NetApp

White Paper

Modern data analytics

Different solutions for different analytics strategies

September 2024 | WP-7348-0924

TABLE OF CONTENTS

LIST OF FIGURES

Modern data analytics

Our customers' strategic use of data is increasing as they strive to operate more profitable, competitive, and differentiated businesses. Artificial intelligence (AI) and analytics are key components to helping customers achieve these business outcomes. The ever-increasing volumes of data used in AI and analytics help improve predictive analysis and give our customers data-driven insights into a more profitable business and improved customer service. With the ever-increasing volumes of data being used for these purposes, NetApp® customers need a modern data architecture for the most efficient, cost-effective, and flexible storage environment.

This paper describes NetApp modern data analytics solution strategies. It includes details about the business outcomes, customer challenges, technology trends, competition, legacy architecture, modern workflows, use cases, industries, cloud, technology partners, data movers, NetApp Active IQ®, NetApp DataOps Toolkit, Apache® Hadoop® to Spark™, software-defined storage with NetApp Astra™ Control, containers, enterprise data management, archiving, and tiering. It also discusses achieving the goals of AI and analytics and how NetApp and customers together are modernizing their data architecture.

Business

Data analytics began with servers with internal and/or direct-attached storage (DAS) to minimize investments and overcome the network bottleneck that included challenges to scaling the infrastructure, application disruption, difficult maintenance, multiple data copies, and data management during media failure. Recent digital transformation and real-time data analysis mean that modern data analytics are built on machine-generated unstructured or semi structured data such as IoT sensors, chatbots, website click-through rate, online retail, and healthcare imaging. Business drivers include the following:

- **Democratization of AI and analytics.** Proliferation of AI and analytics for customer service and insights require large volumes of data for increased accuracy.
- **Accelerate innovation.** Data, and lots of it, is of foremost importance to AI and analytics. IDC reports that the top AI business driver is to accelerate innovation.
- **Machine-generated data.** Drives unprecedented data volumes across all verticals: autonomous driving, healthcare imaging, manufacturing quality control, retail transactions, chatbot and customer service logs, IoT, and more.
- **Data is the new oil.** Valuable data assets require modern, sustainable, and flexible data architectures, essentially making data modernization just as important as the previous application modernization, but with containers.
- **Harvesting unused data.** Harvest untapped and siloed data for processing by a range of analytic tools, and also share the data across an organization.
- **Hadoop market transition.** Some customers are transitioning to Spark or want to move data out of Hadoop Distributed File System (HDFS) to reduce costs and server sprawl.
- **Object storage acceptance.** Object storage in cloud and on premises is increasing, for simplicity and scale.
- **Stream processing access.** The use of event streaming platforms for modern applications, such as Kafka, is growing.

NetApp solutions are based on customer use cases. They support businesses in the following ways:

- Cost-effective, high-performance storage for high-capacity unstructured and semi structured data
- Flexible performance at scale with multiple workloads
- Same experience for on-premises, cloud, and hybrid cloud environments
- Multisource and shared infrastructure support to meet customers' business requirements

Technology trends

In the early days, structured data provided input for predictive analysis. The cost was high for the data formation, and most of the historical structured data provided limited predictability. In addition to structured data, unstructured and semi structured high-volume data offer more insights for better forecasting. In driving business in the current market, data is the new oil.

Big data analytics platforms with many compute nodes process the structured, semi structured, and unstructured data for analysis. Analysis started with distributed infrastructures on a nonshared file system using MapReduce. However, In MapReduce, management and performance were unpredictable and data was siloed and untapped. Then in-memory engines such as Apache Spark™ were introduced for machine learning (ML). These engines process the data in memory and overcome the MapReduce disk I/O-related issues. Spark offers good performance with improved scalability and support for hardware accelerators.

Spark requires increased computation overhead, which can be accomplished by increasing the number of nodes or introducing support GPU, starting with Spark version 3.

Deep learning (DL) as part of AI is the next level of data analytics through hardware accelerators such as GPU-accelerated compute and neural networks. In the past, data storage used a relational database management system (RDBMS) with a specific format (schema on write), but in modern data analytics the data needs to be stored in various formats (schema on read). NoSQL databases, such as MongoDB, Apache Cassandra®, and HBase, offer the flexibility to store data in multiple formats as needed. NoSQL databases store structured JSON documents and key-value pairs in the most reliable way, with distributed format across several nodes for scaling.

Most data analytics deployments need their own infrastructure through physical servers. Customers have started using virtualization platforms such as VMware for data analytics processing to share their infrastructure. Both physical and virtualized infrastructures require up-front capital expenditures.

