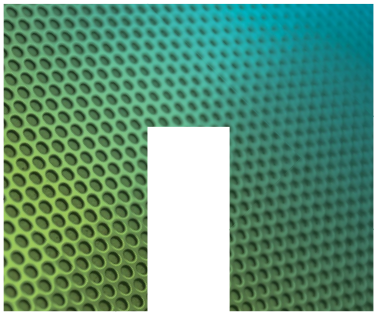

LÖSUNGSÜBERBLICK

NetApp Lösungen für Hadoop

Implementieren einer Hadoop Plattform der Enterprise-Klasse mit höherer Performance, geringerer Cluster-Ausfallzeit und linearer Skalierbarkeit

DIE WICHTIGSTEN VORTEILE

Bessere Performance
- Verringerung der Laufzeit von Wochen auf Stunden
- Performance-Steigerung um über 500 % bei Datenwiederherstellung mit DDP-Technologie im Vergleich zu herkömmlichem RAID 5[1]

Keine Cluster-Ausfallzeit
- Verfügbarkeit von mindestens 99,999 % mit den äußerst zuverlässigen NetApp E-Series Storage-Systemen

Lineare Skalierbarkeit
- Skalierung von Workloads ohne Performance-Einbußen

Senkung der Betriebskosten
- Schnelle Implementierung einer vordefinierten und vorab validierten Storage-Lösung
- Verbesserung der Storage-Effizienz um 33 % und Senkung der Betriebskosten

Flexibilität und Wahlfreiheit
- Unterstützung von mit Apache kompatiblen Distributionen, einschließlich Cloudera, Hortonworks und MapR

Die Herausforderung

Ausschöpfen des Potenzials von Big Data

Mithilfe von Apache Hadoop und seinem stetig wachsenden Produkt-Ecosystem können Unternehmen nützliche Informationen aus großen Volumes diverser Daten extrahieren, die nicht mit herkömmlichen relationalen Datenbanken analysiert werden können. Dank dieser Informationen können Mitarbeiter im gesamten Unternehmen die richtigen Fragen stellen und dann fundiertere Entscheidungen treffen sowie die geschäftliche Transformation vorantreiben.

Da die ersten Hadoop Implementierungen meistens auf Standard-Servern mit internen Laufwerken vorgenommen wurden, können Unternehmen aufgrund der robusten Infrastruktur und der geringen Flexibilität nicht das volle Potenzial ihrer Hadoop Implementierung ausschöpfen. So beeinträchtigt beispielsweise der Ausfall einer einzigen Festplatte die Performance des gesamten Clusters. Ersatzfestplatten müssen kontinuierlich gemanagt werden. Dabei treten leicht Fehler auf. Die dreifache Datei-Replizierung und die Umverteilungsmodelle bei Ausfällen tragen zu steigenden Kosten und komplexeren Netzwerken bei. Standard-Server mit internen Laufwerken unterstützen Nutzungsfälle, die unterschiedliche Verarbeitungsleistungen und Storage-Anforderungen in der gleichen Infrastruktur erfordern, nur unzureichend.

Die Lösung

Enterprise-Storage für Hadoop

NetApp Lösungen für Hadoop enthalten Storage-Bausteine der Enterprise-Klasse, die unabhängig von den Computing-Servern sind. So kann eine Implementierung der Enterprise-Klasse mit einer geringeren Cluster-Ausfallzeit, höheren Datenverfügbarkeit und linearen Skalierbarkeit bereitgestellt werden. Falls eine Festplatte ausfällt, wirkt sich dies mit NetApp E-Series und der DDP-Technologie (Dynamic Disk Pool) kaum auf die Performance aus. Die Recovery erfolgt zehnmal schneller als mit typischen RAID-Systemen auf Standard-Servern mit internem Storage. Mithilfe dieser Lösungen können Daten-Nodes unterbrechungsfrei hinzugefügt werden. Es ist keine Ausbalancierung oder Migration erforderlich. Die externe Datensicherung sorgt für einen kleinen Storage-Platzbedarf und einen geringeren Datenreplizierungs-Overhead.

