Sponsored by:
NetApp Inc.

Authors:
Ritu Jyoti

February 2018

Shared Storage Offers Lower TCO than Direct-Attached Storage for Hadoop and NoSQL Deployments

TCO Highlights

— Shared Storage Versus DAS

Hadoop:

Shared storage saves
45% in initial costs and
54% in ongoing costs.

NoSQL database:

Shared storage saves
40% in initial costs and
44% in ongoing costs.

EXECUTIVE SUMMARY

IDC forecasts that by 2025, the global datasphere will grow to 163ZB (i.e., a trillion gigabytes). More than a quarter of this data will be real time in nature, and real-time IoT data will make up more than 95% of this. Data is the new basis of competitive advantage. IDC predicts that by 2019, 3rd Platform (cloud, social, mobile, and Big Data) technologies and services will drive nearly 75% of IT spending and help enterprises unlock unique user experiences and a new world of business opportunities.

Today, Big Data deployments are in the early stages of adoption, but with acceleration of digital transformation (DX) initiatives, they are becoming business critical. Storage and data management of these enormous data sets is key to an organization's longevity and success.

IDC interviewed two dozen enterprise customers currently running Hadoop and NoSQL databases (MongoDB, Cassandra, Couchbase, etc.) in production over direct-attached storage (DAS) — digital storage directly attached to the computer accessing it — to gain quantitative and qualitative insights into their deployments. With the deployments expanding in scope and business criticality, enterprises need a powerful data management platform that provides flexibility, resiliency, performance, efficiency, and integration with the cloud. The DAS solution not only lacks the necessary enterprise features but also has a higher total cost of ownership (TCO) than the alternative approach of using enterprise-grade shared storage (detailed in the Configuration TCO Analysis section in this white paper). Note that the quantitative insights for enterprise-grade shared storage were sourced from NetApp.

Situation Overview

Based on the in-depth interviews and broader IDC research on infrastructure requirements for Big Data and analytics (BDA) workloads, the key BDA deployment trends are as follows:

- **Data.** Data is increasingly being valued as an asset, and organizational data strategy aims to eliminate silos, improve quality, and support timely access of data. Because data is diverse, dynamic, and distributed, data-related challenges vary from security and compliance to data access, quality, and timely analysis.
- **Use cases.** Big Data solutions are becoming business critical and are being employed for a wide variety of use cases including ensuring adherence to regulatory compliance, time to market with new products/services, top-line revenue growth, and customer satisfaction.
- **Maturity.** Hadoop is more mature in deployment and has been running in enterprises as production instances for an average of 3.5 years. In contrast, NoSQL deployments have been running in production for an average of 2 years. Both Hadoop and NoSQL are being used by organizations. They are typically used for separate use cases and deployed as separate clusters. Hadoop-based data lakes are the majority as the core data analytics platforms. Spark, Drill, and Tableau are widely used for analytics.
- **Location.** While almost half of BDA deployments are on-premises for security, performance, and cost reasons, public cloud is being actively explored for agility. Hybrid deployments are expected to be the norm for BDA use cases, and easy synchronization across multiple cloud stacks is highly desired.
- **Runtime environment.** BDA deployments are mostly running in physical environments. Hypervisors/containers are in exploratory stages but are expected to gain adoption rapidly over the next couple of years.
- **Infrastructure type.** The majority of BDA deployments are running on DAS. The decision to use DAS is usually driven by the following factors:
  - Faster planning and deployment (including procurement)
  - The prescribed norm of the BDA software distribution vendor

  Enterprise-grade shared storage is used in some cases as the cluster grows. Software-defined storage and converged/hyperconverged storage are being seriously explored — with a decent number of deployments.
- **Media.** From a media perspective, hard drives are most commonly used, but flash usage is growing and all-flash arrays are increasingly being added to the cluster.

- **Data life-cycle management.** Backup/recovery, disaster recovery (DR), and archiving adoption, processes, and SLAs vary on a project-by-project basis and are predominantly on-premises. The average number of replicas for those using DAS is 3, but most of the organizations are planning to grow the number of replicas to 5 for resiliency reasons. Public cloud is used sparingly for archiving, yet the longer-term vision is to archive to public cloud or tier to lower-cost storage.
- **Deployment challenges.** Complexity and skill set gaps are reported to be significant challenges for both Hadoop and NoSQL database deployments. Typical preparation and planning duration is 9 months for Hadoop deployments, whereas it varies from 6 months to 9 months for NoSQL databases. Likewise, the deployment duration varies from 9 months to 12 months for Hadoop and from 4 months to 6 months for NoSQL databases. As the deployments mature, data life-cycle management becomes equally challenging.

