

Future-proofing ecommerce:
**Unbx's scalable
and resilient
infrastructure for
peak performance** ✨



Highly adaptable MACH architecture ✨

Microservices

Our functionalities and business logic are built, deployed, and managed as decoupled microservices using Kubernetes. Each service can be scaled and updated independently, ensuring high flexibility and resilience.

API-first

All of Netcore Unbx's services are accessible via APIs. We ensure seamless integration and interaction with external services and systems. Inter-service communication is handled using gRPC with Protobuf for efficient, high-performance messaging.

Cloud-native SaaS

Netcore Unbx is a cloud-native and cloud-agnostic solution that leverages the full potential of cloud computing. Deployed using Kubernetes, we utilize Argo, a CNCF project, to automate and manage workflows. Our communication layer is powered by NATS.io, another CNCF project, enabling high-performance messaging. Additional services such as FluxCD for continuous



delivery, KubeFlow for machine learning workflows, Envoy for edge and service proxy, and Prometheus for monitoring and alerting enhance our cloud-native capabilities.

Headless

Netcore Unbx employs a headless architecture where the front end and back end are decoupled. The front end, a combination of React and native JavaScript, communicates exclusively with the backend through REST APIs. This separation allows for greater flexibility in developing and deploying user interfaces.

Salient features of Netcore Unbx infrastructure ✨

Multi-tenant

Netcore Unbx's infrastructure supports multi-tenancy, allowing multiple businesses to share the same infrastructure while maintaining data isolation and security. This design ensures cost efficiency and streamlined resource management without compromising on performance or privacy.

Highly available

We guarantee high availability and golden uptime, ensuring that your ecommerce platform remains operational even during peak traffic periods and unexpected disruptions.

Our infrastructure is designed to offer continuous uptime, with automated failover mechanisms and self-healing capabilities that keep your services running smoothly 24/7.

Note: Netcore Unbx is the only search provider with zero downtime over the past 10 years.

Scalable and elastic

We leverage Kubernetes for auto-scaling clusters that dynamically adjust to traffic demands, ensuring optimal performance. Our platform is designed to handle and scale effortlessly to accommodate sudden surges during the sales season and steady traffic increases.

60_{ms}

response time for 95%
of queries

150+_{Mn}

requests served per
day

5000+

peak RPS

500+_{Mn}

events processed per
day

Global coverage and multi-region infrastructure ✨

Data center locations

We have strategically positioned five data centers in key regions such as the US, UK, Singapore, and Australia. This strategic placement allows us to cater to customers and users worldwide, providing low-latency access and optimal performance regardless of their geographic location.

Edge locations and VPC peering

In addition to our primary data centers, we have established ten edge locations with inter-region VPC peering. This architecture enhances our network's resilience and ensures high availability by distributing traffic efficiently.

Co-located index with geo load balancing

We implement geo-load balancing techniques to distribute traffic intelligently across our co-located indexes.

Last-mile coverage with Cloudflare CDN

To further optimize content delivery and user experience, we leverage Cloudflare CDN for last-mile coverage. This integration ensures that content is delivered swiftly to end-users, regardless of location, while enhancing security and mitigating potential cyber threats.

