Rakuten Cloud

The key enabler for AI/ML core and edge solutions

WHITEPAPER

MAY 2025

Table of Contents

Introduction

AI/ML workloads at the core and edge present distinct operational paradigms, each with unique advantages and challenges. In core applications, artificial intelligence and machine learning (AI/ML) models typically leverage extensive computational power, storage capacity, and robust infrastructure of centralized cloud data centers. This makes them ideal for training complex models, handling large datasets, and performing intensive analytics. This centralized approach benefits from high availability, scalability, and the ability to integrate diverse data sources.

In contrast, edge AI/ML focuses on real-time data processing and decision-making at the data source, such as IoT devices, sensors, and local servers.

A decentralized approach reduces latency, conserves bandwidth, and enables immediate responses, which are critical for applications like autonomous vehicles, smart cities, and industrial automation. However, edge deployments face challenges such as limited computational resources, intermittent connectivity, and the need for efficient data management. Balancing these two paradigms involves leveraging the strengths of both core and edge environments to create a cohesive, efficient, and responsive AI/ML ecosystem.

Use Cases

Edge analytics, utilizing AI/ML, are revolutionizing numerous industries by facilitating real-time data processing and decision-making directly at the point of data generation, through cameras, point of sale (PoS) terminals and other internet of things (IoT) devices. This is propelling significant advancements across various sectors, resulting in smarter and more responsive systems. In this section we will explore some of the key growing use cases.

Smart Cities and Disaster Management

In urban environments, edge analytics plays a crucial role in managing traffic flow, reducing congestion, and enhancing public safety. By analyzing data from traffic cameras, sensors, and connected vehicles, city planners can optimize traffic signals and reroute traffic in real-time. This not only improves urban mobility but also reduces emissions and enhances the overall quality of life for residents. Edge analytics can be used to monitor air quality, manage waste collection, and ensure efficient energy use in public buildings.

Healthcare

The healthcare sector is undergoing a transformation through the use of wearable devices and remote monitoring systems powered by edge analytics. These devices collect and analyze patient data on the spot, allowing for immediate detection of anomalies and timely medical interventions. This capability is particularly crucial for managing chronic conditions such as rehabilitation, blood pressure, diabetes and heart disease, where continuous monitoring can prevent complications. In remote or underserved areas, edge analytics enables telemedicine solutions, providing critical healthcare services to populations that might otherwise lack access.

Smart Manufacturing

In manufacturing, edge analytics is pivotal for predictive maintenance and quality control. By processing data from machinery and production lines locally, manufacturers can detect equipment malfunctions before they lead to costly downtime. This proactive approach ensures that machinery operates efficiently and reduces the risk of unexpected failures. Additionally, edge analytics helps in maintaining high-quality standards by monitoring production processes in real-time, identifying defects early, and ensuring that products meet stringent quality requirements.

Retail

Retailers are leveraging edge analytics to enhance customer experiences and streamline operations. In-store sensors and cameras analyze shopper behavior, enabling personalized marketing, optimized store layouts and surge pricing. For instance, retailers can use data to understand which products, or live orders in the case of fast-food dining, that attract the most attention and adjust displays and pricing accordingly. Edge analytics also helps in managing inventory more effectively by providing real-time insights into stock levels and demand patterns. This reduces the risk of overstocking or stockouts, ensuring that customers find what they need when they need it.

Private and Public Energy Distribution

In the energy sector, edge analytics is essential for managing smart grids and integrating renewable energy sources. By analyzing data from distributed energy resources such as solar panels and wind turbines, utilities can balance supply and demand more efficiently. This capability is particularly important as the energy landscape becomes increasingly decentralized and reliant on renewable sources. Edge analytics helps in reducing energy losses, ensuring a stable power supply, and optimizing the performance of energy storage systems.

Additionally, energy applies to private use cases found in telcos and enterprises. Modern CPUs have the capability to throttle power consumption and even go to sleep, saving organizations millions. Edge analytics can provide the CPU usage data necessary for the device to be throttled back or slept.

