Getty Images

Managing Data at Petabyte Scale and Beyond

As data grows to the petabyte and even exabyte level, organizations must rethink how they manage and protect that data.

Karen D. Schwartz | Oct 29, 2019

If your company is like most, you have noticed a steady increase in the amount of data that must be stored and managed. This increase is due to many factors, including a huge influx of machine data and website interaction data, a greater emphasis on operational analysis, a focus on developing more complex services, That rate of exponential data growth isn't going anywhere but up. According to one recent report, companies dealt with an average of 9.7 petabytes of data in 2018, representing a 569% increase from 2016. For reference, a petabyte equals 1,024 terabytes, or roughly 500 billion pages of documents. And sooner rather than later, some organizations — if they haven't already — will reach exabyte levels (one exabyte is 1,000 petabytes). One report finds that by 2025, the world will be creating 463 exabytes of data each day.

Often, that means that what was manageable at a smaller scale simply isn't all that manageable at a larger scale.

"Once you reach the petabyte level, you've reached a tipping point," said Randy Kerns, senior strategist and analyst at Evaluator Group. "You may not be able to use the same processes, and you may have to supplement the tools you have."

Dealing with petabytes or even exabytes of data also changes how an organization must manage and protect that data. Data protection processes and technologies, for example, may have been developed when the organization had less data. The methodology around protecting the data may have to change to accommodate these new data levels. Another issue is finding a way to manage vast amounts of data that are often distributed across different environments and systems. And then there are the challenges around availability, scalability, resiliency and cost.

Before even considering new tools or processes, it's important to understand the types of data your organization is collecting, the forms it comes in, how it is classified, the relative importance of different types of data, the requirements for and uses of that data, and your organization's path forward.

"It's always good to step back and look at your workloads and services at a more holistic level, and organizations I've seen be successful with this have done that, clearly defining methodologies around data classification, data governance and

Everything Changes at Petabyte Scale

With exabyte- or petabyte-scale data, the approach to resiliency and business continuity should change as well. At the terabyte level, for example, it is much easier to replicate one box to another or one data set to another, but that scenario changes at higher levels.

Making sure that very large data sets, which are often in distributed environments, can be updated and distributed in a way that ensures continuous access may require a different approach. At this scale, it’s no longer about making sure you can replicate a specific application or discrete data set. It’s more about developing a methodology that will ensure that services can remain running even if you experience a large network outage or a data center power failure.

Performance is another area that can take a big hit with massive data sets, but that is unacceptable in today’s environment. “Millisecond performance is no longer good enough. Performance needs to be sub-millisecond,” said Doc D’Errico, chief marketing officer of Infinidat. “You need storage capable of performing at microseconds in order to get to that sub-millisecond latency.”

Scalability, reliability and availability requirements and processes also can change in petabyte-scale environments. That often means finding a reliable infrastructure that is designed and architected for these types of environments.

“The older five-nines type infrastructures just aren't good enough; four to five nines of reliability translates to minutes or hours of downtime per year,” D’Errico said. “You need something that's going to be measured in low numbers of seconds. You’re after a zero impact in terms of data loss and downtime. Your applications need to continue to run.”

“You can’t be constantly dealing with problems of data placement or you will never catch up. That’s why technologies from 20 years ago that were trying to do automated storage tiering are no longer relevant — because by the time you do the analysis on where things need to go and then move things around your infrastructure, it's too late. The need for that information [has] passed,” D’Errico said. Instead, it should be done in real time. That requires enough density to be able to keep the information together, along with the agility to be able to move closer to the cluster where the applications are needed, he added.

With massive amounts of data, your approach to data security also should change. These environments require highly scalable, highly distributed networked systems, which often necessitate a change from scalable Fibre Channel SANs to complex IP-based systems with different security constructs. What’s more, users are accessing different data layers, dealing with more data sources and using the data for more complex processes. That makes managing permissions and authentication more complicated.

“When designing applications and services at this scale, you really need to build security in from the ground up using techniques like encryption, Kerberos, multitenancy and authentication,” Osborne said. “Security really has to be built in all the way through the stack.”

Does that mean that reaching this scale requires a rip-and-replace approach to data storage, security and management? Not necessarily, D’Errico said. If you architect applications so they are compartmentalized and in segments, you can leverage existing infrastructure by expanding it and moving your workloads around, he said, although that creates other problems, especially at the management level. It only makes sense to build it from the ground up if you work for a startup or are ready for a full refresh, but organizations with legacy infrastructures have to pick and choose, he said.

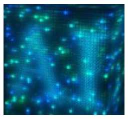
Transitioning to newer infrastructure will take time, but it has to happen, Kerns said.

“If you really want to optimize your environment, you should continually be re-evaluating it,” he said. “You should be doing it both by looking forward at what you expect to happen over the next five years, and always be working on a tactical plan to implement new technologies that allow you to be more effective and accommodate the growth.”

TAGS: BACKUP IDENTITY MANAGEMENT AND ACCESS CONTROL BIG DATA

0 COMMENTS

RELATED



AI-Based Storage Helping Companies Get More Out of Their Data

MAR 10, 2020

Why Computational Storage May Be the Next Big Thing

JAN 17, 2020

Storage Briefings Roundup: Computational Storage, Micro Edge

JAN 13, 2020

2019 Data Storage Trends: Cloud-Based Storage Shone

DEC 19, 2019

tech