NetApp Hybrid-Cloud-Datenlösungen – Spark und Hadoop basierend auf Kundenanwendungsfällen

NetApp Solutions

NetApp
April 25, 2024

This PDF was generated from https://docs.netapp.com/de-de/netapp-solutions/data-analytics/hdcs-sh-solution-overview.html on April 25, 2024. Always check docs.netapp.com for the latest.

Inhalt

NetApp Hybrid-Cloud-Datenlösungen – Spark und Hadoop basierend auf Kundenanwendungsfällen 1
TR-4657: NetApp Hybrid-Cloud-Datenlösungen – Spark und Hadoop auf Basis von Kundenanwendungsfällen 1
Data Fabric von NetApp für Big Data-Architekturen 2
Hadoop Datensicherung und NetApp 6
Übersicht über Anwendungsfälle für die Datensicherung in Hadoop 6
Anwendungsfall 1: Sichern von Hadoop Daten 7
Anwendungsfall 2: Backup und Disaster Recovery aus der Cloud in On-Premises-Systeme 11
Anwendungsfall 3: Enabling DevTest on Existing Hadoop Data 13
Anwendungsfall 4: Datensicherung und Multi-Cloud-Konnektivität 15
Anwendungsfall 5: Schnellere Analyse-Workloads 16
Schlussfolgerung 18

NetApp Hybrid-Cloud-Datenlösungen – Spark und Hadoop basierend auf Kundenanwendungsfällen

TR-4657: NetApp Hybrid-Cloud-Datenlösungen – Spark und Hadoop auf Basis von Kundenanwendungsfällen

Karthikeyan Nagalingam und Sathish Thyagarajan, NetApp

In diesem Dokument werden Hybrid-Cloud-Datenlösungen beschrieben, die NetApp AFF und FAS Storage-Systeme, NetApp Cloud Volumes ONTAP, NetApp Connected Storage sowie NetApp FlexClone Technologie für Spark und Hadoop einsetzen. Mit diesen Lösungsarchitekturen können Kunden eine geeignete Datensicherungslösung für ihre Umgebung wählen. NetApp hat diese Lösungen basierend auf der Interaktion mit Kunden und eigenen Business-Anwendungsfällen entwickelt. Dieses Dokument enthält folgende detaillierte Informationen:

- Warum wir Datensicherung für Spark und Hadoop-Umgebungen sowie auf Kundenherausforderungen benötigen
- Die Data Fabric basiert auf der NetApp Vision und den Bausteinen und Services.
- Wie diese Bausteine zur Erstellung flexibler Datensicherungs-Workflows eingesetzt werden können
- Vor- und Nachteile verschiedener Architekturen, die auf realistischen Kundenanwendungsfällen basieren. Jeder Anwendungsfall umfasst die folgenden Komponenten:
  - Kundenszenarien
  - Anforderungen und Herausforderungen zu bewältigen
  - Verbessern Lassen
  - Zusammenfassung der Lösungen

Warum Hadoop Datensicherung?

In einer Hadoop und Spark-Umgebung müssen die folgenden Aspekte berücksichtigt werden:

- **Software oder menschliches Versagen.** menschliches Versagen bei Software-Updates während der Durchführung von Hadoop-Datenoperationen kann zu fehlerhaftem Verhalten führen, das unerwartete Ergebnisse aus dem Job verursachen kann. In diesem Fall müssen wir die Daten schützen, um Fehler oder unangemessene Ergebnisse zu vermeiden. Als Ergebnis eines schlecht ausgeführten Softwareupdates auf eine Traffic-Signal-Analyse-Anwendung, eine neue Funktion, die nicht richtig Analyse von Verkehrssignaldaten in Form von Klartext. Die Software analysiert weiterhin JSON und andere nicht-Text-Dateiformate, was zu einem Echtzeit-System zur Verkehrssteuerung führt, in dem Vorhersageergebnisse vorliegen, die keine Datenpunkte enthalten. Dies kann zu fehlerhaften Ausgängen führen, die zu Unfällen an den Verkehrssignalen führen können. Datensicherheit kann dieses Problem lösen, indem sie die Möglichkeit bietet, schnell ein Rollback zur vorherigen Applikationsversion durchführen zu können.
- **Größe und Umfang.** die Größe der Analysedaten wächst Tag für Tag aufgrund der stetig wachsenden Zahl von Datenquellen und Volumen. Social Media, mobile Apps, Data Analytics und Cloud-Computing-

Plattformen sind die wichtigsten Datenquellen des aktuellen Big Data-Marktes, der sehr schnell zunimmt und daher die Daten zur Sicherstellung akkurater Datenabläufe geschützt werden müssen.

- **Der native Datenschutz von Hadoop.** Hadoop hat einen nativen Befehl, um die Daten zu schützen, aber dieser Befehl bietet keine Konsistenz der Daten während des Backups. Es unterstützt nur Backups auf Verzeichnisebene. Die von Hadoop erstellten Snapshots sind schreibgeschützt und können nicht zur direkten Wiederverwendung der Backup-Daten verwendet werden.

Herausforderungen bei der Datensicherung für Hadoop und Spark Kunden

Kunden in Hadoop und Spark stehen vor der Herausforderung, die Backup-Zeit zu reduzieren und die Backup-Zuverlässigkeit zu steigern, ohne während der Datensicherung die Performance im Produktions-Cluster zu beeinträchtigen.

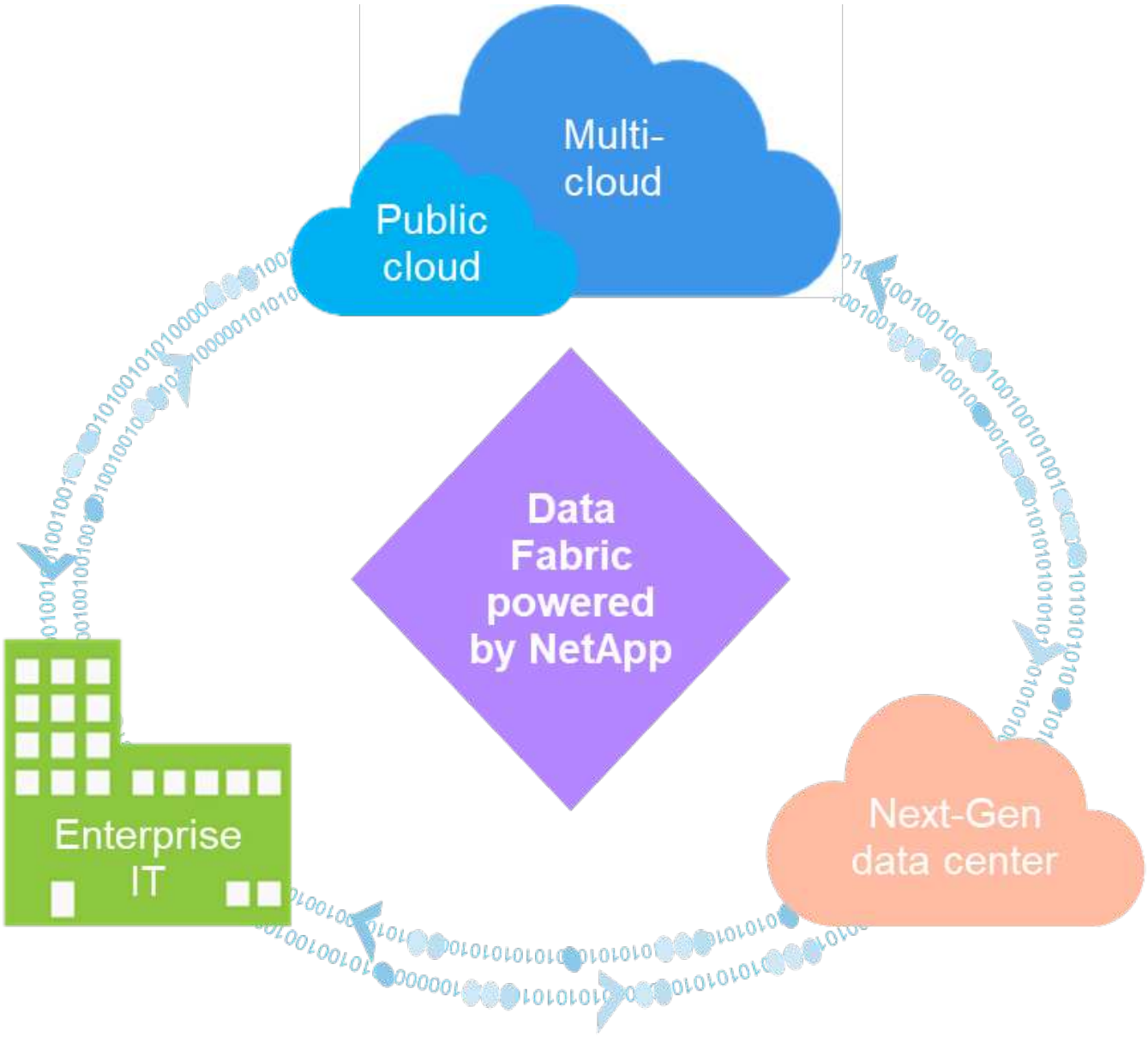
Kunden müssen außerdem Ausfallzeiten bei Recovery Point Objective (RPO) und Recovery Time Objective (RTO) minimieren und ihre Disaster-Recovery-Standorte vor Ort und in der Cloud kontrollieren, um optimale Business Continuity zu erzielen. Diese Kontrollfunktion ergibt sich in der Regel aus Management Tools der Enterprise-Klasse.