Bessere Performance

Bei einem Durchsatz von bis zu 12 GB/s und bis zu 825.000 IOPS unterstützen die NetApp E-Series Storage-Systeme bandbreitenintensive Vorgänge. Bei einer kürzlich durchgeführten ESG Lab-Validierung wurde ein Hadoop Cluster mit zehn Nodes auf einem E-Series System getestet. In dieser Konfiguration sank die Laufzeit für eine Abfrage von 24 Milliarden unstrukturierten Datensätzen von vier Wochen auf 10,5 Stunden. Das ist eine Beschleunigung von mehr als 94 %. Eine zweite Abfrage von 240 Milliarden unstrukturierten Datensätzen, bei der zuvor eine Zeitüberschreitung auftrat, wurde in 18 Stunden durchgeführt.

Auf den E-Series Systemen können Sie die Daten je nach Zugriffspriorität (heiße, warme, kalte und eingefrorene Daten) auf den entsprechenden Storage-Medien – Solid State Drive, SAS und NL-SAS-Laufwerken – in der gleichen Datenmanagement-Architektur speichern. Dank dieser Funktion wird die Umgebung erheblich effizienter und kostengünstiger.

1. ESG Lab-Validierung für NetApp Lösungen für Hadoop, November 2015

NetApp®

cloudera®
NetApp Alliance-Partner

Hortonworks
NetApp Alliance-Partner

MAPR
NetApp Alliance-Partner

Keine Cluster-Ausfallzeit

Festplattenlaufwerke sind die fehleranfälligsten Komponenten in der Hadoop Architektur. Es ist lediglich eine Frage der Zeit, bis eine Festplatte ausfällt. Je größer das Cluster, desto wahrscheinlicher wird ein Festplattenausfall. In einer herkömmlichen Implementierung von Hadoop Distributed File System (HDFS) wird bei einem Festplattenausfall ein Job neu initiiert, was wiederum zu einer längeren Ausfallzeit führt. Mit NetApp Lösungen für Hadoop ist der Neustart eines Jobs bei einem Festplattenausfall dank der DDP-Sicherung der E-Series Systeme nicht erforderlich.

Lineare Skalierbarkeit

In herkömmlichen Hadoop Implementierungen, die Standard-Server mit internem Storage verwenden, wird die Performance beim Hinzufügen weiterer Daten-Nodes beeinträchtigt. Da das NetApp E-Series Design mit Bausteinen der Enterprise-Klasse die Computing- von den Storage-Instanzen trennt, können die Kapazität und die Performance separat skaliert werden. ESG hat die Skalierbarkeit der NetApp Lösungen für Hadoop anhand des E5660 Systems getestet. Bei der Skalierung von vier Daten-Nodes und 60 Laufwerken auf acht Daten-Nodes und 120 Laufwerke stellte ESG keine Beeinträchtigung der Ladezeit der Daten fest. Die Zeit zum Sortieren der Daten wurde sogar um bis zu 11 % verkürzt. Damit wurde nicht nur die erforderliche Kapazität der Workload-Anforderungen erreicht, sondern auch die Performance gesteigert.

Senkung der Betriebskosten

Mit vordefinierten, vorkonfigurierten und vorab validierten NetApp Lösungen für Hadoop können Sie Hadoop schnell implementieren und sofort Erkenntnisse aus Ihren Daten gewinnen.

FlexPod Select für Hadoop vereinfacht die Implementierung noch weiter und ermöglicht die zukünftige Skalierung. In einer vorkonfigurierten Lösung werden Storage, Networking und Server kombiniert, die für Hadoop Umgebungen der Enterprise-Klasse validiert sind. FlexPod Select nutzt NetApp E-Series Storage, der mit Cisco UCS C-Series Servern verbunden ist. So werden Hochverfügbarkeit, nahtlose Skalierbarkeit und bessere Storage-Effizienz sichergestellt.

Gleiche Verfügbarkeit mit weniger Storage-Hardware

Dank der DDP-Technologie muss Hadoop nur zwei statt drei Datenkopien erstellen und speichern. Dadurch wird weniger Storage benötigt. Zudem werden die Kosten für die Stromversorgung und Kühlung sowie die Betriebskosten gesenkt.

