WHITE PAPER

Improve Data Analytics with AWS and NetApp Cloud Volumes Service for AWS

NetApp®

Executive Summary

One of the major benefits of moving to Amazon Web Services® (AWS) is the all-encompassing range of fully managed cloud services that becomes available. For example, for big data analytics, you do not need to provision low-level compute and storage resources to build an Apache™ Hadoop® cluster, avoiding all the administrative and maintenance overhead that this task entails. Instead, you can take a much simpler approach: Use Amazon Elastic MapReduce® (EMR) and NetApp® Cloud Volumes Service for AWS.

Amazon EMR® is a ready-to-use, dynamically scalable Apache Hadoop cluster that acts as the platform for launching a wide range of big data frameworks, including Apache Spark™, Apache HBase™, Apache Hive™, and many others. Amazon EMR clusters can be created in minutes, as opposed to the time that it takes to manually deploy a cluster. Amazon EMR clusters are also optimized to work with Apache Hadoop and can be grown to thousands of compute nodes.

NetApp Cloud Volumes Service of AWS is a fully managed NFS and SMB file-sharing service that is tightly integrated with AWS. So, you can access, use, and pay for Cloud Volumes Service in the same way that you do for Amazon EMR. Cloud Volumes Service delivers industry-leading levels of I/O performance together with advanced storage management features, such as point-in-time NetApp Snapshot™ technology and instant volume cloning.

In this white paper, we describe how you can use Amazon EMR and NetApp Cloud Volumes Service to create a powerful platform for big data analytics in AWS.

What Is Cloud Volumes Service for AWS?

Cloud Volumes Service of AWS is a fully managed platform for shared file system storage from NetApp that is tightly integrated with AWS. With Cloud Volumes Service, you can create, manage, and pay for Cloud Volumes Service just as you would for any other cloud service, such as Amazon S3™ or Amazon Elastic File System™ (Amazon EFS). Because it is a fully managed service, you are not required to manage any of the lower-level infrastructure that makes the service work. High availability, scalability, and I/O performance are all taken care of automatically by Cloud Volumes Service. You just create a volume and start using it.

Cloud Volumes Service delivers very high performance NFS and SMB file systems, enabling the IOPS of each file system to be governed through its service level. Just as the capacity of a file system can be changed instantly at any time, the service level can also be modified dynamically. The Standard service level delivers up to 1,000 IOPS per terabyte (16K I/O) and 16MBps of throughput per terabyte. Premium service gives you up to 4,000 IOPS per terabyte (16K I/O) and 64MBps of throughput per terabyte. For the highest levels of performance, you can also select the Extreme service level. Extreme service delivers up to an industry-leading 8,000 IOPS per terabyte (16K I/O) and a massive 128MBps of throughput per terabyte. Cloud Volumes Service provides simply the highest-performing storage that is available for analytics projects anywhere.

Service Levels for Cloud Volumes Service

- **Standard**: Up to 1,000 IOPS per terabyte (16K I/O) and 16MBps of throughput per terabyte.
- **Premium**: Up to 4,000 IOPS per terabyte (16K I/O) and 64MBps of throughput per terabyte.
- **Extreme**: Up to 8,000 IOPS per terabyte (16K I/O) and 128MBps of throughput per terabyte.

The use of shared file services, such as NFS or SMB, enables hundreds or thousands of clients to read and write to the same files simultaneously. NetApp's vast experience in building enterprise, on-premises NAS solutions makes Cloud Volumes Service a truly cloud-scale platform. This scale is difficult for custom-built NAS services that use native cloud compute and storage to achieve. Though such custom solutions might work initially, there are numerous obstacles to overcome to create an enterprise solution.

Custom NAS Solutions

NetApp Cloud Volumes Service use the advanced data protection features of NetApp RAID-DP®. RAID-DP proivdes replication and allows all the data to be available to all availability zones, maximizing redundancy and availabilty with less cost and complexity normally associated with building a reliable NAS solution, and goes a long way toward safeguarding data without sacrificing performance.