Customers started looking at cloud infrastructure on an on-demand basis to reduce capital costs. However, operating expenses increased over time. Most customers use the cloud for proof of concept in the early stages of a project. When the proof of concept is successful, customers might move their deployment from the cloud to on premises to scale into production. Currently, serverless operations through containers are the trend for most new applications.

NetApp offers solutions for physical, virtualized, on-premises, cloud, and hybrid environments, with customer production use cases as well as proofs of concept, performance validation, and integration with partner products to meet customers' unique requirements.

Innovating with changes in technology, NetApp focuses on software in addition to hardware products. Recently, customers have been focusing on the data lake house, which is the most cost-effective solution for data storage and the most useful approach for data scientists and business intelligence. NetApp's strategy, products, and solutions are aligned with the data lake house infrastructure.

Legacy architecture and modern workflows

About a decade ago, relational database management systems (RDBMS) provided historical analysis of structured data. MapReduce offered predictions based on structured, unstructured, and semi-structured data. Data is the foundation of modern data architecture. For example, NetApp healthcare customers use health data lakes for clinical, pharmacy, claims, and EHR/EMR patient data systems.

In the first 5 to 10 years, both RDBMS and MapReduce technologies were used for data analytics. However, over the past 5+ years, Spark has provided better performance results with its in-memory engine and its enhancements in data structures from resilient distributed datasets and DataFrame to dataset APIs. Similarly, Kafka has evolved for distributed streaming and real-time data pipelines.

A new generation of apps and people in technology means a need for a different analytics solution. NetApp's products and solutions engage with customers to meet their requirements and help them take advantage of current technology trends. Using the modern analytics architecture, NetApp customers identify new product bundles to help increase sales. These predictive analytic models are created from many years of siloed and untapped sales transaction data.

Use cases

NetApp offers data analytics case studies for customers in the following industries and more:

- Cyber analytics for an aerospace company
- Big data queries for a research organization
- In-place analytics for an aviation company
- Loss and fraud prevention in real time for a chain of stores
- Data protection for NoSQL for the banking sector
- Hadoop data protection for a banking customer
- Backup and disaster recovery for a broadcasting customer
- Hadoop DevTest for an online music company
- Data protection and data movement for multicloud connectivity for an IT service provider

Based on the challenges and requirements of the current market, NetApp offers the following solutions for these use cases:

- Data lake with NetApp StorageGRID® object-based storage solution and NetApp ONTAP® S3
- Hadoop HDFS NFS gateway to move analytics data
- MinIO gateway for secure access to NFS data and multiprotocol access to data
- NetApp XCP software to accelerate data movement from HDFS to NetApp StorageGRID object-based storage solution, ONTAP NFS, and ONTAP S3
- Spark 3 with NetApp Unified Storage Architecture, as well as Spark 3 GPU enabled with NVIDIA RAPIDS
- NetApp NFS for Apache Kafka and best practice guidelines and certification with Confluent Kafka, for deployment of Kafka with Spark and other analytics solutions
- Global namespace with NetApp FlexCache® software for remote caching of most actively read data, serving customers with hot data across multiple locations

Industries

Different industry verticals use data analytics to measure various KPIs like cost analysis, demand forecast, customer satisfaction, quality control, and revenue growth for key business decisions and outcomes. For example, the search/advertising enterprise that has driven e-commerce on the web over the past decade uses predictive analytics and recommendation engines to deliver meaningful information to potential customers.

Customers must embrace new analytic engines to remain competitive and to ensure that they are not left behind in the search engine noise. For example, the online shopping platform has improved the top-line financials. This improvement is due to an increase in sales and a decrease in customer selection times, and a decrease in IT costs for optimizing selections and managing large volumes of data.

Here are some examples that are based on our interactions with various industry customers:

- **Financial.** NetApp ONTAP ingests 1TB per 10 Splunk indexers and stores cold and frozen data. NetApp E-Series storage systems ingest 1TB a day to meet customer SLAs. NetApp ONTAP AI provides fraud detection and predictive analytics for a large financial institution.

- **Healthcare.** A NetApp solution offers virtualization, containers with NVIDIA GPUs for massive data collection, and an AI model to automate the segmentation in medical imaging. This solution reduces the workload for radiologists from hours to minutes and thereby reduces patient wait times.

- **Automotive.** NetApp has built an end-to-end solution for a major automotive customer that minimizes data movement, maximizes compute utilization, and consolidates the workloads from siloed, high-performance computing, big data, R&D, and autonomous vehicle deployments.

- **Telecommunication.** NetApp helped a large telecommunication customer distribute their data lake licensing across their premises and the cloud, and also built their dev/test environment in just minutes.

