

Massive Scalability with InterSystems IRIS® Data Platform

Technology Guide





InterSystems IRIS offers a unique, massive data and process scaling, with massive concurrency across data models and types.

Introduction

Faced with the enormous and ever-growing amounts of data being generated in the world, software architects need to pay special attention to the **scalability** of their solutions. They must design systems that can, when needed, handle many thousands of concurrent users. While not easy, designing for massive scalability is an absolute necessity today for most organizations. Whether supporting real-time analytics, machine learning, large language models, retrieval-augmented generation (RAG), or networked applications, data systems must handle growing volumes, velocity, and complexity.

Software architects can design scalable systems in several ways. They can **scale vertically** by using bigger machines with dozens of cores. They can use data distribution (replication) techniques to **scale horizontally** for a growing number of users. And they can scale data volume horizontally by partitioning their data. In practice, software architects will employ several of these techniques simultaneously, trading off hardware costs, code complexity, and ease of deployment to suit their particular needs.

This guide focuses on the mechanics of vertical and horizontal scaling of both user and data volumes, with consideration of performance, cost, and architecture. It outlines several options for distributing and partitioning data and/or user volume, giving scenarios in which each option would be particularly useful. The guide explains how **InterSystems**, a global leader in creative data technology, and **InterSystems IRIS®** can simplify the configuration, provisioning, and operation of distributed systems through scaling.

Vertical Scaling

The simplest way to scale is to scale “vertically” – “scale up” and deploy on a bigger machine with **more CPU cores** and **memory**. Most modern data platforms support parallel processing of critical applications (like SQL) and include technology for optimizing the use of CPUs in multi-core machines. Vertical scaling expands an existing infrastructure without changing the computing architecture. It offers simplicity, lower latency (deployed on a single machine, with limited network demands), and easier compatibility with legacy systems.

However, there are practical limits to what can be achieved through vertical scaling alone. For one thing, even the largest available machine may not be able to handle the enormous data volumes and workloads required by modern applications and will eventually face hard limits in how much it can be upgraded. Also, “big iron” can be prohibitively expensive. Many organizations find it more cost-effective to buy, say, four 16-core servers than one 64-core machine.

Capacity planning for single-server architectures can be difficult, especially for solutions that are likely to have widely varying workloads. Having the ability to handle peak loads may result in wasteful underutilization during off hours. On the other hand, having too few cores may cause performance to slow to a crawl during periods of high usage. In addition, increasing the capacity of a single server architecture implies buying an entire new machine. Adding capacity “on the fly” is impossible. And a single server is a single point of failure for a whole system.

In short, although it is important for software to leverage the full potential of the hardware on which it is deployed, vertical scaling alone is not enough to meet all but fairly static workloads.

Horizontal Scaling

For all of the above reasons, most organizations seeking massive scalability will deploy on networked systems, “scaling out” workloads and/or data volumes “horizontally” by distributing the work **across multiple servers**. Typically, each server in the network will be an affordable machine, but larger servers can also be used, if needed, to take advantage of vertical scalability as well (see “Hybrid Vertical-Horizontal Scaling,” below).

Horizontal scaling offers multiple benefits that are almost unavoidable in modern large data applications. Such elastic systems can grow or shrink dynamically and economically to meet variable demand, with commodity hardware and cloud resources scaled as needed. And horizontal scaling adds resilience: a multinode system doesn’t fail when one node goes down. Horizontal scaling is foundational to cloud-native and distributed systems, enabling massive scaling and fault tolerance.

While essential to the Age of Big Data, horizontal scaling is nonetheless not without challenges. It requires careful orchestration, load balancing, and data distribution. Maintaining consistency of data across nodes and minimizing network latencies can pose exceptional difficulties, especially in real-time systems. Horizontal distributed architecture is constrained by the CAP (Consistency-Availability-Partitioning) theorem. Trade-offs are inevitable and mandate careful consideration of primary versus subordinate goals.

Software architects will recognize that no two workloads are the same. Some modern applications may be accessed by hundreds of thousands of users concurrently, racking up very high numbers of small transactions per second. Others may only have a handful of users, but query petabytes worth of data. Both are very demanding workloads, but they require different approaches to scaling. We consider each scenario as distinct.





Horizontal Scaling by User Volume: **Caching**

To accelerate while scaling, databases make frequent use of caching, adding a high-speed, temporary layer for frequently accessed data and avoiding repeated database queries for the same data. To support a huge number of concurrent users or transactions and to scale by user volume, InterSystems introduced our own unique implementation called **Enterprise Cache Protocol (ECP)**.

Within a network of servers, one will be configured as the data server where data is persisted. The others will be configured as application servers. Each application server runs an instance of InterSystems IRIS and presents data to the application as though it were a local database. Data is not persisted on the application servers. Instead, these servers provide cache and CPU processing power.