Another consideration is that balancing storage costs and I/O performance requires very careful management when you build a custom NAS solution in the cloud. If you need to create a file system for infrequently accessed data, your administrator must manually create and scale the cloud storage. If you later are required to move this storage to faster disks, your storage administrator would have to plan and carry out the process of allocating new storage, migrating existing files, and decommissioning the old storage. With Cloud Volumes Service for AWS, file systems can be grown instantly at any time; you simply adjust the volume size. The service level for each file system, which determines the level of I/O performance that it delivers, can also be changed dynamically.

The complexity of building a solution that supports advanced data management features such as snapshots, clones, data synchronization, and variable service levels (quality of service) requires an organization to have a high degree of storage expertise. Cloud Volumes Service makes instantly creating a read-only, point-in-time copy of any file system a trivial operation, with the additional capability of instantly creating writable clones of the file system based on that snapshot. With the data synchronization features of Cloud Volumes Service, data can be incrementally synchronized from both on-premises systems and other cloud systems, without the need for custom scripts and processes.

Amazon EMR and Cloud Volumes Service

Amazon EMR is a managed service that helps quickly deploy Apache Hadoop compute clusters by using Amazon Elastic Compute Cloud (Amazon EC2). Big data workloads can be processed by using any frameworks that run on top of Apache Hadoop, including Apache Spark. The advantage of using Amazon EMR is the simplicity with which you can reliably create and operate clusters and scale them up or down as you need. Amazon EMR actively monitors the deployed cluster to ensure that failed nodes are replaced immediately.

Amazon EMR Benefits

- Reliably create, scale, and operate big data compute clusters.
- Process big data workloads that run on Hadoop, Spark, and others.
- Directly process data from a Cloud Volumes Service NFSshare.

Compute for Amazon EMR clusters can use Amazon EC2 spot instances, which can reduce by more than half the costs of operating a cluster. This approach can be more suitable for long-running clusters or when running costs are more important than the time it takes to process a workload. You can build Amazon EMR clusters by choosing the specific type and number of Amazon EC2 instances to use, which is useful when you need guaranteed performance. You can also mix the two approaches and use spot instances to handle peak loads.

Amazon EMR compute nodes use EMR File System (EMRFS) to read and write data from remote cloud storage services, such as Amazon S3. To access your data from NFS, will require the use of the NetApp In-Place Analytics Module. You can add this module to the Amazon Linux AMI that you use to create Amazon EMR cluster nodes by using a bootstrap action. You can then specify an NFS file system simply by prefixing the URI location of the file with `nfs://`.

With NFS as the primary storage for analytics data, all users of the data can access it from the same location and by using the same protocol. By using the NetApp In-Place Analytics Module, Amazon EMR can process the data directly from the NFS share. This direct processing eliminates the need to copy the data to a secondary location, such as to Amazon S3 or into the Amazon EMR compute cluster, improving your time to get results.

Analytics Solutions with Cloud Volumes Service

In addition to delivering high-performance NFS, Cloud Volumes Service also features a range of advanced data management technologies that simplify working with the allocated storage volumes. Snapshots enable instant, point-in-time copies to be made of any file share, which can help data professionals version the data that they are processing. By knowing the exact version of the data that they use for an analytics job, your data engineers can perform more thorough regression testing. Users can access snapshots as if they were a read-only file system.

Any Cloud Volumes Service snapshot can also be used to instantly create a writable clone of the source file share. Clones are ideal for testing new transformations or data enrichments and can simply be dropped when they are no longer required, all without adversely affecting the active file system. Cloud Volumes Service supports multiple concurrent clones of the same source file system.

Building a centralized repository for analytics data requires consolidating information from various data sources that might be on the premises or reside in the cloud. Copying this data and keeping it efficiently synchronized over time can be challenging, and not having a solution in place for this task slows down the process of onboarding new systems. To support this requirement, Cloud Volumes Service comes with a built-in data synchronization feature that can incrementally synchronize data both into and out of its hosted file systems.