- **Banking.** NetApp works with banking customers for data protection, performance, and cloud with containers for in-memory engines such as Apache Spark with GPU.

- **Oil and gas and retail.** The NetApp ONTAP with NVIDIA DGX solution for virtual assistants and chatbots assists in customer service, voice search, and customer behavior prediction for a large oil and gas and retail customer.

- **Media.** NetApp partners with movie production companies to make the cloud a complement by bringing the right data to the right artists at the right time.

- **Research and development.** NetApp worked with a large research lab company to size their Splunk environment and architect their deployment to execute more research samples than was possible with their earlier architecture.

- **Service provider.** NetApp partners with an ISV service provider to build an architecture for Splunk and data protection for NoSQL databases.
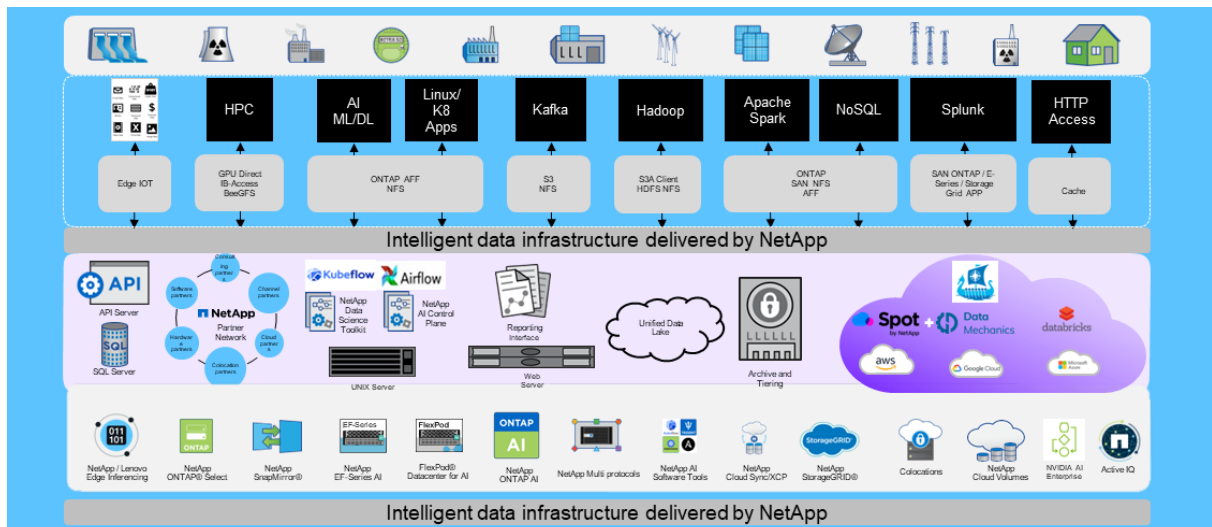
Figure 1) Modern data analytics.

Figure 1 is a high-level overview of a modern data analytics solution with NetApp products, partners, and applications and their interaction with vertical industry applications.

Edge to core to cloud

NetApp enables data movement from the edge to the core to the cloud and back. The IoT devices at the edge produce huge amounts of data, where NetApp customers can benefit from data mover products such as NetApp XCP and NetApp BlueXP™ copy and sync. The main operations happen at the core, including data preparation such as data aggregation; normalization at the data lake; training such as exploration using algorithms on the sample data; and deployments based on the trained model. NetApp AIPod™ with NVIDIA DGX systems and NetApp cloud-connected all-flash storage is used in the core, to scale and integrate data pipelines at a customer data center.

Customers who run dev/test and proof of concept in the cloud can use NetApp cloud offerings such as Cloud Volumes ONTAP, Azure NetApp Files, Amazon FSx for NetApp ONTAP, and Google Cloud NetApp Volumes. Most customers do large-scale production deployments at the core. To help them with seamless data mobility, NetApp offers a combination of products and features such as NetApp SnapMirror®, NetApp XCP software, BlueXP copy and sync, FlexCache, and NetApp BlueXP to classify, move, back up, and serve data across any hybrid multicloud environment.

Cloud

A data fabric powered by NetApp simplifies and integrates data management across the cloud and on the premises to accelerate digital transformation. NetApp partners with the three major cloud service providers: AWS, Microsoft Azure, and Google Cloud. Kafka ingests data from IoT devices like sensors to an in-memory Spark engine to process it in cloud with Cloud Volumes ONTAP and native cloud services based on ONTAP, then replicates it from cloud to on-premises using SnapMirror.

Some customers process some of their data on their premises and some in the cloud. They manage their data by using NetApp solutions and products that offer easy management and save costs. Enterprise customers do most of their work on their premises and send subsets of the data to the cloud to leverage compute power or analytics services from the cloud, as well as archiving through NetApp tiering.