Industrial, Oil, Gas & Mining

In industrial settings, edge analytics is used to monitor and control complex systems and processes. For example, in oil and gas operations, edge analytics can analyze data from drilling equipment and pipelines to detect leaks, optimize production, and ensure safety. In chemical plants, edge analytics can monitor chemical reactions in real-time, ensuring that processes remain within safe and optimal parameters. This leads to increased efficiency, reduced risk, and improved safety in industrial operations.

Logistics Tracking

In the transportation and logistics industry, edge analytics is used to optimize fleet management and improve supply chain efficiency. By analyzing vehicle data, such as location, speed, and fuel consumption, companies can optimize routes, reduce fuel costs, and improve delivery times. Edge analytics also plays a crucial role in predictive maintenance for vehicles, helping to prevent breakdowns and extend the lifespan of the fleet. Additionally, in warehouses, edge analytics can be used to monitor inventory levels, track the movement of goods, and ensure that operations run smoothly.

Telecommunications and Cloud Providers

Telecommunications and Cloud Providers use edge analytics to enhance network performance and operations delivering better service to customers. By analyzing data from network nodes and user devices, telecom providers can detect and resolve issues such as signal interference and bandwidth congestion in real-time. This ensures a more reliable and high-quality service for users. Additionally, edge analytics helps in optimizing network resources, reducing latency, and improving the overall efficiency of the network.

Financial Services

In the financial sector, edge analytics is used to detect and prevent fraud in real-time. By analyzing transaction data at the PoS or and automated teller machines (ATM), financial institutions can identify suspicious activities and take immediate action to prevent fraudulent transactions. This not only protects customers but also reduces financial losses for banks. Edge analytics also helps in providing personalized financial services by analyzing customer behavior and preferences, enabling banks to offer tailored products and services.

Agriculture

Precision agriculture is another area where edge analytics is making a significant impact. Farmers use sensors and drones to collect data on soil conditions, crop health, and weather patterns. By processing this data locally, edge analytics provides real-time insights that help farmers make informed decisions about irrigation, fertilization, and pest control. This leads to more efficient use of resources, higher crop yields, and reduced environmental impact. For example, edge analytics can help determine the optimal time for planting and harvesting, ensuring that crops are grown under the best possible conditions.

Edge cloud challenges

Running Kubernetes at the edge for AI/ML workloads is significantly more challenging than operating it at the core, due to its many constraints and complexities. Edge servers or appliances typically have limited compute power, storage, and memory compared to the robust resources available in centralized cloud data centers. This limitation extends to power and cooling capabilities, which can further restrict the performance and scalability of AI/ML workloads at the edge. Network connectivity issues also pose a significant challenge; edge environments often suffer from intermittent connectivity and limited bandwidth, complicating the consistent communication and data transfer necessary for AI/ML tasks.

Managing a distributed network of edge nodes introduces additional layers of complexity. Unlike centralized clusters, edge deployments span numerous domains that can require sophisticated orchestration tools to provide lifecycle management, including: Deployment, monitoring, updates, and scaling across numerous, geographically dispersed nodes.

Data management also becomes particularly challenging, as AI/ML workloads often necessitate large datasets that must be distributed and synchronized across edge devices. Ensuring data privacy and security in these distributed environments adds another layer of difficulty, especially when dealing with sensitive information. Additionally, since edge resources can be in short supply, data footprint optimization is a must.

Latency and real-time processing requirements further complicate edge deployments. Many edge applications demand real-time or near-real-time processing, which is difficult to achieve given the resource and network constraints. While inference can often be performed at the edge, training typically requires more computational power and is usually conducted in the core, necessitating a careful balance between these needs. Operational challenges such as high availability and fault tolerance are also more pronounced at the edge due to the lack of redundant infrastructure or the ability to staff the location.

Ensuring that AI/ML frameworks and tools are optimized for edge environments can be challenging, as not all software designed for the cloud will run efficiently on edge devices. Packaging AI/ML models in containers and ensuring they operate effectively on resource-constrained edge devices requires meticulous planning and optimization. This challenge requires a solution with advanced workload and storage placement capabilities to ensure that the AI/ML workloads are optimized, given the resources available at the edge. Additionally, regulatory and compliance issues can vary by region, complicating data storage, processing, and transmission at the edge.

