NetApp Spark solutions overview

NetApp Solutions

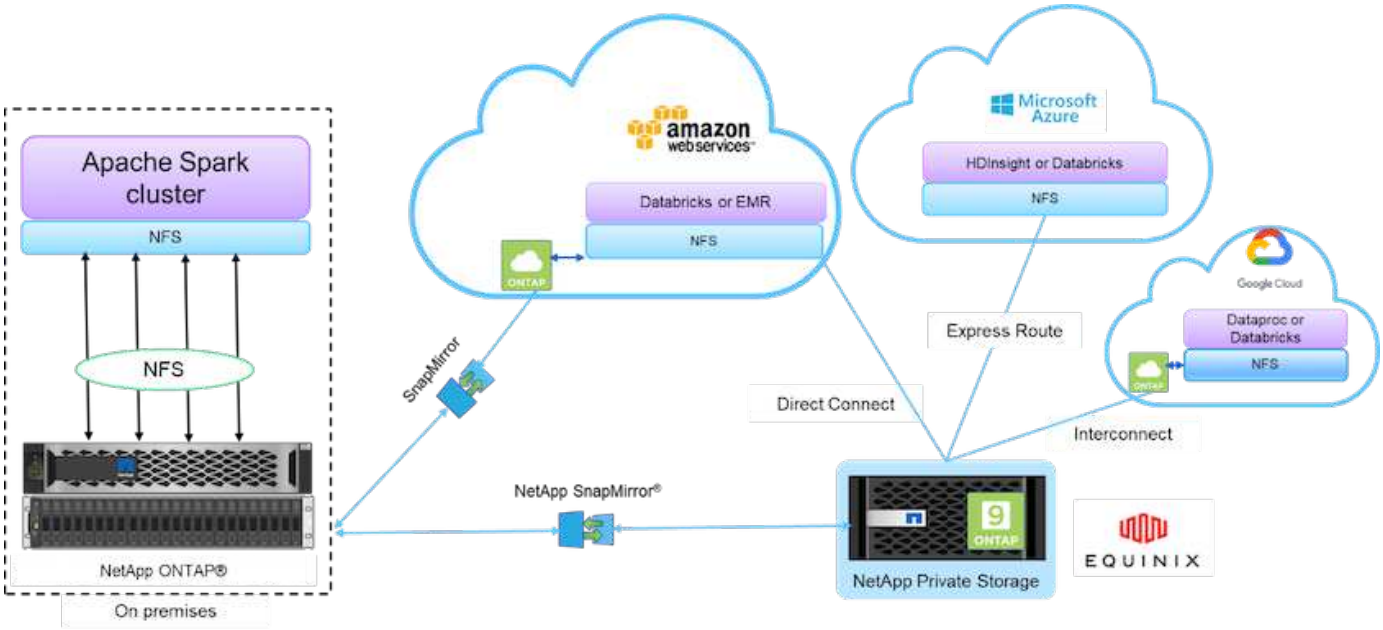
NetApp
October 20, 2023

Table of Contents

NetApp Spark solutions overview

Previous: Solution technology.