NetApp hybrid cloud solutions help to build the analytics platform in minutes. Benefits include instantaneous cloning of data for QA preproduction; efficiency in moving data from one location to another; inexpensive and easy access to GPUs in the cloud; and avoiding lock-in to the cloud.

A recent study by 451 Research shows that more than 50% of their customers have their data lake in a public cloud. NetApp works with major cloud providers to offer flexible options and different service levels. For example, Azure NetApp Files offers basic, premium, and ultra levels so that customers can choose the right service level to meet their performance requirements. NetApp offers integration and solutions with cloud services such as infrastructure as a service (IaaS)—for example, Amazon EC2, Azure Virtual Machines and containers, and Google Compute Engine. Other integrations include platform as a service (PaaS), such as AWS Elastic Beanstalk, Google Application Engine, and Heroku; software as a service (SaaS), such as YAHOO email and Microsoft Office 365 ; and function as a service (FaaS)/serverless, such as AWS Serverless, AWS Lambda, Azure functions, Google Cloud functions, and building blocks for analytical workloads.

Technology partners

As shown in Figure 2, NetApp partners with AI and data analytics technology and solution partners. Together, we build integration and solutions with unique values to solve customers' challenges and generate business value. NetApp and our partners validate and integrate customers' products to build solutions to address their challenges and requirements.

Figure 2) NetApp partners with analytics and AI solution leaders.
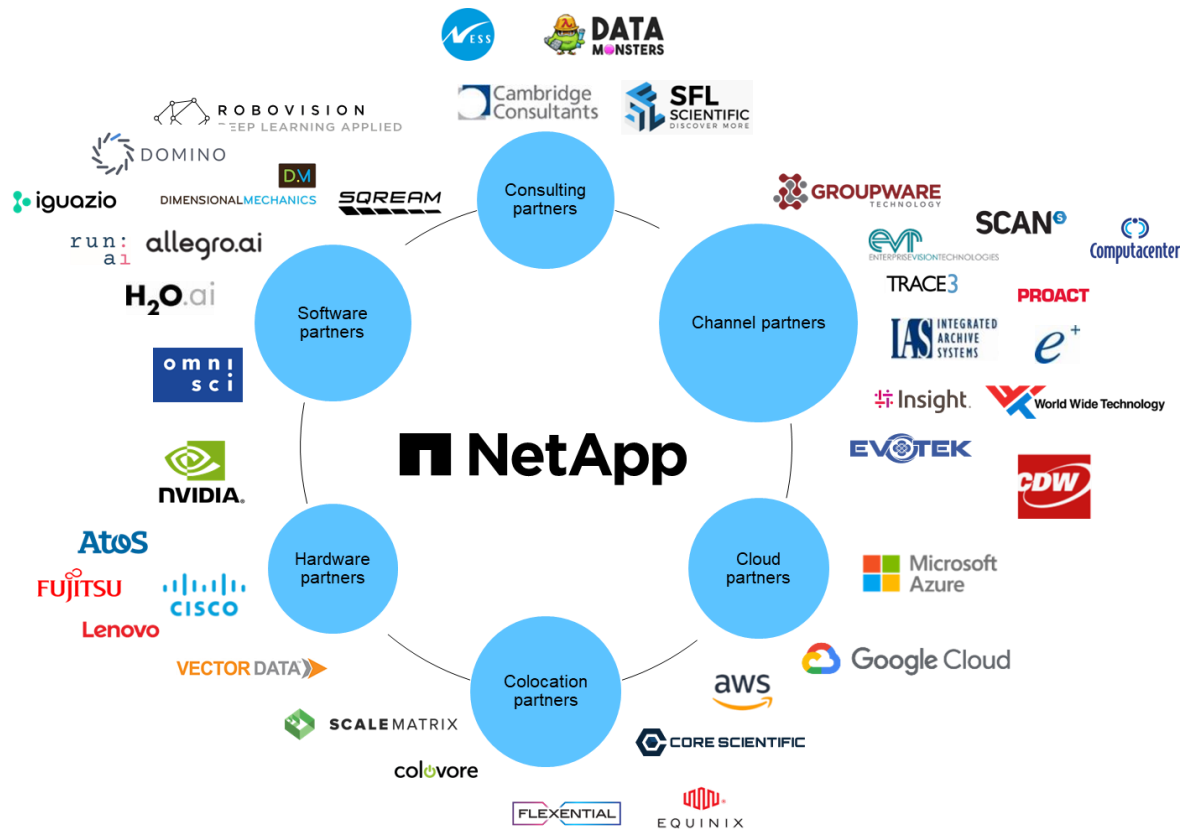
Customer challenges

