



Cloud-Native AI Systems for Industry-Specific Decision Optimization

Katarzyna Nowak
Asst Professor

Abstract

Scalable cloud-based intelligent decision systems are investigated for industry-specific problems in different sectors that can be optimized by AI-assisted big-data analytics. Despite the growing popularity of cloud-based AI and big-data-driven decision systems, evolution in the required cloud infrastructure and the actual data pipelines have not yet been undertaken in detail. In addition, the current research has not gone into sufficient detail on cloud-based systems for specific high-scaling industry problems such as for the healthcare and life sciences sector or those supported by the manufacturing or supply chain domain. Cloud decision systems also have specific requirements for management and governance due to the nature of data belonging to different parties, and these aspects require further consideration. Consequently, the focus is on cloud systems for scalable manufacturing or supply chain operation decision management, with practical industry examples clearly described for evaluation and cross-validation of accuracies.

The study serves teleological domain-focused decision-systems deployment in generic governance-integrity-demand-supporting environments. For exemplary industry-use domain-focused problems, demand forecasting, inventory/supply, production scheduling, logistics, or resiliency-enhancement of supply chains, patient-flow management, clinical-decision-support allocation of resources in healthcare, and integration of multiple-organization datasets for results comparison and research in life-sciences decision support are highlighted. The encoding and timing aspects of teleological risks are also emphasized at the decision-system level, while general privacy and security requirements specify the data-handling and incident-reaction measures needed for protection against unauthorized disclosure or service unavailability.

Keywords: Cloud-Based Intelligent Decision Systems, AI-Assisted Big-Data Analytics, Scalable Cloud Architectures, Industry-Specific Decision Support, Manufacturing and Supply Chain Optimization, Demand Forecasting Systems, Inventory and Production Scheduling, Logistics and Supply Chain Resilience, Cloud-Native Data Pipelines, Governance of Cloud Decision Platforms, Teleological Decision-System Design, Healthcare Resource Allocation, Clinical Decision Support Systems, Life Sciences Data Integration, Cross-Organizational Analytics, Privacy and Security in Cloud AI, Risk-Aware Decision Management, High-Scaling Industry Applications, Cloud Governance and Integrity Frameworks, Intelligent Enterprise Operations.

1. Introduction

Advances in cloud computing and the results of progress in artificial intelligence (AI) and big data technology have led to promises of intelligent decision support systems able to aid human decision making in business and other domains. These developments cater for the deployment of intelligent

decision support systems as a cloud-based service, enabling outsourcing of the underlying big data storage and processing, creating an environment for analyst users to devise intelligent residual models, as well as alleviating the problems relating to data management, privacy security and service level agreement (SLA) concerns. However, while a



generalised cloud-enabled intelligent decision support system exists, these systems have not yet provided decision systems for industry-specific intelligent optimisation tasks. Such optimisation tasks are prevalent in the manufacturing and supply-chain domain (Datta et al. 2022).

The following sections present proposed industry-specific cloud-based scalable intelligent decision support systems designed to fulfil such objectives across a number of important domains. The results can be used by domain experts to identify suitable machine learning predictive models that enable these specific tasks to be solved accurately, efficiently and at a reasonable cost. The aim is to contribute to the design of cloud-based intelligent decision support systems specifically for the manufacturing, healthcare or life-science sectors, capable of optimising domain-relevant decision surrogates that act as input to optimising models of longer-term investment, expenses, performance or risk for a given organisation (Wang et al. 2022). The sub-objectives are to design a manufacturing and supply-chain system that addresses intelligent modelling of demand forecasting, inventory management, production scheduling, logistics network design and resilience optimisation, a healthcare-focused system that provides patient flow predictions, clinical-decision-support system recommendations, resource-allocation strategies and integration support for research-data-collection and analysis consolidation, and a life-sciences system that assists in the investigation and mining of research data.

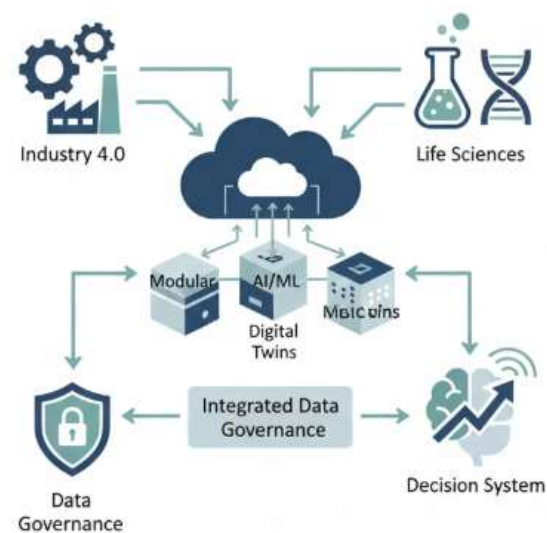


Fig 1: Modular Intelligence: A Governance-First Framework for Cloud-Based Optimization and Decision Support in Manufacturing and Life Sciences

1.1. Purpose and Scope of the Study

Delivering intelligent decision systems specifically aimed at solving industry-relevant optimization problems. Industry 4.0 and the COVID-19 pandemic have created an unprecedented demand for accurate forecasting and reliable decision systems in the manufacturing sector and the supply chain supporting it. The importance of effective decision-making, especially in optimizing scarce resources while satisfying demand during surge periods, has stimulated research in the field. Cloud-based AI and big data-enabled intelligent decision systems are being developed to address optimization problems in the manufacturing and life sciences sectors.

