Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Integrating Machine Learning with Cloud Analytics to Enhance Real-Time Business Intelligence

Uday Surendra Yandamuri
Independent Researcher,
yudaysurendra@gmail.com,
0009-0003-8655-9322

ARTICLE INFO

Received: 03 Nov 2024

Revised: 15 Dec 2024

Accepted: 26 Dec 2024

ABSTRACT

Although business intelligence (BI) solutions historically analyzed near-real-time information, organizations increasingly seek to exploit real-time data. Dramatic reductions in the cost of data storage, cloud-enabled analytics, and investments to deliver streaming-ready information have created the potential to change the latency profile of BI systems. The introduction of machine learning (ML) in a cloud context represents another important opportunity—cloud infrastructure provides a family of services with rapidly decreasing cost and increasing ease of use that are optimized for ML and ML-related workloads. The concurrent desire to optimize the data-to-decision loop and the complementary nature of cloud analytics infrastructure and ML facilitate turning insights from active data into business actions, should that be required. However, these changes are not without challenges and require addressing the following questions: What architectures support integration of ML-driven information with BI? What considerations govern the design and operation of these architectures? Which real-world scenarios have achieved measurable performance improvements, shortening time to insight and supporting real-time data-driven decisioning? A range of publicly available real-time implementations across multiple industries demonstrate that these questions can be addressed, either wholly or in part, and that shortening time to insight improves BI.

Keywords: Real-time BI, cloud analytics, machine learning, data pipelines, streaming, inference, governance, security, ethics,Long-range volcanic pressure waves; concrete; micro-cracking; mechanical integrity; repaired surfaces

1. Introduction

Business Intelligence (BI) refers to the set of tools, techniques, and processes that help organizations capture, analyze, and present business information. Historical application areas have included sales and financial analysis, supply chain management, manufacturing performance, budgeting, and forecasting based primarily on historical data. The emergence of machine learning (ML) technologies—more specifically, the combination of supervised, unsupervised, and reinforcement learning—and their application in cloud environments create new opportunities to integrate real-time decisioning with BI dashboards in a governed, scalable, accessible manner. Early days of BI and analytics relied almost completely on hindsight and historical data to drive decisions.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

As organizations increasingly capture and process mobile, sensor, Internet of Things (IoT), and Web data, new applications and operational strategies, such as customer engagement and personalization, proactive operational intelligence, supply chain and social media monitoring, require timely decisions based on live data. The ability to use ML models in the cloud to support real-time behavioral predictions and other decision-triggering analytics has improved and new architectures for BI that use these features for online data-to-decision-to-action improvements are emerging. Business process/product/service innovations driven by these real-time data insights have demonstrated measurable business value and constitute an active area of investment for organizations.

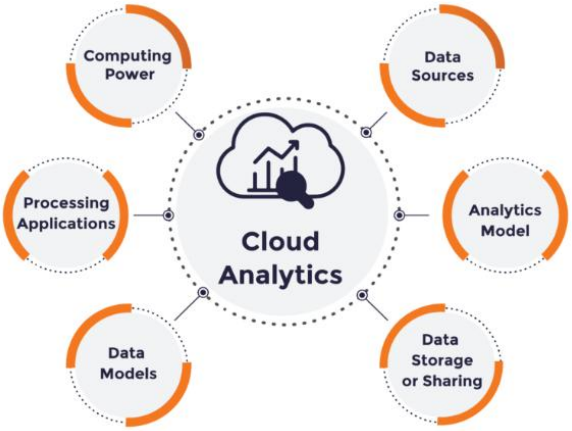


Fig 1: AI-Enabled Data Analytics in the Cloud

1.1. Background and Significance: Historically, business intelligence (BI) has relied on static snapshots of data sources periodically extracted, transformed, and loaded into dedicated data warehouses. With the introduction and broad adoption of cloud-based analytics platforms, the need to extend BI beyond traditional report generation tasks into the constantly evolving realm of operational intelligence has emerged. Today, organizations can access customer and operational data in a real-time manner or with only a few seconds delay and leverage cloud-based analytics to monitor business activity. These low-latency workloads are all candidates for real-time BI, which allows participating organization to leverage data freshness for proactive and prescriptive decision making. This extension changes the governing success factors, key performance indicators, and use cases compared to traditional BI.

Organizations that implement and manage customer analytics in the cloud—especially real-time customer engagement experiences, segmentation, and churn prediction—stand to benefit tremendously by applying machine learning (ML) models to operational intelligence. Modeling combines historical data for training with current and fresh data for inference. Moving ML training and inference to the cloud enables such pipelines to be integrated with the data ingestion flow from business operations. As business events occur in near real time, low-latency data-driven insights can be presented directly to business users through integrated BI dashboards or applied automatically through decision orchestration engines. The key is ensuring that the ML data preparation and model training phases can keep pace with the data ingestion rate, producing a continuously evolving set of features capable of supporting low-latency production-quality inference at a scale that matches operational activity.

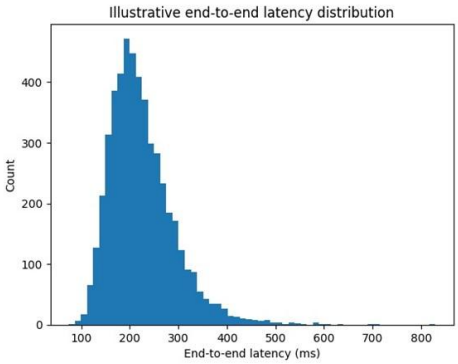
Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

2. Foundations of Real-Time Business Intelligence

Real-time business intelligence (BI) facilitates rapid decision-making based on continuously flowing information. Such near-continuity constitutes a powerful advantage, enabling seamless integration into business operations. However, realizing it requires investments, trade-offs, and the right technology for the job—not just technology that happens to be available or the art of the possible. Paradoxically, cloud analytics might narrow the V that for decades represented the promise of business intelligence. Data for traditional BI has historically come from batch ETL loads into data warehouses, but as a new generation of systems interacts with customers, partners, suppliers, and employees using social media, mobile hardware, and the Internet of Things, organizations have come to expect near-continuous information not just from their external facing systems, but throughout their entire operations. Delivering this capability is hard but, everywhere it is successful, it is a huge competitive advantage.