24 K/sec

observed peak indexing rate

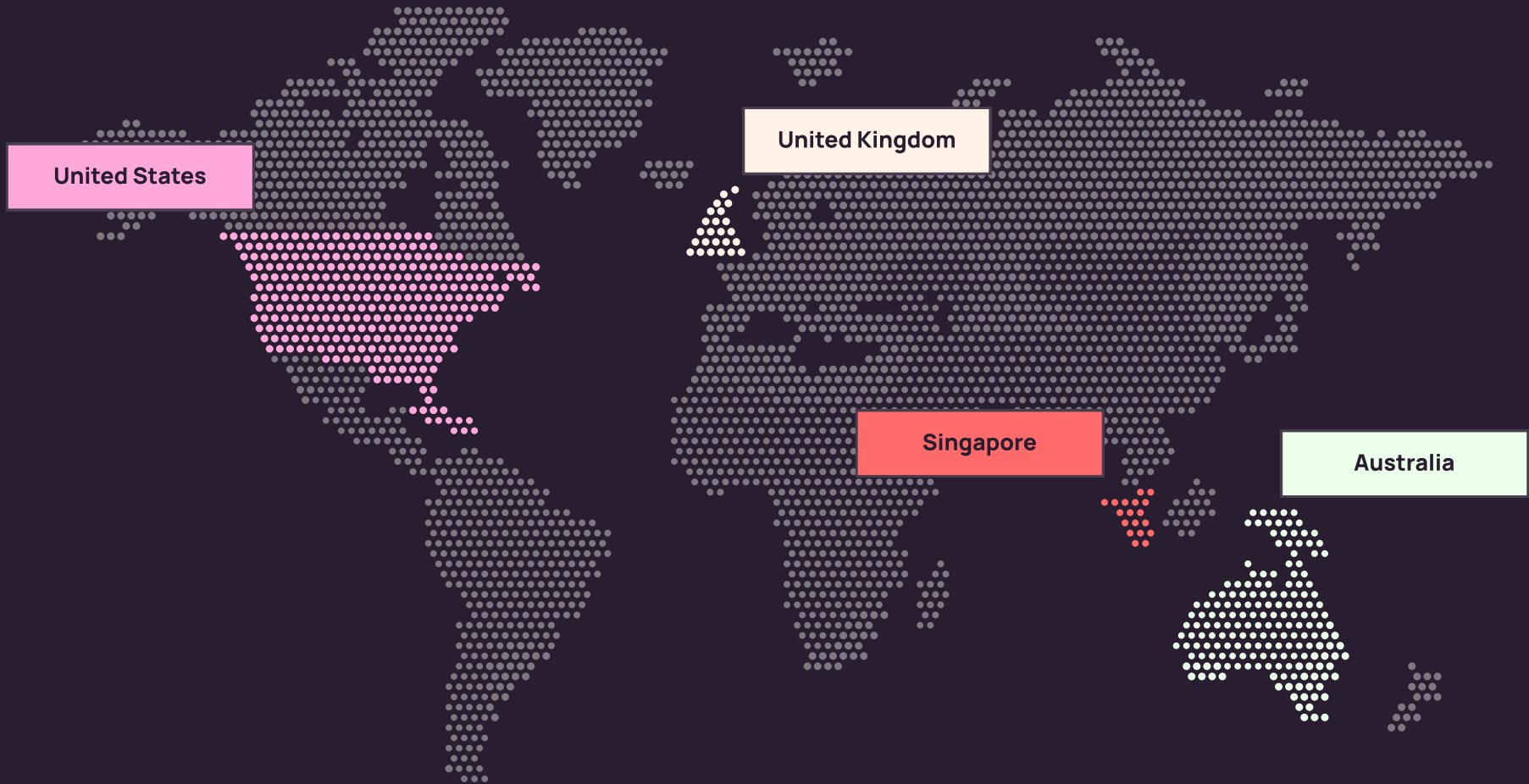
35+ Bn

active products in the Index

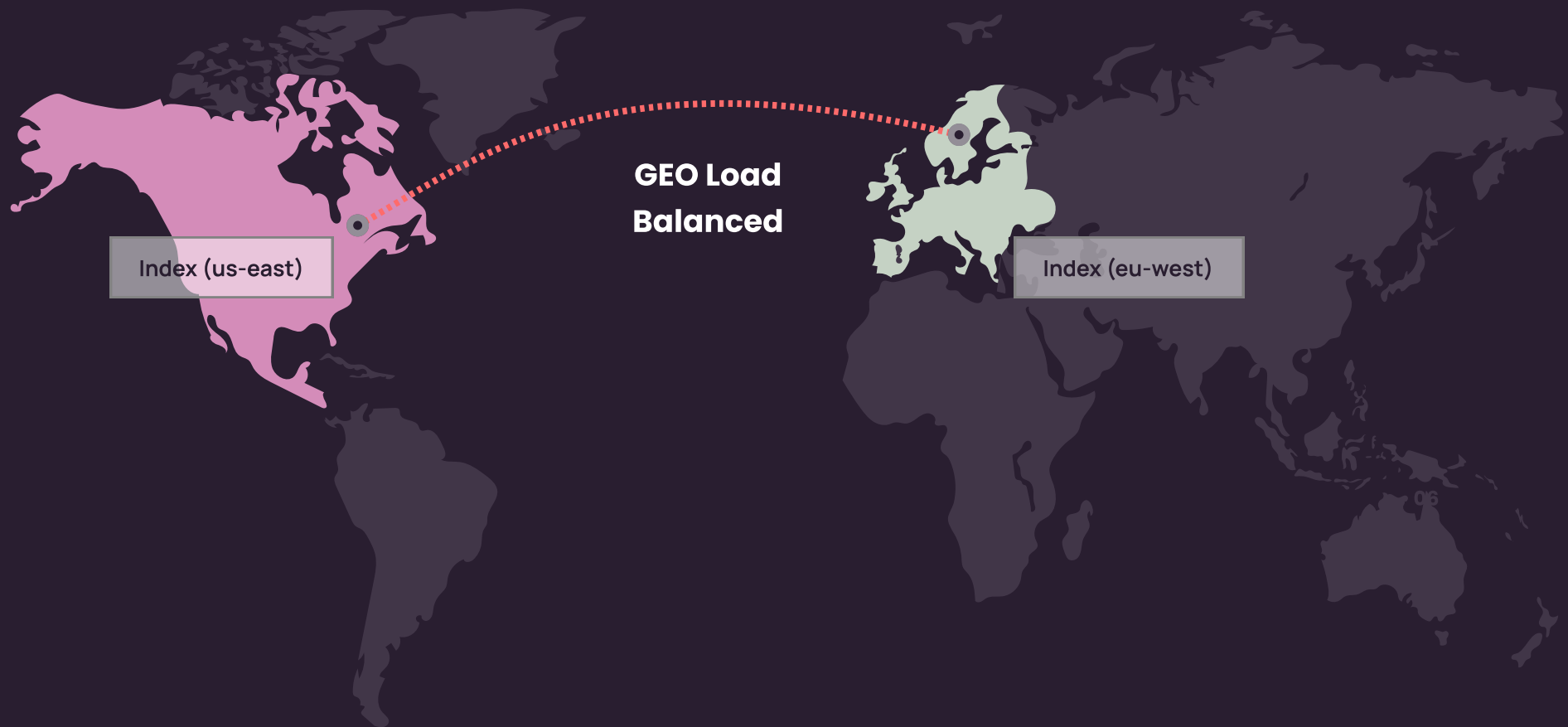
1.1+ Bn

documents processed per month





Co-located index with geo load balancing



Catalog pipeline ✨

Powered by Argo and Kubernetes

Performance and scalability

The catalog pipeline is tested to handle 24,000 operations per second per workflow, showcasing its ability to process many tasks concurrently. It employs an elastic architecture, allowing it to adapt seamlessly to varying workloads and catalog sizes.

Customizable processing workflow

The pipeline offers a customizable processing workflow, enabling businesses to tailor operations according to their specific requirements and data processing needs.

Workflow stages

The workflow comprises several stages, including analysis, aggregation, enrichment, indexing, merging, and cleanup. Each stage is meticulously designed to handle specific tasks in the catalog processing journey, ensuring the accuracy and completeness of data transformations.

Fault tolerance and auto failure detection

With built-in mechanisms for auto failure detection and retry at every workflow stage, we ensure operation continuity even in case of unexpected failures.