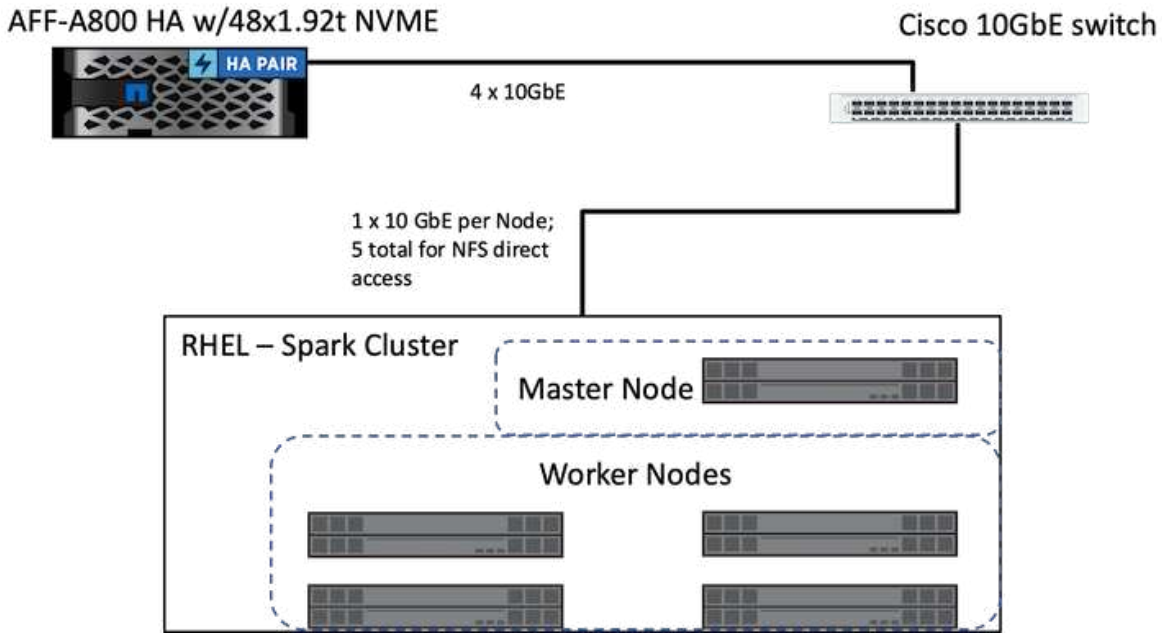
NetApp has three storage portfolios: FAS/AFF, E-Series, and Cloud Volumes ONTAP. We have validated AFF and the E-Series with ONTAP storage system for Hadoop solutions with Apache Spark. The data fabric powered by NetApp integrates data management services and applications (building blocks) for data access, control, protection, and security, as shown in the figure below.



The building blocks in the figure above include:

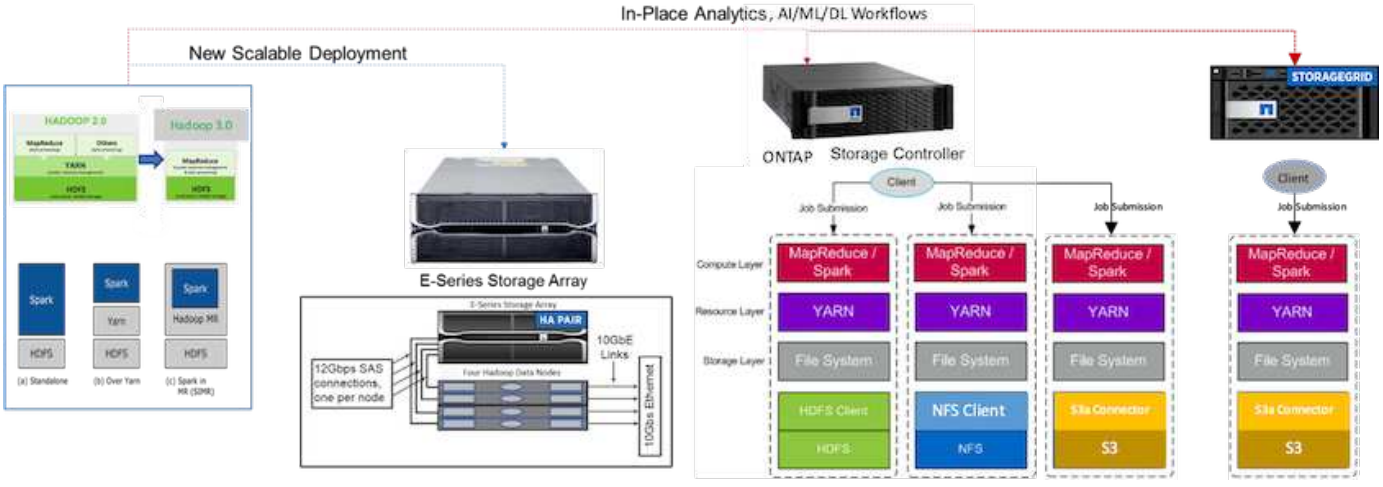
- **NetApp NFS direct access.** Provides the latest Hadoop and Spark clusters with direct access to NetApp NFS volumes without additional software or driver requirements.
- **NetApp Cloud Volumes ONTAP and Cloud Volume Services.** Software-defined connected storage based on ONTAP running in Amazon Web Services (AWS) or Azure NetApp Files (ANF) in Microsoft Azure cloud services.
- **NetApp SnapMirror technology.** Provides data protection capabilities between on-premises and ONTAP Cloud or NPS instances.
- **Cloud service providers.** These providers include AWS, Microsoft Azure, Google Cloud, and IBM Cloud.
- **PaaS.** Cloud-based analytics services such as Amazon Elastic MapReduce (EMR) and Databricks in AWS as well as Microsoft Azure HDInsight and Azure Databricks.