Flexibilität und Wahlfreiheit

Jede NetApp Lösung wurde für den Einsatz in den primären mit Apache kompatiblen Distributionen – Cloudera, Hortonworks und MapR – zertifiziert. Dank dieser Flexibilität ergibt sich eine bessere Interoperabilität im Hadoop Framework und Hadoop kann in Kombination mit bereits vorhandenen Big-Data-Tools oder mit Tools anderer Analyseplattformen genutzt werden.

Aufgrund der rasanten Entwicklung im Analysebereich ist es wichtig, schnell und einfach zwischen Produkten wechseln zu können, damit Sie stets die optimale Lösung für Ihre Geschäftsanforderungen verwenden. Da NetApp Lösungen in Produkte von proprietären und Open-Source-Anbietern integriert werden können, vermeiden Sie eine Anbieterbindung. Diese ist häufig der ausschlaggebende Punkt für den Wechsel zu Open-Source-Technologien. Es stehen mehrere Storage-, Server- und Networking-Optionen zur Verfügung, sodass Sie die optimale Lösung für Ihr Unternehmen auswählen können.

Weitere Lösungen dazu

NetApp FAS NFS Connector für Hadoop

Nutzen Sie den NetApp FAS NFS Connector für Hadoop für Big Data-Analysen älterer NFSv3-Daten – ohne die Daten zu verschieben, ein separates Analysesilo zu erstellen oder ein Hadoop Cluster einzurichten. Mithilfe des NFS Connector können Sie von HDFS zu NFS wechseln oder NFS zusammen mit HDFS ausführen. Der NFS Connector funktioniert mit MapReduce für Computing- oder Verarbeitungsaufgaben und unterstützt weitere Apache Projekte wie HBase (spaltenbasierte Datenbank) und Spark (mit Hadoop kompatible Verarbeitungs-Engine).

Durch diese Funktionen kann NFS Connector noch viele weitere Arten von Workloads unterstützen – zum Beispiel Batch, In-Memory und Streaming. Weitere Informationen finden Sie online unter http://www.netapp.com/de/solutions/big-data/nfs-connector-hadoop.aspx.

FAS Systeme stellen NFS für den Namens-Node Storage bereit, wodurch das Managen des Clusters und das festplattenlose Booten vereinfacht werden. Zum besseren Schutz der Metadaten wird eine Kopie wichtiger NameNode Daten auf dem FAS System gespeichert. Sollte der NameNode Server ausfallen, können die Daten innerhalb weniger Minuten wiederhergestellt werden. Es dauert also nicht mehr Stunden oder Tage wie bei internen SATA-Laufwerken. Festplatten-Shelfs, die im laufenden Betrieb hinzugefügt werden können, verbessern ebenfalls die Verfügbarkeit, da für das Hinzufügen oder Ersetzen keine Cluster-Ausfallzeit geplant werden muss.