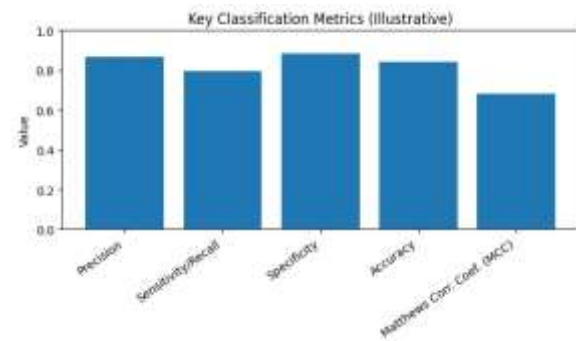
Past studies have concentrated on specific optimization problems within these sectors using advanced methods. For instance, digital twins or meta-learning techniques have been developed for optimizing semiconductor manufacturing equipment, while machine learning approaches have been adopted for demand forecasting along the supply chain. While cross-domain cloud-based infrastructure for data-driven decision support exists, there



tends to be a lack of focus on the governance aspects required for cloud-enabled data management of compliance-critical applications. As a result, the proposed infrastructure covers all aspects required for building scalable intelligent decision systems within cloud environments and attempts to combine the advanced methods available for individual problems in a modular fashion to meet the corresponding decision-making need.

2. Background and Motivation

The continuous evolution of cloud computing and the emergence of demand-side applications have made cloud services a promising way of achieving ubiquitous availability of AI and Big Data-enabled intelligent decision-making. Industry, however, has not yet succeeded in fully realizing the potential of these services, in spite of the high investments made. A detailed examination of past developments shows how gradually-evolving, cloud-based, AI- and Data-enabled intelligent decision support are becoming possible. Simultaneously, it also highlights the crucial gaps that still need to be filled for industry-specific optimization in the manufacturing and life sciences sectors. Cloud architectures have developed considerably over the past two decades, and particularly the last five years have seen significant advances in deep learning, computer vision, natural language processing, reinforcement learning, and data analytics in general. Yet, these advances, together with the expansion of cloud service markets and demand-side smart applications, have added little to the portfolio of available AI and Big Data-enabled decision support in the private sector. A major shortcoming has been the persistent isolation of industry-specific optimization problems, the lack of general-purpose intelligent support and the resulting difficulties in making intelligent decisions in a cost-effective and timely manner. These issues underline the pressing need for the development of scalable cloud-based Intelligent Decision Systems capable of addressing optimization problems in selected domains.



Equation A) Step-by-step derivation of the evaluation-metric equations

A1) Start from the confusion matrix

For a binary classifier:

- **TP:** true positives
- **FP:** false positives
- **TN:** true negatives
- **FN:** false negatives

Total samples:

$$N = TP + FP + TN + FN$$

A2) Precision

Definition: “Of everything predicted positive, how many were actually positive?”

$$\text{Predicted positives} = TP + FP$$

So,

$$\text{Precision} = \frac{TP}{TP + FP}$$



A3) Sensitivity / Recall (True Positive Rate)

Definition: “Of everything actually positive, how many were caught?”

Actual positives = $TP + FN$

So,

$$\text{Sensitivity} = \text{Recall} = \frac{TP}{TP + FN}$$

A4) Specificity (True Negative Rate)

Definition: “Of everything actually negative, how many were correctly rejected?”

Actual negatives = $TN + FP$

So,

$$\text{Specificity} = \frac{TN}{TN + FP}$$

2.1. Historical Context and Rationale

Decision systems enable organizations to tackle complex decision problems involving many variables. Over the years, such systems for a variety of domains have emerged and evolved, with many systems now publicly available in the cloud, integrating capabilities from Artificial Intelligence (AI), Big Data/Data Management, and Cloud Computing. The growing interest in AI and Big Data Technologies, particularly in the last 10 years, has led to the emergence of many academic and industry projects and products. Nevertheless, no cloud-based, scalable, intelligent decision system pervading and supporting a comprehensive set of decision problems across an industry vertical and decades of research effort exists. Consequently, to optimize the architecture of the cloud infrastructure as well as the entire solution for a specific industry and decision problem, the architecture and components of a Scalable Cloud-based Intelligent Decision System are described. The proposed architecture enables integration of emerging AI learning methods and Big Data Technologies with decision systems, extending their applications to problems involving massive

volumes of data and providing the decision-maker a decisive edge.

The evolution of Cloud Computing architectures, AI-based learning and Big Data-enabled Decision Systems is presented, followed by an analysis of the limitations associated with existing research and practice. Industry-specific Decision Problems for which the architecture is being designed are then identified, together with the Optimization Criteria for the Decision Problems in the Manufacturing and Supply Chain, as well as Healthcare and Life Sciences domains. Finally, the need for Scalable Cloud-enabled AI and Big Data-based Intelligent Decision Systems within industry-specific contexts is established.