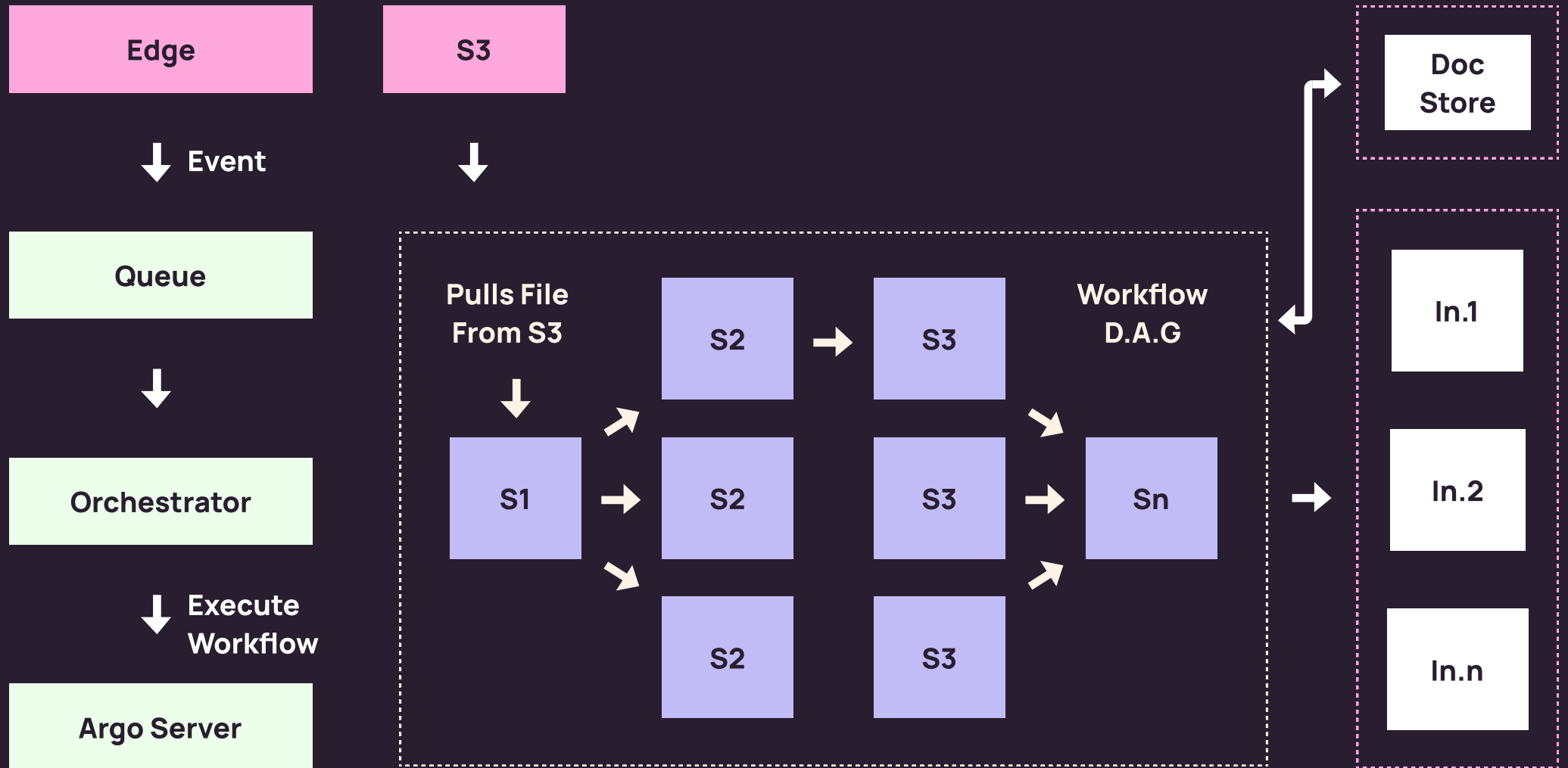
Last-mile coverage with Cloudflare CDN

To further optimize content delivery and user experience, we leverage Cloudflare CDN for last-mile coverage. This integration ensures that content is delivered swiftly to end-users, regardless of location, while enhancing security and mitigating potential cyber threats.

Catalog ingestion speed

Our infrastructure can process over 30,000 items in 3 minutes, 300,000 items in 6 minutes, and 3 million items in 10 minutes. This rapid ingestion capability ensures that catalog updates are processed swiftly, minimizing data availability delays.





Argo + Kubernetes



On-demand scaling ✨

Netcore Unbx's infrastructure features dynamic scaling for both nodes and containers. Nodes, the underlying computing units, scale by adding or removing resources based on workload needs. Containers, encapsulating applications, adjust their numbers to match demand, ensuring efficient resource utilization and responsiveness to workload changes.

Whenever there's a requirement, additional computational resources are provisioned as needed to handle workload spikes or increased processing requirements.

Horizontal Pod Autoscaling (HPA)

HPA is a key feature that allows microservice containers to scale dynamically based on resource utilization metrics such as CPU or memory usage. It ensures that a service's number of running pods (containers) adjusts automatically to meet current demand, maximizing resource efficiency and maintaining performance levels during varying workloads.

Node-level autoscaling

Our cluster architecture incorporates node-level auto-scaling, where the underlying infrastructure scales nodes on demand,

based on workload requirements. This ensures that the cluster can handle increased processing demands without manual intervention, optimizing resource allocation and reducing operational overhead.

Instance flexibility

Based on workload characteristics and cost optimization strategies, our infrastructure's workflows can leverage different instances, including Spot, Reserved, and On-Demand.

GPU instance provisioning

Our infrastructure supports on-demand provisioning of GPU instances for AI and ML workflows that require GPU acceleration. This capability enables high-performance computing for AI and ML tasks, enhancing model training, inference, and data processing capabilities.



