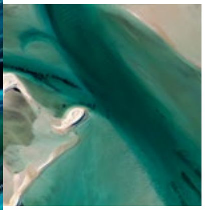

Object Storage: Data Management of Unstructured Data at Extreme Scale

NetApp®

Overview

The term enterprise architecture is often used to describe large-scale IT operations. Scale is an often-used term in business and can have very different meanings depending on the context. For example, in the phrases "we are scaling our team" and "can our business scale?" it has entirely different meanings.

In 2013, Gartner introduced the term web-scale to describe how large cloud services firms such as Google, Amazon, Netflix, Facebook, and others achieved extreme levels of service delivery compared to their enterprise counterparts.

In a research note (subscription required), Gartner identified six elements of web-scale, but noted that while the term scale usually refers to size, smaller IT organizations could still benefit from a web-scale IT approach. Gartner made the point that even modestly sized organizations could achieve some of the capabilities of Google and similar companies.

In this paper, and in the context of Gartner's web-scale definition, we'll look at five examples that bring extreme service delivery capabilities to both small and large IT organizations through the use of object storage.

The Basics of Data

Fundamentally, IT organizations maintain two basic types of application data: structured and unstructured. Structured data applications (that is, databases and other systems of record) require fast data I/O operations to process sophisticated data algorithms and provide useful information quickly.

Unstructured data applications, in contrast, prefer a system of folders and directories to store and locate unrelated text and media files, as well as other types of distributed content, such as those generated by Internet of Things (IoT) sensors. Historically, unstructured data applications were less concerned with fast I/O and more concerned with the ability to store files in a hierarchical structure.

Of the two, unstructured data composes the vast majority of enterprise data. According to many sources, structured data growth has remained relatively unchanged since 2000, whereas unstructured data is growing at the rate of 55% to 65% per year and today composes more than 80% of all enterprise data.

An emerging trend that is driving the collection of unstructured data is data analytics. In this field, data scientists refer to the "data-munging" process as the collection, cleaning, and organizing of unstructured data: in other words, giving structure to unstructured data to model and eventually harness the information in the data for some useful purpose.

Because of the prolific growth of unstructured data and the rising use of data analytics applications, providing large-scale storage platforms that can support billions (or trillions) of unstructured files has become challenging. Traditional file-based storage arrays utilizing NFS and SMB file system protocols simply were not designed to operate at this scale and to efficiently perform data munging and complex analytical operations.

Object storage, which was first proposed in 1995 and delivered in the early 2000s, removes capacity barriers of traditional file-based storage through the use of a global namespace and unique user IDs (UUIDs). Data objects are stored in a flat address space that eliminates the hierarchy (and the capacity limitations) of a traditional file system. For this reason, virtually every large cloud service provider today uses object storage as the underlying architecture in which to store and retrieve data: an essential component in any web-scale architecture.

Object Storage Web-Scale Examples

Managing Distributed Rich Media

Distributed rich media content often consists of information received from a variety of sources. For a website developer, for instance, this might include content pushed from other websites as well as localized content developed specifically for regional websites.

To add more complexity, organizations working with sensitive data (such as pharmaceutical companies or academic research facilities) might need to enforce geo-fencing of certain information to comply with geopolitical requirements.

By implementing a web-scale content management system, organizations can eliminate the time and potential errors introduced when users manually enter content in multiple places, with different priorities, and under changing conditions.

NetApp® StorageGRID® is a distributed object storage system that stores, protects, and preserves fixed-content data across multiple data centers in multiple geo zones. Employed in a grid architecture, object data can be selectively distributed throughout the system, creating a highly automated system where data is continuously available where it should be, but nowhere else.

In a multiple–data center site deployment, the infrastructure of the StorageGRID system can be asymmetrical across data centers and proportional to the needs of each data center site. Typically, data centers are located in geographically different locations. Through the use of a single namespace, applications have worldwide access to all data, but secure policies define where this data actually lives.

Creating an IoT Data Repository

The IoT is producing vast stockpiles of data for the purpose of developing meaningful analytics. Weather patterns, traffic hotspots, and consumer behavior are some of the items analyzed by data generated by humans and machines alike.

Collecting and analyzing data from millions of points around the globe are a daunting web-scale challenge. A case in point is the automobile industry. As explained in a recent blog post, the automotive industry is shifting its emphasis, like other industries, from hardware to software platforms and from products to services.

A single autonomous test car creates between 5TB and 50TB of data in a day. As you might expect, the immediate processing of this data occurs inside the car, but what you might not have considered are the deep analytics and learning algorithms, which are constantly performed from data aggregated by all test cars and processed in various data centers. For example, an automotive company in Germany is using the policy engine in the StorageGRID policy engine to make its IoT data repository available to development teams located in different countries so that each team has local, high-performance access to relevant datasets.

According to the blog post, having IoT data assets immediately available enables faster development cycles and is therefore is a key aspect of delivering innovative products and services.

Defining Hybrid Cloud Workflows

A hybrid cloud workflow contains on-premises and public cloud resources that work together to produce the best possible agility for storing, sharing, and processing application data.