3. Architectural Framework for Scalable Cloud-Based Intelligent Decision Systems

The architectural framework supports the proposed architecture for scalable cloud-based intelligent decision systems. It comprises the core components of the decision systems and their main interfaces, defined as a set of interoperable components that exchange requests, commands, data, and information and deliver the expected outcomes within a cloud environment.

Cloud computing provides a robust and scalable infrastructure that allows organizations to expand their operations by taking advantage of dynamically provisioned resources delivered as services over the Internet or private networks. Cloud infrastructure can be deployed as a public, private, or hybrid cloud. In a public cloud, resources are made available to multiple external customers, while in a private cloud, facilities and services are operated solely for a single organization. Such elasticity empowers organizations to plan for demand and provision resources according to projections, paying only for resource consumption, and addresses unexpected peaks in demand by provisioning resources outside the organization. Cloud infrastructure services can be offered across three layers: infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS). Scalable cloud computing infrastructure, based on IaaS and public cloud



deployment, provides the elasticity required to support the proposed systems and enable testing, validation, and production deployment in a real-world cloud environment.

3.1. Cloud Infrastructure and Deployment Models

A suitable cloud infrastructure, deployment model, and set of services should be chosen to satisfy the scalability and reliability requirements of intelligent decision systems with AI and big data capabilities. Elastic and scalable cloud infrastructures lend themselves well to enabling intelligent systems that are combined with AI and big data capabilities for decision-making optimisation and other objectives in various sectors. When considering individual systems, although such infrastructures can be public, private or hybrid, when provisioning dedicated systems for scalable decisioning solutions, a dedicated private or hybrid cloud infrastructure enables improved reliability during actual operation. Public cloud infrastructures, however, are often seen as lower-cost options. For many intelligent decisioning systems currently under development, either approach can support dynamic elasticity and the associated cost-control benefits.

Intelligent decision systems developed for offering decisioning services that can exploit the contextual aspect of these services — such as Cloud Databases-as-a-Service, Cloud Data Warehousing-, or Data Analytics Services — can benefit from a natively-developed data-management-as-a-service architecture that accommodates the big data age, its principles and the services and tools that accompany it. Intelligent decisioning systems that need to operate at a big data scale can exploit the scalability, elasticity and economic dynamics of public cloud infrastructures. Decision systems that use different types of data stores and storage engines can dynamically choose the database-engine type best suited for the specific workloads currently running. For a particular load, the decision system can provide a Data Management-as-a-Service that supplies Data-Management-as-a-Managed-Service capabilities and competes against the Cloud Database-as-a-Service providers.

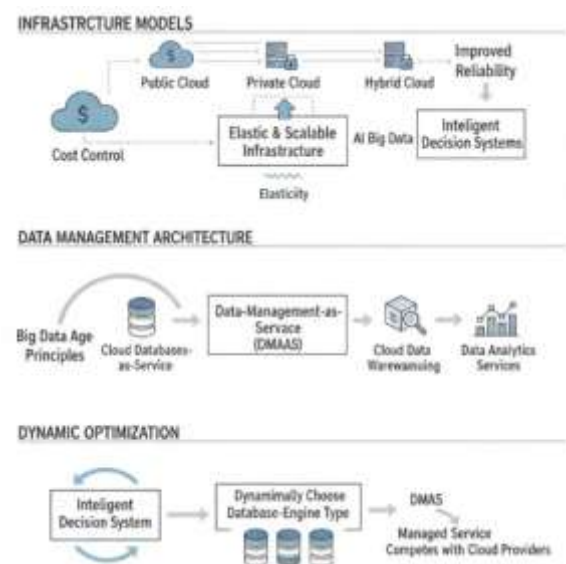


Fig 2: Scalable Cloud Architectures for Intelligent Decisioning: Integrating AI, Big Data, and Managed Data-Management-as-a-Service (DMaaS)

3.2. Data Management and Big Data

The data management architecture facilitates business operations and decision system development, enabling the storage, processing, data analytics, and archiving of both historical and real-time data. Big data characteristics of volume, velocity, variety, and veracity must be considered along with data governance policies for data management, quality, security, and privacy.

Data management within the decision system architecture is realized through a layered structure comprising data pipelines, storage and early-processing components, data processing and analytics engines, and both data management process and data quality management process modules. Data pipelines incorporate data sources, data ingestion components for bulk and streaming data, and early-stage data processing components for data cleansing, normalization, enrichment, and aggregation prior to entry into data storage components. Data management components provide a unified solution for processing,



warehouse, and analytics data management activities in accordance with data governance requirements. Data processing and analytics engines combine processing pipelines for batch and real-time data along with both batch and near-real-time analytics capabilities.