Most customers experience at least some of the following data analytics challenges:

- The customer starts the data lake with DAS architecture to save on capital expenses, but operating expenses are high due to media failures and management issues.
- In DAS architecture, storage and compute are tied together, which limits the customer's ability to scale compute and storage independently.
- Most data lake clusters have HDFS as the default file system. That data gives good insights for data science operations. However, those operations are running in AI systems, and they are not aware of how to access and manage the HDFS and MapRFS data that resides in the data lake.
- Due to vendor lock-in, it's difficult to move the analytics data from one storage vendor to another.
- Customers need to reduce backup windows for large analytical data, and especially for banking sector data.
- Based on recent technology changes, customers need a data lake for multiple workloads such as analytics, data science workloads, and operation workloads.

NetApp DataOps Toolkit

The NetApp DataOps Toolkit (Figure 3) provides multitenant data management capabilities in a simple, easy-to-use interface that's designed for data scientists and data engineers. It's a set of functions based on ONTAP REST APIs in a Python library that makes it easy to perform various data management tasks without having to learn about storage management or call on storage admins for help. These tasks include near-instantaneous provisioning or cloning of a new data volume or JupyterLab workspace and making a near-instantaneous NetApp Snapshot™ copy of a data volume or JupyterLab workspace for traceability and base lining. This library can be used as a command line utility or imported into a Python program or Jupyter Notebook. The toolkit comes in two types, one for Kubernetes-based environments and one for virtual machine and bare-metal environments.

The latest version of the toolkit offers the following features and benefits. (See also the underlined release notes and the blog Accelerating the AI training workflow with the NetApp Data Science Toolkit.)

- Data scientists can store their training data directly in their JupyterLab workspace, which is transparently backed by a NetApp volume on the back end.
- Added support for BeeGFS CSI driver.
- E-Series storage systems enable data scientists to provision, list, and delete JupyterLab workspaces and/or Kubernetes-native persistent volumes on E-Series with parallel file systems.
- Provides access to advanced features without a storage administrator.

Figure 3) NetApp MLOps solutions, high level.
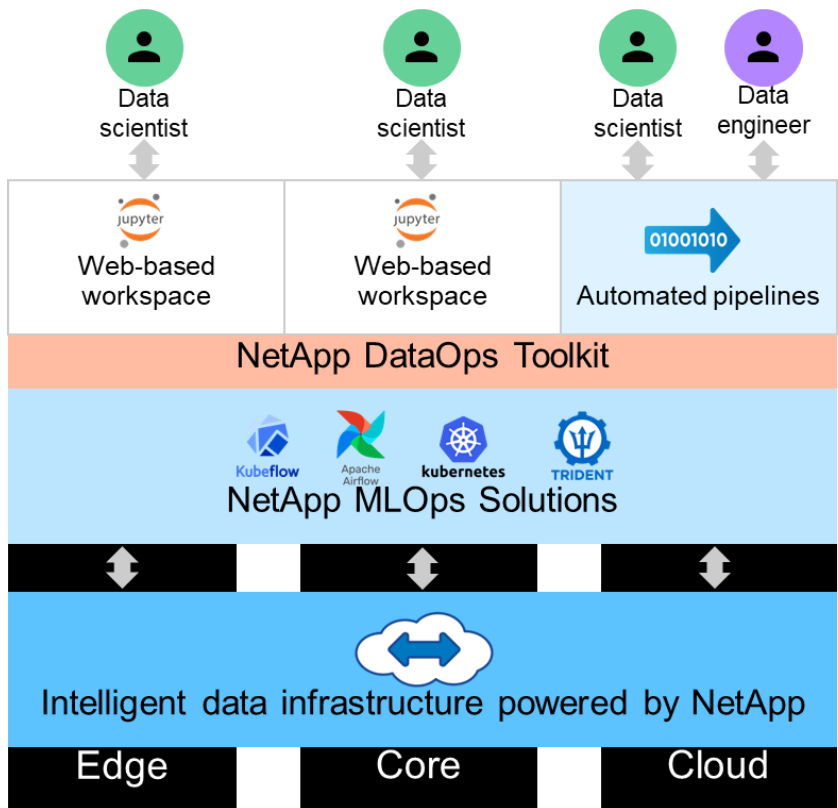
Hadoop to Spark

Traditional Hadoop clusters run on servers that experience data sprawl. Data locked into a data lake Hadoop Distributed Filesystem (HDFS) requires 1.5 to 3 times the storage required for the data. AI/DL training systems and other broad ranges of analytics systems don't understand HDFS. For example, healthcare, financial, and automotive customers need a solution to share HDFS data.

