

Scaling Enterprise Authorization

SUMMARY

A global manufacturing organization needed a faster, more scalable way for internal applications to evaluate organization-level permissions, feature toggles, and licensing. Existing systems struggled with high request volumes, resulting in service failures impacting downstream applications, and preventing new development. We delivered a platform that centralized business rule evaluation and reduced or eliminated dependency on slower legacy services.

APPROACH

Our team developed a centralized API that enabled applications to evaluate permissions and licensing without directly querying multiple backend systems. The focus was on improving performance, scalability, reliability, and simplifying secure access while supporting growing cross-team demand.

- **High-Performance Caching:** Developed a distributed caching strategy with automated cache invalidation to reduce dependency on downstream services while maintaining data accuracy.
- **Scalable Rule Evaluation:** Built a centralized rules engine that evaluated permissions and licensing on behalf of consuming applications, eliminating the need for most downstream applications to directly access sensitive dependency data.
- **Performance and Resiliency:** Optimized caching, scaling, and service architecture to support high request volumes while minimizing the impact of dependency outages and latency.
- **Cloud Native Architecture:** Leveraged AWS, Kubernetes, Java Spring, Open Policy Agent, and Redis-based caching to deliver a resilient, highly scalable solution.

SOLUTION

The team delivered a centralized API that provides fast, reliable access to licensing, permissions, and feature toggle information from multiple enterprise systems. The platform securely accesses the underlying data and evaluates it against centralized policies, allowing most consuming applications to receive the decisions they need without requiring direct access to sensitive systems. Intelligent caching and automated cache updates reduced reliance on slower backend services while simplifying development and supporting future growth.

Results

77.5%

77.5% reduction in P95 response time, improving from 91.8 ms to 20.7 ms

18%

18% faster than the fastest existing dependency while outperforming slower services

Up to 14x

Up to 14x faster than existing enterprise services used for permissions and licensing