Buyer's Guide for Enterprise Caching

Criteria and methodology to help you find the right enterprise caching solution

redis

Table of Contents

What Makes Cache Enterprise-Grade?

Why do you need real-time enterprise caching?

Caching improves application response time by storing copies of the most frequently used data on ephemeral but high-speed storage (RAM). In-memory caching solutions, which hold the working set in fast RAM instead of slow spinning disks, can effectively achieve these goals, often achieving response times under one millisecond.

While caching is commonly used to improve application latency, a highly available and resilient cache can also help applications scale. Offloading responsibilities from the application's main logic to the caching layer frees up compute resources to process more incoming requests. Enterprise caching will provide a sub-millisecond response under an extremely high workload with high volumes of data across global deployments that require five-nines availability. Since it's an integral part of mission-critical application architectures, it has a higher consistency, availability, performance, scalability, and geo-distribution capabilities than standard caching.

Caching has a proven track record of helping companies of all sizes improve application performance. In fact, many companies depend on caching to meet their customers' expectations and SLAs. Caching should be considered mission critical when you and your company rely on caching to meet key business goals or deliver key business services. When your caching is mission critical, enterprise-grade caching is required.

Research into human response indicates applications have roughly 100 milliseconds (ms)—one-third of the time it takes to blink—before users feel like they're waiting for a response. To be considered real time, a feature or service needs to send a request, have it processed, receive the response, and return it in less than 100ms.

Databases need to respond in under one millisecond to ensure the application's end-to-end round trip response time is 100 milliseconds.

What Makes Cache Enterprise-Grade?

Why and when enterprise cache is important

Digital applications and online businesses are hardly new, but recent events have accelerated the growth of digital transactions and online interactions. When operating online, whether through a web browser or mobile device, the user experience is vitally important. And a critical pillar that shapes the experience is response time.

The importance of response

In a Salesforce study, 83% of consumers stated that the experience a company provides is just as necessary as its products and services. On top of that, when page load times get above one second, users lose interest, making it easy for them to leave. According to research conducted by Unbounce, 70% of consumers surveyed declared that page loading times directly impact their willingness to buy from online retailers. For instance, consumers expect sites providing "instant calculations" of real estate property market values, credit card risk, or loan commitments to occur within a second.

The stakes mount with mobile experiences. According to a Deloitte Digital study of mobile sites conducted across a sample of 37 retail, travel, luxury, and lead generation brands, a mere 100-millisecond improvement in load time resulted in conversions growing by 8% for retail sites and 10% for travel sites. For obvious reasons, mobile gaming depends on real-time interaction, requiring responses within 100 milliseconds to make the experience immediate.

While your business may not go down completely if the cache is unavailable, the examples above clearly highlight how your business will be negatively affected by direct revenue loss or by a hit to your business' reputation.

What Makes Cache Enterprise-Grade?

Tenets of enterprise-grade cache

The ability to cost-effectively enable high availability, high performance, simple scalability, and geo-distribution for your cache is critical. To truly be considered enterprise-grade, your cache needs to meet these criteria:

High availability

If your cache goes down and your applications don't provide the performance expected, it means broken SLAs and potentially huge negative business impacts. Digital applications today need to be available 24/7, 365 days a year, which means your cache also needs to be highly available.

Performance

Users expect high-throughput and sub-millisecond performance as a matter of course. Not only do you need to meet these standards, you need to meet them with every service across the board. As shown above, there is a tremendous risk to your business' reputation if your cache goes down and you no longer provide real-time responses.

Scalability

A high-performance cache should easily scale across zones, regions, and geographies and meet the demand arising from business growth or sudden surges such as Valentine's Day and Black Friday, or from unexpected shifts in user behavior such as those caused by natural disasters or pandemics. Additionally, scalability should be done dynamically without causing downtime or offline migrations—without increasing response time. And this scalability should be linear and infinite and as simple as adding shards and nodes.

Geo-distribution

Digital applications not only need to be available 24/7 but available from anywhere in the world. This is further complicated because your customers may not be stationary. For example, your customers could be shopping, banking, or playing games while crossing countries or regions on a high-speed train, causing the online connection to require a shift in location or risk a significant performance degradation—or worse, a loss of responsiveness all together.

Cost efficiency

The ability to flexibly grow with your demand and ensure a highly available, globally distributed application that consistently provides real-time responses can become cost prohibitive if they are not anticipated.

You should consider costs related to expanding your databases for both new and existing projects. You can also grow your business by taking advantage of new opportunities more quickly while recouping revenue previously lost to poor database performance.