Die Umgebungen mit Hadoop und Spark sind kompliziert, da nicht nur das Datenvolumen sehr groß ist und immer größer wird, sondern auch die Geschwindigkeit, mit der die Daten ankommen. Bei diesem Szenario ist es schwierig, schnell effiziente, aktuelle DevTest- und QA-Umgebungen aus den Quelldaten zu erstellen. NetApp kennt diese Herausforderungen und bietet die in diesem Whitepaper vorgestellten Lösungen.

Data Fabric von NetApp für Big Data-Architekturen

Die Data Fabric von NetApp vereinfacht und integriert Datenmanagement für Cloud- und On-Premises-Umgebungen, um die digitale Transformation zu beschleunigen.

Die Data Fabric von NetApp bietet konsistente und integrierte Datenmanagement-Services und -Applikationen (Bausteine) für Datentransparenz und -Einblicke, Datenzugriff und -Kontrolle sowie Datensicherung und -Sicherheit, wie in der folgenden Abbildung dargestellt.



Bewährte Data-Fabric-Anwendungsfälle von Kunden

Die Data Fabric von NetApp bietet Kunden die folgenden neun bewährten Anwendungsfälle:

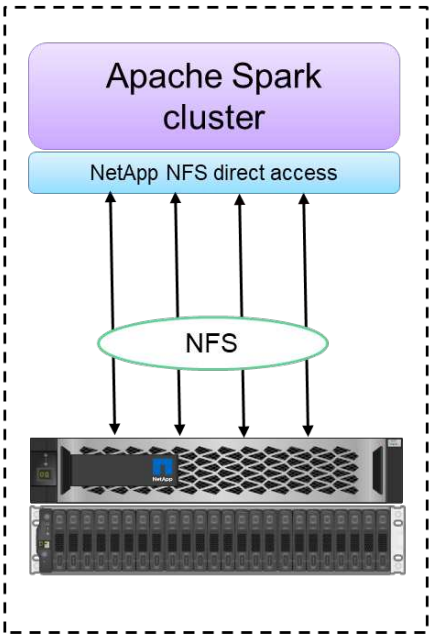
- Schnellere Analyse-Workloads
- Beschleunigung der DevOps-Transformation
- Aufbau von Cloud-Hosting-Infrastrukturen
- Integration von Cloud-Datenservices
- Schutz und Sicherheit für Daten
- Optimierung unstrukturierter Daten
- Datacenter-Effizienzgewinne
- Einblicke aus Daten und Datenkontrolle
- Vereinfachung und Automatisierung

Dieses Dokument umfasst zwei der neun Anwendungsfälle (zusammen mit den zugehörigen Lösungen):

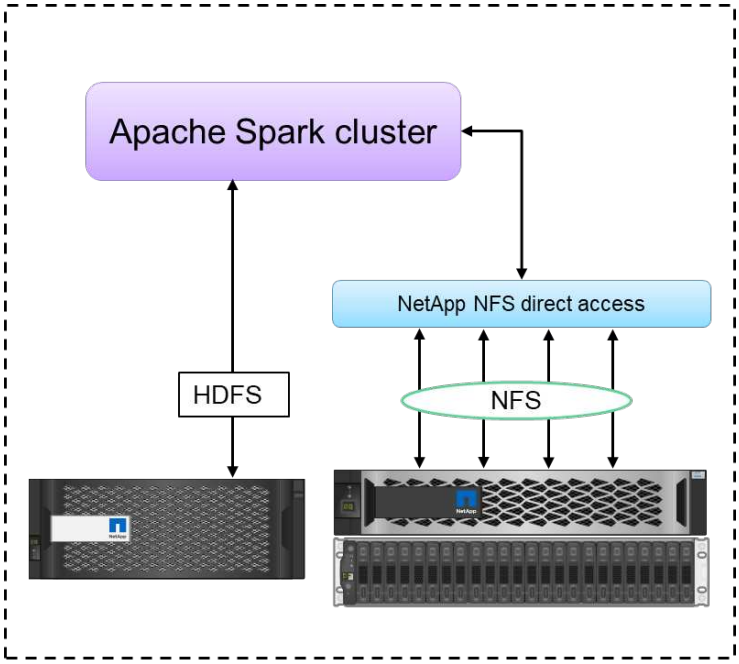
- Schnellere Analyse-Workloads
- Schutz und Sicherheit für Daten

Direkter Zugriff über NetApp NFS

Mit NetApp NFS können Kunden Big-Data-Analysejobs auf vorhandenen oder neuen NFSv3- oder NFSv4-Daten ausführen, ohne die Daten zu verschieben oder zu kopieren. Es verhindert mehrere Datenkopien und macht die Synchronisierung der Daten mit einer Quelle überflüssig. Im Finanzsektor beispielsweise muss der Wechsel von Daten von einem Ort zu einem anderen Ort gesetzlichen Verpflichtungen nachkommen, was keine leichte Aufgabe ist. In diesem Szenario analysiert der direkte NetApp NFS-Zugriff die Finanzdaten vom ursprünglichen Standort. Ein weiterer entscheidender Vorteil besteht darin, dass der NetApp NFS Direct Access die Sicherung von Hadoop Daten mithilfe nativer Hadoop Befehle vereinfacht und Workflows zur Datensicherung mithilfe des umfassenden Datenmanagement-Portfolios von NetApp ermöglicht.



Configuration 1: NFS as primary storage



Configuration 2: HDFS and NFS in single Spark cluster

Der direkte NetApp NFS-Zugriff bietet zwei Arten von Implementierungsoptionen für Hadoop/Spark Cluster:

- Die Hadoop/Spark-Cluster verwenden standardmäßig Hadoop Distributed File System (HDFS) für den Datenspeicher sowie das Standard-Filesystem. Der direkte NetApp NFS-Zugriff ersetzt das Standard-HDFS durch NFS-Storage als Standard-Filesystem, sodass direkte Analysevorgänge bei NFS-Daten möglich sind.
- Bei einer weiteren Implementierungsoption unterstützt der direkte NetApp NFS-Zugriff die Konfiguration von NFS als zusätzlichen Storage zusammen mit HDFS in einem einzelnen Hadoop/Spark-Cluster. In diesem Fall kann der Kunde Daten über NFS-Exporte teilen und gemeinsam mit HDFS-Daten vom selben Cluster aus darauf zugreifen.

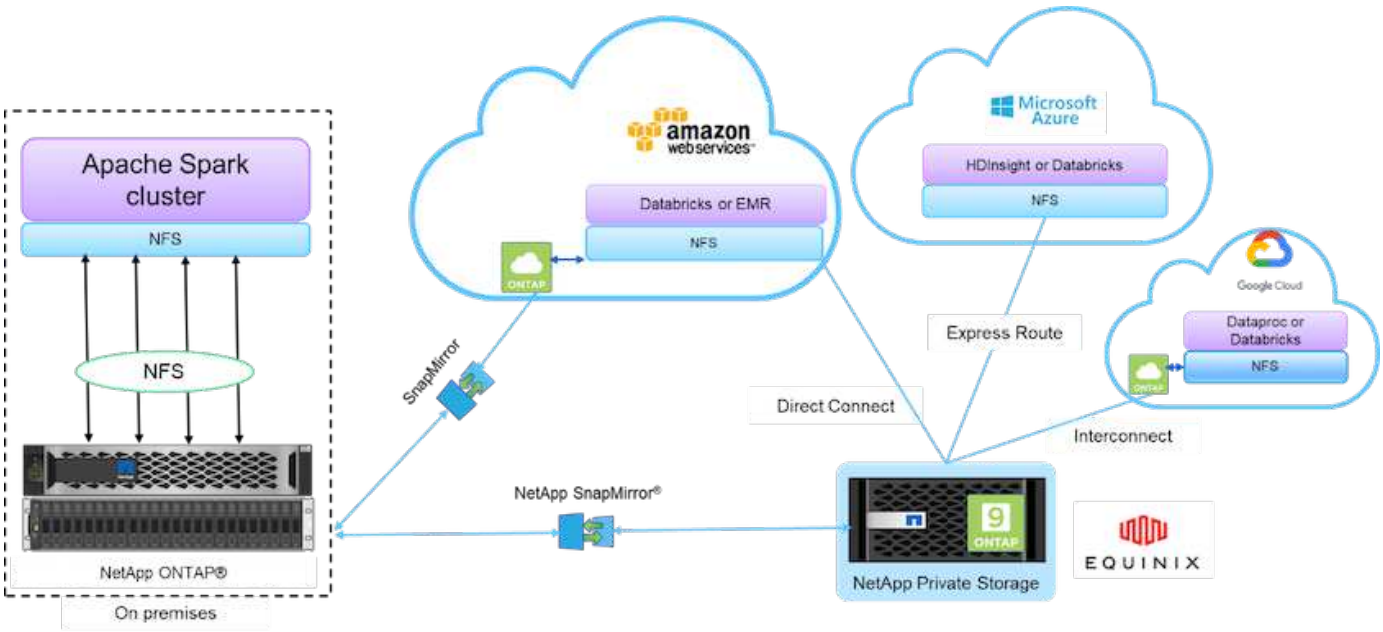
Zu den wichtigsten Vorteilen des direkten NetApp NFS-Zugriffs gehören:

- Analysiert die Daten vom aktuellen Speicherort, um das Zeit- und Performance-zeitaufwendige Verschieben von Analysedaten in eine Hadoop Infrastruktur wie HDFS zu vermeiden.