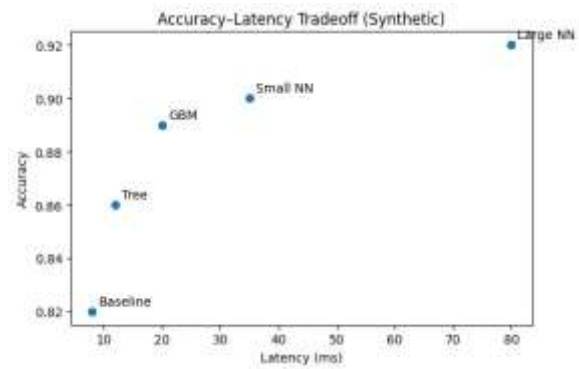
Metric	Value
Specificity	0.8823529411764706
False Positive Rate	0.11764705882352941
False Negative Rate	0.20408163265306123
Accuracy	0.84
Matthews Corr. Coef. (MCC)	0.6815519627720663

4. Industry-Specific Optimization Objectives

Industry-specific optimization objectives are framed in terms of domain-agnostic decision problems pertaining to intelligent systems and scalable solutions. Individual problem - optimization pairings tackle decision issues encountered by organizations within sales, operations, finance, strategy, and information technology (IT) divisions of companies and by healthcare service providers and manufacturers of life sciences providers. These pairings also address decision objectives relating to the system-specific types of intelligent decisions contained within the architectural framework, namely operational execution, asset and offering management, resource allocation, customer and patient experience enhancement, product and service resilience, business support, and regulatory compliance.

Optimization domains are selected on the basis of practical relevance for widely adopted AI and Big Data predictive and prescriptive applications in both industry and academia. The manufacturing and supply chain optimization domain concentrates on key decision areas associated with demand forecasting, inventory management, production planning and scheduling, logistics management, and supply chain resilience. The healthcare and life sciences optimization domain encompasses areas

such as patient flow optimization, clinical decision support system accuracy enhancement, resource allocation efficiency, and integration of external patient cohorts into research-related data-gathering activities.



Equation B) Step-by-step derivation of “scalability / latency / cost-efficacy” equations

B1) Latency

If a request arrives at time t_{req} and response completes at t_{resp} :

$$L = t_{resp} - t_{req}$$

B2) Time-to-completion (batch / complex query)

If a job starts at t_s and finishes at t_f :

$$T_{complete} = t_f - t_s$$

B3) Throughput

If J jobs finish in window Δt :

$$\text{Throughput} = \frac{J}{\Delta t}$$



B4) Parallel scalability (classic)

If runtime with 1 worker is T_1 , with p workers is T_p :

Speedup

$$S(p) = \frac{T_1}{T_p}$$

Efficiency

$$E(p) = \frac{S(p)}{p}$$

B5) Cost-efficacy (typical)

If benefit/impact is I (e.g., savings, service-level improvement points, revenue uplift) and cloud cost is C :

$$\text{Cost-efficacy} = \frac{I}{C}$$

(Or use $\frac{\Delta I}{\Delta C}$ for marginal efficiency.)

4.1. Manufacturing and Supply Chain

Manufacturing and supply chain processes have long presented decision systems with important optimization objectives, motivated by growth or service level requirements as well as the reduction of associated costs. Key issues are demand forecasting, inventory positioning, production scheduling, logistics and supply network design, and risk-averse or -aware planning and scheduling. Addressing these problems leads to substantial, realisable performance improvements in system performance and profitability. Supporting intelligent decision management in these application areas with a scalable cloud-based approach is crucial. Recent automated demand forecasting capabilities, such as in large global-model installations, are poised to generate new large-scale, comprehensive inventories. With both infrastructure and business platforms moving into cloud environments, demand management requirements and cloud-based statistical bases

have to combine for effective system control.

Corresponding demand sides ultimately drive a comprehensive manufacturing and supply chain process assessment, either initiating a detailed analysis of the entire production-distribution-inventory complex or concentrating on the expenditure area with systematic sources of supply disruption. Importance in those areas is heightened within the current global context, characterised by severe disruption from the pandemic combined with the growing imperative of carbon footprint reduction.

Development therefore has to address changes in the propitious demand specification. As success or failure greatly depends on the ability to forecast demand accurately and realise global data-analytic potential within the cloud, the first task is to meet demand expectation with due consideration of uncertainties.

4.2. Healthcare and Life Sciences

Optimization of systems in the healthcare and life sciences sector can enhance performance and serve decision-makers in addressing complex challenges. Fundamental aspects include optimizing patient flow and managing resources within healthcare facilities, integrating and interpreting clinical data using artificial intelligence for decision support, allocating conventional and time-critical resources, and merging databases for research.

In healthcare and life sciences, optimization enables decision-making that improves operational efficiency and health policy support. Hospital systems are confronted with limited resources required to cater for patients whose arrivals vary over time. Efficiently managing patient flow through a healthcare facility minimizes waiting times in emergency departments, increases resource availability for all patients, and reduces the risk of diverting patients seeking emergency treatment to other hospitals. Other decision-making tasks focus on allocating time-critical limited resources, such as intensive care units and transport between hospitals for organ transplants. Finally, the increasing number of clinical studies that share data across institutions requires robust methodologies for integrating and interpreting large-scale datasets. Using state-of-the-art methods for data analysis, data from multiple large studies



involving different criteria, patients, and measures can be integrated and interpreted to generate more reproducible and reliable conclusions.

5. Data Governance, Privacy, and Security

Proper governance, privacy, and security are fundamental for all enterprise data initiatives, including those hosting cloud-based decision systems. Analytics provide crucial insights to support manufacturing and supply chain resilience and to fill a gap in delivering key healthcare services. But availing Big Data to various stakeholders also creates risk. Consequently, risk management and compliance must naturally follow the above elements, forming a unified hierarchy for privacy and security in the services.