Your cache should also ensure you can optimize the cost of your infrastructure by providing you with solutions that allow you to utilize 100% of your infrastructure assets and an automatic way to manage very large data sets cost efficiently.

Enterprise-grade support

Enterprise-grade support is essential for business-critical applications. It's crucial to have someone who stands behind your caching solution and provides enterprise-grade support to quickly resolve issues and challenges and help tune the solution for optimal performance.

What Makes Cache Enterprise-Grade?

The value of enterprise-grade cache

Localized caching is typically a reactive first step—you implement it after you discover that you're not meeting the customers' expectations due to the shortcomings of backend data infrastructure. It's a point solution, typically confined to a single instance, for addressing bottlenecks in the last mile of customer engagement. Localized caching solutions pick up where backend operational and transaction databases leave off. But when caching is used in a critical application software stack, it should be considered mission critical.

The most obvious way to determine if your cache is mission critical is to understand how important caching is to meeting customer expectations. If you cannot meet customers' expectations of real-time experiences without the cache running, then it's mission critical. Even if your applications are still running, only slower, this will still result in, at minimum, business reputation damage or an abandoning customer. With customers demanding real-time responsiveness from their digital applications, any loss in performance can lead to a potential loss of customers and revenue.

When determining if your cache is mission critical, you should ask yourself a few key questions:
- How confident are you in meeting your SLAs?
- What is the maximum downtime you can afford per week/month/year?
- Does your caching need to be replicated in multiple regions to ensure availability?

If your cache is down, you should also consider how your backend database will perform.
- Can your backend database handle the added workload that will occur when the cache is not there to run interference?
- Can your backend database handle the added pressure of rehydrating the cache when it does come back up?
- How will this increase in traffic impact your network?
- Is your network able to handle the additional load and still provide high throughput?

Finally, when your cache goes down, and you have a disaster recovery solution in place, does it have automatic failover and failback, or will you need to intervene manually? If manual intervention is required and your digital apps are always expected to be available, you probably need to constantly monitor your cache to ensure you can react quickly.

Buying Criteria for Enterprise Cache

So you're already using cache. You determined that you needed to speed up the primary database to meet your customers' ever-increasing expectations of real-time response. You need to consider the rest of the business requirements associated with digital applications. Does it need to be highly available? Do you need to distribute your applications across zones, regions, or the world? Do these distributed applications need to provide local sub-millisecond latency to their users?

Also, you should consider the economics of running your digital applications as you scale. Do you have a large data store that needs real-time responses but exceeds your cache, yet you can't justify increasing it due to the cost of RAM? Or are you struggling with the opposite problem, where you have a small data store with significant unused space?

Having the flexibility to deploy your applications isn't just a theoretical discussion. IDC research shows that most enterprises favor a hybrid cloud approach for their IT and business strategies (IDC Perspective, Oct 2021). Also, according to Gartner, 81% of organizations are working with two or more public cloud providers. This gives you the flexibility, when a failure happens, to roll over to another cloud vendor in the same region or geography, enabling you to meet your compliance requirements. A multicloud strategy gives companies the freedom to use the best possible cloud for each workload. The ability to run your applications on-premises, in a hybrid environment, or multicloud is becoming an expectation. Being hybrid or multicloud gives you the ability to optimize costs, security, and reliability and use various services to speed your software delivery. And since your applications are running in multiple environments, this requires that your caching also has the flexibility to run in various environments.

Many of these requirements are likely in your service-level agreements, so now is the time to ensure you can always meet your SLAs with the enterprise-grade cache.

Buying Criteria for Enterprise Cache

Criteria for evaluating enterprise cache

As with any IT solution, there are many options to help you solve a specific problem. The challenge is to select the best solution for your environment without adding unnecessary complexity. To the right are some critical criteria that you can use to evaluate the enterprise-grade caches available in the market.
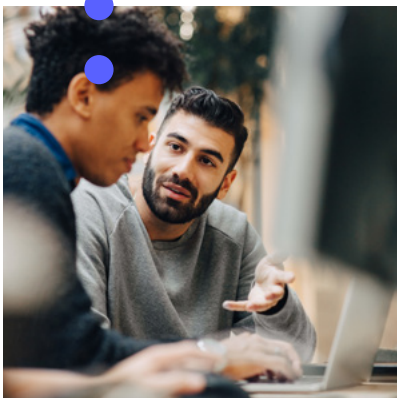
Enterprise-grade cache evaluation feature checklist