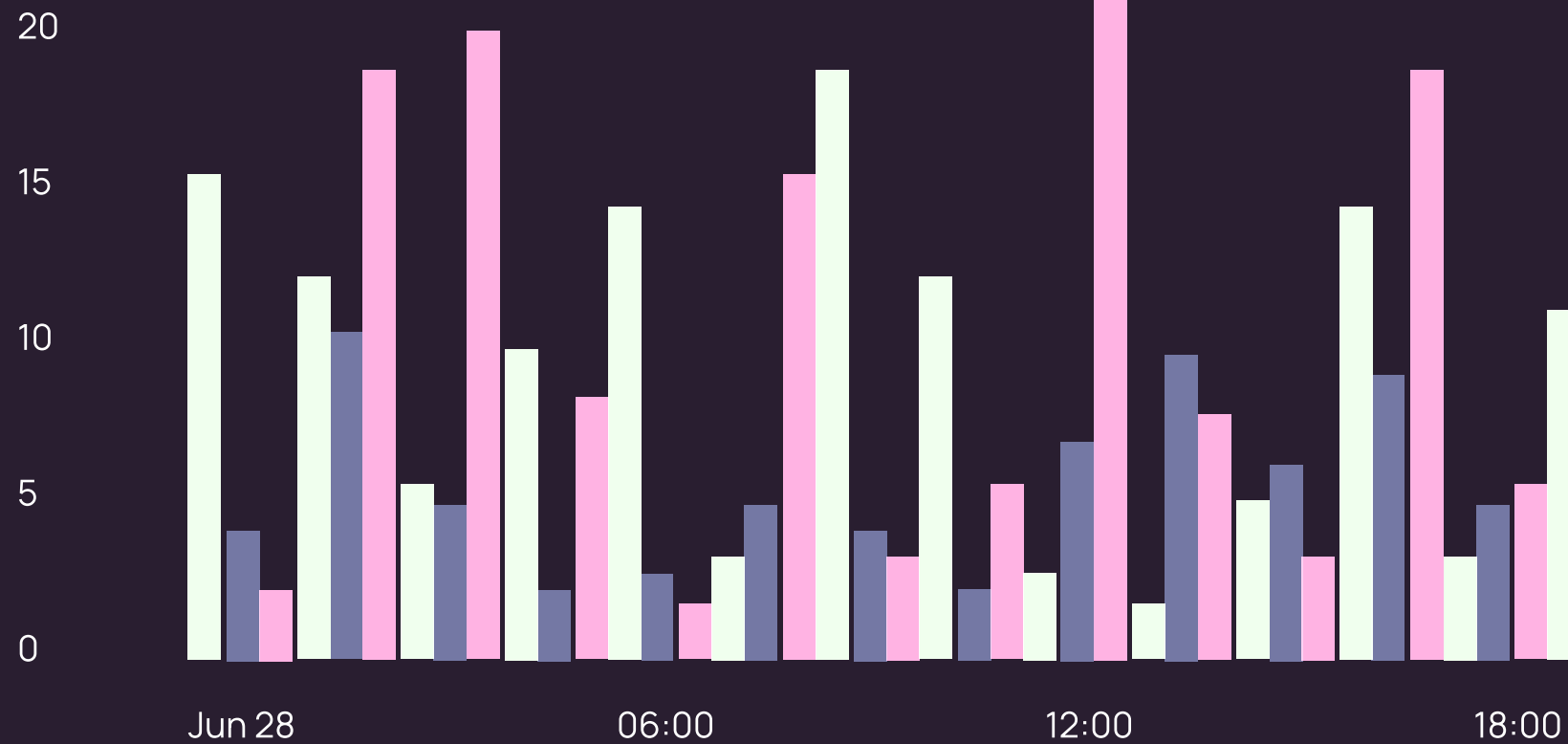
Node Activity

6 hours

1 day

7 days

- Reserved Running
- On-Demand Running
- Spot Running
- Stopped Instances



Index pipeline ✨

The Index Pipeline within Netcore Unbx's infrastructure is a critical component designed to efficiently handle indexing requests across various stages, ensuring robust performance and scalability. Powered by Argo and Kubernetes, the pipeline is designed to scale horizontally based on,

Product count

As the product count increases, the pipeline dynamically allocates resources to efficiently process and index the growing volume of products.