Real-time BI does not come at zero cost. To deliver data in a timely fashion, organizations must automate much of the work historically performed by analysts. They must com-mit to monitoring data pipelines, setting service-level agreements (SLAs) for latency and freshness, and ensuring that service levels can be met and exceeded. They must investigate every alert and, messy though this work will be, must document and share its insights. They must continually improve trusted models. They must provide analysts with the tools they need to embed high-quality, high-velocity data; minimize the data management burden; and, above all, optimize business outcomes. They must improve customer understanding using segmentation, churn prediction, and contact strategy models that generate real-time offers, and apply these models to achieve the desired result: being able to treat every customer as an individual while still doing so at scale. They must monitor operations, spotting trends, anomalies, and outliers early, generating alerts, presenting root-cause analysis to decision makers, and thus ensuring smooth and mistake-free operations. They must ensure common quality, timing, and content of data throughout their supply chain, enabling steady operations and a fast response to any changes in supply or demand. Delivering such results will more than repay the requisite investment.



2.1. Definitions and Objectives : Real-time BI can be defined as a business intelligence solution that offers outcome-oriented, decision-relevant information with a latency limited by a previously defined service level agreement (SLA). It can involve operations management, customer engagement or any other analytics category as long as the process of ingesting, cleansing and processing the data is sufficiently fast and the displayed information is suitable for automated decision making, for example updating official financial trading partners that a certain glass or electric panel are still needed, suggesting to an overall

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

business partner that a transaction should occur to reduce cost or delay, and alerting internal agents that there an anomaly in the production process. When slack or soft budget to achieve such SLAs is temporally allocated, it can combine with mechanisms from the reinforced lerning field to automatize reviews for different applications and agents that require different levels of precision and latencies.

Latencies that succeed in yet achieving sufficiently stable real-time solutions are usually expressed as by seconds or minutes and will seem slower as directly perceptible by humans. Such latencies are usually acceptable by commercial organizations whose BI department can provide technical agent-ready insights for frequent and minor decisions such as transitioning preparing a business partner for an upcoming transaction or monitoring a production process pathogenically driven by mimeographs. Such solutions can naturally teturally structurating to the growing demand for data such as the McKinsey Supply Chain Visibility 1 and 2 studies and the Dun & Bradstreet Customer Experience Cloud Strategy whitepaper clusterized under Operational Intelligence and Supply Chain Optimization.

Equation 1: End-to-end latency (data → dashboard) with stage-by-stage derivation

Let an event flow through a real-time BI pipeline:

- t_0: event emitted by the source system
- t_1: ingested into streaming layer (Kafka/Kinesis/etc.)
- t_2: stream processing complete (clean/transform/aggregate)
- t_3: features retrieved/computed (feature store)
- t_4: ML inference completed
- t_5: dashboard updated / BI layer refreshed

Each stage latency is a time difference:

$L_{\text{ingest}} = t_1 - t_0$ $L_{\text{proc}} = t_2 - t_1$ $L_{\text{feat}} = t_3 - t_2$ $L_{\text{infer}} = t_4 - t_3$ $L_{\text{dash}} = t_5 - t_4$

End-to-end latency is:

$$L_{\text{e2e}} = t_5 - t_0$$

Insert intermediate timestamps (telescoping sum):

$$t_5 - t_0 = (t_5 - t_4) + (t_4 - t_3) + (t_3 - t_2) + (t_2 - t_1) + (t_1 - t_0)$$

So:

$$\boxed{L_{\text{e2e}} = L_{\text{dash}} + L_{\text{infer}} + L_{\text{feat}} + L_{\text{proc}} + L_{\text{ingest}}}$$

This stage decomposition is what the paper alludes to when discussing **per-stage percentiles** and managing bottlenecks to meet SLAs.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

2.2. Architectural Considerations for Real-Time BI Availability and integration tooling are key for cloud-enabled real-time BI. Multi-layered architectural styles accommodate both streaming/trickle data coming from external sources and from batch processes feeding data at intervals into a data lake for subsequent consumption. Data is ingested and processed along multiple paths to support a range of use cases with different pacing and freshness needs, while minimizing latency in the fastest paths. Tolerance to stalled processing and service interruptions relies on distributed, horizontally scalable, fault-tolerant technologies. These can sustain high rates of ingestions and extractions, facilitate seamless integration of data from heterogeneous sources with diverse update frequencies, and support hybrid implementations of data lakes and warehouses.

A core strength of cloud can be found in the elasticity of provisioning and in the extensibility of many services. Not only can appropriate resources be automatically assigned for intensive single-user workloads, but such pipelines or analysis flows can also be run periodically on an ad-hoc basis whenever extra capacity becomes available or with some delay, for instance, to support the determination of the next promotional offer for a holiday season. Another significant advantage is the limited upfront investments required to put in place an extensive pool of services, possibly leveraging the shared use of partially or fully managed cloud-native solutions, which are regularly updated and improved by the cloud vendor. These aspects help to democratize the access to advanced analytics, bringing such capabilities within reach of every type of organization, from large enterprises to small companies without in-house data engineering or data science expertise.

3. Machine Learning in Cloud Analytics

Real-Time Business Intelligence (BI) is an emerging area that focuses on accelerating the data-to-decision and data-to-action loops within an organization. It relies on continuous ingestion of event streams and real-time processing of the information to feed BI dashboards with the freshest possible data. The historical evolution of BI from decision support to advanced analytics, operation intelligence, and subsequently real-time BI is hence an extension of trend data in BI products. Increasingly, organizations are sending all their operational data to the cloud. Although significant parts of this data are not intended for long-term storage, the mere fact that nearly all operational data is sent to the cloud opens doors for novel capabilities.