Configuration TCO Analysis

Hadoop on DAS Versus Shared Storage

From the customer in-depth interviews, we selected a large financial institution for our configuration TCO analysis. This represents a sweet spot Hadoop deployment over DAS. The cluster stores live market data — about half a petabyte currently and growing at an annual rate of about 8–10%. The cluster instances (production, test/dev, backup, etc.) are all housed on-premises.

See Table 1 for cost specifics. The source for DAS costs is customer input via in-depth interview. The source for the same deployment on shared storage is NetApp. As shown in Table 1, using shared storage can help achieve 45% savings in initial costs and 54% cost savings on an annual basis.

The key difference in cost structure stems from the following basics used for calculation in the shared storage deployment:

- Compute costs for shared storage deployment are one-third the compute costs for DAS deployment due to independent scaling of compute and storage, leading to fewer servers to support compute needs, which in turn leads to lower software licensing costs. Because enterprise-level shared storage does not require servers with a large number of storage bays, these servers are also less expensive than those required for DAS.
- Capacity costs for shared storage deployment are one-third the capacity costs for DAS deployment due to reduction in the overall capacity needed due to storage of a single protected copy versus three copies of data in the DAS setup.

- The ratio of NetApp enterprise storage cost to raw media cost is 3:1. This includes the cost of the NetApp array in addition to any switching infrastructure required for NetApp storage versus just raw media in the servers.
- IT staff deployment, training, and ongoing support and lost productivity costs are lower due to a smaller setup of compute and storage infrastructure, centralized and intuitive enterprise-ready data management and governance, and more reliable and consistent infrastructure performance.

In addition, note that the capacity costs reduction will vary depending on the type and the source of data and how much the data can be deduped and compressed. For use cases such as IoT and log analytics, total reduction in capacity can be as high as 10 times because the data is quite repetitive, leading to a TCO that is even lower than the TCO detailed in Table 1 when dedupe and compression technologies are available and utilized.

TABLE 1 Example of Hadoop Cluster

	DAS	Shared Storage	Savings
Initial costs			
Software license	$330,000	$110,000	67%
Hardware — server, media, and switches	$212,000	$104,000	51%
IT staff for deployment at $100,000 including consulting support and operations	$1,142,857	$700,000	39%
Training (IT staff)	$94,862	$72,000	24%
Initial costs total)	$1,779,719	$986,000	45%
Annual costs			
Big Data solution software support	$300,000	$100,000	67%
Big Data solution hardware support	$120,000	$60,800	49%
Facilities (power, cooling, and rack space)	$529,400	$220,000	58%
Backup, archiving, and DR software and hardware	$544,444	$300,000	45%
Annual IT staff at $100,000	$8,467	$4,500	47%
IT user lost productivity at $70,000	$17,376	$8,000	54%
Annual costs total	$1,519,687	$693,300	54%

Source: IDC, 2018

NoSQL Database on DAS Versus Shared Storage

For NoSQL database deployment TCO analysis, we picked a sweet spot MongoDB deployment running on DAS at a large financial institution included in our in-depth interviews. The cluster stores research articles in MongoDB — about 50TB currently and growing at an annual rate of about 5%. The cluster instances (production, test/dev, backup, etc.) are all housed on-premises and serve hundreds of users worldwide.

See Table 2 for cost specifics. The source for DAS costs is customer input via in-depth interview. The source for the same deployment on shared storage is NetApp. As shown in Table 2, using shared storage can help achieve 40% savings in initial costs and 44% cost savings on an annual basis.

The key difference in cost structure stems from the following basics used for calculation in the shared storage deployment:

- Compute costs for shared storage deployment are one-third the compute costs for MongoDB DAS deployment enabled by independent scaling of compute and storage, leading to fewer servers, which in turn leads to lower software licensing costs. Because enterprise-level shared storage does not require servers with a large number of storage bays, these servers are also less expensive than those required for DAS.
- Capacity costs for shared storage deployment are half the capacity costs for MongoDB DAS deployment enabled by reduction in the overall capacity needed due to storage of a single copy versus multiple copies of data in the DAS setup.
- The ratio of NetApp enterprise storage cost to raw media cost is 3:1. This includes the cost of NetApp array and any switching infrastructure required for NetApp storage versus just raw media in the server.
- IT staff deployment, training, and ongoing support and lost productivity costs are lower due to a smaller setup of compute and storage infrastructure, centralized and intuitive enterprise-ready data management and governance, and more reliable and consistent infrastructure performance.