In broad terms, preventive and detective controls must be in place to enable timely reactions throughout the lifecycle of sensitive data. Preventive measures include appropriate checking and tracing of each event on data, allowing detection of malicious behaviour. An integrity and compliance framework comprises attributes that describe the internal and external compliance needs of data and their lineage. The former is essential for enabling controls around auditability, data quality, and access control, while the latter is important for strategy compliance. Industry-specific market modelling provides input to the lineages through explicit rules and thresholds.

5.1. Framework for Ensuring Data Integrity and Compliance

Cloud-based intelligent decision systems are widely exposed to integrity, privacy, and security risks that impact their reliability, and the establishment of an adequate integrity and compliance framework in this regard responds to the growing need to organize the validation of these systems. Data lineage, access control, audit paths, and conformance to established standards and regulations form the core dimensions of infrastructure management, against which a structured process should be designed to ensure

both compliance with established requirements and end-user communications for the adoption of increasingly sensitive systems.

Establishing responsible data management practices that guarantee compliance with established integrity control and auditing regulations is crucial for deploying cloud-based intelligent decision systems that process data from different sources, such as companies' operational platforms or government health agencies. Integrity and compliance control operations aim to minimize data integrity risks through a structured process that continuously attests to the quality and management of the data generated or received by the system for use or interaction in subsequent processes.



Fig 3: Ensuring Trust in the Cloud: A Multidimensional Framework for Continuous Integrity Attestation and Regulatory Compliance in Intelligent Decision Systems

5.2. Strategies for Safeguarding Data Privacy and Security

A comprehensive privacy and security plan is essential for cloud architecture to manage sensitive data safely and meet governance compliance and risk management standards. First, diverse sensitive data categories, data flows, and threat models must be analysed to identify associated considerations and mitigation steps. Potential areas of risk



include data capture, data sharing, external media support, existing software vulnerabilities, user access (authentication and authorisation), and incident handling. An all-hazards approach, considering the full range of potential crises, can inform systematic planning for scenarios such as natural disasters, technological failures, man-made hazards, and political threats.

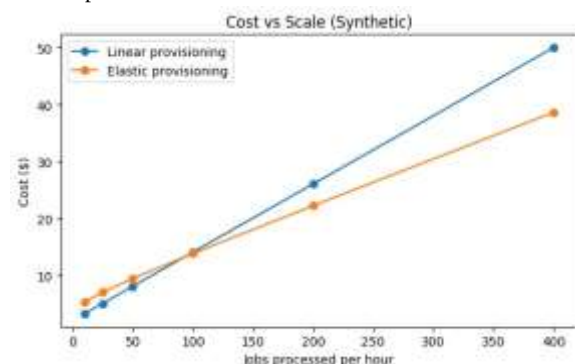
Data anonymisation or pseudonymisation is recommended for protecting user privacy in public clouds. The cloud provider can deploy anonymisation mechanisms in its processes before storing the data, minimizing the chances of privacy invasion from data leaks. Sensitive data should be encrypted during transmission and storage. Furthermore, an encrypted version of sensitive cloud data must be kept offline on external media using a system password that is not store with the cloud service or provider, to mitigate the risks from potential cloud data leakage. Audit logs should be also securely encrypted or compressed prior to transfer to cloud storage. The risk from abnormal user behaviour can be reduced through a combination of threat modelling and an intelligent security incident response mechanism that can automatically detect data-sensitive operations not performed by human actors and create a digital picture to facilitate the analysis of such suspected attacks.

Model	Accuracy	Latency_ms
Baseline	0.82	8
Tree	0.86	12
GBM	0.89	20
Small NN	0.9	35
Large NN	0.92	80

6. Evaluation Metrics and Validation Methodologies

The proposed cloud-based intelligent decision systems for industry-specific optimization using AI and big data need to be validated thoroughly in order to ascertain their performance and correctness. Evaluation/assessment metrics for individually tested components, as well as the

complete system, need to be defined clearly to ensure accurate and faster adoption, moving the research into industrial setups. The proposed validation plan utilizes quantitative and qualitative measures to evaluate the implemented solutions. In addition to the evaluation metrics and validation plan specified for the usages of optimization-focused cloud-based intelligent decision systems, standard metrics related to specific business processes/domains being optimized are also put forth. Accuracy and latency performance metrics are used to assess the modelling phase of the library and are based on the ML methods selected for implementation (e.g., neural networks). Scalability and cost-efficacy of solutions are computed alongside their real-world implementation. Scalability is governed by a combination of the number of jobs processed iteratively and parallel job execution supported by the cloud infrastructure. Cost-efficacy considers the amount spent using the cloud platform indications in relation to the impact achieved. The components related to production networks are benchmarked against other widely utilized production scheduling models, while the resource-allocation-based component is cross-validated with the insights achieved from experts in the healthcare domain.