The following figure depicts the Spark solution with NetApp storage.

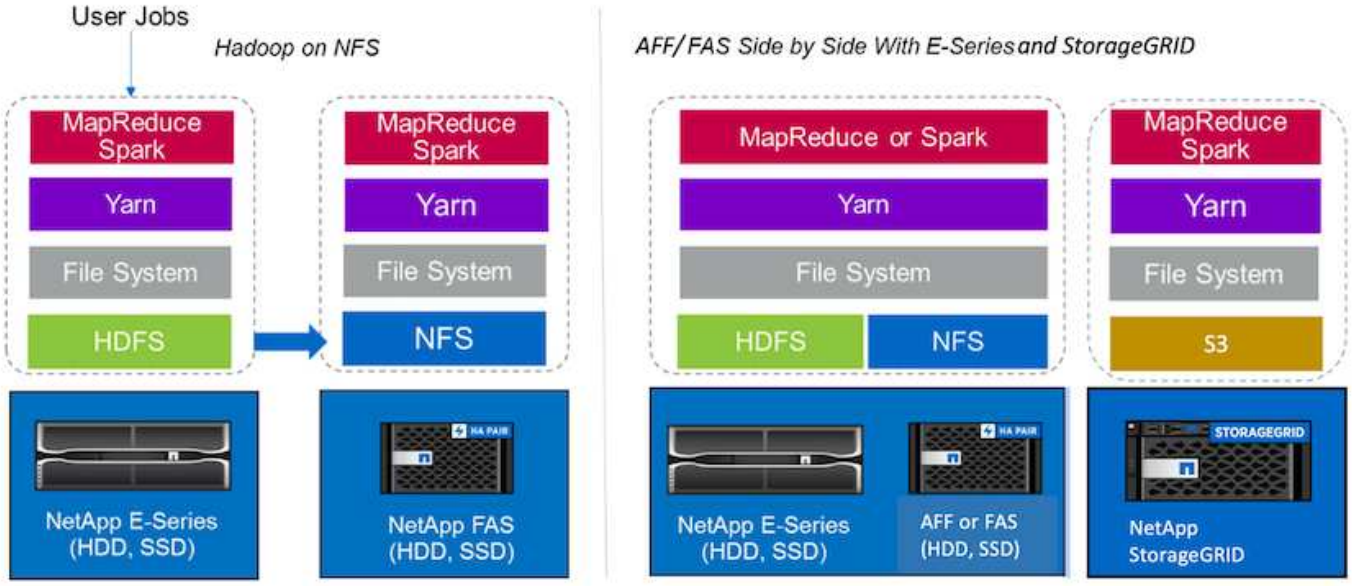


The ONTAP Spark solution uses the NetApp NFS direct access protocol for in-place analytics and AI, ML, and DL workflows using access to existing production data. Production data available to Hadoop nodes is exported to perform in-place analytical and AI, ML, and DL jobs. You can access data to process in Hadoop nodes either with NetApp NFS direct access or without it. In Spark with the standalone or `yarn` cluster manager, you can configure an NFS volume by using `file:///<target_volume`. We validated three use cases with different datasets. The details of these validations are presented in the section “Testing Results.” (xref)

The following figure depicts NetApp Apache Spark/Hadoop storage positioning.



We identified the unique features of the E-Series Spark solution, the AFF/FAS ONTAP Spark solution, and the StorageGRID Spark solution, and performed detailed validation and testing. Based upon our observations, NetApp recommends the E-Series solution for greenfield installations and new scalable deployments and the AFF/FAS solution for in-place analytics, AI, ML, and DL workloads using existing NFS data, and StorageGRID for AI, ML, and DL and modern data analytics when object storage is required.



A data lake is a storage repository for large datasets in native form that can be used for analytics, AI, ML, and DL jobs. We built a data lake repository for the E-Series, AFF/FAS, and StorageGRID SG6060 Spark solutions. The E-Series system provides HDFS access to the Hadoop Spark cluster, whereas existing production data is accessed through the NFS direct access protocol to the Hadoop cluster. For datasets that reside in object storage, NetApp StorageGRID provides S3 and S3a secure access.

Next: Use cases summary.

Copyright information

Trademark information