Understood in such a way, these developments also create the ideal environment for most machine learning (ML) workloads: an abundant amount of data for training, a virtualized environment for on-demand scalability, and a cloud-based ML-as-a-Service available for both training and inference. This paper shows how cloud enables the use of ML and data in a BI environment at an unprecedented scale and speed, resulting in enhanced customer and partner experience across multiple longitudinal data sources and applications. Many BI systems today are still based on semi-structured and historical data stored in data warehouses, with data freshness often taking hours or days. When latency targets are reduced to seconds, additional challenges appear, as a completely new architectural and operational stack is needed.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

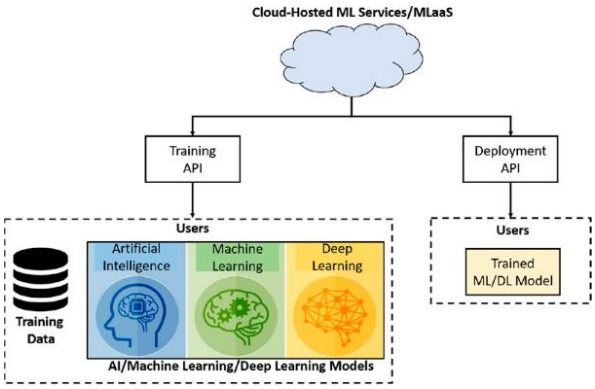


Fig 2: Securing Machine Learning in the Cloud

3.1. ML Paradigms and Their Cloud Implementations Machine learning comprises numerous paradigms and subcategories for modeling behaviors and predicting future events on the basis of historical data. The two main categories are supervised learning, where each training sample comes with a label, and unsupervised learning, whereby the objective is to discover patterns based on input space characteristics without predefined labels. Both categories are complemented by reinforcement learning, a third class of ML models mainly employed in situations where decision-making is sequential and temporal dependencies should be taken into account, with its application being limited because of the need for both exploration capabilities to discover reward systems and exploitation skills to accentuate previously acquired knowledge.

These paradigms are natively supported by several components and tools provided by cloud service providers. Serverless functions can be employed to build the decisioning layer for online serving and inference, whether dealing with low-latency requests or asynchronous batch-based predictions. Container-based deployments offer more flexibility for other types of solutions, with auto-scaling capabilities facilitating horizontal scaling to meet demand peaks while maintaining minimal costs during off-peak periods. Managed ML services abstract the complexity of infrastructure setup and provide a more coherent development and execution experience by encapsulating ML model training, validation, and prediction into a common environment where the solution-building team can focus mainly on the application logic. These managed services constitute the de facto choice for simple batch-based predictions.Unsupported unsupervised learning solutions are generally not cloud-native but can be integrated into the platform through scheduled jobs or pipelines, especially when running in a multi-cloud environment.

3.2. Data Pipelines, Feature Stores, and Model Serving in the Cloud

Accommodating the full breadth of the data-to-decision loop relies on dedicated cloud resources for every underlying operation: data ingestion, ML training, continuous feature engineering, model training, and model inference. These resources are typically organized as data pipelines that establish a contiguous, automated process for moving and transforming information from sources to targets. Within the ML context, pipelines handle online or offline feature engineering, model training from base or auxiliary features, or both, while feature stores support governance, accessibility, and shared reliability for the data needed to support successful model inference.

Feature engineering pipelines transform and combine available raw data into the specific features required for prediction, creation of ML training sets, and ongoing model retraining. At minimum, these pipelines

can utilize the underlying data infrastructure to calculate new features for ML training. Resource-expensive transformations, such as predictive content for CTR models or complex spatial calculations, can be precomputed and cached in a feature store for subsequent direct access. Further, semi-automated ML Integrations can trigger these pipelines for full end-to-end operation.

Feature elasticity can justify hybrid online-offline ML training architectures. Updated base features flow directly into model training. Where additional or more complex feature calculations are required, retraining frequencies can mirror update cadences or Slack priorities, thus minimizing unwanted model drift. As with raw data, base and auxiliary feature versions should be recorded and governed to avoid out-of-sync conditions in feature storage.

The training activity can be decoupled from the inference Executors owing to the nature of cloud-based components. Batch inference models can be registered with source signature, enablement specification, and associated drift dependencies. Executives can automate scheduled retraining, release updates to the inference Executors, and undergo exposure testing. Model-ensemble pipelines can draw from available models across the cloud environment. Feature stores specialize content for predictive serving, reduce latency on layered transformations, and permit fit-for-use control for usage at scale. Model-ensemble Executors address limitations from individual model capabilities by orchestrating inference across logical ensembles.

4. Integration Frameworks for Real-Time Insights

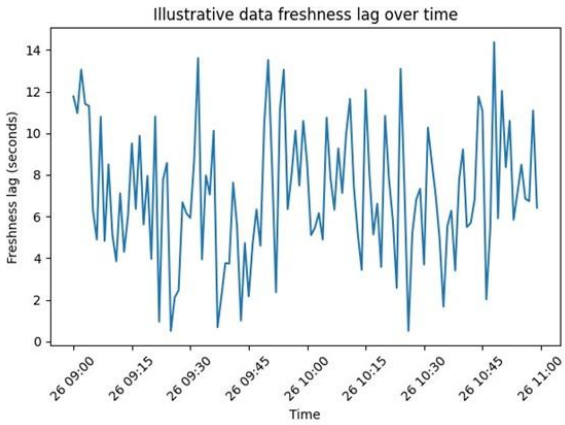
Generic integration frameworks in Figure 1 support end-to-end architectures for real-time business intelligence that connect data sources, machine learning workloads, and business intelligence dashboards. They include specialized products, like streaming data management platforms, as well as broader industry families, such as data or event streaming integrations for data ingestion and transmission and service mesh for microservice-based workflows. A variety of solutions are also available from leading cloud providers and ecosystem vendors for more specific integration needs.

Streaming platforms with processing and analytics functions can serve not only for data ingestion from change data capture, event-bus-style producers, and mobile devices, but also as a transmission and mediation layer for in-flight transformation into other data spaces, whether batch or streaming. Factors like data freshness guarantees and streaming quality of service can help aggregate the most important characteristics of different platforms. Even if they can be addressed separately within a framework, schema management or governance features warrant a degree of attention.