In addition, note that the deployment over NetApp shared storage provides an excellent user experience with the same backup and recovery tools (via SnapCenter) as with Oracle databases. NetApp All Flash FAS (AFF) also delivered sub-1ms sustained performance at scale and near-zero RPO and greatly reduced RTO, providing faster recovery while preventing data loss. All this contributes to lower data management and lost productivity costs.

TABLE 2 Example of NoSQL Database (MongoDB) Cluster

	DAS	Shared Storage	Savings
Initial costs			
Software license	$70,000	$24,500	65%
Hardware — server, media, and switches	$52,000	$46,950	10%
IT staff for deployment at $100,000 including consulting support and operations	$56,548	$33,000	42%
Training (IT staff)	$35,714	$25,000	30%
Initial costs total)	$214,262	$129,450	40%
Annual costs			
Big Data solution software support	$65,000	$22,750	65%
Big Data solution hardware support	$14,000	$13,351	5%
Facilities (power, cooling, and rack space)	$14,286	$7,000	51%
Backup, archiving, and DR software and hardware	$80,000	$55,000	31%
Annual IT staff at $100,000	$8,632	$5,000	42%
IT user lost productivity at $70,000	$15,747	$7,000	56%
Annual costs total	$197,665	$110,101	44%

Source: IDC, 2018

Considering NetApp for Hadoop and NoSQL Databases

The NetApp data platform unifies insight across various data sources and multiple cloud deployments. NetApp offers an enterprise storage architecture with validated storage building blocks stretching across new deployments as well as in-place analytics on existing data, guaranteeing lower total cost of ownership and risk than DAS.

The NetApp approach provides:

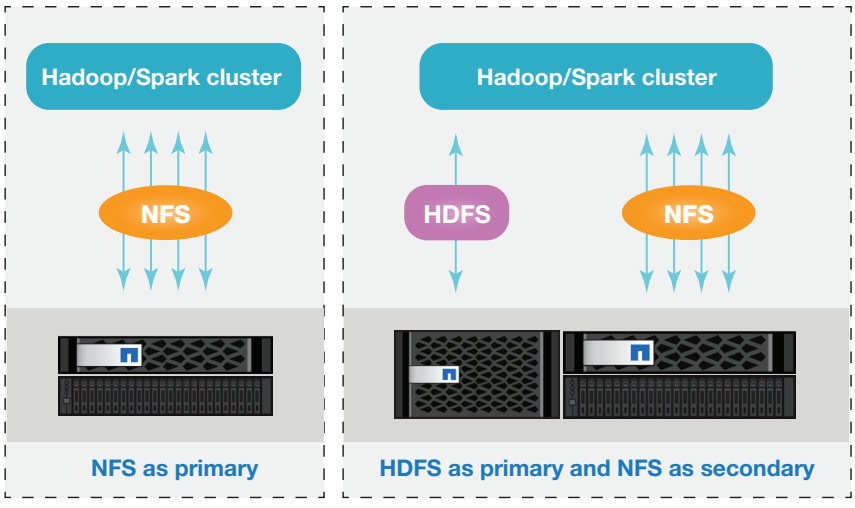
- **A mature and scalable solution architecture that includes validated designs, technical reports, and complete runbooks, which shortens time to value, increases deployment stability, and reduces consulting expenses.** See Figures 1 and 2 for reference architectures provided by NetApp for Hadoop and NoSQL database deployments, respectively.