- Reduziert die Anzahl der Replikate von drei auf eins.
- Ermöglicht Benutzern die Trennung von Computing und Storage, um sie unabhängig voneinander zu skalieren.
- Datensicherung der Enterprise-Klasse durch Nutzung der umfassenden Datenmanagementfunktionen von ONTAP
- Ist für die Hortonworks Datenplattform zertifiziert.
- Ermöglicht Implementierung von hybriden Datenanalysen
- Verringerung der Backup-Zeit durch Nutzung der dynamischen Multithread-Funktion

Bausteine für Big Data

Die Data Fabric von NetApp integriert Datenmanagement-Services und -Applikationen (Bausteine) für Datenzugriff, Kontrolle, Sicherung und Sicherheit, wie in der Abbildung unten dargestellt.



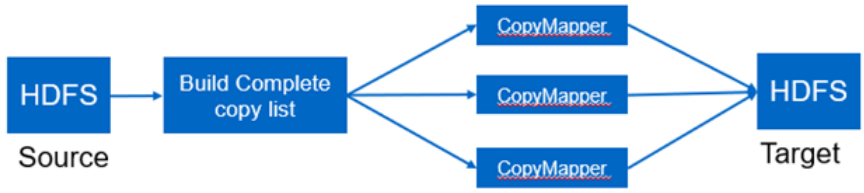
Die Abbildung oben beinhaltet folgende Bausteine:

- **NetApp NFS Direct Access.** bietet die neuesten Hadoop und Spark Cluster mit direktem Zugriff auf NetApp NFS Volumes ohne zusätzliche Software- oder Treiberanforderungen.
- **NetApp Cloud Volumes ONTAP und Cloud-Volume-Services.** softwaredefinierter vernetzter Storage auf Basis von ONTAP, der in Amazon Web Services (AWS) oder Azure NetApp Files (ANF) in Microsoft Azure Cloud-Services ausgeführt wird.
- **NetApp SnapMirror Technologie**. Bietet Datensicherungsfunktionen zwischen lokalen und ONTAP Cloud oder NPS Instanzen.
- **Cloud-Service-Provider.** zu diesen Anbietern gehören AWS, Microsoft Azure, Google Cloud und IBM Cloud.
- **PaaS.** Cloud-basierte Analyseservices wie Amazon Elastic MapReduce (EMR) und Databricks in AWS sowie Microsoft Azure HDInsight und Azure Databricks.

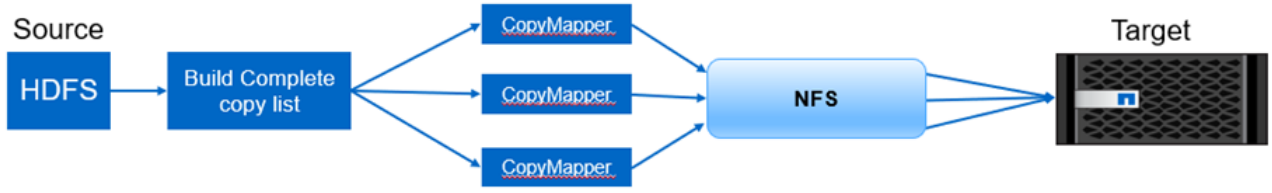
Hadoop Datensicherung und NetApp

Hadoop DiCp ist ein natives Tool für große Intercluster- und Intracluster-Kopien. Der in der Abbildung unten dargestellte grundlegende Hadoop DistCp-Prozess ist ein typischer Backup-Workflow, in dem native Hadoop Tools wie MapReduce zum Kopieren von Hadoop Daten von einer HDFS-Quelle auf ein entsprechendes Ziel verwendet werden.

Durch den direkten Zugang zu NetApp NFS können Kunden NFS als Zieladresse für das Hadoop DistCp Tool einstellen, um die Daten von HDFS Quelle in eine NFS-Freigabe über MapReduce zu kopieren. Der direkte NetApp NFS-Zugriff fungiert als NFS-Treiber für das DistCp Tool.



Hadoop Distcp Basic Process



Hadoop Distcp and NetApp

Übersicht über Anwendungsfälle für die Datensicherung in Hadoop

Dieser Abschnitt enthält eine ausführliche Beschreibung der Anwendungsfälle für die Datensicherung, die den Schwerpunkt dieses Dokuments bilden. Die verbleibenden Abschnitte enthalten ausführlichere Angaben zu den einzelnen Anwendungsfällen, z. B. zum Kundenproblem (Szenario), den Anforderungen und Herausforderungen sowie zu Lösungen.

Anwendungsfall 1: Sichern von Hadoop Daten

Für diesen Anwendungsfall half das NetApp NFS Volume einem großen Finanzinstitut, die lange Backup-Zeit von über 24 Stunden auf knapp unter ein paar Stunden zu reduzieren.

Anwendungsfall 2: Backup und Disaster Recovery aus der Cloud in On-Premises-Systeme

Durch den Einsatz der Data Fabric von NetApp als Bausteine konnte ein großes Rundfunkunternehmen seine Anforderung erfüllen, Backups von Cloud-Daten auf seinem lokalen Datacenter zu erstellen, je nach unterschiedlichen Übertragungsmodi wie On-Demand, unmittelbar Oder auf Basis der Hadoop/Spark-Cluster-Last

Anwendungsfall 3: Enabling DevTest on Existing Hadoop Data

NetApp Lösungen halfen einem Online-Musikdistributor, schnell mehrere platzsparende Hadoop Cluster in verschiedenen Niederlassungen aufzubauen, um mithilfe von geplanten Richtlinien Berichte zu erstellen und tägliche DevTest-Aufgaben auszuführen.

Anwendungsfall 4: Datensicherung und Multi-Cloud-Konnektivität

Ein großer Service-Provider nutzte die Data Fabric von NetApp, um seinen Kunden aus verschiedenen Cloud-Instanzen Multi-Cloud-Analysen zu liefern.

Anwendungsfall 5: Schnellere Analyse-Workloads

Eine der größten Finanzdienstleistungsbanken und Investmentbanken setzte die Network-Attached Storage-Lösung von NetApp ein, um die I/O-Wartezeit zu verringern und die Plattform für quantitative Finanzanalysen zu beschleunigen.

Anwendungsfall 1: Sichern von Hadoop Daten

Szenario

In diesem Szenario verfügt der Kunde über ein großes Hadoop Repository vor Ort, für Disaster-Recovery-Zwecke sollten die Daten gesichert werden. Die aktuelle Backup-Lösung des Kunden ist jedoch kostspielig und mit einem langen Backup-Fenster von mehr als 24 Stunden verbunden.

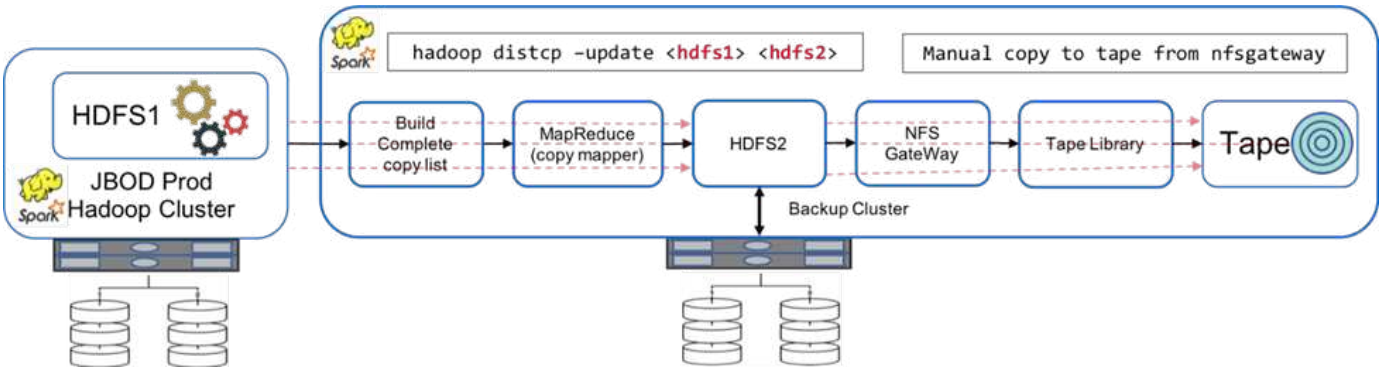
Anforderungen und Herausforderungen zu bewältigen

Zu den wesentlichen Anforderungen und Herausforderungen dieses Anwendungsfalls gehören:

- Abwärtskompatibilität der Software:
  - Die vorgeschlagene alternative Backup-Lösung sollte mit den derzeit verwendeten Softwareversionen im Hadoop-produktiven Cluster kompatibel sein.
- Zur Einhaltung der verpflichteten SLAs sollte die vorgeschlagene alternative Lösung sehr niedrige RPOs und RTOs erzielen.
- Das mit der NetApp Backup-Lösung erstellte Backup kann sowohl in dem lokalen Hadoop Cluster im Datacenter als auch in dem Hadoop Cluster verwendet werden, der am Disaster-Recovery-Standort am Remote-Standort ausgeführt wird.
- Die vorgeschlagene Lösung muss kostengünstig sein.
- Die vorgeschlagene Lösung muss die Performance-Auswirkungen auf die derzeit laufenden, in der Produktion laufenden Analysen während der Backup-Zeiten verringern.