A temporal coupling approach allows an ML system to detect when both the online and offline training tables contain samples for the same time period. Once different Databricks workspaces are created to generate the online features and host the ML model training, for the features being generated daily, the training table can be selected every hour for online testing. Each of these online-ready features can be the prediction target in an ML model predicting for a control time period. Results are sent to an API, and the decision is taken by a dedicated business logic either through automatic scoring or by controlling the monitoring dashboards.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article



4.1. Data Ingestion and Streaming Platforms: Dedicated ingestion infrastructure and streaming technologies play a crucial role in connecting data sources with real-time processing and decisioning capabilities. While batch-oriented pipelines often utilize a message queue or bus for conveying intermediate artifacts, the requirements of analytical workloads with low-latency and high-throughput objectives necessitate a different approach. Data freshness guarantees make it possible to support dashboards and other BI channels directly fed by the event stream, obviating the need for intermediate storage in a database. At the same time, the use of streaming queues or event buses enables a publish-subscribe notification mechanism that decouples source systems from BI subscribers. This decoupling design principle also supports intrinsic fault tolerance; if the system routing a data source fails, consumers can keep functioning as long as historical copies of the data remain in the streaming layer.

In addition to message queueing, specific platforms exist to absorb and transform streams of events. Apache Kafka is a prominent open-source project that enables the capture, storage, and transformation of events in a fault-tolerant way. By retaining data in Kafka for a configurable period, both the write and read sides may resume their operation after short-lived outages. The Confluent extension to Kafka goes a step further; in addition to mere data routing capabilities, it provides functionality to register a common message schema, define transformation logic in a familiar SQL-like syntax, and connect with popular databases. For environments based on other cloud providers, managed versions of data streaming platforms (e.g., Amazon Kinesis for AWS, Azure Stream Analytics for Microsoft Azure) can also support ingesting and transforming streams of events.

4.2. Feature Engineering and Model Training at Scale Scalable real-time feature engineering supports both batch and streaming ML pipelines, while hybrid online/offline training balances model freshness against resource constraints. Temporal coupling minimizes latency; version-control automation enforces governance.

Data freshness influences the degree of automation of training processes in any ML system. Executing large training workloads once a month doesn't need automation, but taking new training samples from all ML-ready features every hour is better done through automation. This holds particularly true when the ML framework is pod-based in a Kubernetes cluster, as the training process can scale according to the available resources. If the trained model is not consumed every hour, costs can be optimally managed. To enable continuous model refresh while optimizing resources, two tables can be created in the feature repository

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

and coupled temporally: one table can contain recent production serving features, while the other stores the sample features used for the last retraining.

Equation 2: Percentile latency (p95 / p99) and SLA compliance

Suppose you record n end-to-end latencies:

$$\{L^{(1)}, L^{(2)}, \dots, L^{(n)}\}$$

1. Sort them ascending:

$$L_{(1)} \le L_{(2)} \le \cdots \le L_{(n)}$$

2. For percentile q (e.g., $q = 0.95$ for p95), compute index:

$$k = \lceil q \cdot n \rceil$$

3. The empirical percentile is:

$$\boxed{\mathrm{p}q = L_{(k)}}$$

The paper frames real-time BI as having latency **limited by an SLA**.

Let SLA threshold be S (e.g., $S = 300$ ms). A natural SLA requirement is:

$$\boxed{\Pr(L_{\mathrm{e2e}} \le S) \ge \alpha}$$

Commonly, $\alpha = 0.95$ (p95) or $\alpha = 0.99$ (p99). Then:

- SLA passes if $\mathrm{p95} \le S$ (for $\alpha = 0.95$)
- SLA breach rate estimate from samples:

$$\widehat{\Pr(\mathrm{breach})} = \frac{\#\{i : L^{(i)} > S\}}{n}$$

4.3. Real-Time Inference and Decisioning Low-latency inference engines supply rapid predictions or recommendations to downstream applications such as customer-facing solutions (e-commerce, marketing) or operational support systems (alerts, corrective actions). Typical latency requirements are at the sub-second level, although there are also scenarios such as batch decisioning that can tolerate higher latency. These endpoints should ideally be hosted close to the user community in order to minimize network delays—hence many cloud providers allow the deployment of applications on edge devices for latency-sensitive scenarios. To address the need for real-time responses, low-latency inference engines are supported by models and ensembles designed not only for accuracy but also for speed.

Decisioning logic built on live inference can orchestrate the triggering of actions based on expected business impact (e.g., offer acceptance), service quality (e.g., SLA breaches), acceptable risk levels (e.g., counter-fraud measures), or financial or resource costs (e.g., automatic call routing). Such decision engines should be governed by business rules that define the criteria for each action, as well as access to functions that

carry out the associated activity. As with any solution built on predictions, caution is needed. Models exhibit drift over time due to their reliance on patterns in data and cannot always be relied on implicitly—hence steering logic for the expected response should encompass monitoring and alerts, as well as human oversight for truly critical decisions, thereby ensuring that model outputs can be treated responsibly and appropriately.

5. Case Studies and Applications

The preceding material outlined the integration of streaming and batch processing, feature stores for real-time retrieval of historical data, and the orchestration of low-latency inference to link ML workload results to business dashboards. To ground these concepts in practice, the present section discusses domains that achieve quantifiable business value by delivering upstream data directly from the cloud-enabled BI pipelines.

Customer Analytics and Personalization Integration architectures targeting customer analytics and personalization often commence by defining segments based on key attributes, such as high-value customers, followed by monitoring members through machine learning to determine their likelihood of churning. Once probabilities reach a predefined threshold, BI dashboards in the marketing department can trigger alerts to inform the marketing team using specially tailored offers for these customers to prevent them from terminating their relationship with the company. Subsequently, when the predictive ML model is sufficiently trusted, it can be used to manage churn in real time by automatically invoking the marketing communication APIs and executing offers on an active-active basis.

Operations Intelligence and Anomaly Detection Generic BI integration architectures can be applied to monitoring critical qualitative and quantitative operational components or services and triggering alerts whenever any of the monitored indicators reveal an out-of-bounds condition. Cloud BI solutions further support event-driven architectures in which potentially latent anomalies can automatically initiate a root-cause analysis process aimed at determining the underlying cause—from technical components (e.g., systems and networks) to business processes (e.g., supply chain or product quality). By employing connectivity with advanced analytics platforms, the system can automatically connect with all components relevant to the symptoms observed and execute shallow investigations on each of them, returning insights on what requires the greatest attention.