Equation C) Step-by-step derivations for key “industry optimization”

C1) Demand forecasting loss functions

If actual demand at time t is y_t , forecast is $\hat{y}_t$, and there are n points:

Error

$$e_t = y_t - \hat{y}_t$$

MAE

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n |e_t|$$

RMSE

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n e_t^2}$$

MAPE (when $y_t \neq 0$)

$$\text{MAPE} = \frac{100}{n} \sum_{t=1}^n \left| \frac{e_t}{y_t} \right|$$

C2) EOQ (Economic Order Quantity) — full derivation

Let:

- annual demand D
 - order quantity Q
 - fixed ordering cost per order S
 - holding cost per unit per year H
1. Orders per year = D/Q
 2. Annual ordering cost = $S(D/Q)$

3. Average inventory $\approx Q/2$

4. Annual holding cost = $H(Q/2)$

Total relevant cost:

$$TC(Q) = S \frac{D}{Q} + H \frac{Q}{2}$$

Minimize $TC(Q)$:

$$\frac{dTC}{dQ} = -S \frac{D}{Q^2} + \frac{H}{2} = 0$$

Rearrange:

$$S \frac{D}{Q^2} = \frac{H}{2} \Rightarrow Q^2 = \frac{2SD}{H} \Rightarrow Q^* = \sqrt{\frac{2SD}{H}}$$

6.1. Assessment Criteria and Validation Techniques

Rigorous evaluation of believable and usable solutions is essential. The technical effectiveness or accuracy of decision outputs is fundamental, achieved through measures of precision, sensitivity, specificity, false positive and false negative rates, Matthews correlation coefficients, or similar methods for predictive function over data with known ground truth. A second, critical functional property is the score latency or delay in receiving a decision, often measured in microseconds or milliseconds. For scalable, cloud-based architectures, time-to-completion for complex queries is a further important factor. In addition, the assumptions underlying any specific solution and degree-of-fit to a population of users must be examined through benchmarking against industry-standard implementations, consistency checks of advisors’ scores across use cases, and sensitivity analysis.

Cloud-based instantiations offer substantial infrastructure, platform, and software cost rewards. The aim within the domain is thus for cost-efficient optimization with good accuracy and low server-side latency. Scalability of installed solutions remains paramount: “The lowest cost—considered as the total of construction, commissioning,



operation, and geographic footprint”, with capacity significantly above that needed to meet the foreseen future peak. An alternative perspective focuses on data quality, quantity, and circulation. These nuance the canonical engineering-mode bias towards performance, with information-latency (the time from urgent information need to response deployment) and the expected volume drop-out of useful data enhancing decision support for non-critical-or-non-evolving tasks.

7. Conclusion

Overarching drivers—cloud technologies, the advance of AI and big data, and the demand for ever-greater automation—have significantly transformed software development in recent years. At the same time, these technologies have made it possible to design systems for new product areas or even create entire classes of systems that were previously considered infeasible. Scalable cloud-based intelligent decision systems have been developed for a wide range of areas, from enterprise resource planning to healthcare. Nevertheless, decision problems and associated objectives specific to particular industry sectors have not received the same scalability focus. Likewise, no integrated architecture has yet been proposed that exploits these technologies for industry-specific decision problems that can benefit from focused accuracy and reduced execution delays.

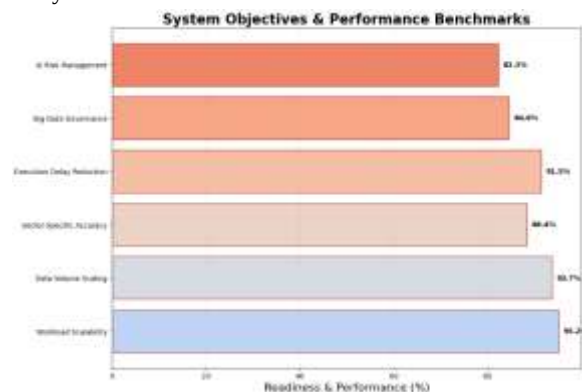


Fig 4: System Objectives & Performance Benchmarks

To address these gaps, an architectural design that scales with the increase in user workload and data volume, while incorporating the intelligence required for a particular area and supporting decisioning, is presented. The design strives to optimize five domains—manufacturing, supply chains, healthcare, life sciences, and security and risk management—and an end-to-end solution is initially described for the manufacturing and supply chain domains, covering processes from demand forecasting to logistics management. Integration of AI and big data governance capabilities supports the management of risks associated with the adoption of these technologies and their application across any of the five specific domains.

7.1. Summary of Findings and Future Directions

Broadly defined, the study proposes and demonstrates a framework for scalable cloud-based intelligent decision systems that enable optimization of domain-specific objective functions using AI and big data. The framework consists primarily of three interrelated elements. First, cloud infrastructure and data management architectures specifically designed to meet the scalability and reliability requirements of cloud environments. Second, an outline of a set of objective functions that address common optimization problems in industries such as manufacturing, supply chain, healthcare, and life sciences. Finally, a robust governance, privacy, and security framework that integrates risk management with compliance across industries.

These findings contribute to the nascent field of cloud-based intelligent decision systems aimed at the optimization of industry-specific objectives. The motivation for research in this area stems from the realization of how current global disruptions, accelerated by the COVID-19 pandemic, have exposed vulnerabilities in highly complex, tightly coupled, globalized systems. Resilience-driven optimization of manufacturing and supply chains, together with improved management of healthcare and life sciences, have come to the fore as critical priorities. Deploying scalable AI- and big-data-enabled systems in cloud environments offers the ability to manage complex optimization requirements for such



domains and thereby significantly improve decision-making. Empirical validation of the architectural framework, which is situated at the crossroads of cloud computing, AI, and big data, represents a first step in this direction.