Vorhandene Backup-Lösung des Kundensx

Die Abbildung unten zeigt die ursprüngliche native Hadoop Backup-Lösung.



Die Produktionsdaten sind über das Zwischenbackup Cluster auf Band geschützt:

- HDFS1-Daten werden durch Ausführen des auf HDFS2 kopiert `hadoop distcp -update <hdfs1> <hdfs2>` Befehl.
- Der Backup-Cluster fungiert als NFS-Gateway und die Daten werden manuell über Linux auf Tapes kopiert `cp` Befehl über die Tape Library.

Zu den Vorteilen der ursprünglichen nativen Hadoop Backup-Lösung gehören:

- Die Lösung basiert auf nativen Hadoop-Befehlen, sodass der Anwender sich nicht mit neuen Verfahren vertraut machen muss.
- Die Lösung nutzt eine branchenübliche Architektur und Hardware.

Zu den Nachteilen der ursprünglichen nativen Hadoop-Backup-Lösung zählen:

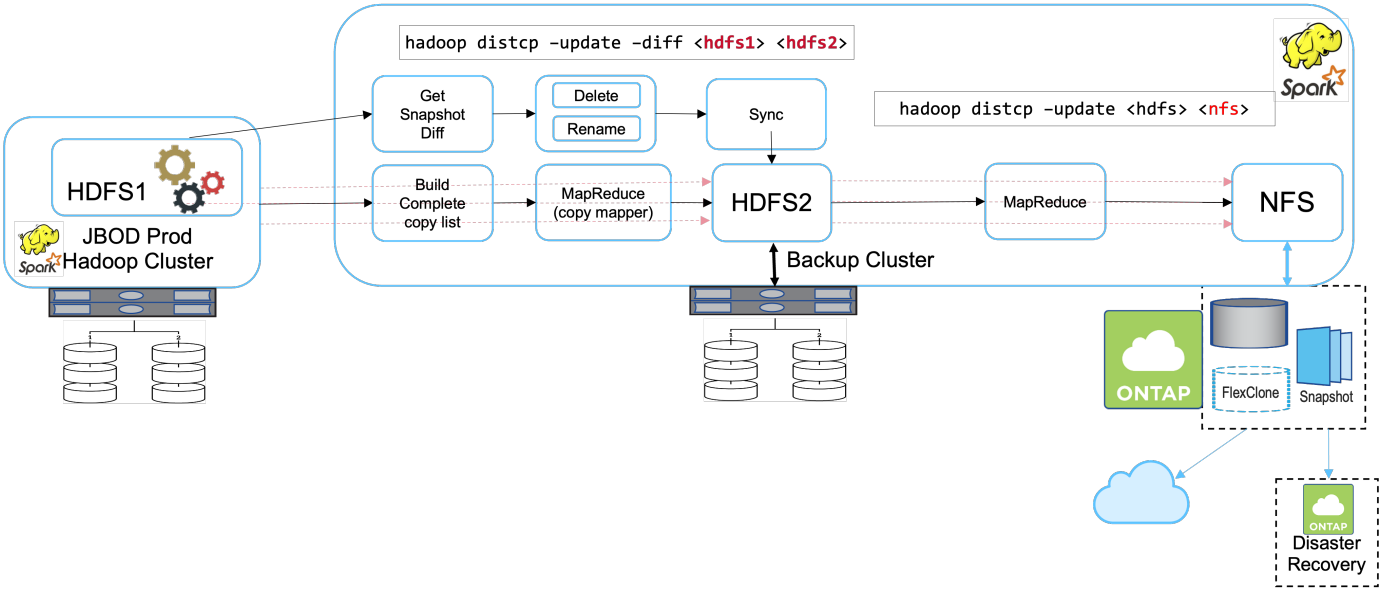
- Die lange Backup-Zeitdauer beträgt mehr als 24 Stunden, wodurch die Produktionsdaten gefährdet werden.
- Erhebliche Performance-Verschlechterung des Clusters während der Backup-Zeiten
- Das Kopieren auf Tape ist ein manueller Vorgang.
- Die Backup-Lösung ist in Bezug auf die erforderliche Hardware und die Arbeitsstunden, die für manuelle Prozesse erforderlich sind, teuer.

Backup-Lösungen

Auf der Grundlage dieser Herausforderungen und Anforderungen und unter Berücksichtigung des vorhandenen Backup-Systems wurden drei mögliche Backup-Lösungen vorgeschlagen. In den folgenden Abschnitten werden die drei verschiedenen Backup-Lösungen beschrieben, die als Lösung A bis Lösung C. bezeichnet wurden

Lösung A

In Lösung A sendet der Hadoop Backup-Cluster die sekundären Backups an NetApp NFS-Storage-Systeme, wodurch die Tape-Anforderungen, wie in der Abbildung unten dargestellt, eliminiert werden.



Zu den detaillierten Aufgaben für Solution A gehören:

- Das Hadoop Produktions-Cluster verfügt über die zu schützenden Analysedaten des Kunden in HDFS.
- Der Hadoop Backup Cluster mit HDFS fungiert als Zwischenstandort der Daten. Nur ein paar Festplatten (JBOD) bieten den Storage für HDFS sowohl bei Produktions- als auch bei Backup-Hadoop-Clustern.
- Der Schutz der Hadoop Produktionsdaten ist durch Ausführen des vor dem Produktions-Cluster HDFS und dem Backup-Cluster HDFS geschützt `Hadoop distcp -update -diff <hdfs1> <hdfs2>` Befehl.

Der Hadoop Snapshot schützt die Daten vor der Produktion auf das Hadoop Backup-Cluster.

- Der NetApp ONTAP Storage Controller liefert ein NFS-exportiertes Volume, das dem Hadoop Backup-Cluster bereitgestellt wird.
- Durch Ausführen des `Hadoop distcp` Mit MapReduce und mehreren Mappern sind die Analysedaten vom Backup-Hadoop-Cluster zu NFS gesichert.

  Nachdem die Daten im NFS auf dem NetApp Storage-System gespeichert wurden, werden mithilfe von NetApp Snapshot, SnapRestore und FlexClone Technologien die Hadoop Daten nach Bedarf gesichert, wiederhergestellt und dupliziert.

Hadoop Daten lassen sich mit SnapMirror Technologie in der Cloud sowie an Disaster Recovery-Standorten sichern.

Die Vorteile der Lösung A umfassen:

- Die Produktionsdaten von Hadoop sind vor dem Backup-Cluster geschützt.
- Die HDFS-Daten werden durch NFS geschützt, sodass sie in der Cloud und bei der Disaster-Recovery-Wiederherstellung gesichert sind.
- Verbessert die Performance durch Auslagerung von Backup-Vorgängen auf das Backup-Cluster.
- Keine manuellen Tape-Vorgänge mehr nötig
- Ermöglicht Enterprise Management-Funktionen über NetApp Tools.
- Erfordert minimale Änderungen an der vorhandenen Umgebung.

- Ist eine kostengünstige Lösung.

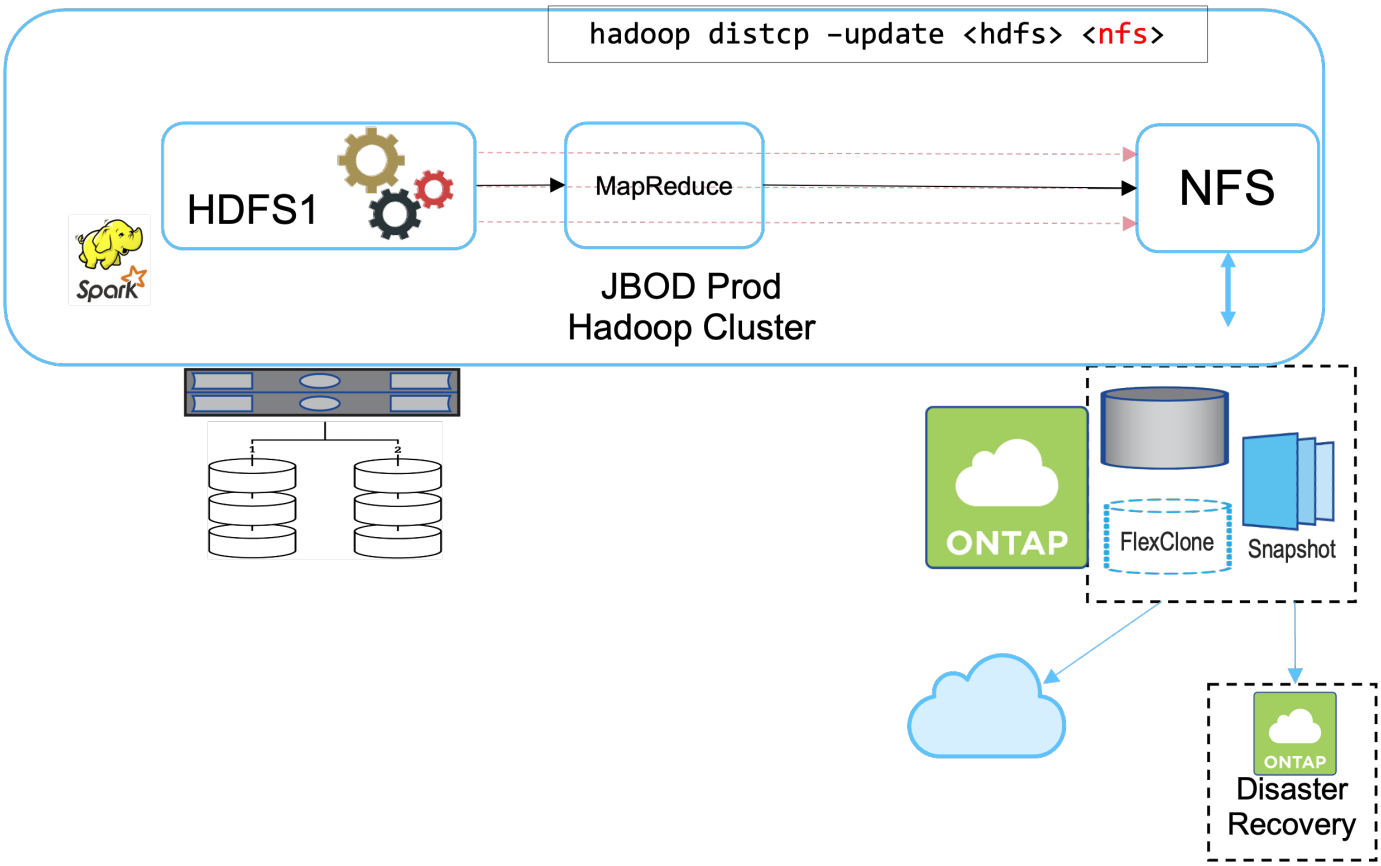
Der Nachteil dieser Lösung ist, dass ein Backup-Cluster und zusätzliche Mappern benötigt werden, um die Performance zu verbessern.