User sessions are distributed among the application servers, typically through a load balancer, and queries are satisfied from the local application server cache, if possible. Application servers will retrieve data from the data server only if necessary. ECP automatically synchronizes data between all cluster participants.

With the compute work handled by the application servers, the data server can be dedicated mostly to persisting transaction outcomes. Application servers can easily be added to, or removed from, the cluster as workloads vary. For example, in a retail use case, you may want to add application servers to deal with the exceptional load of Black Friday shopping and switch them off again after the holiday season has finished. Application servers are most useful for applications where large numbers of transactions must be performed, but each transaction only affects a relatively small portion of the entire data set.

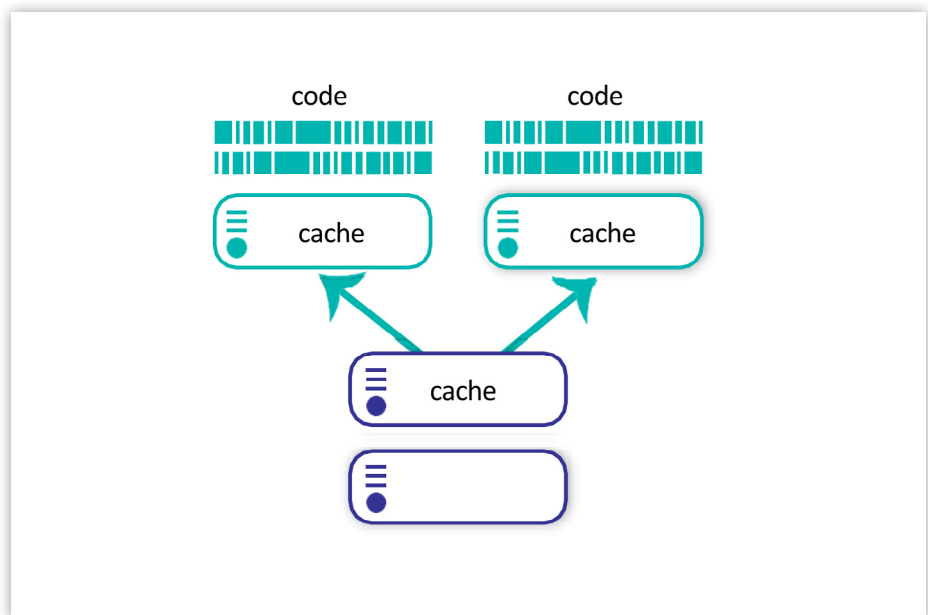


Figure 1: Database Workload Distribution with Enterprise Cache Protocol

InterSystems IRIS and ECP

InterSystems IRIS and IRIS for Health implement ECP as an integral part of their data architecture. Deployments that use application servers with ECP have been shown to support many thousands of concurrent users in a variety of industries.

Horizontal Scaling by Data Volume: **Sharding**

When queries — usually analytic queries — must access a large amount of data, the “working dataset” that needs to be cached in order to support the query workload efficiently can exceed the memory capacity on a single machine. A powerful technique to cope with such large datasets is **sharding**, which physically partitions large database tables across multiple server instances. Applications still access a single logical table on an instance designated as the shard master. The shard master decomposes incoming queries and sends them to the shard servers, each of which holds a distinct portion of the table data and associated indices. The shard servers process the shard-local queries in parallel and send their results back to the shard server for aggregation.

