WHITE PAPER

Analytics Workloads with Azure NetApp Files and NetApp Cloud Volumes Service

Executive Summary

Data analytics requires processing large volumes of data from multiple sources. This data must be stored on a fault-tolerant platform that can sustain high levels of performance to facilitate access by analytics solutions such as Apache Hadoop and Apache Spark. Moving the source data to the cloud enables services such as Amazon EMR to process the data with nearly unlimited horizontal scalability. It also reduces the complexity of setting up an analytics compute cluster, removing the barrier to entry for many organizations.

Azure NetApp® Files and NetApp Cloud Volumes Service for Amazon Web Services (AWS) and Google Cloud Platform (GCP) are fully managed cloud storage solutions that use cloud compute and storage resources to provide an on-demand service for allocating, synchronizing, and managing highly available and scalable file shares in the cloud. Users can employ Azure NetApp Files and Cloud Volumes Service just as they would any other native cloud service, without the need to manage any of the lower-level infrastructure.

This white paper looks at how Azure NetApp Files and Cloud Volumes Service can be used to facilitate analytics in the cloud. It examines the benefits of using Cloud Volumes for synchronizing data from multiple data sources, facilitating access to the data from compute clusters, and provisioning data test environments.

Analytics in the Cloud

Data in large organizations grows organically, normally spreading across various databases and other repositories. One of the first tasks in setting up a data analytics platform is to consolidate all relevant data into a single repository, or data lake. This repository can then be accessed by a cluster of compute nodes that apply different algorithms to the data in an attempt to find patterns and gain insights. Because the source data is usually historical in nature and very sizable, data storage solutions must provide high levels of throughput performance.

Data lakes are accessed by hundreds or thousands of compute nodes simultaneously, which requires the host storage service to guarantee scalable and predictable I/O performance. This can be difficult to achieve with a custom NAS solution built on cloud compute and storage resources, because managing the capacity and performance characteristics of the underlying disks becomes more and more challenging as the deployment grows.

Another major challenge for creating a centralized data repository is keeping the data synchronized with the source systems after the initial baseline copy is created. As data in the source systems continues to change, updates must be applied efficiently to the analytics data repository. Transferring the updates incrementally removes the need to copy over the entire dataset every time and dramatically reduces the time it takes to complete synchronization operations. As well as ingesting data, there is usually a requirement to synchronize the results of analytical processing out of the data repository to other systems, which may reside on the premises or in the cloud.

Once the raw source data is being synchronized with a data lake on a regular schedule, a certain amount of preprocessing may be required to optimize the data for processing by downstream analytics processes. Data engineers require independent copies of the data to work with in order to develop these data transformation routines. Due to the huge volumes of information in a data lake, it's very difficult to maintain multiple, up-to-date test copies of the data. Possible workarounds for this scenario include data sampling, other forms of test data, or the use of a common data staging environment.

With the data lake ready to serve out data, a compute cluster can be used to target the repository and execute analytics workloads. Apache Hadoop is a framework for building distributed data processing solutions that natively supports cluster scheduling, management and Map Reduce operations. The foundation of Apache Hadoop is HDFS, a clustered file service for horizontally scaling out data storage with fault tolerance. Each node in the HDFS cluster is also an Apache Hadoop compute node that can operate on its local data, as well as the data in the rest of the cluster. In this way, an Apache Hadoop cluster is able to support a wide range of other data processing platforms, such as Apache Mahout, Apache Hive, and Apache Spark.

Creating a data lake in the cloud means that it is no longer necessary to set up an Apache Hadoop cluster manually. Cloud services such as Amazon EMR and Amazon RedShift provide ready-to-use solutions for working with Apache Hadoop and data warehousing, respectively. These services in particular make use of data stored in Amazon S3. Amazon EMR uses EMRFS to provide an HDFS interface for Apache Hadoop directly over Amazon S3 data, thereby removing the need to copy data. Amazon RedShift is able to pull in Amazon S3 data and combine it with its data warehouse for unified querying across the repositories.