Supply Chain Visibility and Optimization Data pipelines spanning the supply chain can be engineered to support business viability and risk optimization from a qualitative and quantitative standpoint. Systems often deploy separate MIW to predict future demand or expected throughput across the various supply chain paths. By connecting this output with appropriate optimization engines, cloud BI becomes a valuable enabler to minimize working capital and avoid possible disruptions on the service provided to clients.

5.1. Customer Analytics and Personalization: Illustrative use cases demonstrate segmentation, churn prediction, and real-time offers, reflecting improvements in latency and accuracy. Providing the right offering to the right user at the right time is often the holy grail of marketing, but these segments tend to change over time as customer preferences and external factors (such as macroeconomics and politics)

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

influence their behavior. A poorly targeted offer can result in lower conversion or even damage brand equity. However, a near real-time customer view can help identify the best offering for each of them.

Several businesses now have dashboards that display segments identified through clustering or other unsupervised methods that run periodically. However, these dashboards are often inadequate for daily operations, as the business partners rely on last-minute promotions or A/B testing. A common pattern is to predict customer happiness on a continuous scale and provide that score as an additional feature along with the latest promotions and recommendations. These scoring ML models can run much more frequently than previous unsupervised jobs and can even consider input variables that are close to real time, such as the effect of the last five promotions in the customer transaction history. Marketers and salespeople can then use that score to reinforce their normal marketing strategies, such as specialized promotions, upselling, or cross-selling. It is also crucial to conduct model monitoring to make sure that such real-time preparation is worth the effort and investment.

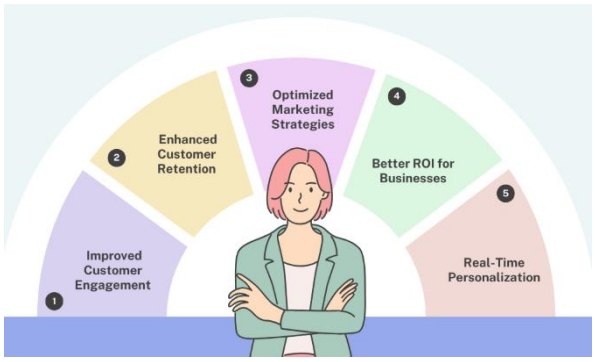


Fig 3: The Role of Data Analytics in Customer Personalization

5.2. Operational Intelligence and Anomaly Detection Illustrative use cases demonstrate how ML-driven real-time enhancements to Cloud Analytics increase the utility of Business Intelligence services. Typical applications employ ML as a complement to traditional BI – augmenting human decision-making through, for example, customer segmentation and churn prediction – or aim at operationalizing BI phases that traditionally suffer from latency, relying on ML for automatic handling of parts of the process that cannot be efficiently served or managed by people. Along these lines, the services explored include customer analytics and personalization, operational intelligence and anomaly detection, and supply chain visibility and optimization. For these operational-oriented use cases, Cloud Analytics pipelines bridge between ML and BI, thus enabling the Cloud architecture to support both short-term decisioning/tactical operations and longer-term strategic decisions.

Ongoing monitoring, alerting, and root-cause analysis of IT systems and operations characterize Service Operation processes and are essential components of an Operational Intelligence approach. When correctly managed, the key data sources that feed these processes are monitored applications, services, infrastructure (cloud and on-premises), transactions, cybersecurity events, business processes, and microservices. For these sources, and especially for the infrastructure and service monitoring data, Cloud Analytics solutions often use dashboards, whereas typical alerts generated by monitoring systems trigger actions taken by operators or by systems – for example, auto-healing functionalities or notifications sent to an internal ticketing system. Complementarily, Business Intelligence tools examine these data sources to assess the

past and patterns of systems' operation, service performance and quality, application-related transactions, and so on.

5.3. Supply Chain Visibility and Optimization: Demand forecasting can be important for businesses operating in a Just-in-Time (JIT) environment as poor forecasts can lead to dissatisfied customers and additional costs for expedited shipments. Machine Learning (ML) can improve forecasting accuracy and allow businesses to better manage the tradeoff between forecast accuracy and inventory holding costs for other customers. Improved forecasting accuracy can help predict inventory depletion for any product, so that production and/or procurement can be adjusted accordingly. While process optimization requires prediction of forecasted demand, actual throughput or customer arrival cannot be predicted accurately in advance as very short-term variations cannot be captured easily. Businesses therefore require near-real-time monitoring capability to track how the process is performing with respect to forecasted values through a control dashboard.

To retain supply chain visibility in near-real time, MetricStream was able to establish a cloud-based Business Intelligence (BI) dashboard by integrating ML and data from multiple sources, including supplier websites and third-party vehicle location data. The dashboard provided visibility into the real-time status of inventory levels in various warehouses, alerting the operations team when the stock of any product fell below a predefined threshold. Low stock status was combined with other features to generate real-time alerts for on-time delivery of customer orders. These capabilities helped internal teams act faster on stock replenishment activities and improve on-time delivery. Machine learning pipelines deployed in the cloud environment analysed demand patterns for products serviced by specific suppliers and forecasted future demand including freshness levels for the forecasts.

6. Governance, Security, and Compliance

Governance frameworks ensure trustworthy, reliable, and responsible organizational data. Key elements include data quality processes (such as cleansing and contextualization), data lineage tracing and auditing, access control and data-sharing procedures, and risk assessment and monitoring. Attention to these aspects improves the reliability of data used by business apps and ML workloads alike.

In a cloud environment, data provenance refers to the ability to track the origin and life cycle of data, enabling organizations to understand where the data comes from, how it has been transformed, and how it is being used, including data-sharing practices. Data provenance helps organizations (1) improve decision making, (2) model and predict data quality, (3) compute and manage cloud costs, and (4) audit and ensure compliance. Data provenance encompasses elements such as data security and privacy, support for business strategy execution, decision-making enhancement, and alerting for potential issues.