Spark provides similar functionality to the Hadoop cluster and excels at iterative computation with a high-quality algorithm that is 100 times faster than MapReduce. Spark includes two AI/ML libraries: MLlib and SparkML. MLlib algorithms are iterative and provide better performance than MapReduce with NetApp flash storage. Spark is much faster for AI for Hadoop services due to in-memory processing and

NetApp high-bandwidth network shared file system. Spark works with multiple sources such as HDFS, S3, NFS, and WASB.

Spark with object storage is having an impact on the current market. Spark eliminates the disk I/O issues of Hadoop. It is compute intensive, which can be addressed by adding more Spark worker nodes or GPUs. Apache Spark workloads with a NetApp storage solution offers customers options for the right unified NetApp storage.

Figure 4) NetApp analytics values.

Figure 5) NetApp solution for Apache Spark with archive.
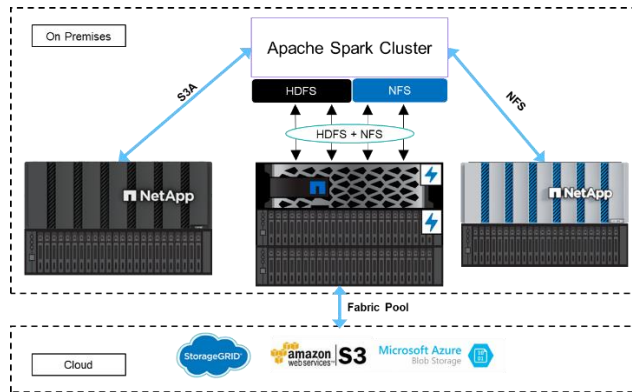
Figure 6) NetApp Astra Trident open-source dynamic storage orchestrator.
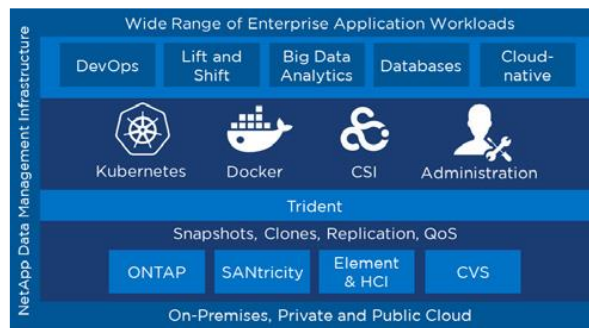
Archiving and tiering

Archiving and tiering decide the locations of data. Customers need to analyze their data usage and its growth. Currently, data lakes need simplicity, millions of objects, and bucket categorization. Based on the age of the data, they decide the data location. NetApp FabricPool moves the data to StorageGRID based on its age. Hadoop storage tiering enables multiple storage media, including disk, SSD, RAM disk, and archive.

NetApp solutions and products are integrated with data lake storage autotiering such as StorageGRID (object-based storage) for archiving, performance, and management. Data is moved automatically based on age. The tired data can be moved to the cloud or to the premises. NetApp StorageGRID can be used for cold storage, accessed on demand. The latest flash- based StorageGRID version offers improved performance.

Competition

According to a Coldago Research industry study, NetApp is #1 in file storage. The NetApp portfolio addresses file, block, and object storage across premises and cloud. It also differentiates based on data fabric mobility and hybrid multicloud data analytics solutions and positioning. Our competitors are years behind NetApp in offering consistent data management and services across premises and cloud that seamlessly support hybrid cloud workflows and analytics use cases. NetApp supports almost all

protocols, including NFS, SMB, NVMe/FC, iSCSI, HTTP, HDFS, and S3. This support means that NetApp provides a unified experience rather than disparate complex architectures.

NetApp offers rich features with AFF and FAS, such as NVMe media support, guaranteed efficiency, data reduction, QOS, predictive analytics, cloud tiering, replication, cloud deployment, and security. Most of our competitors have limited features, with separate products for each feature.

To help customers meet their requirements, NetApp offers features such as ONTAP File System Analytics, quotas, and on-box load balancing with no additional license costs. Our competitors might offer similar functionality with additional license costs. NetApp has better performance in the number of concurrent jobs, lower latency, simpler operations, and higher gigabytes per second throughput. NetApp Cloud Volumes ONTAP runs on all three major cloud providers. Our competitors support specific hardware and run on one or two clouds.

Containers

According to the CNCF Survey 2020, 55% of stateful applications such as file workloads, big data analytics, machine learning, media processing, and databases are running in containers in production. Customers need continuous availability of applications and data; NetApp Astra Trident solves persistence for stateful applications. It provides persistent volumes (PVs) for containers, a container storage interface that manages PVs with fully automated data services for Kubernetes, manageable by K8s, apps, or storage admins. That means one provisioner for all storage systems and services and native integration with K8s and storage systems.

