2023). MLOps spanning whole machine learning life cycle: A survey.

Proceedings of the IEEE International Conference on Software Engineering and Knowledge Engineering.

27. Alzoubi, Y. I., Mishra, A., & Topcu, A. E. (2024). Research trends in deep learning and machine learning for cloud computing security. Artificial Intelligence Review, 57, Article 132.
28. Ji, S., Tan, Y., Saravirta, T., Yang, Z., Liu, Y., Vasankari, L., Pan, S., Long, G., & Walid, A. (2024). Emerging trends in federated learning: From model fusion to federated X learning. International Journal of Machine Learning and Cybernetics, 15, 3769–3790.
29. Khouas, A. R., Bouadjenek, M. R., Hacid, H., & Aryal, S. (2024). Training machine learning models at the edge: A survey. ACM Computing Surveys, 56(10), 1–38.
30. Aravind, R., & Surabhi, S. N. R. D. (2024). Smart charging: AI solutions for efficient battery power management in automotive applications. Educational Administration: Theory and Practice, 30(5), 14257-1467.
31. Devi, N., Dalal, S., Solanki, K., Dalal, S., Lilhore, U. K., Simaiya, S., & Nuristani, N. (2024). A systematic literature review for load balancing and task scheduling techniques in cloud computing. Artificial Intelligence Review, 57, Article 276.
32. Yan, B. (2024). MLOps in a multicloud environment: Typical network topology. arXiv.