Der Kunde hat vor Kurzem eine Lösung A implementiert, weil es einfach, kostengünstig und insgesamt Performance bietet.

In dieser Lösung können SAN-Festplatten von ONTAP anstelle von JBOD verwendet werden. Diese Option entlastet den Storage für das Backup-Cluster an ONTAP, allerdings sind die SAN-Fabric-Switches erforderlich.

Lösung B

Lösung B fügt dem Hadoop Cluster in der Produktionsumgebung NFS Volume hinzu, wodurch der Hadoop Cluster für Backups nicht mehr benötigt wird, wie in der Abbildung unten dargestellt.



Zu den detaillierten Aufgaben für Lösung B gehören:

- Der NetApp ONTAP Storage Controller stellt den NFS-Export in das produktive Hadoop Cluster bereit.

  Hadoop als native Cloud `hadoop distcp` Mit dem Befehl werden Hadoop Daten aus dem Produktionscluster HDFS zu NFS gesichert.

- Nachdem die Daten im NFS auf dem NetApp Storage-System gespeichert wurden, werden die Snapshot, SnapRestore und FlexClone Technologien verwendet, um die Hadoop Daten nach Bedarf zu sichern, wiederherzustellen und zu duplizieren.

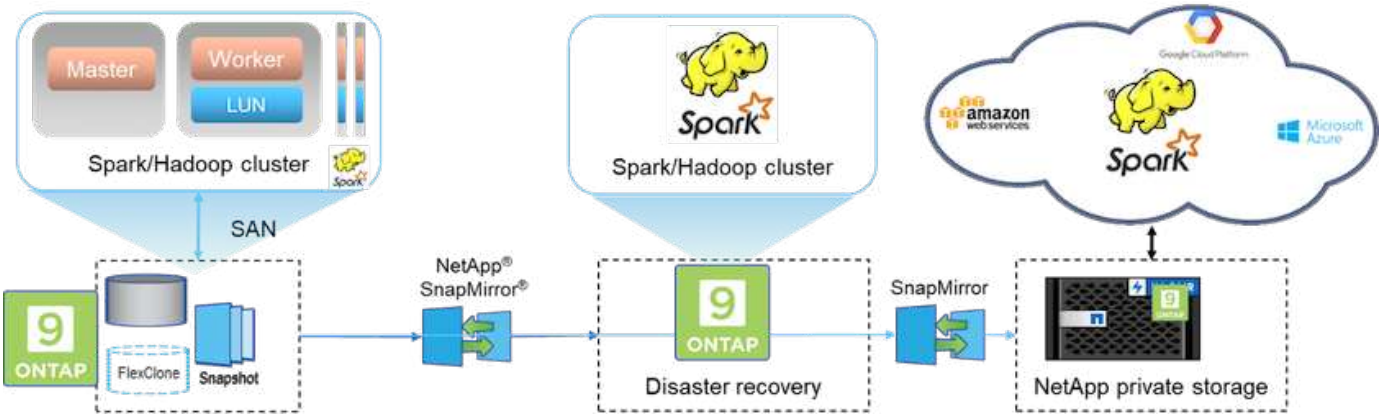
Zu den Vorteilen von Lösung B gehören:

- Das produktive Cluster wird für die Backup-Lösung leicht modifiziert, wodurch die Implementierung vereinfacht und die zusätzlichen Infrastrukturkosten gesenkt werden.
- Ein Backup-Cluster für den Backup-Vorgang ist nicht erforderlich.
- HDFS-Produktionsdaten werden bei der Umwandlung in NFS-Daten geschützt.
- Die Lösung ermöglicht Enterprise Management-Funktionen über NetApp Tools.

Der Nachteil dieser Lösung ist, dass sie im Produktionscluster implementiert wird, was zusätzliche Administratoraufgaben im Produktionscluster hinzufügen kann.

Lösung C

In Lösung C werden die NetApp SAN-Volumes für HDFS-Storage direkt in dem Hadoop Produktions-Cluster bereitgestellt, wie in der Abbildung unten gezeigt.



Zu den detaillierten Schritten für Lösung C gehören:

- Der NetApp ONTAP SAN-Storage wird im Hadoop Cluster in der Produktionsumgebung für HDFS-Storage bereitgestellt.
- Mit NetApp Snapshot und SnapMirror Technologien werden die HDFS-Daten aus dem Hadoop Cluster in der Produktionsumgebung gesichert.
- Für den Hadoop/Spark-Cluster während des Backup-Prozesses mit Snapshot-Kopien werden keine Performance-Auswirkungen auf die Produktion erzielt, da sich das Backup auf Storage-Ebene befindet.

Die Snapshot Technologie ermöglicht Backups, die innerhalb von Sekunden abgeschlossen werden, unabhängig von der Größe der Daten.

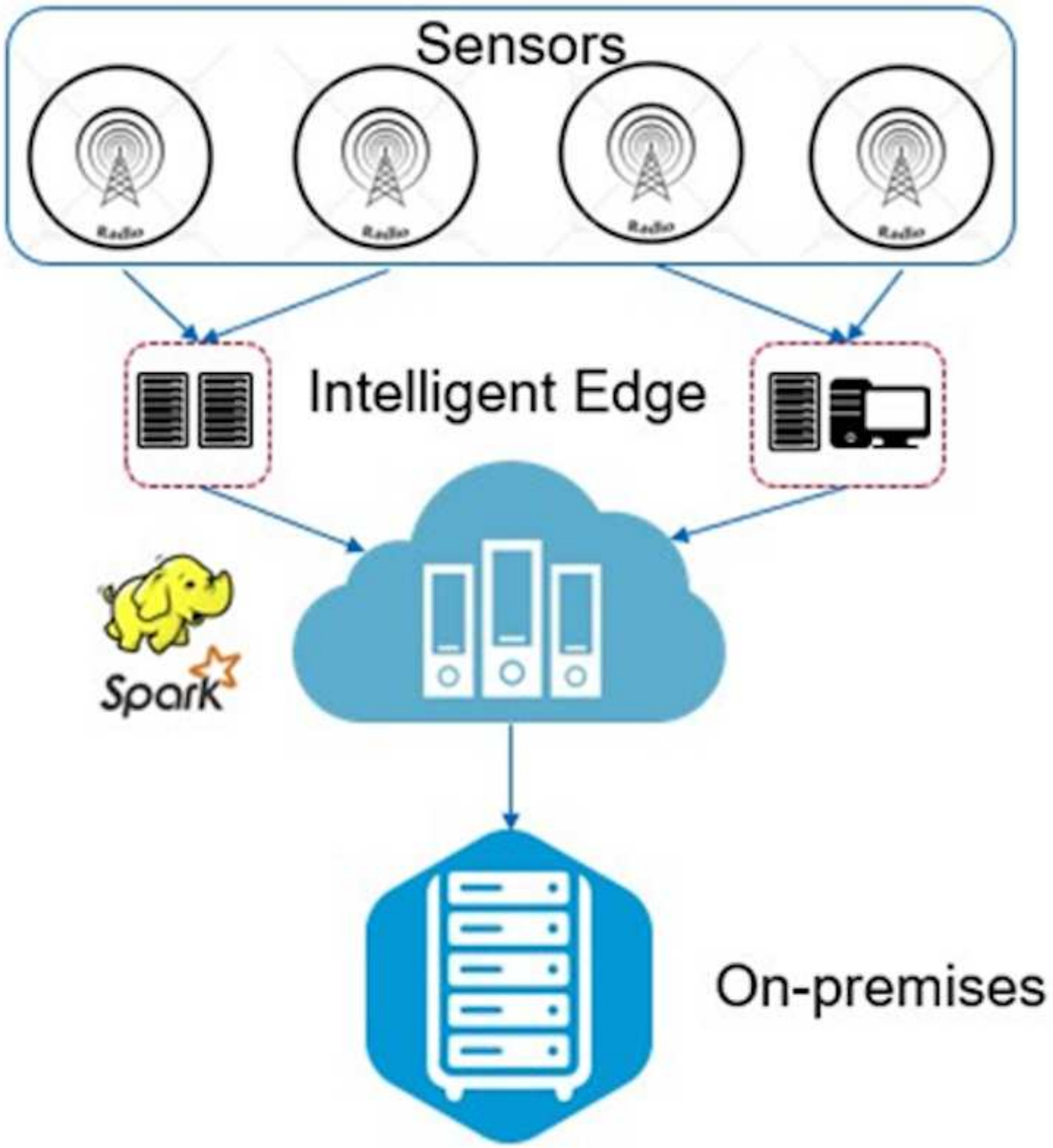
Lösung C bietet u. a. folgende Vorteile:

- Platzsparende Backups können mithilfe der Snapshot Technologie erstellt werden.
- Ermöglicht Enterprise Management-Funktionen über NetApp Tools.

