IAI-Assisted API Traffic Management and Anomaly Detection in High-Scale Commerce Platforms

Viplove Goswami

goswamiviplove@gmail.com

Abstract:
The quick development of high-scale commerce platforms makes us change from a fundamental to a more sophisticated way of traffic management. As these platforms evolve their architectures towards decentralized microservices, Application Programming Interfaces (APIs) are rapidly becoming the primary data delivery vehicles, which makes them prime targets for advanced automation and volumetric assault. This study investigates the infusion of Intelligent Artificial Intelligence (IAI) into the API management lifecycle to overcome the limitations of static thresholding capabilities and manual intervention. IAI-assisted frameworks utilize advanced machine learning models, Long Short-Term Memory (LSTM) networks for predictive scaling, and Random Forest ensembles for multi-layer anomaly detection to improve resilience of systems in a proactive manner. We look into context-aware monitoring systems that use dynamic knowledge graphs to correlate application-layer dependencies with infrastructure-layer resource constraints. Moreover, the study explores how service meshes and sidecar proxies can enable low latency edge inference using WebAssembly. By analyzing what industries do during the peak traffic event like a flash sale, this paper proves that IAI-driven optimization can help reduce infrastructure costs and enhance service availability. The results highlight the need for hybrid security orchestration, which combines the interpretability of deterministic logic with the adaptive analytic power of IAI to secure the future of global digital commerce.

Keywords: API Management, Intelligent Artificial Intelligence, Anomaly Detection, E-commerce Scalability, Predictive Auto-scaling, Bot Mitigation, Microservices, Zero-Trust Architecture.

INTRODUCTION

At this time, the global e-commerce sector is dramatically expanding digitally, where high-scale platforms are processing millions of concurrent transactions on a diverse range of devices and geographies. In this world, the API went from a technical integration tool to the architectural foundation of modern commerce. Nevertheless, traffic management and attack surface management are getting more complex and wider with an API-first, microservice-based architecture approach. The distributed systems' internal and external traffic patterns are getting increasingly volatile. As a result, traditional security perimeters no longer suffice.

E-commerce platforms face two challenges as they grow. They need the ability to quickly increase and decrease resource allocation in response to consumer demand. Furthermore, they need to identify threats automatically and in real time to prevent automated sophisticated attacks. Over 50% of all internet traffic is now from automated bots as malicious actors, now using AI to simulate human browsing patterns, circumvent age-old rate limiters that can block these bot actions and exploit business logic flaws. Regular defensive systems like WAFs and GACs largely depend on static human-programmed rules that are unable to keep up with the stochastic attack traffic that is common with modern applications. According to the specialist firm Arkose Labs, 1 in every 4 eCommerce transactions in 2022 has been identified as fraudulent.

To overcome these challenges, Intelligent Artificial Intelligence (IAI) has emerged as the new paradigm for managing API traffic. AI frameworks, unlike narrow AI systems that operate in silos with specific tasks, bring together deep learning, natural language processing, and context execution analytics to adaptively provide platform health and security as a whole. Systems reinforced by IAI move from "reactive" to "proactive", the

system expects traffic surges and unusual occurrence based on a learnt behaviour profile degrading from the learnt behaviour rather than any based threshold value.

The aim of the research is to deliver an extensive study of IAI on API traffic management and anomaly detection on the high-scale commerce site. We will first look at the API economy and the threat it currently poses to e-commerce. Next, we discuss the theory and implementation of machine learning models for traffic classification and predictive scaling. A detailed discussion of popular architectural patterns such as the service mesh and sidecar proxy follows, and how these are used to deploy IAI models at the network edge. Ultimately, we turn to the impact of IAI on operating speeds and costs, infrastructure costs and other regulatory issues. This paper suggests a conceptual and technical roadmap towards resilient, intelligent and secure commerce infrastructures from academic evidence and industrial studies.

The Evolution of API Architecture and the Rise of IAI

The evolution of computer science and distributed networking is where API development dates back to. Early on in computing, it was recognised that making machines more 'intelligent' would be one of the major tasks of computing. This would often require creating complex programs and design of higher-level machine languages. In the late 1990s, the industry began to no longer require monolithic systems but were opting for a "Virtual Enterprise" model. Therefore, the demand for an IT infrastructure that was standardized, flexible and extensible increased. Projects like the ESPRIT venture VEGA focused on architectural basements of open distributed frameworks for heterogeneous enterprise environments.