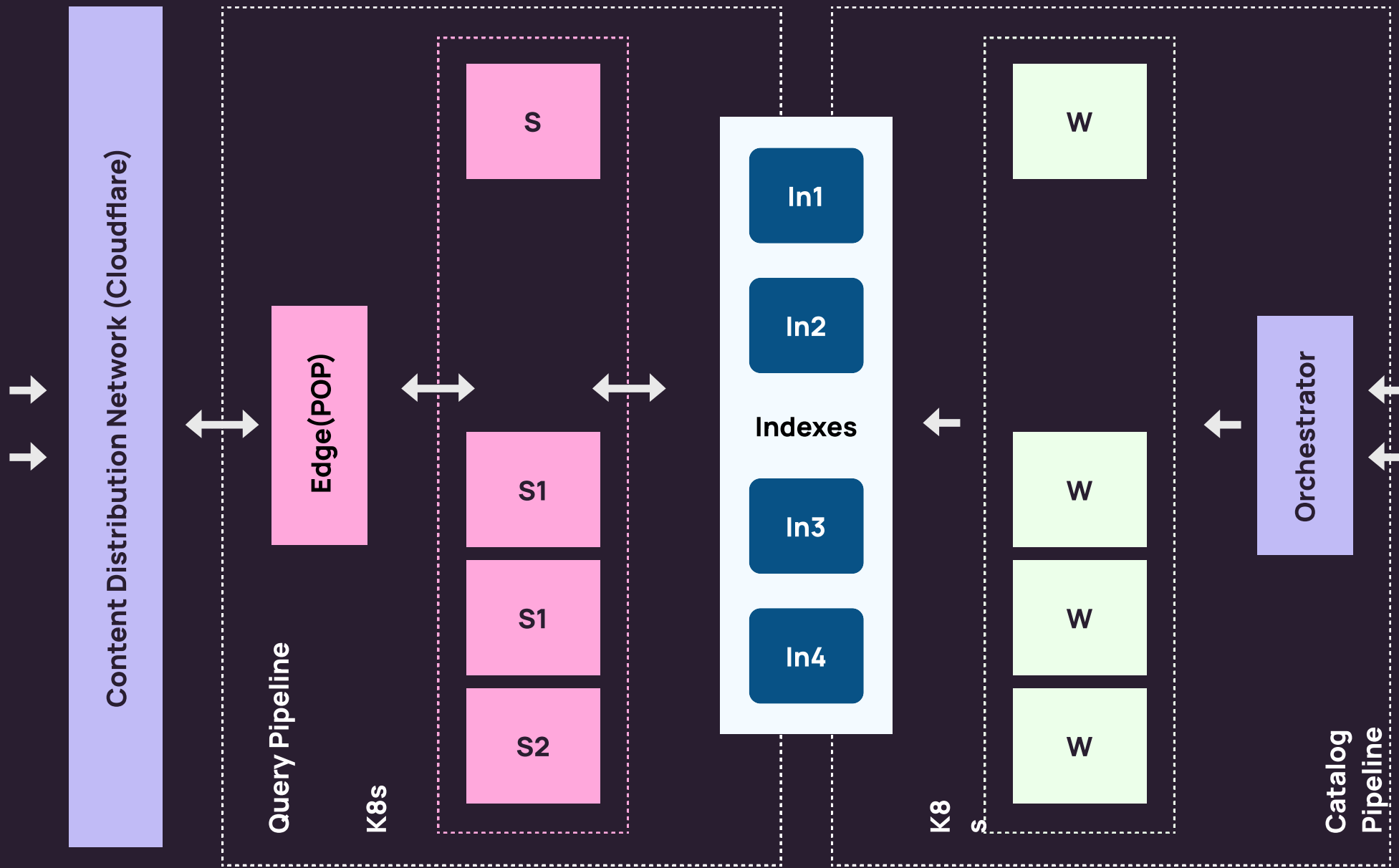
Number of variants per document

Another factor influencing the scaling of the pipeline is the number of variants associated with each product document. Variants can include different sizes, colors, or other product attributes. The pipeline intelligently adjusts its processing capacity based on the complexity introduced by these variants.

Size of catalog files

The size of the catalog file also plays a crucial role in determining the pipeline's scaling requirements. Large catalog files require more computational resources for indexing, and the pipeline dynamically scales up or down to accommodate these varying file sizes.





Query pipeline ✨

Service DAG with built-in parallelism

By organizing tasks into a directed graph without cycles, we can orchestrate parallel processing of queries and data operations (using the Service-Directed Acyclic Graph (DAG) structure). This parallelism optimizes resource utilization within the pipeline, allowing multiple tasks to execute concurrently. As a result, query response times are significantly reduced, leading to faster and more responsive search experiences for users.

Pluggable Modules for AI, Merchandising

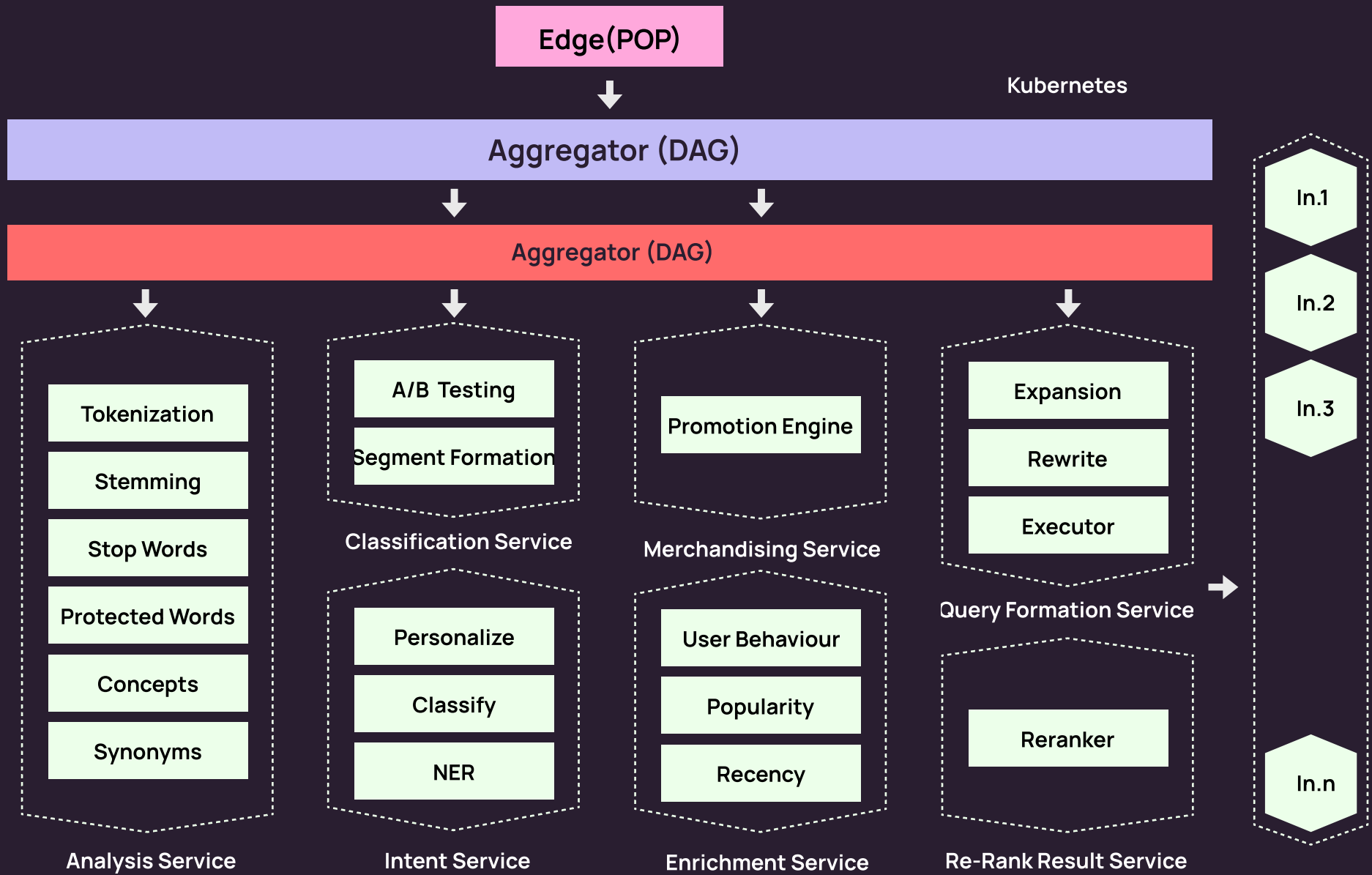
Our query pipeline incorporates pluggable modules specifically designed for AI-driven capabilities and merchandising strategies. These modules are flexible and extensible, enabling businesses to seamlessly integrate advanced AI algorithms for query understanding, semantic search, and personalized recommendations.