- ✔ Provides consistent high throughput and sub-millisecond latency
- ✔ Highly available (offers 5-9s availability)
- ✔ Geographically distributed and available across time zones with local low-latency
- ✔ Has automatic failover and failback
- ✔ Linearly scales to optimize infrastructure resources
- ✔ Scales with no performance degradation
- ✔ Backup with no performance degradation
- ✔ Provides intelligent tiering to manage large data sets economically
- ✔ Multi-tenancy to enable efficient infrastructure utilization
- ✔ Runs on-premises, in a hybrid environment, and multiple clouds without data transfer issues
- ✔ Enterprise-grade support
- ✔ Offers the innovation of an open source foundation and community

Buying Criteria for Enterprise Cache

How enterprise caching compares to basic caching

Basic caching meets the minimum requirements of temporarily storing data so that repeated database requests can be served faster, reducing the latency of the response. Enterprise caching is basic caching plus the enterprise-grade functionality to ensure that the caching solution can infinitely and linearly scale, is highly available, is secure, can be geo-distributed with data consistency, and works in hybrid and multicloud environments so it can be deployed where needed.

The chart on the right highlights the differences between basic caching, such as AWS ElastiCache and Google Cloud Memorystore, and enterprise-grade caching, such as Redis Enterprise for caching. Use this chart to help you select the right solution level for your business. As you can see, many features need to be considered when looking for highly available, sub-millisecond responses at a geo-distributed scale. If your cache needs to be mission critical, you should consider all enterprise-grade features.

	Basic Caching	Enterprise-Grade Caching
High throughput	✗	✔
Low latency	✔	✔
Cloud DBaaS	✔	✔
Hybrid deployment	✗	✔
Multicloud deployment	✗	✔
Five-nines availability	✗	✔
Geo-distribution	✗	✔
Guaranteed local low-latency	✗	✔
Data consistency with no performance degradation	✗	✔
Linear scaling	✔	✔
Infinite scaling	✗	✔
Basic clustering	✗	✔
Advanced clustering	✗	✔
Intelligent tiering	✗*	✔
Multi-tenancy	✗	✔
Eventual consistency	✗	✔
RBAC support	✗	✔
Support for Kubernetes as a native service (AKS, GKE, EKS, OpenShift)	✗	✔
Upgrade with no downtime	✗	✔
Auto failure detection	✗	✔
Enterprise-grade support	✗	✔

* Partial for AWS

Why Choose Redis Enterprise for Enterprise-Grade Caching?

Redis: Core to the real-time revolution

Redis addresses the need for real-time performance, both as a cache and as a data platform designed for real-time interactions. As shown in table 1, it's been well recognized as one of the most popular, most launched, and most used databases among developers.

Most developers are familiar with Redis as a cache. It's a compact, self-contained, high-performance in-memory engine that's proved ideal for databases requiring a frontend for accelerating performance. Redis Enterprise is the enterprise-grade version of open source Redis.

Popular use cases for Redis Enterprise as cache tier include:
- Queuing of backend batch processing jobs
- Publish/subscribe engine for distributing data on demand
- Complementing backend transactional databases with its sub-millisecond latency and ability to persist data to disk
- Counters, which involve designating a key and then issuing simple increment commands
- Scaling legacy databases

Table 1. Redis popularity

Source	Year	Category
Db-Engines	2021	Top 10 database
Docker Hub	April 2021	7.2M launches/day
Sumologic	November 2020	Most popular database deployed on AWS
DatDocker Hubadog	November 2020	The largest share of container images running in Kubernetes Stateful Sets
Stack Overflow	2017-present	"Most loved" database

Why Choose Redis Enterprise for Enterprise-Grade Caching?

Developers often use Redis as a cache under one of the top cloud vendors' versions, such as AWS ElastiCache or Google Cloud Memorystore. Redis Enterprise as a cache provides a robust, enterprise-grade, real-time cache users can deploy anywhere: on-premises, in a public or private cloud, hybrid, or multicloud.

However, many developers aren't aware of the differences between cloud-provided Redis services, based on open source Redis, a basic caching solution, and Redis Enterprise Cloud, an enterprise-grade solution. Offered by Redis, Inc., Redis Enterprise Cloud delivers consistent sub-millisecond performance and auto-scaling across all major public clouds. The need for Redis Enterprise as a cache becomes apparent when workload volumes grow, and enterprises require the consistent SLAs, scaling, automation, and security necessary for supporting their real-time commitments.