Anwendungsfall 2: Backup und Disaster Recovery aus der Cloud in On-Premises-Systeme

Dieser Nutzungsfall basiert auf einem Rundfunkkunden, der Cloud-basierte Analysedaten in seinem On-Premises-Datacenter sichern muss, wie in der folgenden Abbildung

dargestellt.



Szenario

In diesem Szenario werden die IoT-Sensordaten in die Cloud aufgenommen und mithilfe eines Open-Source-Apache Spark-Clusters in AWS analysiert. Die Anforderung besteht darin, die verarbeiteten Daten aus der Cloud in On-Premises-Systeme zu sichern.

Anforderungen und Herausforderungen zu bewältigen

Zu den wesentlichen Anforderungen und Herausforderungen dieses Anwendungsfalls gehören:

- Das Aktivieren der Datensicherung sollte keine Performance-Auswirkungen auf das Produktions-Spark/Hadoop-Cluster in der Cloud haben.

- Cloud-Sensordaten müssen effizient und sicher in On-Premises-Systeme verschoben und geschützt werden.
- Flexibilität bei der Übertragung von Daten aus der Cloud unter unterschiedlichen Bedingungen in On-Demand-Umgebungen, sofort und während Low-Cluster-Ladezeiten

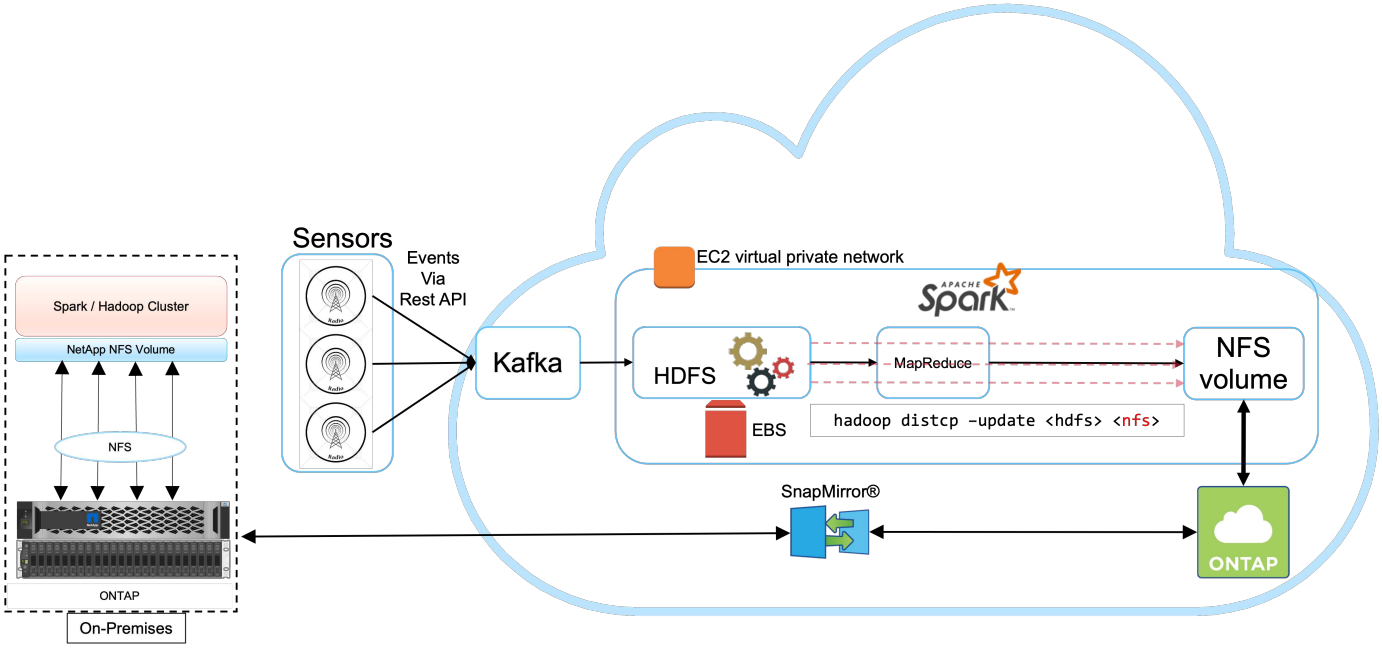
Nutzen

Kunden nutzen den AWS Elastic Block Store (EBS) für ihren Spark-Cluster-HDFS-Storage, um Daten von Remote-Sensoren über Kafka zu empfangen und einzuspielen. Dementsprechend fungiert HDFS-Storage als Quelle für die Backup-Daten.

Um diese Anforderungen zu erfüllen, wird NetApp ONTAP Cloud in AWS implementiert und eine NFS-Share wird als Backup-Ziel für den Spark/Hadoop-Cluster erstellt.

Nach der Erstellung der NFS-Freigabe kopieren Sie die Daten aus dem HDFS EBS Storage in die ONTAP-NFS-Freigabe. Nachdem sich die Daten in ONTAP Cloud im NFS befinden, können Sie mit SnapMirror Technologie die Daten sicher und effizient aus der Cloud in den On-Premises-Storage spiegeln.

Dieses Bild zeigt Backup und Disaster Recovery von der Cloud zu einer On-Premises-Lösung.



Anwendungsfall 3: Enabling DevTest on Existing Hadoop Data

In diesem Anwendungsfall muss der Kunde schnell und effizient neue Hadoop/Spark-Cluster erstellen, die auf einem vorhandenen Hadoop-Cluster basieren und große Mengen an Analysedaten für DevTest und Reporting-Zwecke im selben Datacenter und an Remote-Standorten enthalten sind.

Szenario

Bei diesem Szenario wurden mehrere Spark/Hadoop-Cluster aus einer umfangreichen Hadoop-Data-Lake-Implementierung vor Ort und an Disaster-Recovery-Standorten erstellt.

Anforderungen und Herausforderungen zu bewältigen

Zu den wesentlichen Anforderungen und Herausforderungen dieses Anwendungsfalls gehören:

- Erstellen Sie mehrere Hadoop Cluster für DevTest, QA oder jeden anderen Zweck, für den Zugriff auf dieselben Produktionsdaten erforderlich ist. Die Herausforderung besteht hier darin, einen sehr großen Hadoop Cluster mehrmals unmittelbar und äußerst platzsparend zu klonen.
- Synchronisieren Sie Hadoop Daten mit DevTest- und Reporting-Teams, um betriebliche Effizienz zu erzielen.
- Verteilen Sie Hadoop Daten mit denselben Anmeldedaten über Produktions- und neue Cluster.
- Verwenden Sie geplante Richtlinien, um effizient QA-Cluster zu erstellen, ohne das Produktions-Cluster zu beeinträchtigen.

