International Journal of General Engineering and Technology (IJGET)
ISSN (P): 2278–9928; ISSN (E): 2278–9936
Vol. 11, Issue 2, Jul – Dec 2022; 175–198
© IASET

International Academy of Science, Engineering and Technology
IASET
Connecting Researchers; Nurturing Innovations

IMPROVING LATENCY AND RELIABILITY IN LARGE-SCALE SEARCH SYSTEMS: A CASE STUDY ON GOOGLE SHOPPING

Suraj Dharmapuram[1], Rakesh Jena[2], Satish Vadlamani[3], Dr. Lalit Kumar[4], Prof. (Dr) Punit Goel[5], Dr S P Singh[6]

[1]Suraj Dharmapuram, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213,

[2]Scholar Biju Patnaik University of Technology, Rourkela, Bhubaneswar, Odisha 751024,

[3]Osmania Universit , Amberpet , Hyderabad, Telangana State, India

[4]Asso. Prof, Dept. of Computer Application IILM University Greater Noida

[5]Maharaja Agrasen Himalayan Garhwal University, Uttarakhand,

[6] Ex-Dean, Gurukul Kangri University, Haridwar, Uttarakhand ,

ABSTRACT:

As e-commerce grows, optimizing latency and reliability in large-scale search systems becomes critical for delivering seamless user experiences and maximizing business potential. This case study examines Google Shopping's approach to reducing latency and enhancing reliability across vast datasets and high user volumes. Google Shopping's search infrastructure must address the dual challenges of processing a high volume of queries with low latency while ensuring high availability and minimal service disruptions. This research outlines the technical approaches taken by Google Shopping to tackle these issues, emphasizing a combination of infrastructure optimization, algorithmic advancements, and architectural shifts.

At the infrastructure level, Google Shopping leverages geographically distributed data centers and strategically designed caching mechanisms to ensure data locality and quick response times. These systems employ a hierarchical caching structure to reduce the number of data fetches needed from backend storage, decreasing load on primary databases and reducing user-facing latency. At the software layer, Google Shopping utilizes a combination of query rewriting and ranking algorithms optimized for performance and relevance. By prioritizing popular queries and frequently accessed products in their indexing structure, the system minimizes response time for high-demand items.

The case study also discusses reliability strategies, which encompass fault tolerance and failover mechanisms. Google Shopping employs replication and redundancy protocols to handle traffic spikes and manage potential system failures. Load balancers distribute traffic across multiple replicas, and these replicas are continually monitored to detect anomalies and initiate failover when necessary. The implementation of "graceful degradation" techniques ensures that, even in the event of partial system failures, the search service can continue to operate with reduced functionality rather than failing entirely.

This case study also explores how Google Shopping mitigates the “tail latency” problem, where a small percentage of queries experience significantly higher latency. Using specialized queuing techniques, the search system reroutes high-latency queries through optimized channels to balance load and reduce delays. Additionally, machine learning models are applied to anticipate potential performance bottlenecks based on historical data and usage patterns, allowing for preemptive adjustments to system configurations.

KEYWORDS: *latency, Reliability, Search Systems, Google Shopping, Caching, Load Balancing, Fault Tolerance, Tail Latency*

Article History

Received:* *13 Nov 2022* | *Revised:* *20 Nov 2022* | *Accepted:* *22 Nov 2022