By leveraging these modules, businesses can enhance search relevance, improve product discovery, and deliver tailored customer experiences, ultimately driving higher engagement and conversions.

Service Level Scaling based on System Metrics and Telemetry

Netcore Unbx's query pipeline has intelligent service-level scaling mechanisms that respond dynamically to real-time system metrics and telemetry data. This ensures optimal resource allocation across different pipeline stages, adapting to query volumes and complexity fluctuations. We maintain consistent performance and availability by automatically scaling resources such as computing, memory, and storage based on workload demands, even under varying traffic patterns or peak loads. This scalability and responsiveness are essential for delivering a seamless search experience and handling spikes in user queries during high-traffic periods or promotional events.





Analytics pipeline ✨

Our system is designed to handle a high volume of events while providing real-time insights and data enrichment capabilities for AI models.

High throughput and event capture

The Analytics pipeline has been rigorously tested to handle over 20,000 events per second, ensuring seamless processing of massive data streams. It captures a wide range of signals, including user interactions, clickstream data, purchase behaviors, and more, totaling 120+ signals contributing to comprehensive customer understanding.

Real-time data availability for AI models

Our Analytics Pipeline has the ability to provide real-time data availability for AI models. This means that AI algorithms and machine learning models can access fresh, up-to-date data streams immediately, enabling timely decision-making and personalized user experiences.

Robust session and experiment resolution

Our solution incorporates advanced session management and

experiment resolution capabilities. It precisely tracks user sessions, behavior patterns, and experimental variations, facilitating in-depth analysis, A/B testing, and performance evaluation of different strategies and features.

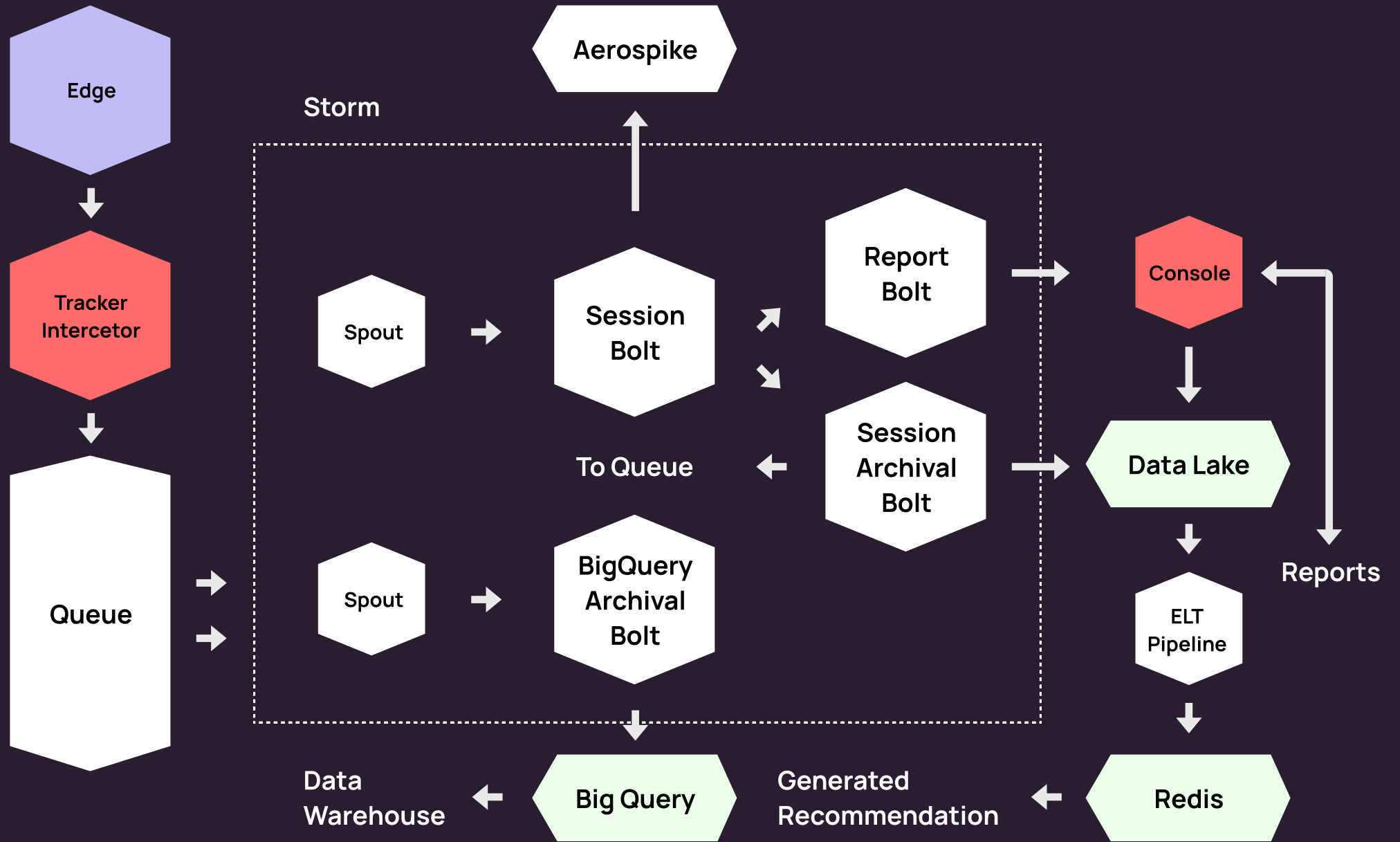
Data warehousing and data lakes

Our analytics integrate with data warehousing solutions for aggregated data storage and data lakes for raw data storage. This architecture allows for efficient data management, storage, and retrieval, supporting both historical analysis and real-time processing needs.