Also, the ability to geo-distribute consistent data with guaranteed local sub-millisecond latency across regions is why many users turn to Redis Enterprise for caching.

Why Choose Redis Enterprise for Enterprise-Grade Caching?

Benefits of Redis Enterprise

There are significant differences between Redis Enterprise (developed by Redis, Inc.) for cache and the open source solution. For starters, Redis Enterprise for cache can run anywhere: on-premises or in the cloud through Redis Enterprise Cloud. On-premises support is critical for organizations where data gravity, data sovereignty, or other policy mandates preclude public cloud deployment.

Additionally, Redis Enterprise for cache uniquely delivers

→ **High availability/ reliability**

→ **Geo-distribution**

→ **High performance and auto-scaling**

→ **Optimized storage cost**

→ **Multicloud deployment**
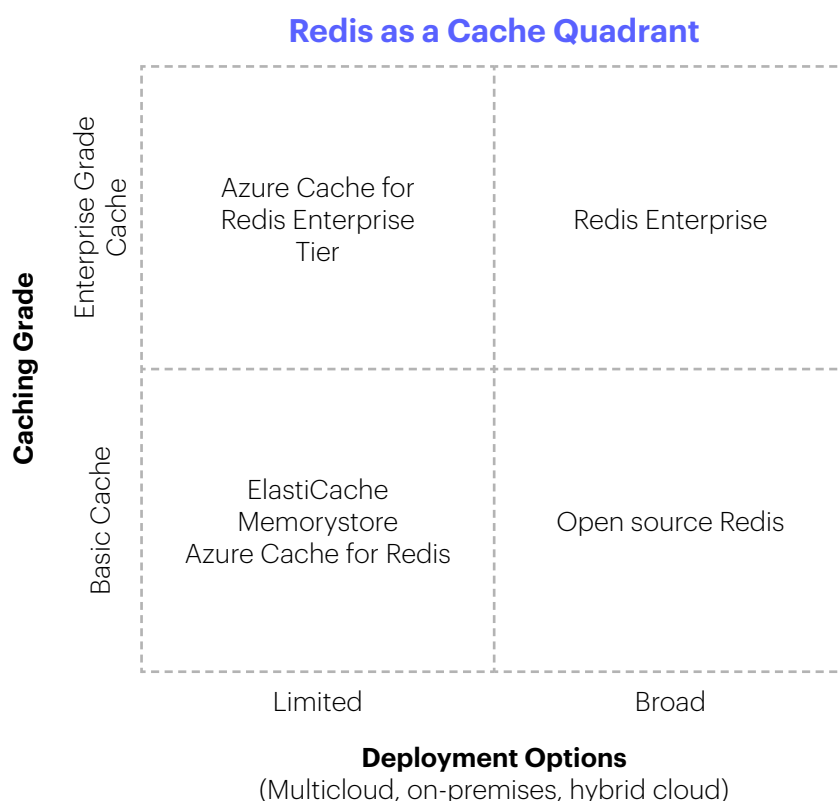
→ **Enterprise-grade support**

Why Choose Redis Enterprise for Enterprise-Grade Caching?

How Redis Enterprise caching compares to other Redis-compatible offerings

Not all versions of Redis in the cloud are created equal. There are significant functional differences (see Table 2 below). Redis is available as a community-based open source through the [3-Clause-BSD license](#), also known as the "Modified BSD" license, which allows the software to be modified by the developer and redistributed. This is the edition of Redis that third-party providers also deliver as a managed cloud service (i.e., ElastiCache and Memorystore).

The quadrant below shows how the various Redis caching editions compare based on two critical requirements: enterprise-grade and deployment choice. Redis Enterprise is the only version that meets the key requirements to be enterprise-grade and enables flexibility in your deployments.

Redis as a Cache Quadrant

	Limited	Broad
Enterprise Grade Cache	Azure Cache for Redis Enterprise Tier	Redis Enterprise
Basic Cache	ElastiCache Memorystore Azure Cache for Redis	Open source Redis

Caching Grade (vertical axis)

Deployment Options
(Multicloud, on-premises, hybrid cloud)

The capabilities needed to provide enterprise-grade caching include high availability and reliability; ability to easily distribute applications across zones, regions, and the world; automated scaling, intelligent tiering, and multi-tenancy for cost-efficient storage; hybrid and multicloud; and enterprise-grade support. The table below provides a detailed comparison of the caching solutions based on open source Redis: Redis open source cache, cloud solutions based on Redis open source, and Redis Enterprise as a cache.

