DECEMBER 2021

4 Tips for Processing Real-Time Data

Processing real-time data in your own data centers or at the edge can be more affordable and effective with the right technology. Learn how to create a scalable solution that reduces costs while increasing revenues.

SPONSORED BY DELL Technologies intel

Real-time data processing is exploding. In fact, industry researchers say that the amount of real-time data in the global datasphere will grow from 9.5 zettabytes in 2020 to 51 zettabytes in 2025. That's more than a five-fold increase in just five years.

No matter what business you're in, chances are good that you're already processing real-time data on a regular basis. Companies rely on real-time data processing for a wide variety of use cases, including fraud prevention, recommendation engines, financial transactions, identity management, customer relationship management, social media monitoring, and more. And the growth in the use of predictive analytics, artificial intelligence, and machine learning makes it likely that the volume of real-time data processed will continue to increase exponentially.

But while real-time data has undeniable benefits, that explosive growth is leading to some significant challenges.

Real-time data processing is exploding.

First of all, storing and processing that data can be expensive—whether the data resides in the public cloud or in an enterprise's own data centers. Enterprises have significant hard costs related to purchasing servers or cloud computing instances, not to mention the related soft costs for IT personnel to manage it all.

As a result of that management burden, many enterprises are migrating more of their data to the public cloud. But not all data can reside in the cloud. In some cases, organizations find it necessary to store data on-premises to comply with government regulations or to align with their own security policies. In Informa's 2021 State of the Cloud survey, 55% of IT decision makers surveyed agreed with the statement "In general, it is a good idea to keep my organization's most sensitive data in on-premises systems, rather than in the cloud." In addition, 52% agreed that "My organization's data is more secure on-prem."

On the other hand, rapidly scaling in on-premises environments is much more difficult than in the cloud. And customers do need to scale quickly. They need to transform data into revenue—and do so faster than their competition—if they want to stay in business.

So what can enterprises do to overcome these real-time data challenges? The following four tips can help you build a scalable, cost-effective environment for processing real-time data in your own data centers or directly on the edge.

1. Integrate a NoSQL database with Kafka or Spark.

If you have a large database (more than 5 TB) and you are processing a high volume of data in real time, you'll get the best performance if you deploy a NoSQL database alongside your other real-time data tools like Kafka or Spark. And those performance benefits will be even more significant if your data volume is scaling quickly and you need a high degree of availability and reliability.

When compared with tradition relational database management systems (RDBMSs), NoSQL databases offer a number of advantages. First, while all data needs to fit into a row-column table structure in a RDBMS, NoSQL databases can store unstructured or semi-structured data, as well as structured data. NoSQL also makes it easier to update your data schema and fields. And most importantly for this discussion, NoSQL can handle much higher volumes of data at high speed, and it scales easily.

Many organizations with applications that make use of real-time data deploy the Apache Kafka event streaming platform. Its distributed architecture allows it to transmit a large volume of data with very low latency. It also scales easily and offers high availability for mission-critical data.

Another popular real-time tool is the Apache Spark unified analytics engine. It supports both SQL queries and advanced machine learning and graph processing algorithms, which is important for many real-time data use cases. And it is exceptionally fast—up to one hundred times as fast as other popular large-scale data processing engines.

When you augment your Kafka or Spark deployment with a NoSQL database, you can magnify the performance benefits of both. You'll streamline your data

processing—allowing you to keep up with high volumes in real time—while ensuring the data consistency that is necessary for applications like financial transactions, identity management, fraud prevention, and similar applications.

2. Match your server components to your use case.

Of course, the software that you use to build your real-time data pipeline is only part of the equation; you also need the right hardware to support real-time performance at scale.

In the past, servers relied on hard drives for storage and dynamic random access memory (DRAM) for memory. This arrangement worked fairly well for a long time because hard drives are slow, but they are relatively inexpensive and can store data even when the system is shut down. DRAM is fast, but it is much more expensive and requires power to keep storing data. Installing both allowed servers to use the DRAM for the small amount data it was accessing frequently with the hard drive providing reliable long-term storage.

Persistent memory (PMEM) is even faster than flash while still retaining data during a power interruption.

Memory-first (i.e. DRAM-first) database systems will struggle to scale given DRAM costs and inability to persist data.

Fortunately, manufacturers have developed alternatives that blur the lines between storage and memory, giving organizations more options. In recent years, flash-based solid state drives have become much more popular for high-volume storage. Flash is faster than hard drives (but not as fast as DRAM). It's also cheaper than DRAM, and it is persistent—meaning that it retains data when the system isn't turned on. If you use flash in servers that are processing and storing real-time data, you can speed up performance, particularly when you have very high volumes of data.

A newer technology called persistent memory (PMEM) is even faster than flash while still retaining data during a power interruption. It's actually about the same speed as DRAM and less expensive. Intel Optane technology allows servers to use PMEM as a caching layer—making very large volumes of data accessible very quickly. It's becoming a more common option for servers processing real-time data.

It's important to note that no one type of server architecture is right for every use case. DRAM, flash, and PMEM each have their own advantages and disadvantages. You'll need to find the right balance between latency, scalability, and cost.

Smart enterprises are matching their data volume, performance requirements, and other needs to the infrastructure that is right for them. And since different use cases often have different needs, that often means having different types of servers for different workloads.

Some NoSQL databases allow organizations to take an even more granular approach to matching hardware to use cases. For example, the Aerospike NoSQL data platform gives you the option of deploying a variety of different storage architectures within the same cluster, and Aerospike's patented Hybrid Memory Architecture™ treats SSDs as memory where one can receive optimal performance at scale with a lower cost. That can make the entire cluster faster and more efficient. Table 1 provides latency and scale variances for DRAM, PMem, and Flash storage architecture combinations.