Data enrichment

Our pipeline includes powerful data enrichment functionalities that can enhance the value of raw data. It supports experimentation, product attribution, and user profile enrichment, enabling businesses to gain deeper insights into customer preferences, behavior trends, and product performance.







Auto failure detection and retry

Built-in mechanisms for auto failure detection and retry ensure robustness and reliability across every stage of the analytics process. In case of failures or disruptions, the pipeline automatically detects issues, initiates retries, and ensures data integrity and continuity without manual intervention.

Reliability

Granular tolerances

Services deployed on Kubernetes

- Guarantees Uptime, Reliability & Fault Tolerance
- Automated Rollouts & Rollbacks
- Self Healing for Services & Cluster itself

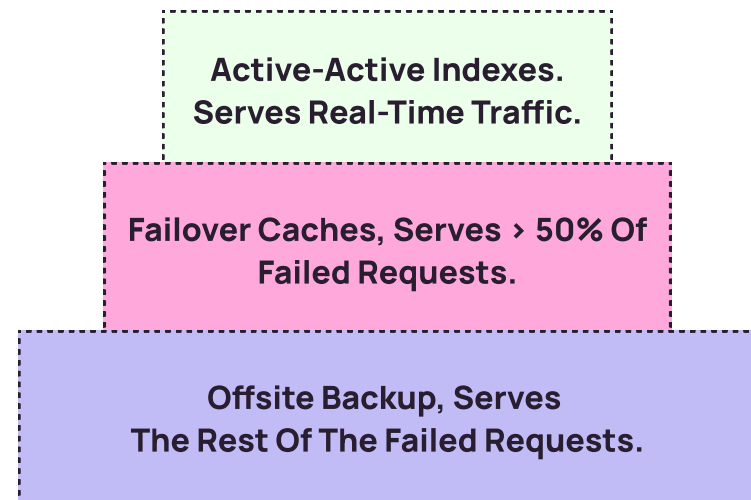
Self healing indexes

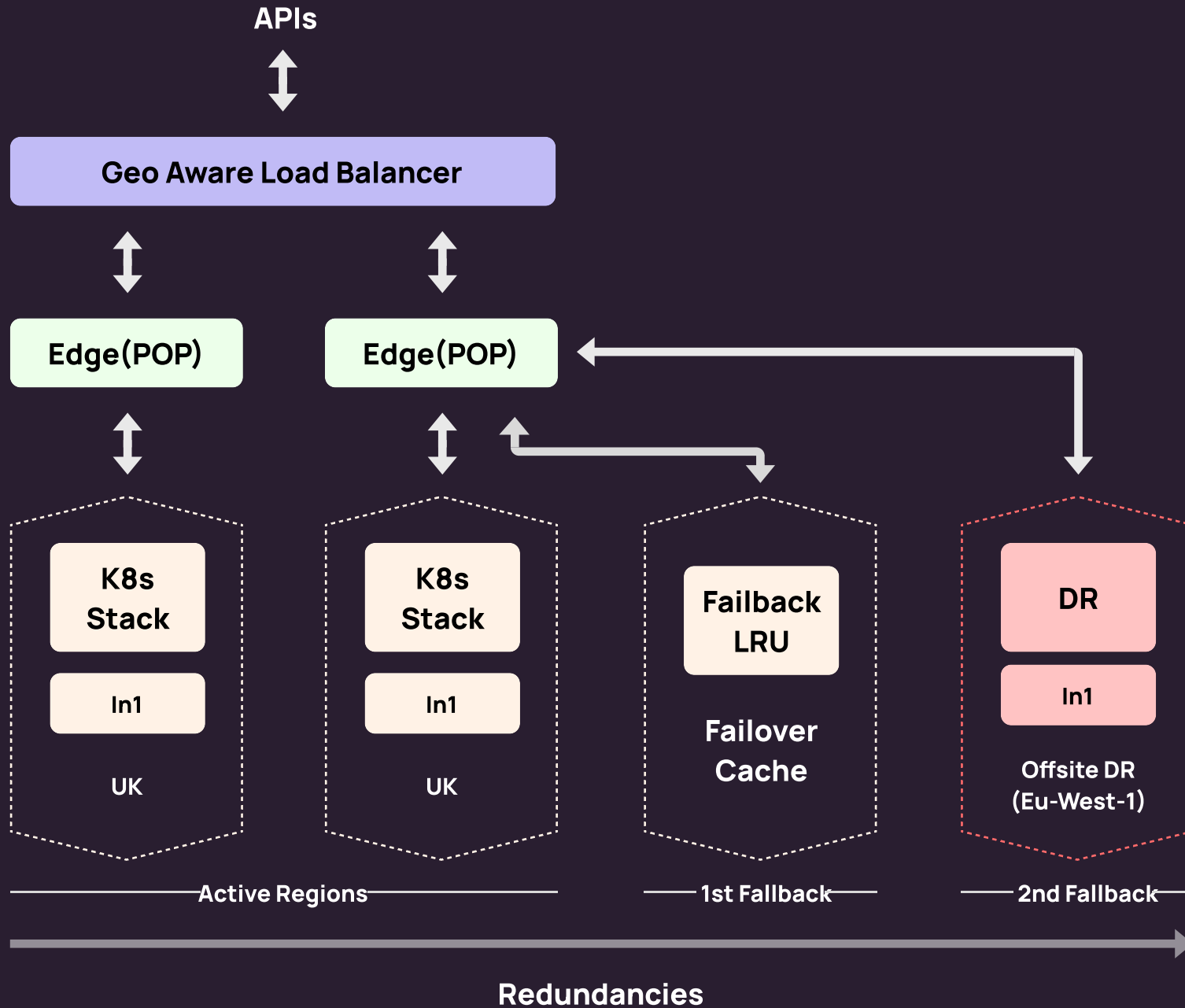
Our self-healing index management solution is a powerful tool designed to handle the complexities of varying index sizes seamlessly. It has the capability to manage indexes ranging from hundreds to millions of documents efficiently.

The system ensures custom index distribution and replication strategies, which not only handle failures but also proactively self-heal to maintain optimal performance. This feature contributes significantly to the reliability and resilience of our infrastructure, ensuring uninterrupted service delivery even under challenging conditions.