Why Choose Redis Enterprise for Enterprise-Grade Caching?

Table 2. Comparing Caching Solutions Based on Redis Open Source

Feature	Capability	Redis open source cache (community edition)	Cloud solutions based on Redis open source	Redis Enterprise as a cache
Product support	**Experts when you need them**	Supported by customer	Supported by the cloud provider with a recommendation to develop a relationship with Redis, Inc.	Supported by Redis, Inc.
High availability	**5-9s HA**	Not supported	3-9s across a single region; no data persistence; snapshot only	5-9s SLA across one or more regions; supports data persistence and backup without impacting performance; automatic cluster recovery; pure in-memory replication
	Active-passive replication	No	Limited	Yes
	Automated failure detection	No	No	Yes
Performance		Difficult to scale	Scaling negatively impacts performance	Sub-millisecond via multicore and shared-nothing architecture
Active-active replication	**Global data distribution**	Single instance only	Single instance only; replication to other availability zones (AZs) is via a single master	Fully distributed with active-active replication support across multiple clouds AZs and regions
Redis on Flash	**Cost-effectively managed large data sets**	Not supported	Not supported on Memorystore; Similar but less functional and less performant tiering supported on ElastiCache	Supported, enabling via intelligent, cost-effective management of large data sets
Multi-tenant support	**100% asset utilization**	Not supported	Not supported	Supported, reducing TCO
Multi-cloud support	**Prevent vendor lock-in**	Users can run on any public cloud	Single cloud only	Fully managed support on AWS, Azure, and Google Cloud***

** Azure Cache for Redis is a native service integrated with Azure Cloud and fully supported by Microsoft.

*** Redis Enterprise on Google Cloud is a service integrated with Google Cloud and available directly from Redis, Inc., or Google Cloud Marketplace.

Why Choose Redis Enterprise for Enterprise-Grade Caching?

Return on investment

Return on investment is key to evaluating enterprise-grade caching solutions. Identifying specific costs and savings can help you understand the true value of enterprise-grade cache.

Recently, Redis commissioned Forrester Consulting to conduct a Total Economic Impact™ (TEI) study and examine the potential return on investment (ROI) enterprises may realize by deploying Redis Enterprise. The purpose of this study is to provide readers with a framework to evaluate the potential financial impact of Redis Enterprise on their organizations. Read Forrester's Total Economic Impact of Redis Enterprise to understand the expected ROI on deploying Redis Enterprise.

Summary

In an economy that's becoming increasingly digital, real-time experiences matter more than ever. Delays of more than a second can prompt players to lose interest in a game, lead customers to abandon shopping carts, or enable fraud to propagate. All too often, in applications the database is the weakest link. Latencies in the database can slow overall app performance up to 100x and are becoming an impediment to delivering the expected real-time experiences. Not surprisingly, enterprises have embraced caching as a workaround to the limitations of their existing data platforms.

Redis, best known as a caching layer and cited by developers as the "most loved" database, has earned its popularity thanks to its simplicity, versatility, and speed. But not all implementations of Redis are alike. Beyond open source and third-party cloud Redis services, which support basic caching use cases, Redis Enterprises offers a complete enterprise-grade solution for caching.

Redis Enterprise delivers the capabilities expected of an enterprise cache that handles business-critical data. Unlike third-party cloud Redis services, you can deploy Redis Enterprise anywhere, whether on-premises or in a public or private cloud of your choice. It delivers a real-time caching solution's high availability and consistent performance, even under heavy workloads. Thanks to tiered storage through Redis on Flash and multi-tenancy, it is more economical to operate. Additionally, Redis Enterprise Cloud delivers a full DBaaS that eliminates the burden of database infrastructure management, freeing enterprises to focus on their core business.

As organizations seek to transform themselves and respond to their customers' ever-increasing digital expectations, they will require a real-time caching solution to provide mission-critical functionality. They will need Redis Enterprise for cache.

About Redis

Data is the lifeline of every business, and Redis helps organizations reimagine how fast they can process, analyze, make predictions, and take action on the data they generate. Redis provides a competitive edge to any business by delivering open source and enterprise-grade data platforms to power applications that drive real-time experiences at any scale. Developers rely on Redis to build performance, scalability, reliability, and security into their applications.

Born in the cloud-native era, Redis uniquely enables users to unify data across multi-cloud, hybrid, and global applications to maximize business potential. Learn how Redis can give you this edge at redis.com.

redis