The shift from the early protocols CORBA, COM/DCOM, Java/RMI to the new RESTful APIs and gRPC has been driven access to company data from all platforms and any location. As the number of API calls grew in the e-commerce boom of the 2000s, manual methods to manage this traffic became impossible. Due to numerous microservices, the API sprawl entails that automatic integration and smart routing is needed.
The Intelligent Artificial Intelligence (IAI) framework is different from ordinary automation. IAI is employing cutting-edge technology to transform API traffic. To create self-healing systems that adapt to changing conditions, IAI is using machine learning and predictive analytics. In an IAI-mediated terrain, the integration platform serves as a 'live responsive central nervous system' that continuously harvests telemetry to constantly improve upon performances and security. This transition means businesses can use data for more than just reporting, and start making decisions based on predictive and prescriptive intelligence rather than just descriptive analytics.

The Modern Threat Landscape in High-Scale Commerce

High-scale commerce platforms operate in a hostile digital environment where the boundary between legitimate and malicious activity is constantly blurred. The surge in automated attack networks has reshaped the threat landscape, making Distributed Denial of Service (DDoS) and bot mitigation core availability challenges.

Bot Traffic and Automated Exploitation

Bot traffic has hit all-time high recently, with automated requests accounting for about 50% of all HTML requests. During busy times, bot traffic may exceed human traffic by as much as 25 percentage points. The bots are classified or categorized based on their intention and sophisticated level. Essential for SEO and market transparency, search engine crawlers and price aggregators are "good" bots. However, the more "bad" bots are 31% of the total traffic to e-commerce.

Malicious automated systems exploit specific weaknesses in the trade process. Bots for price scraping at scale enables competitors to negate pricing strategies in real-time. Scalper bots buy limited-edition products in bulk as soon as they are made available so that real customers can't buy. Moreover, the so-called "sneaky" bots take part in the manipulation of referral programs and cart abandonment which distorts the business's metrics and creates disruptions in the inventory. Defenders face the challenge that almost 60% of these bad bots imitate human behaviour such as non-linear navigation and varying request frequencies, making them extremely hard to detect using the same signature-based methods.

Layer 7 and Economic DDoS Attacks

Modern Distributed Denial of Service (DDoS) attacks are multi-layered. For instance, DDoS attacks are targeting application and API layer. The backend work that is triggered as a result of these attacks is costly as the pages are loaded continuously or require complex authentication flows which eat up all resources and cause the server to slow down or the application to break.

The "economic" or "cost-exhaustion" DDoS attack is growing in popularity. In contrast to the conventional aim of service disruption, these attacks are intended to degrade performance and increase infrastructure costs unnoticed by leveraging auto-scaling to provide additional resources. These attacks are formulated with requests per second that usually deviate only slightly from the expected usage pattern. So, a “normal” traffic baseline is quite hard to establish. As a result, creating a reliable analysis of probable attack traffic is a major bottleneck for defenders. IAI-assisted systems can uniquely address this issue by investigating how requests evolve over time and whether their interaction with particular endpoints strays from the long-term historical average for that service.

Business Logic and Identity Vulnerabilities

Payment information and personal data exposed due to e-commerce API could be subject to authorization injections and authorization attacks. SQL injection continues to be a serious problem. This is when unfiltered customer data is submitted directly to the database. APIs can also be exploited through Server-Side Request Forgery (SSRF) attacks, where malicious actors manipulate the API to access internal resources.

Credential stuffing and Account Takeover (ATO) attacks that utilize botnets to verify stolen credentials against login endpoints. The financial losses and chargebacks from these attacks are compounded by the stakes of eroding customer distrust and regulatory penalties. The traditional implementation of rate limiting based on IP addresses does not work here as the attacker rotates identities using residential proxies. To achieve effective protection, practitioners must begin to shift behaviour-based analysis towards the user’s journey context taken in the platform.

IAI Frameworks for Traffic Management

To manage the volatility and complexity of high-scale commerce traffic, IAI frameworks implement proactive strategies that optimize both performance and security.