Data privacy involves protecting personally identifiable information (PII), ensuring that it is not disclosed to unauthorized individuals, and managing its usage throughout its life cycle. Encryption of data in transit and at rest, as well as access policies such as user and system roles, are means of data privacy assurance. Data governance minimizes the risk of data misuse, helping organizations comply with General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), and similar privacy regulations. Security controls cover both unintentional and deliberate exposure of data.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Equation 3: Drift monitoring → triggering retraining (equations you can use)

Bucket a feature into B bins. For bin b:

- p_b: fraction in **training/reference** data
- q_b: fraction in **serving/current** data

PSI is:

$$\boxed{\text{PSI} = \sum_{b=1}^{B} (q_b - p_b) \ln\left(\frac{q_b}{p_b}\right)}$$

Step-by-step:

2. Compute histograms → proportions $\{p_b\}$ and $\{q_b\}$
3. Compute each bin contribution $(q_b - p_b)\ln(q_b/p_b)$
4. Sum across bins

Then define a trigger rule like:

$$\text{if PSI} > \tau \Rightarrow \text{trigger investigation / retraining}$$

For feature CDFs $F_{\text{ref}}(x)$ and $F_{\text{cur}}(x)$:

$$\boxed{D = \sup_x |F_{\text{ref}}(x) - F_{\text{cur}}(x)|}$$

If D exceeds a threshold (or p-value is small), you flag drift.

6.1. Data Provenance and Lineage: An important aspect of creating trustworthy and actionable data for BI is ensuring that it meets expectations related to quality, freshness, and sources. Data provenance gives a complete history of how the data has been generated and modified. It tracks all alterations made to the data over its life span, starting from the very first stage when it was collected from an external source. Data lineage, on the other hand, provides only the set of next upstream nodes contributing to the final data in the flow. Both these aspects can be relevant depending on the use case. If a new data quality issue comes up, an organization can debug it by going back to find how that specific set of data items was modified during the different stages of their journey. If the goal is to decide whether a certain data set has to be trusted before making an important decision based on a BI dashboard, then it is sufficient to just track the latest provenance nodes involved in the generation process.

In a modern cloud context, where BI are often split into multiple cloud services, it is vital to monitor the data journey across these services. Cloud service providers usually provide a good level of monitoring, such as logs that provide information about input and output data sets and can be automatically parsed. Certain cloud-based BI tools also automatically capture the metadata of execution paths, including source and target data sets. Public cloud-data services can also provide data-lineage information that represents the dependencies among data sets, showing the relationships between raw data and the final BI dashboards. A

proper implementation of provenance and lineage support can thus be obtained by combining these services and tools.

6.2. Privacy, Security, and Compliance Considerations Data privacy is a key concern for organizations because of reputational risk and potential regulatory impacts. Organizations must identify the personal information of customers, employees, and business partners stored or processed within their cloud analytics solution and implement adequate policies and systems to address data privacy. Depending on geographical location and industry, organizations may also be bound by legal privacy regulations, such as the EU General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), Health Insurance Portability and Accountability Act (HIPAA), and Payment Card Industry Data Security Standard (PCI DSS), which set requirements about data collection, processing, storage, sharing, and deletion. Data masking, minimization, and pseudonymization techniques are often appropriate for analytics that include privacy-sensitive information. Additionally, cloud providers typically offer encryption and a range of access policies to mitigate the risk of unauthorized access to sensitive or private data.

Some cloud vendors, including AWS and Google, have made statements against the use of their tools for surveillance or discriminatory purposes. An organization planning to leverage analytics to automatically identify, exclude, or treat people differently should take appropriate care to avoid introducing model bias, and using a human-in-the-loop framework has proven to further mitigate this risk.

6.3. Ethical and Responsible AI in Cloud BI: Within the cloud business intelligence context, adopting machine learning can directly exacerbate ethical and responsible artificial intelligence (AI) challenges and thus must be mitigated or prevented. Safe and trustworthy AI systems can be attained by addressing the issues related to data, processes, and algorithms. Given that data is the lifeblood of machine learning-based systems, data bias mitigation—where training and operational data follow similar distributions—constitutes an essential concern. To achieve a high level of algorithmic transparency, the inclusion of adequate interpretability and explainability mechanisms is essential. Equally important is the governance and auditability of the overall development lifecycle, ensuring that the various formal and informal processes and methodologies are being followed. Furthermore, building accountability through attribution of responsibility for machine learning-based decisions, especially those affecting human beings, is equally important.

From an operational perspective, ensuring that human-driven decisions—and their consequences—are based on machine learning-based insights/decisions can minimize undesirable business outcomes. In effect, making the overall decision chain a human-in-the-loop capability leads to responsible and accountable AI systems. Moreover, requiring human involvement in the decision-making processes also manages possible legal repercussions when using machine learning in cloud business intelligence environments.

7. Evaluation and Metrics

While conventional business intelligence solutions (often with batch analytics) can deliver tremendous insights, the value of real-time integration and latency-sensitive processing is increasingly recognized. Latency is not an objective in itself; it is a means to achieve an outcome. Defining target latencies for data-to-decision loops and translating them into processing SLAs is crucial to delivering real-time actionability.

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Cloud architectures natively designed for streaming analytics can also run machine learning (ML) workloads for scoring and other data-intensive tasks.

A tailored service lifecycle, including drift detection for model accuracy, is key for embedding management into real-time operational delivery. Real-time BI relies on a variety of performance metrics; latency, freshness, and volume are obvious raw indicators, but others—such as the predictive accuracy of a ML model or the cost per decision made—distill into measures that map directly to the top-level business objectives of user organizations.

Real-Time Performance Metrics

Business-driven service organization models establish service levels for the operations within the organization—in effect, the emulation of an external service provider responsible for delivering accurately defined SLAs in terms of content and performance. Relevant metrics include latency, throughput, freshness, and feature availability between data-oriented, processing-oriented, and model-based systems. Per-target percentiles for processing delays at each stage are critical to managing risks of overall service level breaches. The latency requirement established for a real-time BI solution affects not only the overall ML life cycle but also the respective performance SLAs in the data pipeline and feature engineering stages.