- **The ability to handle structured, semistructured, and unstructured data with the portfolio of products and multiprotocol support.**
- **Sustained consistent performance with leadership optimized offerings for BDA workloads.** Because data is externally protected with the NetApp shared storage offering, additional performance and efficiency gains can be realized by reducing the amount of data replication, lightening the load on compute and network resources, and reducing the amount of storage required just for data protection.
- **Flexible deployment with Data Fabric for on-premises, near cloud, hybrid cloud, and multicloud setup**.
- **Accelerated testing and development of Big Data solutions by ensuring seamless data movement between on-premises, near cloud, and private and public cloud environments, along with automated tiering of cold data to the cloud utilizing advanced data life-cycle management.** Alternatively, data can be tiered and replicated to NetApp Private Storage (NPS), which places storage near the cloud while retaining data sovereignty and enabling analytics or — for the most critical use cases — backup and disaster recovery in the cloud.
- **Reduced costs by providing independent scaling of compute and storage resources, enabling higher utilization and fewer, less expensive servers.** This also eliminates the need for rebalancing or migration when new data nodes are added, thereby making the data life-cycle nondisruptive.
- **The ability to perform in-place analytics on existing NAS data using NetApp technologies, thereby reducing infrastructure cost.** It also eliminates the time and cost to unnecessarily duplicate existing data to a data lake and provides faster time to insights.
- **A dramatic reduction in the facilities (power, cooling, and space) costs through the reduced data volume realized by the use of data efficiency services such as deduplication and compression.** Additional savings in facilities costs come from leveraging the built-in NetApp storage resiliency and eliminating the need to double storage with 1:1 mirroring to protect data.
- **Enterprise-grade reliability and availability of data.** NetApp ONTAP provides numerous options for efficiently protecting and managing data. Snapshot creates near-instant point-in-time data copies without impacting performance, making backup and restore quick and efficient. Synchronous and asynchronous replication using NetApp MetroCluster and SnapMirror technologies provides additional protection. Data is deduplicated and compressed automatically — in many cases drastically reducing the

data footprint, saving money, and enabling faster data movement and more cost-effective data protection.

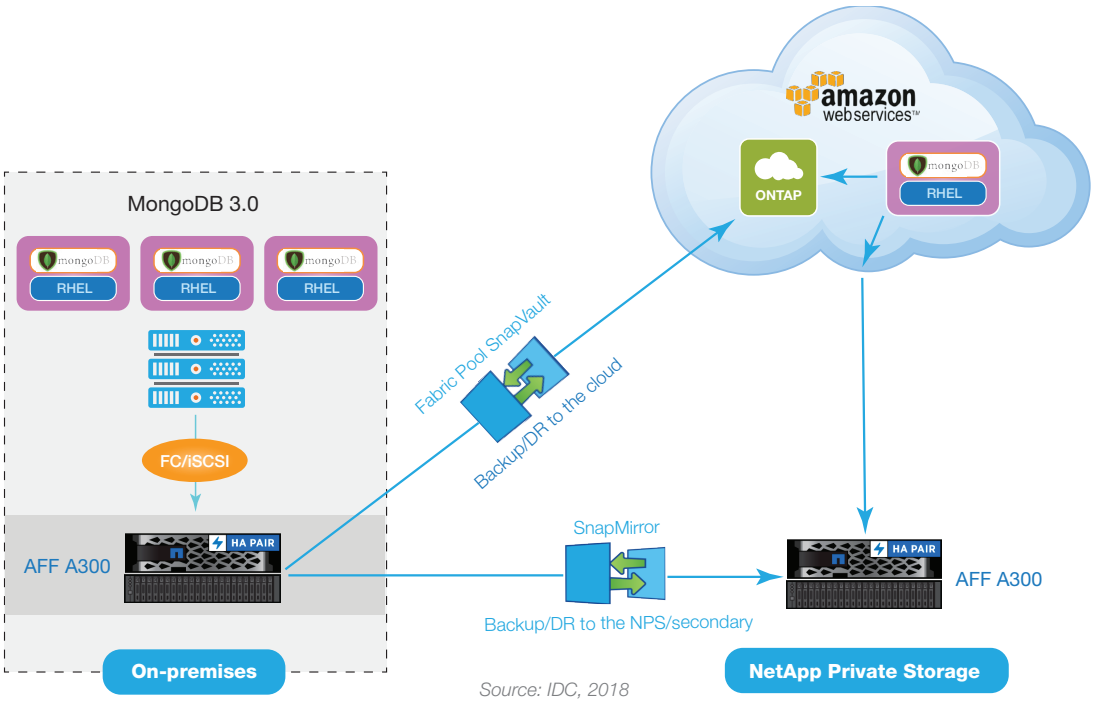
- **Consistent enforcement of data security, privacy, governance, and compliance, plus support for in-flight and at-rest data encryption.**
- **Reduced operational complexity, including speed of provisioning capacity and users with centralized data management, and a single console view of the entire storage infrastructure across on-premises, near cloud, and cloud.** SnapCenter and NetApp plug-ins enable application-specific copies to be automatically created for both traditional and Big Data applications, all with the same user experience. Besides data protection, the lightweight clones allow developers to utilize real production data for testing without impacting production in any way. OnCommand Workflow Automation automates the provisioning of new workflows in minutes and supports more than 60 enterprise applications out of the box, including both traditional and Big Data applications. For example, OnCommand software automates cross-datacenter provisioning for MongoDB sharded clusters. These are complemented by OnCommand Insight for system health checks and NetApp AutoSupport.
- **NetApp ONTAP provides powerful QoS features, allowing administrators to leverage preset policies or create their own policies.** In addition, ONTAP Adaptive QoS automatically adjusts the QoS level to changes in workload. The shared storage topology with ONTAP allows the data to be handled as one large pool of data rather than many distributed smaller pools of data, thereby speeding up access times.