8. References

1. Csalódi, R., Süle, Z., Jaskó, S., Holczinger, T., & Abonyi, J. (2021). Industry 4.0-driven development of optimization algorithms: A systematic overview. *Complexity*, 2021, 1–22.
2. Lee, C., & Lim, C. (2021). From technological development to social advance: A review of Industry 4.0 through machine learning. *Technological Forecasting and Social Change*, 167, 120653.
3. Ganesh, S. K., Subbareddy, K., Gopi, A., Davuluri, P. N., Mannar, B. R., & Karnawat, A. T. (2026, June). Fraudulent Credit Card Transaction Detection Using a Hybrid Ensemble-Anomaly Detection Framework. In *2026 6th International Conference on Intelligent Technologies (CONIT)* (pp. 1-6). IEEE.
4. Chen, J., Du, T., & Xiao, G. (2021). A multi-objective optimization for resource allocation of emergent demands in cloud computing. *Journal of Cloud Computing*, 10, 20.
5. Nawrocki, P., & Osypanka, P. (2021). Cloud resource demand prediction using machine learning in the context of QoS parameters. *Journal of Grid Computing*, 19, 20.
6. Yandamuri, U. S., Loganathan, R., Davuluri, P. S. L. N., Rani, P. R. S., Kolla, S. H., & Nagubandi, A. R. (2026). Adaptive Intelligence Networks for Humancentered Enterprise Automation and Governance. In *2026 Third International Conference on Innovations in Cybersecurity and Data Science (ICICDS)* (pp. 678–683). IEEE. 2026 Third International Conference on Innovations in Cybersecurity and Data Science (ICICDS). <https://doi.org/10.1109/icicds70526.2026.11604640>
7. Singh, H., Bhasin, A., & Kaveri, P. R. (2021). QRAS: Efficient resource allocation for task scheduling in cloud computing. *Discover Applied Sciences*, 3, 474.
8. Sarker, I. H. (2022). AI-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems. *SN Computer Science*, 3, 158.
9. Khan, T., Tian, W., & Buyya, R. (2022). Machine learning (ML)-centric resource management in cloud computing: A review and future directions. *Journal of Network and Computer Applications*, 204, 103405.
10. Feng, X., Guo, C., Jiao, T., et al. (2022). A maturity model for AI-empowered cloud-native databases: From the perspective of resource management. *Journal of Cloud Computing*, 11, 39.
11. Banoth, S., Santoshi Kumari, M., Lalitha, P., Deepthi, P., Kumar Peddi, R., & Kavitha, P. (2026). An Efficient Hybrid K-Means and Random Forest-Based Approach for Cloud Malware Detection and Privacy Protection. In *2026 6th International Conference on Intelligent Technologies (CONIT)* (pp. 1–6). IEEE. 2026 6th International Conference on Intelligent Technologies (CONIT). <https://doi.org/10.1109/conit69683.2026.11621822>
12. Rosati, R., Romeo, L., Cecchini, G., Tonetto, F., Viti, P., Mancini, A., & Frontoni, E. (2023). From knowledge-based to big data analytic model: A novel IoT and machine learning based decision support system for predictive maintenance in Industry 4.0. *Journal of Intelligent Manufacturing*, 34, 107–121.
13. Jan, Z., Ahamed, F., Mayer, W., Patel, N., Grossmann, G., Stumptner, M., & Kuusk, A. (2023). Artificial intelligence for Industry 4.0: Systematic review of applications, challenges, and opportunities. *Expert Systems with Applications*, 216, 119456.
14. Qiu, H., Mao, W., Wang, C., Franke, H., Kalbarczyk, Z. T., Basar, T., & Iyer, R. K. (2023). On the promise and challenges of foundation models for learning-based cloud systems management. *NeurIPS 2023 Workshop*.
15. Kreuzberger, D., Kühl, N., & Hirschl, S. (2023). Machine learning operations (MLOps): Overview,