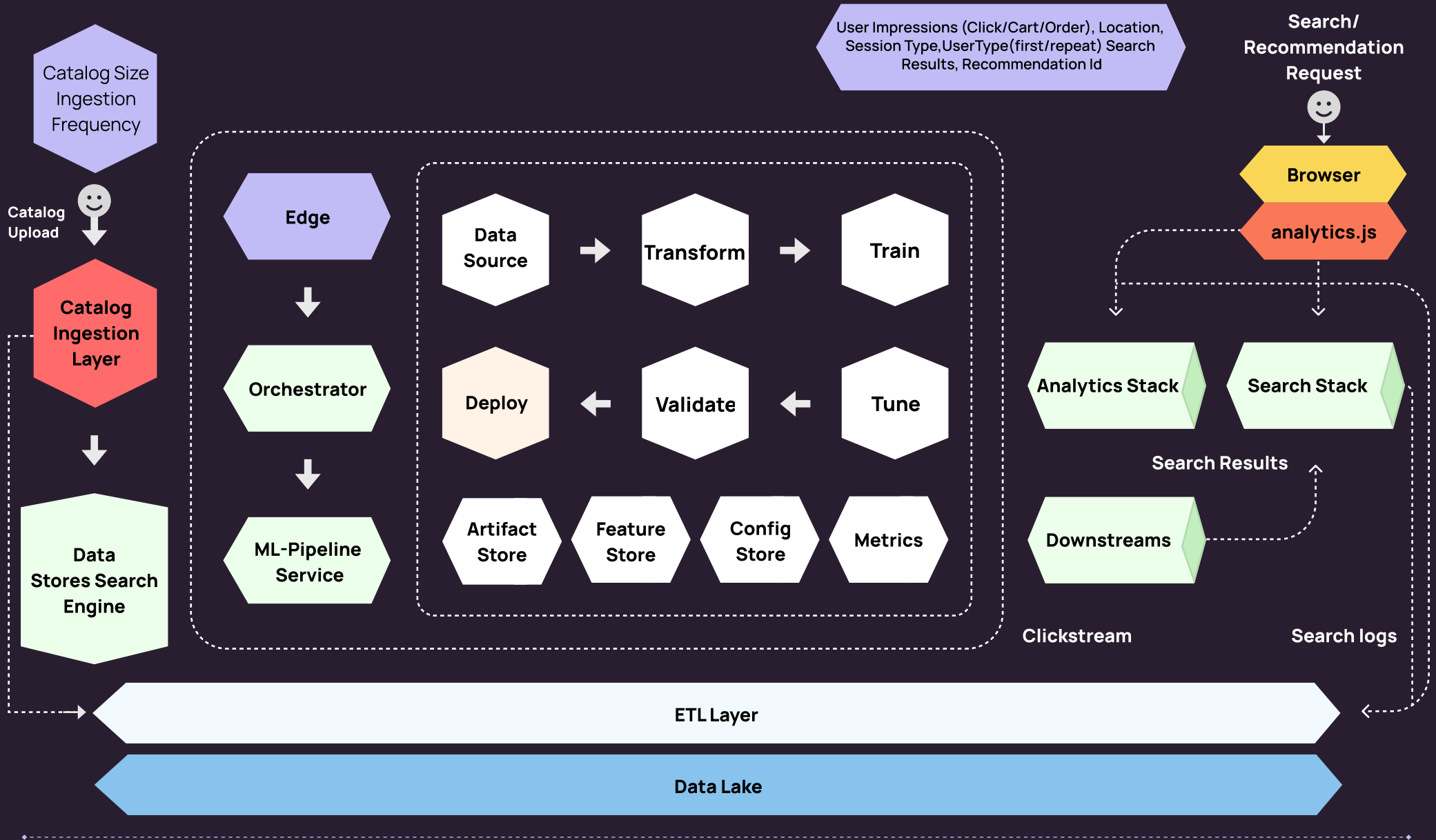
Redundancy

Multiple Levels of Redundancy of Data. Handles failures in real-time





Netcore Unbx AI infrastructure



Feed + Status API

Facilitates the management of product feeds and provides real-time status updates on data ingestion and processing, ensuring data accuracy and timeliness.

Search + Browse API

Enables powerful search and browsing capabilities, empowering users to discover products efficiently through customizable search queries and navigation options.

External Score API (Personalization)

Integrates external scoring mechanisms for personalized recommendations, enhancing customer engagement and conversion rates.

Autosuggest API

Enhances user convenience by providing real-time autocomplete suggestions based on user input, improving search efficiency and user satisfaction.

Analytics API

Offers insights into user behavior and system performance. It enables advanced analytics and reporting functionalities, empowering data-driven decision-making.

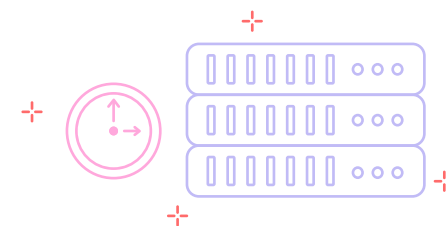
Recommends API

Leverages machine learning algorithms to generate personalized product recommendations based on user preferences and behavior, driving upsells and cross-sells.

Performance metrics

Netcore Unbx.com handled unprecedented traffic levels during the holiday season


- Peak RPS 14% higher than 2021 BFCM sale
- 12% year-over-year increase in traffic during the Black Friday sale
- 100% uptime across all holiday seasons for the past ten years





About Unbx

Netcore Unbx is an AI-powered platform that helps brands provide personalized customer experiences to scale online exponentially. Our commitment to revolutionizing ecommerce experiences has garnered us esteemed recognition, positioning us as a leader in Gartner® 2024 Magic Quadrant™ for Search and Discovery and the Forrester Wave™: Commerce Search and Product Discovery, Q3 2023 report.

Contact Us

 1710 S. Amphlett Blvd
Suite 124 San Mateo,
CA 94402

 sales@unbx.com
support@unbx.com

 +1 (650) 282-5788