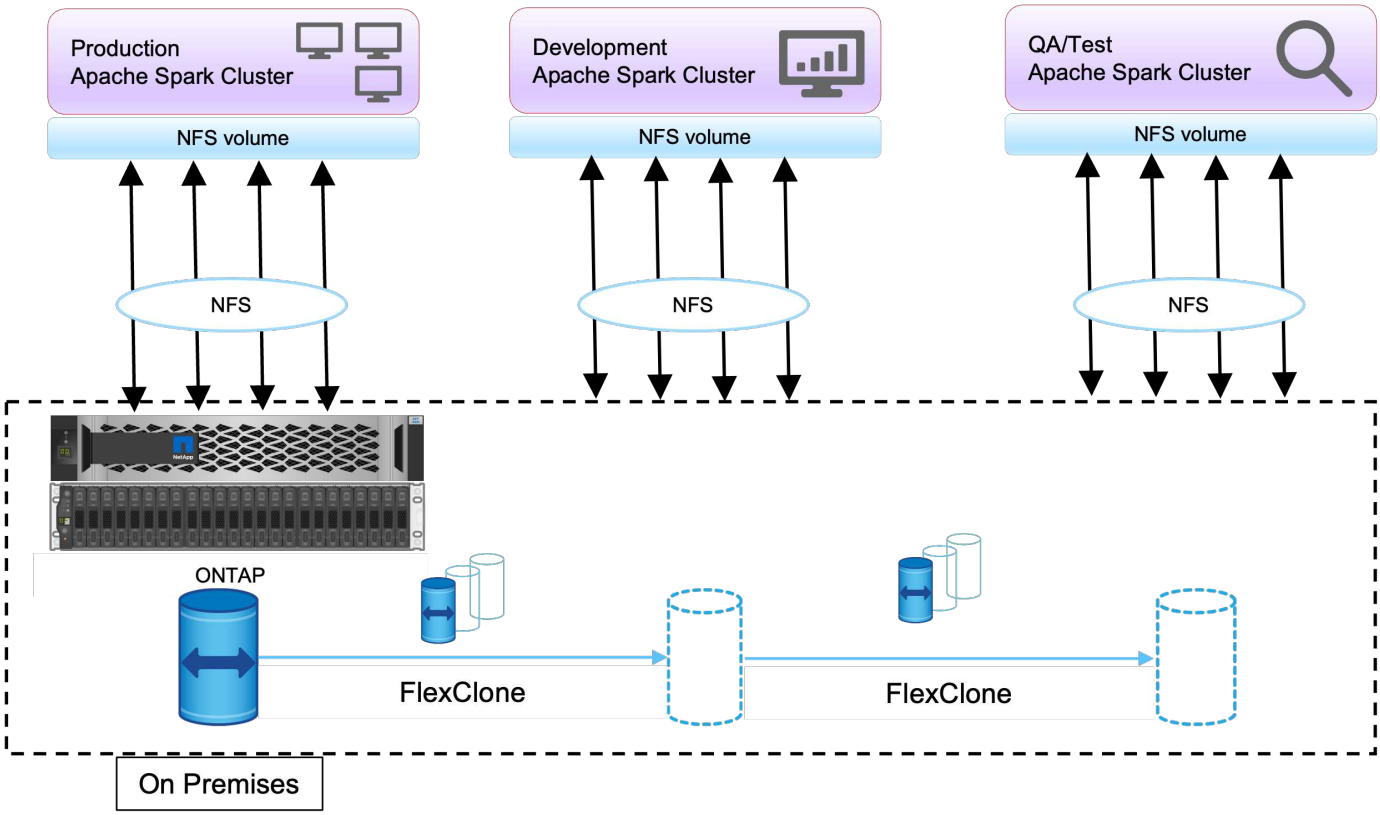
Nutzen

FlexClone Technologie wird zur Erfüllung der soeben beschriebenen Anforderungen verwendet. Die FlexClone Technologie ist die Kopie einer Snapshot Kopie für Lese-/Schreibzugriffe. Die Daten werden aus den Daten der übergeordneten Snapshot Kopie gelesen, und verbraucht nur zusätzlichen Speicherplatz für neue/geänderte Blöcke. Sie ist schnell und platzsparend.

Zunächst wurde eine Snapshot Kopie des vorhandenen Clusters mithilfe einer NetApp Konsistenzgruppe erstellt.

Snapshot Kopien innerhalb von NetApp System Manager oder der Eingabeaufforderung für den Storage-Administrator Die Snapshot Kopien der Konsistenzgruppen sind applikationskonsistente Snapshot Kopien von Gruppen, und das FlexClone Volume wird basierend auf Snapshot Kopien der Konsistenzgruppen erstellt. Es ist erwähnenswert, dass ein FlexClone Volume die NFS-Exportrichtlinie des übergeordneten Volume übernimmt. Nach der Erstellung der Snapshot Kopie muss für DevTest und Reporting ein neuer Hadoop Cluster installiert werden, wie in der Abbildung unten dargestellt. Das geklonte NFS Volume aus dem neuen Hadoop Cluster greift auf die NFS-Daten zu.

Dieses Bild zeigt den Hadoop Cluster für DevTest.



Anwendungsfall 4: Datensicherung und Multi-Cloud-Konnektivität

Dieser Anwendungsfall ist relevant für einen Cloud-Service-Partner, der für Multi-Cloud-Konnektivität für Big-Data-Analysedaten von Kunden steht.

Szenario

In diesem Szenario werden die in AWS empfangenen IoT-Daten aus verschiedenen Quellen an einem zentralen Standort in NPS gespeichert. Der NPS Storage ist mit Spark/Hadoop Clustern in AWS und Azure verbunden, sodass Big-Data-Analyseapplikationen in diversen Clouds ausgeführt werden und auf dieselben Daten zugreifen können.

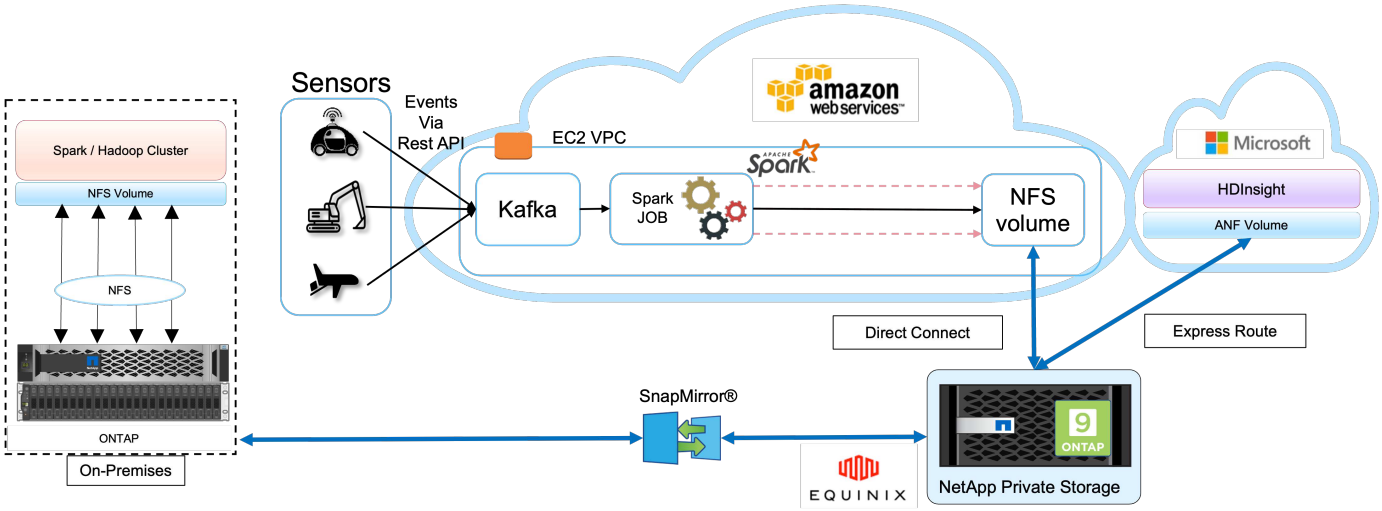
Anforderungen und Herausforderungen zu bewältigen

Zu den wesentlichen Anforderungen und Herausforderungen dieses Anwendungsfalls gehören:

- Kunden möchten Analysejobs in denselben Daten mithilfe mehrerer Clouds ausführen.
- Daten müssen von unterschiedlichen Quellen wie On-Premises und Cloud über verschiedene Sensoren und Hubs empfangen werden.
- Die Lösung muss effizient und kostengünstig sein.
- Die größte Herausforderung ist der Aufbau einer kostengünstigen und effizienten Lösung, die hybride Analyseservices zwischen On-Premises-Systemen und verschiedenen Clouds bereitstellt.

Nutzen

Dieses Bild zeigt die Datensicherungs- und Multi-Cloud-Konnektivitätslösung.



Wie in der Abbildung oben gezeigt, werden die Daten von Sensoren gestreamt und über Kafka in den AWS Spark-Cluster aufgenommen. Die Daten werden in einem NFS-Share in NPS gespeichert, der sich außerhalb des Cloud-Providers in einem Equinix Datacenter befindet. Da NetApp NPS über Direct Connect bzw. Express Route mit Amazon AWS und Microsoft Azure verbunden ist, können Kunden sowohl von Amazon als auch von AWS Analyse-Clustern auf die NFS-Daten zugreifen. Dieser Ansatz ist die Lösung dafür, dass Cloud-Analysen über mehrere Hyperscaler hinweg durchgeführt werden können.

Da sowohl On-Premises- als auch NPS-Storage auf ONTAP Software ausgeführt werden, kann SnapMirror die NPS Daten in das On-Premises-Cluster spiegeln. Dadurch profitieren Hybrid-Cloud-Analysen von On-Premises-Systemen und mehreren Clouds.

Um die beste Performance zu erzielen, empfiehlt NetApp normalerweise, mehrere Netzwerkschnittstellen und direkte Verbindungs-/Express-Routen zu verwenden, um auf die Daten von Cloud-Instanzen zuzugreifen.

Anwendungsfall 5: Schnellere Analyse-Workloads

In diesem Szenario wurde die Analyseplattform einer großen Finanzdienstleistungsbank und einer Investmentbank mit der NetApp NFS Storage-Lösung modernisiert, um erhebliche Verbesserungen bei der Analyse von Investitionsrisiken und Derivaten für ihr Asset Management und ihren quantitativen Geschäftsbereich zu erzielen.