- definition, and architecture. *IEEE Access*, 11, 31866–31879.
16. Faubel, L., Schmid, K., & Eichelberger, H. (2023). MLOps challenges in Industry 4.0. *SN Computer Science*, 4, 828.
 17. Mahadevan, S. (2026). Governed Serverless Automation Ecosystem for AI-Driven Enterprise Integration and Sustainable Cloud Operations. *International Journal of Computer Information Systems and Industrial Management Applications*, 18(16s), 1561-1570.
 18. Kosińska, J., & Zieliński, K. (2023). Enhancement of cloud-native applications with autonomic features. *Journal of Grid Computing*, 21, 44.
 19. Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, 616, 259–265.
 20. Diaz-de-Arcaya, J., Torre-Bastida, A. I., Zárate, G., Miñón, R., & Almeida, A. (2024). A joint study of the challenges, opportunities, and roadmap of MLOps and AIOps: A systematic survey. *ACM Computing Surveys*, 56(4), Article 84.
 21. Mangalampalli, B. M., Peddi, R. K., Kolla, S. K., Reddy, V. A. R., Mangala, N., & Seenu, A. (2026, June). Explainable Clinical Graph Intelligence Framework for Longitudinal Risk Modeling and Care Pathway Optimization. In *2026 Third International Conference on Innovations in Cybersecurity and Data Science (ICICDS)* (pp. 1-6). IEEE.
 22. Alshahrani, R., Yenugula, M., Algethami, H., Alharbi, F., Goswami, S. S., Naveed, Q. N., Lasisi, A., Islam, S., Khan, N. A., & Zahmatkesh, S. (2024). Establishing the fuzzy integrated hybrid MCDM framework to identify the key barriers to implementing artificial intelligence-enabled sustainable cloud system in an IT industry. *Expert Systems with Applications*, 238, 121732.
 23. Xu, Z., Elomri, A., Baldacci, R., Kerbache, L., & Wu, Z. (2024). Frontiers and trends of supply chain optimization in the age of Industry 4.0: An operations research perspective. *Annals of Operations Research*, 338, 1359–1401.
 24. Sun, X., Yu, H., Solvang, W. D., & Govindan, K. (2024). A two-level decision-support framework for reverse logistics network design considering technology transformation in Industry 4.0: A case study in Norway. *The International Journal of Advanced Manufacturing Technology*, 134, 389–413.
 25. Bhavani, B. D., SR, S., Loganathan, R., & Nagaraj, S. (2026, April). Evolutionary Gravitational Neocognitron Neural Network, Snow Leopard Optimization and Deep Graph Reinforcement Learning for Routing Protocol in WSN. In *2026 2nd International Conference on Intelligent Systems and Computational Networks (ICISCN)* (pp. 1-8). IEEE.
 26. Vhatkar, K., Kathole, A. B., Kshirsagar, A. P., & Katti, J. (2024). Improved optimization algorithm for resource management in cloud applications with performance monitor of VM provisioning, placement and recycling. *Journal of Health and Social Sciences*, 30(4).
 27. Soori, M., Karimi Ghaleh Jough, F., Dastres, R., & Arezoo, B. (2024). AI-based decision support systems in Industry 4.0: A review. *Journal of Economy and Technology*, 4, 206–225.
 28. Phillips-Wren, G., & Virvou, M. (2025). Issues and trends in generative AI technologies for decision making. *International Journal of Information and Decision Sciences*, 19(2).
 29. Vakkalagadda, T., & Rajamanickam, V. (2026). Agentic AI Architectures for Next-Generation Investment Advisory and Portfolio Intelligence. *International Journal of Computer Information Systems and Industrial Management Applications*, 18(9s), 1206-1219.
 30. Bandhana, A., & Vokřínek, J. (2025). AI-driven manufacturing: Surveying for Industry 4.0 and beyond. *Operations Research Forum*, 6, 145.
 31. Page, S. E., & Kallapur, A. (2026). Replace, augment, disrupt: AI & organizational decision-making. *Journal of Organization Design*, 15, 19–26.



32. Soori, M., Karimi Ghaleh Jough, F., Dastres, R., & Arezoo, B. (2026). AI-based decision support systems in Industry 4.0: A review. *Journal of Economy and Technology*, 4, 206–225.
33. Piyush Kumar Pareek. (2026). Human-Centric Machine Learning Frameworks for Scalable Software Quality Prediction. *Journal of Intelligent Decision Making and Information Science*, 3(5s), 1525–1543. <https://doi.org/10.59543/jidmis.v3.1284>
34. Al-Ababneh, H. A., Alkhaldy, I., Altarawneh, H. M., Popova, O., Azam, A., & Zawaideh, F. H. (2026). How can a cloud architecture unify data processing and AI-assisted decision support for marketing automation? *Journal of Theoretical and Applied Information Technology*, 104(14).
35. Piyush Kumar Pareek. (2026). Self-Evolving Analytics Pipelines for Reliable AI-Augmented Software Systems. *Journal of Intelligent Decision Making and Information Science*, 3(5s), 1558–1578. <https://doi.org/10.59543/jidmis.v3.1286>
36. Nguyen, S. T., & Tulabandhula, T. (2023). Generative AI for business strategy: Using foundation models to create business strategy tools. *arXiv*.
37. Loganathan, R. (2026). SaaS Entitlement Governance: A Reproducible Model for Detecting and Reclaiming Over-Provisioned Access at Enterprise Scale. *Journal of Intelligent Decision Making and Information Science*, 3(7s), 2519-2530.
38. Yang, S., Nachum, O., Du, Y., Wei, Y., Abbeel, P., & Schuurmans, D. (2023). Foundation models for decision making: Problems, methods, and opportunities. *arXiv*.
39. Schrepel, T., & Pentland, A. (2024). Competition between AI foundation models: Dynamics and policy recommendations. *Industrial and Corporate Change*, 34(5), 1085–1103.
40. Phillips-Wren, G., & Virvou, M. (2025). Generative AI for decision-making: A multidisciplinary perspective. *Journal of Innovation & Knowledge*, 10(4), 100751.
41. Davuluri, P. S. L. (2023). AI-Augmented Sanctions Screening: Enhancing Accuracy and Latency in Real Time Compliance Systems. *AI-Augmented Sanctions Screening: Enhancing Accuracy and Latency in Real Time Compliance Systems* (December 15, 2023).