FIGURE 1 Hadoop on NetApp Shared Storage (AFF/FAS on ONTAP) — Reference Architecture



Source: IDC, 2018

FIGURE 2 NoSQL Database on NetApp Shared Storage (AFF) — Reference Architecture



Source: IDC, 2018

Challenges/Opportunities for NetApp

In the DX era, where adoption of Hadoop and NoSQL databases has begun in earnest, NetApp has the opportunity to continue to innovate and ensure that its current and new clients are able to fulfill their Big Data analytics strategies and plans:

- NetApp needs to continue to enhance its ability to provide a Big Data platform that can be deployed as readily on-premises as off-premises in private and public cloud environments (Amazon, Azure, Google, etc.) and enable one-click application mobility from any platform to any other platform (physical or virtual).
- To provide an end-to-end solution experience for end users, NetApp needs to maintain its commitment to expand existing ecosystem integration with open source and Big Data solution providers, including Hadoop software distribution providers such as Cloudera, Hortonworks, and MapR and NoSQL database providers such as MongoDB, Cassandra, and Couchbase.

- NetApp also needs to sustain its policy of making its enterprise data platform available through a variety of consumption models and partners. Existing alternative consumption models include ONTAP Cloud, a pay-as-you-go offering on Amazon Web Services and Microsoft Azure; NetApp Private Storage, a "near cloud" monthly subscription storage solution offered by NetApp partners; and ONTAP Select, a software-defined storage solution deployed on commodity hardware and licensed on a capacity basis.
- NetApp needs to continue to embrace a software-defined infrastructure strategy with a rich set of data services including erasure coding and semantic deduplication for varied data formats and hybrid/multicloud deployments in its solution offerings.
- NetApp's vision for data management is a data fabric that seamlessly connects different clouds, whether private, public, or hybrid environments. NetApp Data Fabric simplifies, automates, and evolves the management of data. NetApp's transformation from a hardware focus to software and cloud services will empower customers to unleash data to accelerate digital transformation and address their business imperatives. NetApp's cloud portfolio is an integral part of the NetApp Data Fabric, and the company is expanding key strategic partnerships in the cloud. IDC believes NetApp's forward-looking plans make the company a thought leader within the enterprise infrastructure space and sets the company apart from many other infrastructure vendors of comparable size and age. That said, it's important to note that NetApp will need to execute on its vision at a time of heightened competition from incumbents as well as start-ups and a new and fast-evolving data services landscape.

Conclusion

In evaluating DAS or enterprise-grade shared storage for BDA (e.g., Hadoop or NoSQL databases) deployments, IT professionals are advised to take their long-term plans and objectives into consideration. DAS could be a good starting point for some of the use cases but is challenging as the deployment grows in size and varied data formats and expands in footprint across locations and multiclouds. Enterprise-grade shared storage is beneficial as data life-cycle management, security, compliance, consistent performance, and overall costs become critical.

Enterprise storage solutions pretested with Hadoop distributions, NoSQL databases, and applications such as Splunk and Spark typically provide better and more consistent performance, along with single-interface management across storage environments and locations. Customers particularly value the independent scaling of storage and compute, performance tiering, and space efficiency due to single source of data, no resync, and no copy.

In addition, there is a TCO advantage stemming from lower software license costs (reduction due to use of fewer servers driven by independent scaling of compute and storage), savings on storage capacity by reduction in replication factor and utilization of technologies such as deduplication and compression, and lower facilities costs due to reduction in power, cooling, and space costs. These customers also enjoy better uptime, recoverability, and performance than customers that used internal DAS during data rebuilds.

IDC believes the future of enterprise storage is software defined, server based, and cloud connected, with a full suite of enterprise storage data services. IDC recommends that as BDA deployments become business critical, enterprise storage solution providers should continue to integrate with the ecosystem of BDA software providers and help customers optimize their agility, performance, scale, reliability, and security needs along with the total cost of ownership.

IDC Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-insights-community.com
www.idc.com

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

IDC
ANALYZE THE FUTURE