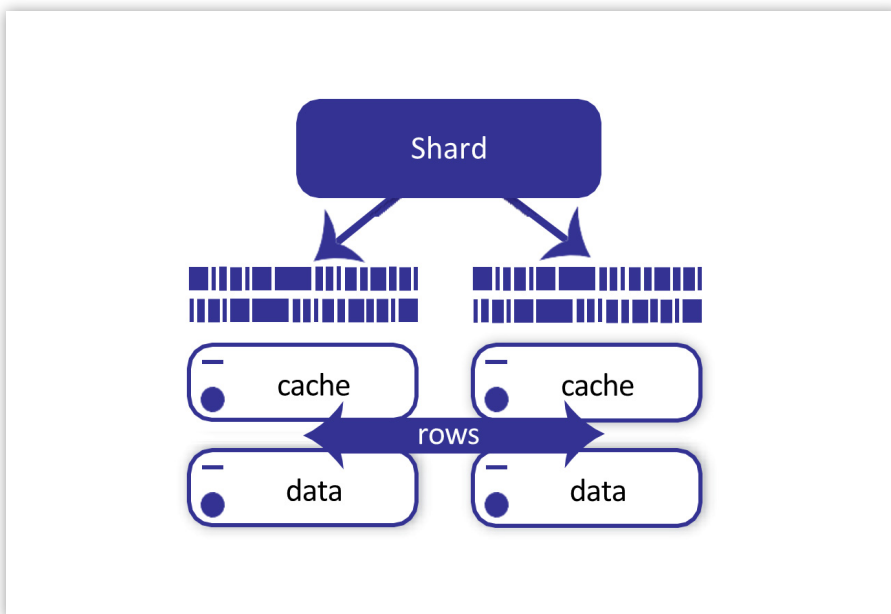


Figure 2: Sharding with Intelligent Inter-Shard Communication

Data is partitioned among shard servers according to a shard key, which can be automatically managed by the system, or defined by the software architect based on selected table columns. Through careful selection of shard keys, tables that are often joined can be co-sharded, so rows from those tables that would typically be joined together are stored on the same shard server. This colocation allows the tables to be joined locally on each shard server, thus maximizing parallelization and performance. As data volumes grow, additional shards can easily be added. Sharding is completely transparent to the application and to users.

Not all tables need to be sharded. For example, in analytical applications, facts tables (e.g., orders in a retail scenario) are usually very large and will be sharded. The much smaller dimension tables (e.g., product, point of sale, etc.) will not be. Non-sharded tables are persisted on the shard master. If a query requires joins between sharded and non-sharded tables, or if data from two different shards must be joined, InterSystems technology uses a highly efficient ECP-based mechanism to correctly and efficiently satisfy the request. In these cases, only the rows that are needed are shared between shards, rather than broadcasting entire tables over the network, as many other technologies would. InterSystems technology transparently improves the efficiency and performance of big data query workloads through sharding, without limiting the types of queries that can be satisfied.

InterSystems architectures enable complex multi-table joins when querying distributed, partitioned data sets—**without** requiring co-sharding, replicating data, or broadcasting entire tables across networks.

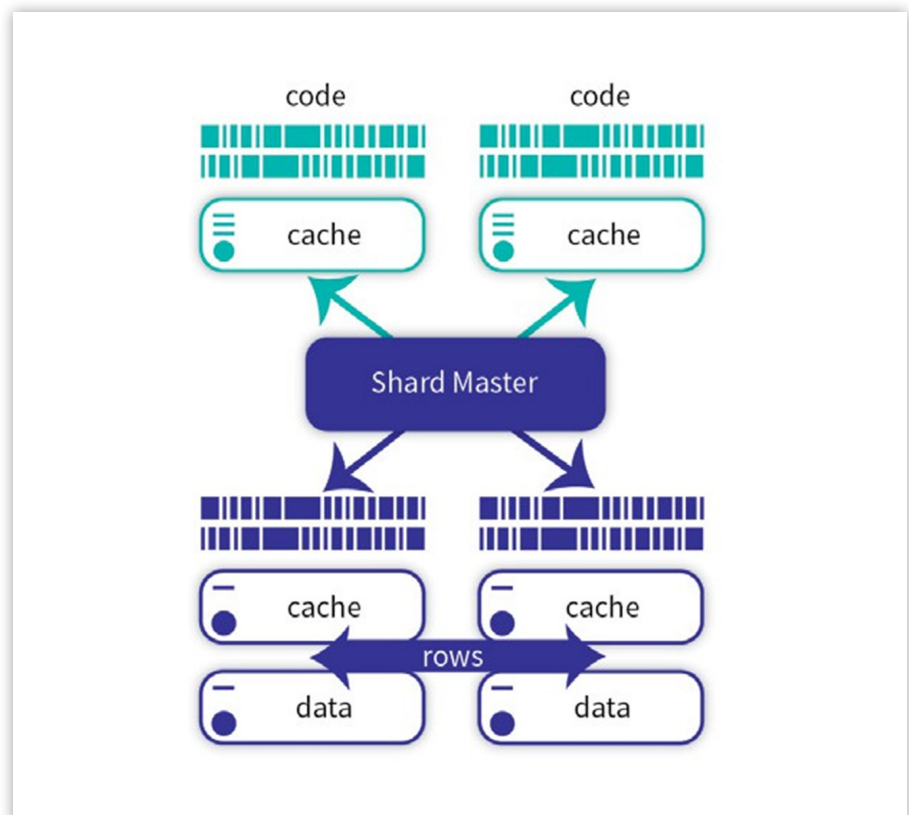


Figure 3: Sharding and Workload Distribution

Hybrid Horizontal Scaling by Both User and Data Volumes

Modern data solutions often must simultaneously support both a **high transaction rate** (user volume) and analytics on **large volumes of data**. An example: a private wealth management application that provides dashboards summarizing clients' portfolios and risk, in real time based on current market data.