Predictive Auto-scaling and Capacity Planning

The prediction of auto-scaling is one of the major applications of IAI in traffic management. A Kubernetes-based system typically leverages a reactive mechanism. For example, the Horizontal Pod Autoscaler (HPA) modifies a workload’s replicas based on current CPU or memory usage. Reactive scaling is often insufficiently prompt to address sudden surges and cause other ‘cold start’ delays, and SLO violations.

Predictive scaling assisted by IAI makes use of modelling by ARIMA (Auto-Regressive Moving Average) or Long Short-Term Memory (LSTM) networks utilising prediction of traffic. These models assess historical workload traces. Often they require at least 24 hours of metrics bleed to identify patterns. By preparing for demand fluctuations, the system can provision resources ahead of the traffic hitting, which can help businesses reduce their costs by 30% and provide systems uptime of 99.5%. This difference can mean the difference between stability and outage during major event days such as Black Friday, which sees millions of dollars being spent every minute by consumers.

Intelligent Traffic Shaping and Prioritization

Traffic shaping is the management of requests to prevent congestion and match the infrastructure’s capabilities. In an IAI-enabled environment, it achieves “intelligent degradation,” the slowing down or deferral of non-essential features to maintain the core checkout flow stable. For instance, in the event of database overload the IAI agent may lessen search results or postpone recommendation calculations.

AI-based routing analyzes present conditions in the network. It measures loads on the systems to route API requests dynamically. This guarantees that all processing of business transactions is prioritized based on their

value. Moreover, in the case of organizations that leverage third-party APIs that are priced according to usage, IAI can help in the elimination of unnecessary calls, and instead, batch the requests and minimize operational expenditures.

Context-Aware Anomaly Detection Methodologies

To this end, we propose a novel model that is designed to capture the communication structure of microservice systems. Static thresholds for individual services in traditional methods do not take into account a fault in one service propagating to another.

IAI methods usually make use of dynamic knowledge graphs (KGs) for the modeling of those relationships. By viewing the system as a static graph, each created at time intervals, IAI can keep track of the situation in the monitored environment. Thus, the model can identify whether a service failed because of a code bug or because nodes were unavailable to process the request due to resource shortage at the infrastructure layer. Consistently outperforming non-contextual baselines, this holistic view detects system faults in dynamic settings like autoscaling events.

Machine Learning Models for API Anomaly Detection

The technical implementation of IAI-assisted anomaly detection relies on a suite of machine learning algorithms, each suited to different aspects of traffic analysis.

Multi-Layer Detection Architectures

Research suggests that a two-layered approach to API anomaly detection is highly effective for securing microservice architectures.

Layer 1 operates at a generalized level, targeting large-scale, high-efficiency attacks like zero-day DDoS events. This layer typically employs a RandomForest model, which has demonstrated an accuracy of 91.63% in identifying broad malicious patterns across real-time datasets. Layer 1 focuses on volumetric features such as request frequency, data-out rates, and the number of unique API endpoints accessed per minute.

Layer 2 is application-specific and acts as a fail-safe for attacks that bypass the generalized checks of Layer 1. This layer is tailored to the local application's scale and resources and often uses ensemble methods like Bagging (with Decision Tree classifiers). Experimental results show that Layer 2 can achieve accuracies as high as 97.77% in detecting subtle malicious behaviors within specific API traffic. By combining these two layers, the system provides comprehensive coverage across both the network and application layers.

Recurrent Neural Networks and LSTM for Temporal Patterns

Traffic on commerce platforms is inherently sequential in nature and therefore has temporal dependencies which a simple classifier would not be able to capture. Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) that are capable of learning long-term dependencies. The temporal behavior of users during a session can be effectively tracked with LSTM models. By understanding the normal human shopper's 'tempo' (time spent on the product page before adding to cart), LSTM-based systems can pick up automatic bots that flow too purposefully (and/or too fast) through the conversion funnel.

According to the literature on healthcare IoT and e-commerce traffic, accuracy and precision scores over 97% can be achieved with hybrid deep learning model LSTM. Seasonal traffic spikes create noise in the data. These models can better accommodate that noise as they understand that context.

Isolation Forests and Unsupervised Learning

In many high-scale environments, labeled datasets of malicious activity are not always available, especially for zero-day threats. Isolation Forests provide an unsupervised alternative by focusing on isolating anomalies rather than profiling normal behavior. This algorithm reinterprets intricate commerce interactions into clear, interpretable decision trees. Each tree can correspond to a specific Quality of Service (QoS) dimension, such as responsiveness, availability, or security. This approach is highly scalable and allows businesses to pinpoint unusual patterns in customer feedback or system logs in real-time.