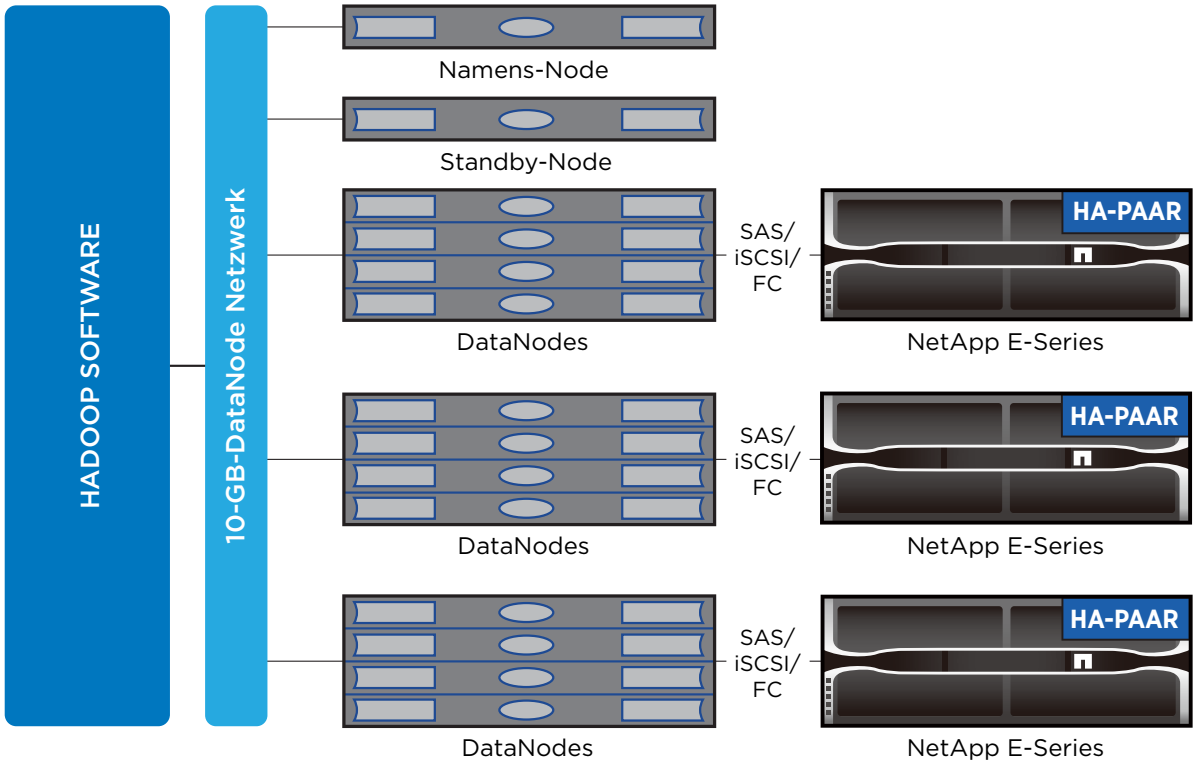

Abbildung 1) NetApp Lösungen für Hadoop – Architekturdiagramm

Management des Daten-Lebenszyklus mit Zaloni
NetApp und Zaloni haben zusammen eine Lösung entwickelt, um die Data-Fabric-Strategie von NetApp auf das Management des Daten-Lebenszyklus für Data Lakes zu erweitern. Mit dieser Funktion können Sie Richtlinien für den Daten-Lebenszyklus festlegen und anwenden, sodass Ihre Mitarbeiter Daten über NetApp Storage Tiers hinweg managen und verschieben können.

Mehr Informationen dank Big Data Analytics

Die bewährten NetApp Lösungen für Hadoop unterstützen jeden bei umfangreichen Datenanalysen – von Unternehmensinhabern über Big-Data-Nutzer und Datenexperten bis zu Entwicklern und Administratoren. Da weniger Hardware benötigt wird und die Kosten für Strom und Softwarelizenzen gesenkt werden können, bieten die NetApp E-Series Storage-Bausteine der Enterprise-Klasse einen größeren Mehrwert und niedrigere Gesamtbetriebskosten.

Dank der besseren Uptime und Performance können Sie jederzeit die gewünschten Informationen abrufen. Aufgrund der Trennung der Computing- und Storage-Instanzen können Sie diese unabhängig voneinander skalieren und besser an die sich ändernden Applikationsanforderungen anpassen. Storage Tiers für heiße, warme, kalte und eingefrorene Daten helfen bei der Optimierung der Infrastruktur entsprechend Ihren Anforderungen. Die zentrale Schnittstelle vereinfacht das Management des gesamten Hadoop Clusters. Wenn Sie bereits Daten in NetApp FAS Systemen gespeichert haben, können Sie mithilfe des NetApp FAS NFS Connector für Hadoop auch Analysen der vorhandenen Daten durchführen.

Sollten Sie Hilfe beim Design oder bei der Implementierung der NetApp Lösungen für Hadoop benötigen, helfen Ihnen die NetApp Services Experten und unsere zertifizierten Partner gerne weiter.

Weitere Informationen zu NetApp Lösungen für Hadoop finden Sie unter
http://www.netapp.com/de/solutions/big-data/hadoop.aspx.

Info zu NetApp

Unternehmen in aller Welt zählen auf die Software, Systeme und Services von NetApp, um ihre Daten zu managen und zu speichern. Kunden schätzen unsere Teamarbeit, unsere Expertise und unser Engagement für ihren Erfolg.

www.netapp.de