Addressing these challenges necessitates a combination of advanced orchestration tools, optimized AI/ML frameworks, robust data management strategies, and careful infrastructure planning. Only through such comprehensive approaches can the potential of running AI/ML workloads on Kubernetes at the edge be fully realized. The tool or suite of tightly integrated tools should support the entire stack, from bare-metal all the way up to services management, not to mention a scripting solution to manage other infrastructure and appliances not typically considered part of the cloud with a unified lifecycle management approach.

Rakuten Cloud: an AI/ML ready suite from core to edge

Rakuten Cloud comprises a suite of products which enables organizations to optimize AI/ML workloads.

Rakuten Cloud-Native Platform is an enhanced Kubernetes platform that exemplifies automation and self-service. It features a no-code, intent-driven interface that resembles an app store, making it user-friendly and accessible. Rakuten Cloud-Native Platform includes workload and storage placement optimizations, VM and container harmonization, and enhanced networking and storage automation. This platform is not only suitable for larger core clouds but also excels in edge deployments due to its low footprint and data optimization features. The product suite includes cluster, application, service, and infrastructure observability, ensuring comprehensive monitoring and management capabilities.

Rakuten Cloud-Native Storage is integrated with Rakuten Cloud-Native Platform but can also be deployed separately. Rakuten Cloud-Native Storage provides block, file, and object storage, and customers have the flexibility to use different cloud-native storage vendors with Rakuten Cloud-Native Platform if needed. Rakuten Cloud-Native Storage is application-aware and software-defined, going beyond mere storage provisioning. It considers the holistic lifecycle of data, including the running Kubernetes application configuration, metadata, and secrets, all managed in a single automated motion. This approach ensures data protection and maintains the exact running state of applications, which is crucial for scenarios like backup, snapshot, clone disaster recovery, and data migration. Traditional setups often lacked application-awareness in storage operations, but Rakuten Cloud-Native Storage bridges this gap effectively.

Rakuten Cloud-Native Orchestrator provides lifecycle management and observability across multiple operational domains from core to edge. It supports bare-metal servers with integrated power management, multi-distribution Kubernetes and K3s cloud platforms, network functions and applications for both VMs and containers, and services lifecycle management and design, including service stitching. Rakuten Cloud-Native Orchestrator also features a methods of procedures (MOPs) scripting engine that enables large-scale management of any device or appliance, such as switches, routers, sensors, cameras, and security appliances. This engine can utilize existing executors like Ansible and other scripting languages, supporting over 500,000 MOPs a day across 800,000 network elements.

All automation within Rakuten Cloud-Native Storage can be triggered automatically with a single click or via a policy engine. Operators can mix and reuse multiple workflow elements from any domain using the programmable and contextually aware built-in workflow engine. This allows for simultaneous preparation of servers, Kubernetes clusters, network devices, network functions/applications, custom service chains, and switch configurations in a GitOps model. Rakuten Cloud-Native Storage also supports full observability, including logging, monitoring, alerting, visualization and comprehensive power management across all domains and the entire solution.

Key technology areas for AI/ML success

GPU Support

Rakuten Cloud-Native Platform provides robust support for GPUs, which are essential for high-performance computing tasks in AI/ML. GPUs are the industry standard due to their superior processing power and efficiency. The platform's support for these GPUs ensures it can handle intensive computational tasks required by modern AI applications. The solution allows for seamless scaling of GPU resources to meet the demands of growing AI workloads.

Automated storage and workload placement

Rakuten Cloud-Native Platform offers an easy-to-use app store experience for deployment and lifecycle management. Everything is modeled and policy-driven, eliminating the need for manual configuration and reducing the risk of errors. The platform is Non-Uniform Memory Access (NUMA) aware, down to every physical node, providing higher granularity than other Kubernetes distributions. The solution can auto-discover and categorize different GPU models/types, ensuring that different workloads receive the appropriate GPU resources. This awareness allows for automated workload placement that takes into account workload placement GPU/CPU resources, storage placement, memory, application persistency needs, and more.