Transformers and Natural Language Processing (NLP)

The use of Transformers the architecture behind today's large language models will start to revolutionize API security with granular NLP. IAI-assisted pipelines do not look only at request metadata for similar requests but instead applies NLP on URIs. This applies especially for traffic that is poorly represented in publicly available datasets, such as traces dominated by JSON over high-entropy fields or Base64 encoding.

Transformers can leverage long sequences of time-series data, making them promising candidates for NPD predictions (next purchase day) and complex business logic exploitation that requires chaining together multiple API calls. A "Returnformer" model is proposed to predict product return behaviour by capturing user-product connections in a bipartite graph. It helps in efficient inventory management for e-commerce platforms.

Architectural Integration: Gateways, Meshes, and Sidecars

The deployment of IAI models within a high-scale architecture must be carefully managed to avoid introducing excessive latency or operational complexity.

API Gateways as Centralized Entrances

An API Gateway serves as the "front door" for external traffic, providing a centralized point for enforcing global policies like authentication and basic rate limiting. Gateways are effective at handling "North-South" traffic—the flow between clients and the platform. IAI can be integrated here to provide edge-level analytics and risk scoring for every incoming request. By analyzing edge metrics, the gateway can track which APIs are called, by whom, and with what frequency, providing a first line of defense against volumetric attacks.

Service Mesh and Sidecar Proxies for Internal Security

As requests traverse the gateway to the microservice cluster, it is the east-west traffic that becomes the focus. A Service Mesh resolves the problem by adding a sidecar proxy to each service replica. Such proxies (often based on Envoy) intercept all the traffic between services within the application, enabling fine-grained manipulation and policy enforcement without changing applications.

Service meshes can create strong workload identities and encrypt the communication between services through mutual TLS (mTLS), ensuring no service trusts another by default. IAI can be inserted into these sidecars, attempting to catch and tag internal API calls for suspicious activity, like lateral movement by an attacker who has compromised a single service. This architectural pattern makes observability uniform across heterogeneous services, allowing platform teams to enforce uniform security policies irrespective of programming languages.

WebAssembly (Wasm) and Edge Extensibility

To perform real-time anomaly detection at the sidecar level, developers are increasingly turning to WebAssembly (Wasm). Wasm provides a lightweight, sandboxed runtime with near-native performance, allowing custom IAI logic to be injected into the proxy as "filters". These Wasm modules can perform complex tasks like payload inspection and behavioral scoring within the network path itself, minimizing the need for expensive "call-outs" to external security services.

The advantage of using Wasm is that it supports polyglot development; IAI models trained in Python or C++ can be compiled to Wasm and run within the Envoy data plane. This facilitates a modular approach to security, where specific "filters" can be dynamically added or removed based on the current threat level or traffic conditions.

Ambient Mesh and Sidecar-less Architectures

Despite its power, the sidecar model introduces significant resource overhead, as each proxy requires dedicated CPU and memory. Emerging trends like Istio's "Ambient Mesh" aim to solve this by moving the data plane to shared, node-level proxies. Ambient Mesh splits functionality into a lightweight L4 "ztunnel" per node and optional L7 "waypoint" proxies for advanced routing.

In some scenarios, this sidecar-free design can reduce mesh CPU and memory overhead by up to 90%, making it inherently more scalable for high-scale platforms. IAI frameworks must adapt to this shift by providing node-level intelligence that can efficiently analyze traffic for multiple application pods simultaneously.

Practical Applications and Industrial Case Studies
The real-world impact of IAI-assisted traffic management is most evident during large-scale commerce events.

Black Friday and Seasonal Surge Management
During events like Black Friday, E-commerce sites get bombarded with traffic. During peak hours, up to 63 million users hit e-commerce sites. Just one minute of downtime can cost a retailer an average of $5,600 not accounting for reputation damage. Companies leveraging IAI for predictive auto-scaling can withstand spikes in workload three times larger than their normal baseline without performance degradation.

IAI's anomaly detection technology is quite effective here since it understands that "normal" behavior changes during holidays. For instance, vacation shoppers take longer to browse, add more items to carts, and make more database queries. IAI models can identify a legitimate shift as opposed to the bot activity trying to camouflage itself in all the noise.