InterSystems technology is unusual in its simultaneous capability for analytic queries and ingestive transactions. It enables such Hybrid Transactional and Analytical Processing (HTAP)—or translytical—applications by allowing application servers and sharding to be used in combination. Application servers can be added to the architecture (pictured in Figure 2) to distribute the workload on the shard master. Workloads and data volumes can be scaled **independently** of each other, depending on the needs of the application.

When applications require the ultimate in scalability (for example, if a predictive model must score every record in a large table while new records are being ingested and queried at the same time) each individual data shard can act as the data server in an ECP model. We refer to the application servers that share the workloads on data shards as query shards. This, combined with the transparent mechanisms for ensuring high availability of an InterSystems IRIS cluster, provides solution architects with everything they need to satisfy their solution's unique scalability and reliability requirements.

Hybrid Vertical-Horizontal Scaling

Many modern systems combine vertical and horizontal scaling. For example, a database might run on powerful vertically scaled nodes while distributing queries across horizontally scaled application servers. Hybrid models offer flexibility and resiliency, and they can be tailored to specific workloads and requirements.

Impact of Architecture

Effective scaling requires thoughtful architectural design, which both influences the scaling and is influenced by it:

- **Data partitioning**, primarily sharding and replication strategies, helps manage large datasets (see “Horizontal Scaling by Data Volume: Sharding,” above).
- **Stateless versus stateful service** choices shape scaling decisions. Stateless services are easier to scale horizontally.
- **Load balancing** distributes traffic evenly across nodes, ensuring no one node is overwhelmed.
- **Tenancy** strategies shape how distinct users share access to a data platform and possibly its infrastructure while maintaining logical separation of each user's data.
- **Consistency models**, constrained by the CAP theorem (consistency, availability, partition tolerance), guide trade-offs in distributed systems.



Applications and Trends

Scalable data platforms are critical to success in the industries where InterSystems has been at work solving customer problems. The most important is healthcare data, where InterSystems works often at very large scale, managing patient records, medical imaging and other diagnostic data, and real-time monitor, among a large number of medical informatics applications.

InterSystems is also active in other data applications, including finance (processing transactions, detecting fraud, and reporting for regulatory purposes) and supply chain (tracking shipments, optimizing delivery, and managing inventory).

InterSystems has played a role in enabling data solutions across all these industries, helping organizations meet their data, performance, and reliability goals at large scale.

The landscape of scalable data platforms continues to evolve as new data applications and systems are introduced:

- **Cloud-native architectures**, like Kubernetes and serverless models, simplify horizontal scaling.
- **AI-driven scaling** with predictive algorithms optimizes resource allocation.
- **Edge computing** distributes processing closer to data sources, reducing latency.
- **Data mesh and federated architectures** promote decentralized data ownership and scalability.
- **Data fabric architectures** overcome data silos and fragmentation.

Massive Scaling with InterSystems

Massive scalability is no longer optional—it's a **necessity** for modern data applications. Such scaling is especially a requirement for Hybrid Transactional and Analytical Processing (HTAP) applications that must handle large workloads and high data volumes simultaneously, possibly with a large number of concurrent users.

Vertical scaling offers simplicity and performance for smaller systems, while horizontal scaling provides elasticity and resilience essential for large-scale applications. Specialized architectures can tailor the system's operation for different combinations of requirements and priorities. By understanding the strengths and limitations of each approach, organizations can design data architectures that meet current needs and adapt to future growth.

About InterSystems IRIS

InterSystems IRIS is a data platform that gives software architects options for efficiently scaling their applications and making effective use of the strategies outlined in the guide. It supports vertical scaling, application servers for horizontally scaling by user volume, and a highly efficient approach (ECP) to sharding for horizontally scaling by data volume that eliminated the need for network broadcasts. These technologies can be used independently or together to tailor an architecture scaled to an application's specific requirements. For more information, see the [InterSystems IRIS documentation on scalability](#).

InterSystems IRIS achieves a unique fusion of **massive scalability**, highly **efficient** and **concurrent transactional-analytical** processing, **multi-model** data representation in a single store without copying or moving, **flexible tenancy** modes, and customizable **local-cloud** deployment.

For more information about InterSystems IRIS, visit [InterSystems.com/IRIS](https://interSystems.com/IRIS) and try it for free!

Why InterSystems

Established in 1978, InterSystems is the leading provider of next-generation solutions for enterprise digital transformations in the healthcare, finance, manufacturing, and supply chain sectors. Its cloud-first data platforms solve interoperability, speed, and scalability problems for large organizations around the globe. InterSystems is committed to excellence through its award-winning, 24×7 support for customers and partners in more than 80 countries. Privately held and headquartered in Boston, Massachusetts, InterSystems has 37 offices in 28 countries worldwide. For more information, please visit [InterSystems.com](https://interSystems.com).