Standard Kubernetes NUMA awareness groups all resources in a worker node into one group without knowing which physical device it is on, which can lead to configuration failures or misconfigurations. In contrast, Rakuten Cloud NUMA awareness ensures precise resource allocation, preventing such issues. The platform supports affinity/anti-affinity across any component, CPU and CPU sibling isolations, multiple networks (overlay and underlay), and multiple IP addresses per pod/network function (NF).

Advanced networking

Rakuten Cloud-Native Platform offers advanced networking capabilities, including IP persistency across start, stop, heal, and migration operations. It supports SR-IOV underlay networks for high throughput, low jitter, redundancy, and network function (NF) interconnect. Open vSwitch underlays extend corporate operations networks to NFs, while Container Network Interface (CNI) plugin customization and per-pod multi-IP network support (Multus) provide flexible networking options. Calico overlays, NIC bonding for redundancy and throughput, IPv4 / IPv6 support, and a built-in MetalLB load balancer further enhance the platform's networking capabilities, ensuring robust and flexible connectivity for AI workloads.

Granular multi-tenancy and RBAC

Rakuten Cloud-Native Platform offers a potent combination of separation based on tenant, cluster, pod, container / VM, namespace, and cloud resource pools. Nodes can be grouped into resource pools for easy-to-use isolation and automated application placement. NUMA resources, CPUs, GPUs, storage, SR/IOV, memory, and other resources are auto-mapped to the pool, providing isolation at both the user and application levels. The platform can auto-place NFs onto nodes with the required resources, segregate nodes for certain applications and users, and map multiple tenants to a resource pool and vice versa. Further segregation within a resource pool can be achieved through Kubernetes tagging. Personalized chargeback for all resources, including GPU, CPU, memory, and storage, is also supported, ensuring efficient resource management and billing.

Monitoring and logging, deployment anywhere

Every operation Rakuten Cloud-Native Platform performs is policy-driven, fully modeled, and discovered, eliminating the need for hunting for the right server configuration, hardcoding, or CLI / coding experience, although these options are available if needed. Rakuten Cloud-Native Platform works in private, public, hybrid, and multi-cloud environments, eliminating both operations and resource silos for VMs and containers. The platform supports the deployment of entire multi-domain workflows at the click of a button, including bare metal server lifecycle management (LCM), NF and supporting application LCM, service stitching and LCM, and the deployment of appliances, switches, routers, and more. This comprehensive approach ensures seamless and efficient deployment and management of AI workloads across various environments.

Highly performant cloud-native storage

High-performing storage is crucial for AI / ML workloads because these applications require rapid access to vast amounts of data to train and infer models efficiently. Fast storage solutions minimize data retrieval times, enabling quicker iterations and real-time processing, which are essential for tasks such as image recognition, natural language processing, and predictive analytics. Additionally, high-performing storage ensures that data-intensive operations do not become bottlenecks, thereby maintaining the overall performance and scalability of AI/ML systems. This leads to more accurate models, faster insights, and the ability to handle larger datasets, ultimately driving better outcomes in AI / ML projects.

Rakuten Cloud-Native Storage performs at near bare-metal speeds across a wide variety of workloads. It supports block, file, and object storage, catering to any type of AI data from core to edge. The platform includes advanced data compression and thin provisioning, per-application replication policies for volumes across hosts and disks, and advanced security policies with volume encryption at rest and in motion. Built-in storage/resource pooling ensures high availability under node failure, providing a reliable and efficient storage solution for AI applications.

Advanced power management

Rakuten Cloud-Native Platform Kubernetes implementation incorporates an advanced power management solution designed to optimize energy efficiency across the entire network by meticulously monitoring and managing CPU power states, specifically C-States. This solution continuously tracks the C-States of CPUs, which represent various levels of power consumption and operational readiness, across all nodes in the Kubernetes cluster and across multiple data centers. By leveraging data-driven policies, the system can dynamically adjust these C-States, transitioning CPUs into lower power states or even sleep modes when they are underutilized. This capability ensures that power consumption is minimized during periods of low demand without compromising the performance and responsiveness of critical applications.

The power management solution is equipped with comprehensive monitoring and logging functionalities, capturing granular data on CPU usage, power states, and transitions. This data is stored and analyzed to provide a detailed view of the power consumption patterns and operational efficiency of the cluster. The system employs advanced algorithms to analyze this data, identifying trends and anomalies that could indicate inefficiencies or opportunities for further optimization.