For web-scale data storage, implementing an on-premises private cloud is often necessary in providing high-speed access for users and for integrating into existing applications. Public clouds, in contrast, are well suited for sharing data with geographically dispersed off-premises users and tapping into cloud-based software services. Instead of operating independently, it is possible to merge public and private clouds and implement a hybrid cloud workflow strategy using policy-based management to automatically notify and move data to the most appropriate resource.

Examples of hybrid cloud workflow services supported by NetApp StorageGRID:

- **Cloud mirror replication** automatically mirrors specified objects to a remote S3 bucket, which can be on AWS S3 or a second StorageGRID system. For example, you could choose to use cloud mirror replication to mirror specific customer records to a sister organization that has its own instance of StorageGRID, essentially creating a cross-organizational hybrid grid.

- **Amazon Simple Notification Service (SNS) event notification** is used to integrate the events occurring on the cloud and events occurring in the data center. An example is a cloud application capturing a significant event and sending data to an in-house application for processing.

- **Elasticsearch integration:** Elasticsearch is a popular open-source search and analytics engine used to perform log analytics, real-time application monitoring, and clickstream analytics. NetApp StorageGRID automatically sends object metadata to an Elasticsearch index, where the metadata can be searched, visualized, and analyzed. Data scientists use this capability to search through billions of object metadata tags. Importantly, object creations, deletions, or metadata updates are automatically sent to the Elasticsearch cluster.

Accelerating Business with Data Analytics

Today's enterprise businesses continuously need to generate business insight at breakneck speed to sustain innovation |and competitive advantages. Business acceleration through data analytics leads to lowered costs, reduced risk, and accelerated innovation, all with the predictive insights gained by data analytics.

One example of this is NetApp's own use of StorageGRID object storage to manage its Active IQ® database. Active IQ is NetApp's largest data analytics application, collecting 7 billion data points from NetApp IoT assets every day.

Specifically, NetApp expanded its one-tier data storage system by adding an archive tier for "cold" data and a more responsive tier for "hot" data, utilizing StorageGRID object storage, NetApp AFF storage, and NetApp E-Series hybrid flash storage.

After implementing this infrastructure, NetApp reported a 300% gain in operational efficiency, with a 40% reduction in licensing fees for archived IoT data, 66% reduction in storage footprint for the most active data, and simplified analytics workflows through utilization of native S3 APIs.

A detailed overview of the Active IQ storage architecture can be found on the NetApp website, authored by the architects of this innovative solution.

Achieving Compliant Data Retention Policies

Data retention is a high priority for enterprise organizations. In most companies, IT is responsible not only for the integrity of the data, but also for adherence to industry policies and government regulations for the retention of data.

NetApp StorageGRID uses a policy-based information lifecycle management (ILM) system to manage object retention. The ILM policy contains rules that filter objects based on their metadata and determine what happens to an object's data after it is ingested: where it is stored, how it is protected from loss, and for how long it is stored.

This ILM policy consists of a set of rules that also describe how object data is managed over time. Depending on how the data is used and the applicable requirements for data retention, unique ILM rules can be defined for archived data different from the ILM rules used for day-to-day operations.

This is important when considering that retention policies for different datasets often require different attention based on sovereignty laws. One user noted, "StorageGRID object storage, with its geo-tagging capability, scalability, and simplified maintenance, was a perfect fit. It provided data governance for the myriad files—from images to video—that [our] customers need to keep safe."

To provide web-scale data retention, NetApp StorageGRID can specify AWS S3 as an archival destination. Archival objects stored to AWS S3 are still managed by StorageGRID, which means that their location is transparent to users who continue to store, retrieve, or update these objects directly from the grid.

In addition, with the recent 11.1 release, StorageGRID provides a web-scale infrastructure for financial and personal data retention as an integrated resource across public and private clouds. For example, StorageGRID users can configure unique WORM data and data flagged for litigation hold to be wholly retained in independent AWS S3 buckets.

Summary

In reviewing these examples, it becomes clear that the needs of next-generation web-scale organizations are much different than those of traditional enterprise customers operating at scale. The ability to utilize hybrid cloud workflows in managing the global distribution of vast data repositories is the ultimate goal of web-scale organizations.

According to a recent report, the three largest cloud service providers—AWS, Microsoft, and Google—are projected to achieve combined cloud revenue of $80 billion by 2020. By no coincidence, these three companies, as well as many other cloud service providers, all use a form of object-based storage in their cloud storage infrastructure.

Object storage, once relegated to niche archival applications, is moving into the mainstream of web-scale architectures due to its ability to remove the restrictions of volumes, folders, and files, instead treating billions of objects as a single pooled capacity.

With object storage, the guesswork of capacity planning is eliminated. Volumes no longer need to be tied to particular servers, applications, or users. If one client unexpectedly grows at 80%, it simply consumes resources from the pool. There is no need to reconfigure and reallocate the other applications to accommodate growth. Simply stated, object storage is the only technology that can satisfy the needs of quickly arriving web-scale environments.

For more information about planning and building your web-scale architecture with object storage, visit:

- Optimizing Unstructured Data
- NetApp StorageGRID Object Storage
- NetApp StorageGRID Resource page

NetApp®