7.1. Real-Time Performance Metrics: Real-time business intelligence hinges on specific metrics that gauge stream processing quality and speed. Typical success indicators encompass latency, throughput, data freshness, and their associated fallback measures, while common service-level agreements delineate performance benchmarks for BI dashboards. Latency denotes the end-to-end processing duration from source emission to BI dashboard update; streaming systems are often assessed on 95th or 99th percentile latency, with extreme values checking the overall contribution of processing bottlenecks. Throughput quantifies processed data volume within a defined time interval; analytic systems usually satisfy a throughput SLA through load balancing among multiple streaming paths. Data freshness signals the lag between dashboard refresh and source event emission, commonly narrowed for alerting systems to a few seconds.

These objective real-time performance markers primarily fulfill SLAs. However, the ultimate indicator of BI architecture remains the improved accuracy and timeliness of BI decision-making, which relate to trackable business parameters like revenue growth in customer analytics and cost reductions in supply chain optimization.

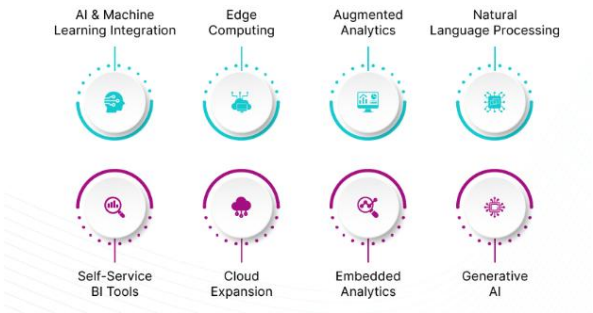


Fig 4: Real-Time Performance Metrics

Journal of Information Systems Engineering and Management
2024, 9(4s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

7.2. Model Monitoring and Lifecycles : Organizational decisioning is inherently tied to the data used for training and serving predictive or prescriptive models. Statistical relationships between prediction variables and outcome variables evolve consistently, resulting in drift. Accordingly, deployed models must be constantly monitored for performance degradation. Business understanding should help define monitoring conditions and performance thresholds. The relevant stakeholders should be notified for any threshold breach with supporting evidence for corrective action. Either a manual review and retraining process can be implemented, or a retraining pipeline can be automatically triggered. Common triggers are data volume or latency, but detection of data drift in the serving set through suitable test statistics is also becoming popular. The drift detection can be achieved by means of validation datasets and cloud-integrated libraries.

The execution of model retraining must consider data freshness. For batch-processing of historical data, models must be retrained before they drift beyond acceptance levels, whereas, for online-processing models, the presence of minor drift is acceptable until a sufficient volume of new data arrives to warrant batch retraining. The architecture must formally impose a governance and lifecycle management on models, fully automating the retraining and deployment while supporting novelty and/or user validation whenever needed.

8. Conclusion

The synthesis and integration of cloud-enabled analytics with the capabilities of machine learning hold the promise of significant improvements in the real-time aspects of business intelligence. Organizations can leverage real-time machine learning models to make rapidly-decaying decisions based on the freshest data. A comprehensive set of architectural considerations guides the integration of both layers at a higher level than that of traditional analytics and inform the development of end-to-end integration architectures.

These architectures were applied to three exemplars. In the area of customer analytics and personalization, real-time segmentation, churn prediction, and offer generation can be scaled to large populations, with pipeline latency and the accuracy of predictions forming two focal points. Maturity in operational intelligence and anomaly detection allow business objectives to be monitored and alerts issued, supported by drill-down analytics that help isolate root causes. Advances in visible supply chains enable demand forecasting, greater throughput in production processes, and real-time risk signals on inventory positions. Key business applications of this emerging space are thus supported by integrated cloud-based approaches.

8.1. Future Trends: The area of real-time cloud-enabled BI is set for significant evolution, characterized by the maturation and convergence of several fields that presently operate independently. For example, advances in data-intensive and open-source frameworks, coupled with growing volumes of data, are encouraging the formation of a vibrant ecosystem of real-time analytics. In addition, the deployment of different types of ML models in production has now become business-as-usual. Probably the most exciting aspect of the future of cloud-based BI lies in the integration of end-to-end pipelines stitching together data and ML workloads, monitoring model performance, and driving cloud-based applications, often in real time. Nevertheless, the overall solution space remains child-like, where each building block behaves in isolation but meets the formal and functional architectural requirements for real-time BI.

The convergence of these different areas offers the possibility of augmenting an increasing array of BI-related processes—from simple market segmentation to more complex customer experience management—using a unique pipeline approach executed in the cloud. Integrating data, ML workloads, and BI systems, while catering to the characteristics of specific business concerns, ensures that these applications benefit from the most appropriate analysis and inference capabilities. Nevertheless, these cloud-enabled solutions must be designed, implemented, and operated carefully to address security, privacy, robustness, quality, and compliance issues in an integrated manner.

9. References

[1] Lahari Pandiri, "AI-Powered Fraud Detection Systems in Professional and Contractors Insurance Claims," International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering (IJIREEICE), DOI 10.17148/IJIREEICE.2024.121206.

[2] Armbrust, M., Das, T., Sun, L., Yavuz, B., Zhu, S., Murthy, M., Torres, J., van Hovell, H., Ionescu, A., Łuszczak, A., & others. (2020). Delta Lake: High-performance ACID table storage over cloud object stores. Proceedings of the VLDB Endowment, 13(12), 3411–3424. doi:10.14778/3415478.3415560

[3] Recharla, M. (2024). Advances in Therapeutic Strategies for Alzheimer's Disease: Bridging Basic Research and Clinical Applications. American Online Journal of Science and Engineering (AOJSE)(ISSN: 3067-1140), 2(1).

[4] Foidl, H., Golendukhina, V., Ramler, R., & Felderer, M. (2024). Data pipeline quality: Influencing factors, root causes of data-related issues, and processing problem areas for developers. Journal of Systems and Software, 209, 111855. doi: 10.1016/j.jss.2023.111855.

[5] Fragkoulis, M., Carbone, P., Katsifodimos, A., & others. (2024). A survey on the evolution of stream processing systems. The VLDB Journal, 33, 507–541. doi:10.1007/s00778-023-00819-8.