Enterprise data management

NetApp offers the following enterprise management features for the customer's environment:

- NetApp products such as XCP software and BlueXP copy and sync can transition petabytes of data per day from a data lake file system to NFS or S3, creating a sharable data lake for analytics in a short period of time.
- NetApp solutions help customers to scale up or scale out storage and compute independently.
- Today's data analytics are required to access data from different sources through multiprotocol access. NetApp unified architecture has supported multiprotocol access for 25 years.
- Enterprise data protections such as backup and restore, cloning, disaster recovery, storage efficiency, QOS, and hybrid cloud are the basic needs of all enterprise customers. These functionalities are built into NetApp solutions and products.
- Data is sharable within the enterprise on the premises and across the cloud and back.

Software-defined storage with Astra in the cloud, analytics customers need software as a service (SaaS) for applications such as Spark, NoSQL databases, and AI instances. NetApp Astra Control Service is a fully managed service for cloud-native applications through end-to-end app and data lifecycle management. (See Figure 7.) This application-aware data management platform protects data with Snapshot copies, enables disaster recovery with remote backups, and manages migration with instant active clones. Using Astra, you can move selected workloads, or all workloads based on capacity or policy needs with full authorization and access control. You can easily control where and when you run your cloud-native applications.

Figure 7) End-to-end app-data lifecycle management for cloud applications.
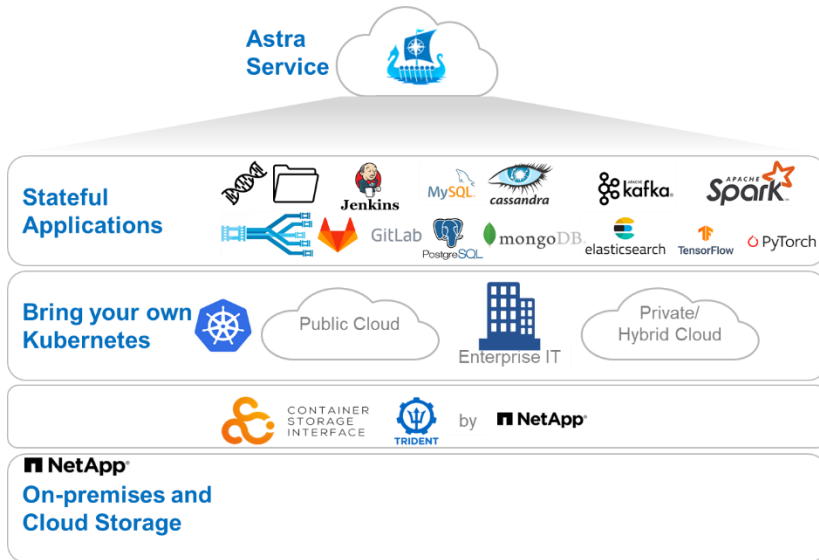
With Astra you can register and autodiscover apps in K8s, back up and recover, clone, and move workloads within and across clouds with just a few clicks. Astra is foundational to a hybrid model (on premises plus cloud), and it can help you to build your own data fabric across your data center and cloud. Astra offers solutions to different roles such as application admin, cloud infrastructure and Kube admin, and IT admin.

- **Application admin.** Run your stateful analytical applications worry-free on your K8s clusters. Astra automatically manages and protects your applications' state, configuration, and data.
- **Cloud infrastructure and Kube admin.** Use NetApp's proven data management and protection functionality to protect and move your analytical workloads running on K8s within and across clouds and on premises.
- **IT admin.** Your one-stop storage and data services provider for containerized workloads such as Spark running on K8s clusters.

Data mover

Data movement is a crucial part of digital transformation—specifically, moving data from edge to core to cloud and back. NetApp has three unique data mover products—NetApp SnapMirror, XCP software, and Cloud Sync—to migrate and move data. (See Figure 8.) And there's a simple way to choose the appropriate data mover product.

Use XCP software to migrate data quickly and easily from any NAS to NAS. The tool transfers files from a data lake file system to ONTAP NFS for AI and syncs data by using an NFS gateway using the official XCP product. Or it directly transfers data from the data lake file system. Because most AI infrastructures are compatible with NFS, XCP software is also used to migrate data from high-performance computing (GPFS and BeeGFS) to ONTAP NFS. In the cloud, the tool is used to move the data from an Azure Blob Storage, HDFS, or data processing file system to NetApp Cloud Volumes Service. XCP is suitable for a million or more file transfers, moving petabytes of data per day.