Incremental data synchronization means that when data is changed in the source file system, only the data that is related to the changes must be propagated to the destination. For large volumes of data, this approach greatly reduces the time and the network traffic that are required to keep two file systems in sync.

Benefits of NFS with Cloud Volumes Service

Following is a summary of the benefits of using Cloud Volumes Service to host analytics data with NFS:

- **High I/O performance.** Analytics platforms process large volumes of data, and therefore require consistent, high-performance I/O systems to keep compute resources busy. With Cloud Volumes Service, you can choose from one of three service levels for each storage volume: Standard, Premium, and Extreme, which deliver 1,000, 4,000, and 8,000 IOPS per terabyte, respectively.
- **Scalability.** Cloud Volumes Service scales client data access to levels. As more nodes are added to analytics clusters, the storage systems must continue to provide the same standard of high performance. This performance level is especially difficult with custom-built NAS solutions.
- **Faster results.** Preparing, transforming, and enriching data require temporary, writable copies of the source files to test these preprocessing operations. With the snapshot and cloning technology that is built into Cloud Volumes Service, volume clones can be created instantly. Cloud Volumes Service enables multiple clones of the same source volume to be active at the same time.
- **Data consolidation.** With the synchronization services that are part of Cloud Volumes Service, data can be incrementally synchronized to and from multiple data sources. Data can be consolidated from different on-premises file shares, cloud-based storage, and even across cloud vendors; for example, you can consolidate data to or from AWS, Microsoft Azure, and Google Cloud Platform. By centralizing data from multiple data sources into Cloud Volumes Service, you can

create data lakes in the cloud, with consistent performance and enterprise data protection.

- **Integration with public cloud analytics.** Cloud Volumes Service can incrementally synchronize data with Amazon S3, which enables your users to access the data from many AWS services. By using the NetApp In-Place Analytics Module, Amazon EMR can also read data directly from an NFS share that is hosted in Cloud Volumes Service, allowing human and machine users to share the same repository.
- **Secure multicloud data mobility.** All network traffic to and from Cloud Volumes Service is encrypted and secure. Establishing secure data communication is a must for building enterprise file services and the data analytics platforms that depend on them.

Conclusion

Amazon EMR is a dynamic platform for building and operating big data compute clusters. Cloud Volumes Service is a fully managed cloud service for hosting enterprise-grade NFS and SMB file shares. When used together, Amazon EMR and Cloud Volumes Service create a superior solution for compute and storage, respectively, when you process big data workloads in the cloud.

With the NetApp In-Place Analytics Module, you can access NFS data directly from an Apache Hadoop cluster. Therefore, the same repository that your data engineers and data scientists use can also be used by Apache Hadoop compute nodes. This benefit speeds up processing of the data and reduces the need to make additional copies. Cloud Volumes Service also allows instantaneous point-in-time snapshots to be made of any file share, which can be vital to support data versioning. Each snapshot can also be used to instantly create writable clones of the source file system for testing.

Cloud Volumes Service is a high-performance, highly available, and flexible platform for sharable file systems in the cloud that provides real value in many application areas, including file services, database systems, and DevOps.

To start making full use of AWS for your big data analytics projects with the fastest, highest-performing data platform in the cloud, register for NetApp Cloud Volumes Service for AWS.

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 like AWS, 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.

Copyright Information
Copyright © 2018 NetApp, Inc. All Rights Reserved. Printed in the U.S. No part of this document covered by copyright may be reproduced in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or storage in an electronic retrieval system—without prior written permission of the copyright owner.

Software derived from copyrighted NetApp material is subject to the following license and disclaimer:
THIS SOFTWARE IS PROVIDED BY NETAPP "AS IS" AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE, WHICH ARE HEREBY DISCLAIMED. IN NO EVENT SHALL NETAPP BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

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).

Trademark Information
NETAPP, the NETAPP logo, and the marks listed at http://www.netapp.com/TM are trademarks of NetApp, Inc. Other company and product names may be trademarks of their respective owners.

WP-7280-1018

NetApp®