Bot Mitigation and Revenue Protection
Industrial implementations of IAI-driven bot management, such as those by Radware or Akamai, use cross-correlation and global attacker feeds to detect sophisticated automation. These tools interrogate the client early—often at the CDN edge—to determine its intent. By differentiating between human sessions and machine behavior, digital teams can remove synthetic traffic from their analytics, ensuring that product teams do not optimize for fake behavior.

Studies show that platforms utilizing AI-powered recommendation systems and bot filtering have seen improvements in conversion rates by up to 31%. By shedding abusive traffic that overloads bandwidth, these systems keep pages fast for genuine shoppers, directly influencing customer satisfaction and loyalty.

Financial Services and High-Stakes APIs
In financial services—including payment processing and trading platforms—the requirements for performance and security are exceptionally stringent. Integrated IAI-driven optimization has been shown to yield substantial improvements over traditional methods, reducing latency and enhancing security. By analyzing vast amounts of operational data, IAI systems can identify patterns and make autonomous decisions that reduce operational costs while maintaining the high reliability required for digital banking.

Implementation Challenges and Best Practices
While the benefits of IAI are clear, its implementation in high-scale environments presents several hurdles.

Data Quality and Pipeline Orchestration
The quality of the training data ultimately limits the performance of any IAI model. A model is likely to make inaccurate predictions if the dataset has duplicates or incomplete entries. Organizations should spend a considerable amount of money on data preprocessing and cleaning. We must remove irrelevant fields (like request to static resources), normalize user-agent strings, and so on.

IAI monitoring should be integrated into an overarching platform for training, validation and interpretability of model. With the help of workflow execution engines like Airflow, we can automate monitoring tasks such as retrieving score distributions from the database and computing performance metrics at suitable intervals.

Explainability and Transparency (XAI)
The "black box" nature of some deep learning models can be a barrier to adoption, especially in regulated industries where analysts must understand why an alarm was triggered. This is where Explainable AI (XAI)

techniques like SHAP and LIME are essential. XAI simplifies complex models into human-readable rules, highlighting the specific factors (like a login from a new IP or a high request frequency) that contributed to a detection. This not only helps analysts validate alerts but also ensures compliance with global data protection regulations like GDPR, which may require companies to explain automated decisions.

Balancing Latency and Precision

Every security check added to the API path introduces latency. Rule-based models exhibit lower computational overhead, whereas IAI-based systems can add a marginal delay due to complex anomaly detection processing. Platform architects must choose the right enforcement point—whether at the CDN edge for quick wins or deep within the service mesh for granular context—to balance security needs with the user experience.

Future Directions: 6G, Autonomous Agents, and Beyond

The future of API traffic management will be shaped by the continued advancement of IAI and the evolution of the underlying network infrastructure.

Security in 6G and IoT Ecosystems

As we move toward 6G environments, the integration of IAI with advanced machine learning models will be necessary to secure complex IoT ecosystems. These environments will require even more resilient postures against adversarial attacks, where IAI must ensure effective decision-making transparency and alignment through recursive feature elimination.

The Rise of Agentic AI

We are entering an era of "agentic" AI, where autonomous agents use APIs to orchestrate end-to-end processes across siloed applications. These agents rely heavily on APIs for real-time decision-making, creating new integration challenges and potential security concerns. IAI frameworks will need to evolve to not only detect malicious bots but also to manage and authorize legitimate AI agents that act on behalf of human users.

Ethical Governance and Global Standards

The increasing adoption of IAI necessitates robust governance frameworks to ensure that AI is developed and deployed ethically and transparently. Regulatory approaches are evolving from industry self-governance toward "hard law," such as the EU AI Act, which requires risk classification and transparency for AI systems. International collaboration will be essential to harmonize standards and address cross-border challenges as AI technologies transcend national borders.

Conclusion

With the high-scale commerce platforms, managing API traffic does not lie in the hand of human-built manual and rule-based systems anymore. The research states that Intelligent Artificial Intelligence (IAI) offers a holistic and effective solution through proactive adaptive and context-aware traffic orchestration. IAI-assisted frameworks provide a notable enhancement in platform resilience against economic DDoS and highly automated threats through deterrence based on behavioural anomaly detection.