In addition to real-time adjustments, the solution generates actionable insights and recommendations based on historical and current data. These suggestions are designed to help administrators fine-tune their power management strategies, balancing the trade-offs between energy savings and performance. For instance, the system might recommend adjusting the thresholds for transitioning between C-States or suggest specific times for scheduling maintenance tasks when the impact on power consumption would be minimal.

By implementing these recommendations, organizations can achieve significant energy savings, reduce operational costs, and enhance the sustainability of their IT operations. The power management solution also supports compliance with energy efficiency standards and regulations, providing detailed reports that can be used for auditing and certification purposes. Overall, this integrated approach to power management not only optimizes the energy efficiency of Kubernetes deployments but also contributes to the broader goals of operational excellence and environmental responsibility.

Platform and data footprint reduction

The edge footprint and data optimization are crucial because they directly impact the efficiency, scalability, and feasibility of deploying AI/ML workloads in edge environments. A smaller edge footprint ensures that AI/ML models can run on devices with limited computational resources, such as IoT sensors, mobile devices, and edge servers, without compromising performance. Data optimization techniques like thin provisioning and data compression reduce the amount of data that needs to be stored and processed, which not only conserves bandwidth and storage space but also accelerates data processing and transmission. This is particularly important in scenarios where real-time decision-making is critical, such as in autonomous vehicles, smart cities, and industrial automation.

In the case where data processing is not being done at the edge, data preprocessing and data migration to the cloud are critical components of modern data management strategies, particularly in environments where real-time decision-making and efficient data handling are paramount. Edge data preprocessing involves the initial filtering, aggregation, and transformation of raw data at the edge of the network, close to the data source.

This process reduces the volume of data that needs to be transmitted to the cloud, thereby minimizing bandwidth usage and latency. By performing tasks such as data normalization, anomaly detection, and preliminary analytics at the edge, organizations can ensure that only relevant and high-quality data is sent to the cloud for further processing and long-term storage.

Once preprocessed, the data is securely migrated to the cloud, where it can be integrated with other datasets, subjected to advanced analytics, and leveraged for comprehensive insights. This hybrid approach not only enhances the efficiency and speed of data processing but also optimizes resource utilization, ensuring that cloud infrastructure is used effectively while maintaining the agility and responsiveness of edge computing.

By optimizing the edge footprint and data handling, organizations can achieve faster, more reliable AI/ML operations, enhance data privacy and security, and reduce operational costs, making edge AI/ML deployments more practical and effective.

Conclusion

Edge analytics, a key component of this solution, is driving transformative changes across various industries by enabling real-time data processing and decision-making at the data source. From smart cities and healthcare to manufacturing, retail, energy distribution, industrial IoT, logistics, telecommunications, financial services, and agriculture, edge analytics is enhancing operational efficiency, reducing costs, and improving service quality.

Rakuten Cloud stands as a comprehensive and robust suite tailored for the demanding needs of AI/ML workloads. By integrating advanced features such as GPU support, automated storage and workload placement, high-performance cloud-native storage, and sophisticated networking capabilities, it ensures optimal performance and efficiency. The platform's granular multi-tenancy and RBAC, coupled with extensive monitoring and logging features, provide a secure and manageable environment for diverse applications.

Key technological areas such as GPU support, automated storage and workload placement, advanced networking, and power management are critical for the success of AI/ML workloads in edge environments. The platform's ability to optimize the edge footprint and handle data efficiently further enhances its suitability for edge deployments.

Rakuten Cloud-Native Platform, Rakuten Cloud-Native Storage, and Rakuten Cloud-Native Orchestrator together form a powerful suite that supports the entire stack from bare-metal to services management. This integrated approach ensures seamless deployment, management, and optimization of AI/ML workloads across core and edge environments, making Rakuten Cloud the ideal AI Kubernetes cloud solution for organizations looking to harness the full potential of AI/ML in a cloud-native ecosystem.

Rakuten Cloud

For more information, visit cloud.rakuten.com
800 Concar Dr, San Mateo, CA 94402
© Rakuten Symphony US