[6] Nandan, B. P. (2024). Semiconductor Process Innovation: Leveraging Big Data for Real-Time Decision-Making. Journal of Computational Analysis and Applications (JoCAAA), 33(08), 4038-4053.

[7] Gupta, R., & others. (2024). Data-centric AI governance: Addressing the limitations of model-centric governance. arXiv. (arXiv:2409.17216).

[8] Mashetty, S., Challa, S. R., ADUSUPALLI, B., Singireddy, J., & Paleti, S. (2024). Intelligent Technologies for Modern Financial Ecosystems: Transforming Housing Finance, Risk Management, and Advisory Services Through Advanced Analytics and Secure Cloud Solutions. Risk Management, and Advisory Services Through Advanced Analytics and Secure Cloud Solutions (December 12, 2024).

[9] Harby, A. A., & Zulkernine, F. Data lakehouse: A survey and experimental study. Information Systems, 127, 102460. doi:10.1016/j.is.2024.102460.

[10] Paleti, S. (2024). Transforming Financial Risk Management with AI and Data Engineering in the Modern Banking Sector. American Journal of Analytics and Artificial Intelligence (ajaai) with ISSN 3067-283X, 2(1).

[11] Horchidan, S.-F., Kritharakis, E., & others. (2024). Crayfish: Navigating the labyrinth of machine learning inference in stream processing systems. In Proceedings of the 27th International Conference on Extending Database Technology (EDBT 2024). doi:10.48786/edbt.2024.58.

[12] Jakubik, J., Křivánek, R., & others. (2024). Data-centric artificial intelligence. Business & Information Systems Engineering. doi:10.1007/s12599-024-00857-8.

[13] Kaulwar, P. K. (2024). Agentic Tax Intelligence: Designing Autonomous AI Advisors for Real-Time Tax Consulting and Compliance. Journal of Computational Analysis and Applications (JoCAAA), 33(08), 2757-2775.

[14] Merli, M., Guo, S., Li, P., Chen, H., & Lu, N. Ursa: A lakehouse-native data streaming engine for Kafka. Proceedings of the VLDB Endowment, 18. (Page range as published in PVLDB volume 18).

[15] Oliveira e Sá, J., Gonçalves, R., & Kaldeich, C. (2024). Benchmark of market cloud data warehouse technologies. Procedia Computer Science, 239, 1212–1219. doi: 10.1016/j.procs.2024.06.289.

[16] Koppolu, H. K. R., & Sheelam, G. K. (2024). Machine Learning-Driven Optimization in 6G Telecommunications: The Role of Intelligent Wireless and Semiconductor Innovation. Global Research Development (GRD) ISSN: 2455-5703, 9(12).

[17] Schneider, J., Gröger, C., Lutsch, A., Schwarz, H., & Mitschang, B. (2024). The lakehouse: State of the art on concepts and technologies. SN Computer Science, 5, 449. doi:10.1007/s42979-024-02737-0.

[18] Vogel, A., Henning, S., Perez-Wohlfeil, E., Ertl, O., & Rabiser, R. (2024). High-level stream processing: A complementary analysis of fault recovery. arXiv. (arXiv:2405.07917).

[19] Singireddy, J. (2024). AI-Enhanced Tax Preparation and Filing: Automating Complex Regulatory Compliance. European Data Science Journal (EDSJ) p-ISSN 3050-9572 en e-ISSN 3050-9580, 2(1).

[20] Werner, S., & Tai, S. (2024). A reference architecture for serverless big data processing. Future Generation Computer Systems, 155, 179–192. doi: 10.1016/j.future.2024.01.029.

[21] Singireddy, S. (2024). The Integration of AI and Machine Learning in Transforming Underwriting and Risk Assessment Across Personal and Commercial Insurance Lines. Journal of Computational Analy- sis and Applications(JoCAAA), 33(08), 3966-3991.

[22] Azeroual, O., Schöpfel, J., Ivanovic, D., & Nikiforova, A. (2022). Combining data lake and data wrangling for ensuring data quality in CRIS. Procedia Computer Science, 211, 3–16. doi: 10.1016/j.procs.2022.10.171.

[23] Sheelam, G. K., & Koppolu, H. K. R. (2024). From Transistors to Intelligence: Semiconductor Architectures Empowering Agentic AI in 5G and Beyond. Journal of Computational Analy- sis and Applications(JoCAAA), 33(08), 4518-4537.

[24] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), Article 44. doi:10.1145/2523813.

[25] Meda, R. (2024). Agentic AI in Multi-Tiered Paint Supply Chains: A Case Study on Efficiency and Responsiveness. Journal of Compu-tational Analysis and Applications (JoCAAA), 33(08), 3994-4015.

[26] Katsifodimos, A., & others. (2023). Stream processing systems in the cloud: Trends and challenges. IEEE Internet Computing.

[27] Rongali, S. K. (2024). Federated and Generative AI Models for Secure, Cross-Institutional Healthcare Data Interoperability. Journal of Neonatal Surgery, 13(1), 1683–1694. Retrieved from https://www.jneonatalsurg.com/index.php/jns/article/view/9558.

[28] Kuhlenkamp, J., & others. (2023). Debunking serverless myths: A practical survey. ACM Computing Surveys.

[29] Marcu, O. C., & others. (2024). Big data stream processing. Doctoral dissertation/monograph.

[30] Guntupalli, R. (2024). Enhancing Cloud Security with AI: A Deep Learning Approach to Identify and Prevent Cyberattacks in Multi-Tenant Environments. Available at SSRN 5329132.

[32] Ormenisan, A. A., Meister, M., Buso, F., Andersson, R., Haridi, S., & Dowling, J. (2020). Time travel and provenance for machine learning pipelines. In USENIX Conference on Operational Machine Learning (OpML '20).

[33] A Scalable Web Platform for AI-Augmented Software Deployment in Automotive Edge Devices via Cloud Services. (2024). American Advanced Journal for Emerging Disciplinaries (AAJED) ISSN: 3067-4190, 2(1).

[34] Zhengxin, F., & others. (2023). MLOps spanning whole machine learning life cycle: A survey. arXiv. (arXiv:2304.07296)