In addition, current architectural patterns, such as service meshes and sidecar-less ambient modes, allow IAI to execute security enforcement at low latency and high speed at the edge. The cost advantages achievable through predictive auto-scaling, as well as the reliability gains offered in peak period events, suggest a measurable economic impact of these systems. For IAI to work successfully, there needs to be a strong emphasis on data quality, and the use of explainable AI among other measures. Plus, more emphasis should also be on intelligent rules with hybridization.

With the accelerating decentralization of commerce as well as the increasing usage of autonomous AI entities, the need for IAI-assisted management will grow. The paper provides a framework for firms to help them in the scaling of digital platforms in an escalating uncertain threat environment. Infrastructures that can learn

and adapt in real-time to threats will have a significant impact on consumer experience. The future of global commerce rests on intelligent infrastructures that will ensure stability and trust.

REFERENCES:

1. Soffer, D. (2026). DDoS Protection Faces Fresh Challenges As Bot Traffic Reaches New Peak. *IT Security Guru*.
2. Zuplo. (2025). API Security in E-commerce APIs. *Learning Center*.
3. Codilar. (2024). The Growing Threat of Bad Bots in eCommerce and How to Fight Back.
4. Milvus. (2025). Difference between rule-based and AI-based anomaly detection.
5. ArXiv. (2025). Paradigms of industrial monitoring: From rule-based to data-driven. *arXiv:2509.15848*.
6. ResearchGate. (2024). Comparative Study of AI-Powered vs Rule-Based Authorization Models for Zero-Trust API Security in Azure.
7. MDPI. (2024). Rule-Based and Signature-Based Approaches to Cloud Service Protection. *Applied Sciences*.
8. Palo Alto Networks. (2024). AI in Threat Detection: Evolution and Core Capabilities. *Cyberpedia*.
9. ResearchGate. (2021). Systematic Review of API Gateway Patterns for Scalable and Secure Application Architecture.
10. ResearchGate. (2023). API Traffic Anomaly Detection in Microservice Architecture.
11. Azati AI. (2024). ML and AI for Seasonal Traffic Scaling in E-commerce.
12. International Journal of Scientific Research and Modern Technology. (2023). AI-Driven Predictive Analytics for Customer Retention in E-Commerce.
13. WJAETS. (2025). Artificial Intelligence Approaches to Optimizing API Platforms in Financial Services.
14. Virtana. (2025). What is Service Mesh? Data Plane and Control Plane Patterns.
15. Boomi. (2024). AI applies technological innovation to API traffic routing.
16. Journal of Theoretical and Applied Electronic Commerce Research. (2025). AI-Driven Anomaly Detection in E-Commerce Services.
17. Flatline Agency. (2025). 20 AI Concepts Shaping eCommerce in 2025.
18. MDPI. (2024). Dynamic Resource Quota Auto-scaling Framework for Kubernetes.
19. International Journal of Multidisciplinary and Current Research. (2021). Optimizing Cloud Resource Management via Dynamic Auto-Scaling.
20. The SAI Organization. (2021). Predictive Scaling for Elastic Compute Resources. *International Journal of Advanced Computer Science and Applications*.
21. Software: Practice and Experience. (2025). K8sidecar: A Modular Kubernetes Chain of Sidecar Proxies.
22. Medium. (2026). Beyond Kubernetes: Platform Engineering Trends for 2026.
23. IJSRED. (2025). Intelligent Artificial Intelligence–Based System for Automated Decision Support.
24. MDPI. (2024). Returnformer: A Graph Transformer for Return Prediction in E-commerce. *Entropy*.
25. PMC. (2024). Computational Framework for Predicting Business Sales Using Transformers.
26. MDPI. (2023). Transformer-Based Model for Predicting Customers' Next Purchase Day. *Computers*.
27. IFIP. (1999). A Framework for Distributed Information Management in the Virtual Enterprise. *Advances in Information and Communication Technology*.
28. IEEE. (2026). Context-Aware Anomaly Detection Methodology using Dynamic Knowledge Graphs in Microservices.
29. ResearchGate. (2023). Experimental Results of Two-Layer API Traffic Anomaly Detection.
30. MDPI. (2024). Context-Aware ML/NLP Pipeline for Real-Time Anomaly Detection in Cloud API Traffic.