Use NetApp Cloud Sync to transfer between NFS,

Amazon S3, Azure Blob Storage, and NetApp StorageGRID. Cloud Sync is good for fewer than a million files. It has autoscaling capability.

Use NetApp SnapMirror to move data from ONTAP to ONTAP at the storage level.

Figure 8) NetApp Cloud Sync and XCP migration overview. Launch a data broker from NetApp BlueXP and manage the relationships you create.
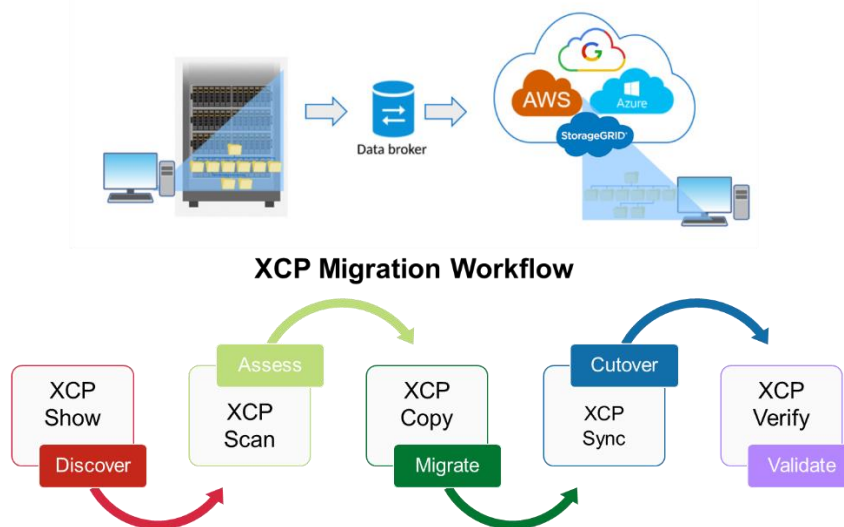
Active IQ

NetApp Active IQ provides predictive analytics for customers based on telemetric data collected from NetApp systems. (See Figure 9.) An Active IQ data lake runs with AFF that offers 11 times storage space reduction, 3 times compute cost reduction, and 30 times performance improvement compared to DAS architecture. We are running a single data lake Hadoop cluster for processing massive amounts of telemetric data to drive predictive analytics. Based on the processed data, we are able to provide input for NetApp Flash Advisor and Cloud Advisor.

Figure 9) NetApp Active IQ uses AI to turn data into actions.
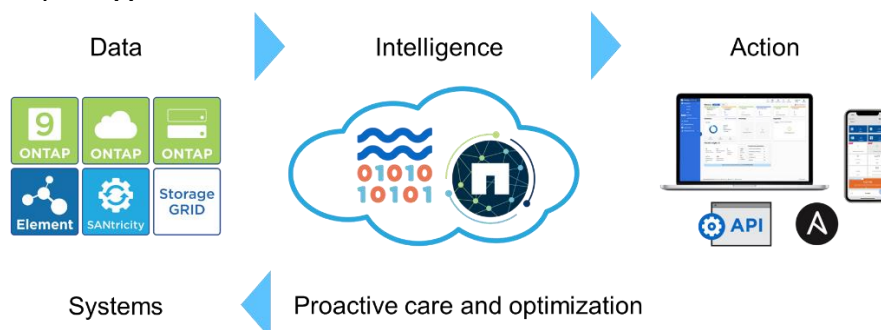
Conclusion

NetApp provides innovative, industry-leading modern data analytics solution strategies. They address business benefits, customer challenges, technology trends, competition, legacy architecture, modern workflows, use cases, industries, technology partners, data movers, BlueXP, Active IQ, DataOps Toolkit, software-defined storage with Astra, Hadoop to Spark, containers, enterprise data management, archiving, tiering, and an intelligent data infrastructure for storage and services across the hybrid multicloud environment. These strategies enable simplicity, availability, performance, scalability, and manageability of data to meet the ever-increasing demands of modern analytics workloads.

About NetApp

NetApp is the intelligent data infrastructure company, combining unified data storage, integrated data services, and CloudOps solutions to turn a world of disruption into opportunity for every customer. NetApp creates silo-free infrastructure, harnessing observability and AI to enable the industry's best data management. As the only enterprise-grade storage service natively embedded in the world's biggest clouds, our data storage delivers seamless flexibility. In addition, our data services create a data advantage through superior cyber resilience, governance, and application agility. Our CloudOps solutions provide continuous optimization of performance and efficiency through observability and AI. No matter the data type, workload, or environment, with NetApp you can transform your data infrastructure to realize your business possibilities.