Table 1 – Memory and flash configurations for relative performance and scale with Aerospike. Note: performance statements below are conservative and are not meant to serve as benchmark references as underlying hardware and database tuning can have positive impacts. See www.aerospike.com/benchmarks/ for more information.

	Pure In-Memory	Pure PMem	Hybrid DRAM/Flash	Hybrid PMem/Flash	All Flash
Index storage location	DRAM	PMem	DRAM	PMem	Flash
Data storage location	DRAM	PMem	Flash	Flash	Flash
Scale	Up to 10s of TB	Up to 10s of TB	Up to 1 PB	10s of PBs or more	10s of PBs or more
Latency	99% <1 ms	99% <1 ms	95% <1 ms	95% <1 ms	95% <10ms
Use Case	Smaller data volumes	Small or medium data volumes where data must persist after a restart	Medium data volumes where longer restarts are unacceptable	High-density nodes where fast restarts are important	Extremely large data volumes where slower performance is acceptable

3. Scale up and scale out.

Most data processing solutions are designed either to scale up or to scale out. As a reminder, scaling up is when you add more computing, memory or storage resources to an existing server or node. For example, if you have a server with 8 GB of RAM and 2 processing cores and you upgrade it to 16 GB of RAM and 4 processing cores, you have scaled up.

Scaling out involves adding additional servers or nodes to a distributed system. For example, if you have 4 nodes in your hyperconverged infrastructure rack and you add two more nodes, you have scaled out.

When it comes to physical infrastructure in a data center, scaling up generally takes longer, but it takes up less space. Scaling out is fast but requires more space.

Ideally, for real-time data you want a database and hardware that will allow you to scale both up and out. That allows you to scale from terabytes to petabytes while maintaining the smallest data center footprint and keeping costs low. For example, Aerospike's hybrid memory architecture allows it to scale both up and out as it makes sense for an organization's needs. Keep in mind that you will also need hardware that supports both kinds of scaling as well. You should look for servers with latest-generation multi-core Intel processors to support parallel processing.

4. Use smart data distribution to reduce latency while increasing resiliency.

While scaling your hardware is necessary to deal with large, growing data volumes, scaling also introduces challenges. As a cluster grows, one of the biggest causes of performance problems in real-time data processing is the existence of "hot spots." Ideally, you want all the resources in your cluster to be doing approximately the same amount of work. If some portions of the cluster get more use than others, the system becomes less efficient. Bottlenecks emerge, and the performance of the entire cluster degrades.

Load-balancing technology can ensure that each node in the cluster is doing the same amount of work and storing the same amount of data. This smart data distribution reduces latency, speeding performance.

As an added benefit, if you use smart data distribution to create clusters that span multiple data centers, you can also increase resiliency. As a result, customers are better able to recover from adverse events, while maintaining incredibly fast performance. This high availability makes Aerospike a good fit for mission-critical data processing.

Plan today for your future needs

The trend towards real-time data shows no signs of slowing down. Companies have seen the opportunities available in data analytics, and as a result, they are

collecting and processing more data than ever before. To remain competitive, they need to do more with their data—and do it more quickly—than the competition.

As companies begin to scale their data operations, the complexity of their infrastructure increases as well. That in turn can cause costs to climb even more quickly than data volumes in some cases.

However, by investing in the right hardware and software, smart organizations can actually reduce their server footprint while improving performance. By deploying a solution with four key capabilities—NoSQL database, server components matched to workloads, the ability to scale up and scale out, and smart data distribution—you can optimize your spending.

Most importantly, that real-time data processing solution will be able to meet your needs now and for years to come, providing the speed and reliability you need to support your use cases.

Investing in the right hardware and software, smart organizations can actually reduce thier server footprint while improving performance.

About Dell

Dell Technologies (NYSE:DELL) helps organizations and individuals build their digital future and transform how they work, live, and play. The company provides customers with the industry's broadest and most innovative technology and services portfolio for the data era.

Dell Technologies and Aerospike: Better Together

More than half of Aerospike on-premises customers choose to run their Aerospike workloads on Dell Technologies hardware. Most of these organizations have hundreds of terabytes in their databases, and they need the latest technology, including Dell NVMe storage and Intel Optane PMEM, to provide the level of performance that they need.

Together, Dell and Aerospike can save customers more than 60% in total cost of ownership compared to other data processing solutions. And Dell Technologies OEM Engineered Solutions has certified solutions available that are proven to provide fast performance for Aerospike clusters.

For more information, reach out to: warren_jackson1@dell.com

About Intel

Intel is an industry leader, creating world-changing technology that enables global progress and enriches lives. Inspired by Moore's Law, we continuously work to advance the design and manufacturing of semiconductors to help address our customers' greatest challenges. By embedding intelligence in the cloud, network, edge, and every kind of computing device, we unleash the potential of data to transform business and society for the better.

About Aerospike

The Aerospike Real-time Data Platform enables organizations to act instantly across billions of transactions while reducing server footprint by up to 80%. The Aerospike multi-cloud platform powers real-time applications with predictable sub-millisecond performance up to petabyte scale with five-nines uptime with globally distributed, strongly consistent data. Applications built on the Aerospike Real-time Data Platform fight fraud, provide recommendations that dramatically increase shopping cart size, enable global digital payments, and deliver hyper-personalized user experiences to tens of millions of customers.