SRE-Ready NDC Shopping: Caching at Scale Without Pricing Drift

Presented by Mukul Kumar Gaur
Independent Contributor

Scaling Challenges

Addressing NDC Shopping Demands

Growing Demand

The increase in NDC shopping requests from multiple channels such as metasearch engines and OTAs places immense pressure on existing platforms, necessitating innovative solutions for management and scalability.

Real-Time Pricing

Implementing full real-time pricing logic significantly impacts latency and infrastructure load, leading to increased cloud costs and potential degradation of user experience, which must be carefully managed.

Latency and Load

Real-Time Challenges

The substantial volume of search requests places immense strain on infrastructure, leading to latency issues. This can severely affect user experience and operational costs.

Caching: Key to Scalability and Performance

Load Reduction

Caching minimizes the Offer Engine’s workload significantly.

Fast Responses

Achieving sub-second response times enhances user experience.

Demand Handling

Caching enables stability during sudden demand increases.

Caching Strategies

Three Approaches for NDC Efficiency

Full Offer Caches

Full offer caches provide complete, precomputed offers, allowing for rapid retrieval. This approach significantly enhances speed, making it ideal for stable market conditions and high-demand scenarios.

Hybrid Anchored Caches

Hybrid anchored caches blend precomputed data with dynamic elements, maintaining fresh pricing while ensuring quick access. This method offers flexibility, balancing performance and data integrity for evolving markets.

Full Offer Caches

Optimizing Performance and Reliability

Fastest Response

Full offer caching enables the quickest retrieval of precomputed offers, drastically reducing response times and improving user satisfaction during the booking process and peak traffic periods.

Simplified Logic

This caching strategy simplifies retrieval logic by serving complete offers directly, minimizing the need for complex backend calculations, which can ultimately lead to improved operational efficiency.

Hybrid and Partial Caches

Balancing Freshness and Speed

Hybrid Caches

Hybrid caching strategies anchor key data points, allowing for improved speed while maintaining freshness, ideal for dynamic pricing environments with frequent updates and fluctuations in demand.

Partial Caches

Partial component caching focuses on specific elements like fare rules and taxes, enhancing granularity and minimizing stale data risks, ensuring accurate pricing information during high traffic periods.

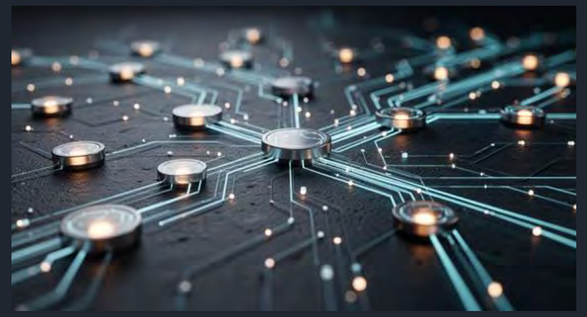
Ensuring Pricing Integrity and Stability

Freshness Thresholds

Set limits to prevent stale data exposure.

Real-Time Invalidation

Trigger updates based on pricing or inventory changes.



High-Elasticity Handling

Manage fluctuations during peak demand periods effectively.

Continuous Pricing Alignment

Ensuring Accuracy and Trust

BI Monitoring

Leveraging business intelligence-driven monitoring dashboards allows teams to track cache hit rates effectively, identifying potential discrepancies in pricing and ensuring data accuracy in real-time.

Automated Alerts

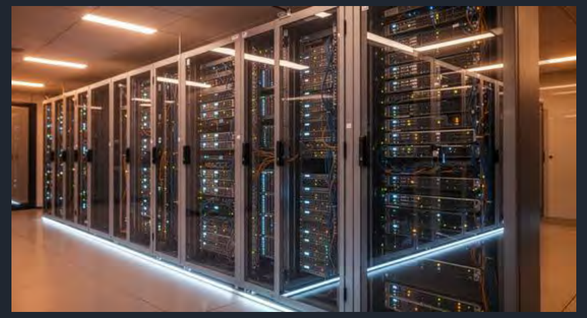
Implementing automated alerts provides immediate notification of any pricing drift, enabling rapid responses to discrepancies and maintaining customer trust while ensuring operational stability during peak demand periods.

Peak Season Promotion

Effective Caching Strategies

During a high-demand promotional event, caching strategies and governance controls ensured seamless airline operations, without pricing errors or downtime, showcasing the system's reliability and effectiveness.

Best Practices for SRE-Ready Caching



Layered Caching

Implement multiple caching strategies to optimize performance.

Strict Policies

Establish clear cache freshness and invalidation guidelines.

Real-Time Monitoring

Utilize BI tools to track performance and alignment.

Thank You!