Using Azure NetApp Files and Cloud Volumes Service for Analytics

As described in the previous section, successfully establishing data lakes in the cloud opens up a world of possibilities for data analysis. Azure NetApp Files and Cloud Volumes Service are part of an easy-to-use, robust, high-performance platform with the precise feature set required to create and support data lake environments.

Azure NetApp Files and Cloud Volumes Service is a high-performance platform for deploying NFS and SMB file services, providing unprecedented levels of I/O throughput. Users can control the level of IOPS delivered by configuring the service level, which can be set to Standard, Premium, or Extreme. The extreme service level provides up to 8000 IOPS per TB and 128MB/s of throughput per TB. Cloud Volumes Service allows service levels to be changed on the fly, which is either very difficult or not possible with other storage platforms.

Cloud Volumes Service Performance Levels

- Standard. Up to 1000 IOPS per TB (16k I/O) and 16MB of throughput per TB.
- Premium. Up to 4000 IOPS per TB (16k I/O) and 64MB of throughput per TB.
- Extreme. Up to 8000 IOPS per TB (16k I/O) and 128MB of throughput per TB.

With Azure NetApp Files and Cloud Volumes Service, NetApp brings to bear decades of experience in building enterprise NAS solutions. This means that Cloud Volumes Service easily scales to meet the most demanding conditions, providing concurrent access to thousands of client hosts and applications. Scalability to this degree is a difficult requirement for large-scale environments and is impossible to achieve with custom-built NAS solutions.

One of the biggest wins for a data lake deployment is to leverage data synchronization capabilities from NetApp. Data from multiple external data sources, with support for a range of different protocols, can be synchronized to a data lake storage volume and kept synchronized through scheduled, incremental updates, whether the source data resides on-premises or in other cloud systems, including systems hosted by other cloud vendors. For example, if on-premises data is to be consolidated in AWS, the Cloud Volumes Service is able to connect and transfer the data securely and efficiently. Data can also be synchronized from the cloud to other systems. This is useful when the results of processing an analytics workload need to be sent to another, possibly on-premises, location.

To support data preprocessing, the service offers sophisticated features for creating NetApp Snapshot™ copies and cloning storage volumes. Using NetApp cloning technology, writable clones of a volume can be created very quickly. The cloned volumes can be easily dropped and re-created to refresh them with up-to-date data. This is possible for a source volume of any size, and gives data engineers the power to work with data across as many concurrent environments as they require.

Data lakes can be used to support custom-built compute clusters, for example using Amazon EC2, as well as public cloud analytics services. NetApp offers the fault tolerance, scalability, and high performance necessary for driving data processing solutions such as Apache Hadoop with the NetApp NFS connector for Hadoop, Hadoop operations can use NFS data directly, without the need to copy the data to HDFS. Apache Spark nodes are also able to operate on NFS data stores directly.

To support public cloud analytics services, Azure NetApp Files and Cloud Volumes Service can use NetApp technology to incrementally synchronize data into and out of Amazon S3, as well as any other object data store that supports the S3 protocol. This makes it possible to simplify data synchronization and preprocessing, while still leveraging the power of solutions such as Amazon EMR and Amazon RedShift.

Azure NetApp Files and Cloud Volumes Service Benefits

As described in the previous sections, Azure NetApp Files and Cloud Volumes Service offer many advantages when working with analytics workloads. Here is a summary of the benefits to build data analytics solutions:

- **High I/O performance.** Processing large volumes of data, as is typical in analytics environments, requires consistent, high-performance I/O systems to ensure that data is readily available to compute resources. Azure NetApp Files and Cloud Volumes Service allows one of three different service levels to be configured for each storage volume, Standard, Premium, or Extreme, which deliver up to 16MB, 64MB, and 128MB of throughput per TB, respectively.
- **Scalability.** Azure NetApp Files and Cloud Volumes Service scales data access to a level that is not possible with other shared file services. As analytics clusters grow in size, the storage environments they depend on must continue to provide predictable high performance. This can be especially difficult to achieve with custom-built NAS solutions.
- **Faster results.** Analytics environments usually require temporary working copies of the data to perform preprocessing operations. Such an environment is required, for example, when testing data transformations that enrich the source data. By using the Snapshot and cloning technology that come with Azure NetApp Files and Cloud Volumes Service, writable cloned volumes can be created in a very short time. And multiple clones can be created of the same source volume concurrently.
- **Data consolidation.** Using the built in synchronization services, data can be seamlessly synchronized to and from multiple data sources. Data can be consolidated from on-premises systems, cloud-based environments, and even across cloud platforms. Consolidating data from multiple sources into Azure NetApp Files or Cloud Volumes Service brings consistent performance with enterprise data protection.
- **Integrate with public cloud analytics.** Azure NetApp Files and Cloud Volumes Service can be used to create data lakes that are incrementally synchronized with object data stores, such as Amazon S3. Using these solutions, data can be consolidated, preprocessed, and then made available to cloud-based analytics services such as Amazon EMR and Amazon RedShift.
- **Secure multicloud data mobility.** Ensuring the security of data traffic is essential for building enterprise systems, such as a data analytics platform. All data in Azure NetApp Files and Cloud Volumes service is encrypted and secure for SMB and NFS connections.

Conclusion

High performance, scalable and highly available shared file storage is crucial to delivering a data analytics platform. Managing data from multiple source systems effectively can be another major obstacle. Azure NetApp Files and Cloud Volumes Service provide cloud-based file service solutions that address the major challenges in creating a repository for data analytics workloads and can be used with custom-built Apache Hadoop clusters or public cloud analytics services.

Azure NetApp Files and Cloud Volumes Service are designed to deliver the highest levels of I/O performance and scalability. End users simply input the size of storage volume they need, choose the appropriate service level for their performance requirements, and NetApp takes care of the rest. This removes the significant burden on organizations to manage in-house NAS solutions.

The synchronization capabilities of Azure NetApp Files and Cloud Volumes Service allow data from multiple systems to be consolidated into a single storage volume. Data can also be synchronized out of Azure NetApp Files and Cloud Volumes Service to provide integration with other external systems. Volume cloning adds to the ability to manage and work with large volumes of data.

To get the most out of your analytics projects, sign up for Azure NetApp Files or Cloud Volumes Service on AWS or GCP today.

About NetApp

NetApp is the data authority for hybrid cloud. We provide a full range of hybrid cloud data services that simplify management of applications and data across cloud and on-premises environments to accelerate digital transformation. Together with our partners, we empower global organizations to unleash the full potential of their data to expand customer touchpoints, foster greater innovation, and optimize their operations. For more information, visit www.netapp.com. #DataDriven

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

NetApp reserves the right to change any products described herein at any time, and without notice. NetApp assumes no responsibility or liability arising from the use of products described herein, except as expressly agreed to in writing by NetApp. The use or purchase of this product does not convey a license under any patent rights, trademark rights, or any other intellectual property rights of NetApp.

The product described in this manual may be protected by one or more U.S. patents, foreign patents, or pending applications.

Data contained herein pertains to a commercial item (as defined in FAR 2.101) and is proprietary to NetApp, Inc. The U.S. Government has a non-exclusive, non-transferrable, non-sublicensable, worldwide, limited irrevocable license to use the Data only in connection with and in support of the U.S. Government contract under which the Data was delivered. Except as provided herein, the Data may not be used, disclosed, reproduced, modified, performed, or displayed without the prior written approval of NetApp, Inc. United States Government license rights for the Department of Defense are limited to those rights identified in DFARS clause 252.227-7015(b).

NetApp®