Szenario

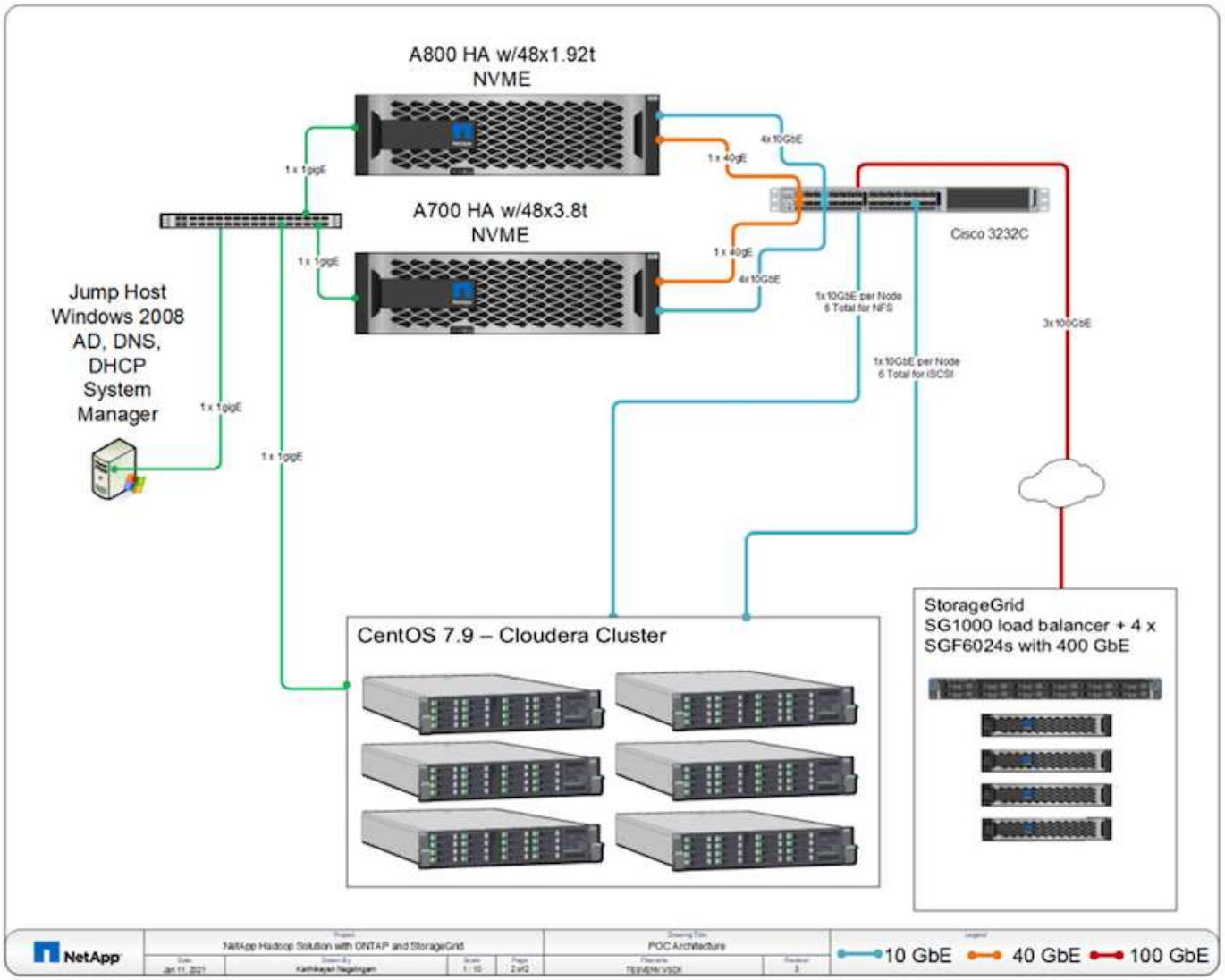
In der bestehenden Umgebung des Kunden nutzte die für die Analyseplattform verwendete Hadoop Infrastruktur den internen Storage der Hadoop Server. Aufgrund der proprietären JBOD-Umgebung konnten viele interne Kunden innerhalb des Unternehmens ihr quantitatives Modell von Monte Carlo nicht nutzen, eine Simulation, die auf den wiederkehrenden Proben von Echtzeitdaten beruht. Die suboptimale Fähigkeit, die Auswirkungen der Unsicherheit auf die Marktbewegungen zu verstehen, diente dem Geschäftsbereich quantitative Vermögensverwaltung als ungünstig.

Anforderungen und Herausforderungen zu bewältigen

Die quantitative Geschäftseinheit der Bank wollte eine effiziente Prognosemethode, um präzise und zeitnahe Vorhersagen zu treffen. Das Team erkannte daher, dass eine Modernisierung der Infrastruktur nötig ist, die bestehende I/O-Wartezeit reduziert und die Performance bei Analyseapplikationen wie Hadoop und Spark verbessert wird, um Investitionsmodelle effizient zu simulieren, potenzielle Gewinne zu messen und Risiken zu analysieren.

Nutzen

Der Kunde hatte JBOD die vorhandene Spark-Lösung. Anschließend wurden NetApp ONTAP, NetApp StorageGRID und Minio Gateway auf NFS genutzt, um die I/O-Wartezeit für die quantitative Finanzgruppe der Bank zu verringern, die Simulationen und Analysen zu Investitionsmodellen ausführt, die potenzielle Gewinne und Risiken bewerten. Zeigt die Spark-Lösung mit NetApp Storage.



Wie in der obigen Abbildung gezeigt, wurden AFF A800, A700 Systeme und StorageGRID implementiert, um über die NFS- und S3-Protokolle auf Parkett-Dateien in einem Hadoop-Cluster mit sechs Nodes mit Spark sowie YARN- und Hive-Metadaten-Services für den Datenanalyse-Betrieb zuzugreifen.

Eine Direct-Attached Storage-Lösung (das) in der alten Kundenumgebung hatte den Nachteil, dass Computing und Storage unabhängig voneinander skaliert werden mussten. Mit der NetApp ONTAP Solution für Spark

konnte die Geschäftseinheit der Bank für Finanzanalyse Storage von Computing-Ressourcen entkoppeln und Infrastrukturressourcen nach Bedarf nahtlos effizienter bereitstellen.

ONTAP mit NFS wurden die Computing-Server-CPUs nahezu vollständig für Spark SQL-Jobs genutzt, sodass die I/O-Wartezeit um fast 70 % reduziert wurde. Dadurch verbessert sich die Computing-Leistung und die Performance für Spark-Workloads. Mit der höheren CPU-Auslastung konnte der Kunde dann auch GPUs wie GPUDirect nutzen, um eine weitere Plattformmodernisierung zu ermöglichen. Darüber hinaus bietet StorageGRID eine kostengünstige Storage-Option für Spark-Workloads, und Minio Gateway bietet über das S3-Protokoll sicheren Zugriff auf NFS-Daten. Für Ihre Daten in der Cloud empfiehlt NetApp Cloud Volumes ONTAP, Azure NetApp Files und NetApp Cloud Volumes Service.

Schlussfolgerung

Dieser Abschnitt enthält eine Zusammenfassung der von NetApp zur Erfüllung verschiedener Hadoop Datensicherungsanforderungen bereitgestellten Anwendungsfälle und Lösungen. Mit der Data Fabric von NetApp profitieren Kunden von folgenden Vorteilen:

- Dank der umfassenden Datenmanagement-Funktionen und der Integration in native Hadoop Workflows können die passenden Datensicherungslösungen flexibel gewählt werden.
- Reduzieren Sie das Backup-Fenster für Hadoop Cluster um beinahe 70 %.
- Keine Performance-Auswirkungen durch Hadoop Cluster-Backups
- Datensicherung und Datenzugriff über mehrere Clouds gleichzeitig von verschiedenen Cloud-Providern für eine zentrale Quelle der Analysedaten
- Schnelle und platzsparende Hadoop Cluster-Kopien mit FlexClone Technologie

Wo Sie weitere Informationen finden

Weitere Informationen zu den in diesem Dokument beschriebenen Daten finden Sie in den folgenden Dokumenten bzw. auf den folgenden Websites:

- NetApp Big-Data-Analyselösungen

  "https://www.netapp.com/us/solutions/applications/big-data-analytics/index.aspx"

- Apache Spark Workload mit NetApp Storage

  https://www.netapp.com/pdf.html?item=/media/26877-nva-1157-deploy.pdf

- NetApp Storage-Lösungen für Apache Spark

  "https://www.netapp.com/media/16864-tr-4570.pdf"

- Apache Hadoop on Data Fabric basiert auf NetApp

  "https://www.netapp.com/media/16877-tr-4529.pdf"

Danksagungen

- Paul Burland, Vertriebsmitarbeiter, Vertrieb Australien/Victoria District, NetApp

- Hoseb Dermanilian, Business Development Manager, NetApp
- Lee Dorrier, Director MPSG, NetApp
- David Thiessen, Systems Engineer, ANZ Victoria District SE, NetApp

Versionsverlauf

Version	Datum	Versionsverlauf des Dokuments
Version 1.0	Januar 2018	Erste Version
Version 2.0	Oktober 2021	Aktualisiert mit Anwendungsfall #5: Analytics-Workloads beschleunigen
Version 3.0	November 2023	NIPAM-Details wurden entfernt

Copyright-Informationen

Copyright © 2024 NetApp. Alle Rechte vorbehalten. Gedruckt in den USA. Dieses urheberrechtlich geschützte Dokument darf ohne die vorherige schriftliche Genehmigung des Urheberrechtsinhabers in keiner Form und durch keine Mittel – weder grafische noch elektronische oder mechanische, einschließlich Fotokopieren, Aufnehmen oder Speichern in einem elektronischen Abrufsystem – auch nicht in Teilen, vervielfältigt werden.

Software, die von urheberrechtlich geschütztem NetApp Material abgeleitet wird, unterliegt der folgenden Lizenz und dem folgenden Haftungsausschluss:

DIE VORLIEGENDE SOFTWARE WIRD IN DER VORLIEGENDEN FORM VON NETAPP ZUR VERFÜGUNG GESTELLT, D. H. OHNE JEGLICHE EXPLIZITE ODER IMPLIZITE GEWÄHRLEISTUNG, EINSCHLIESSLICH, JEDOCH NICHT BESCHRÄNKT AUF DIE STILLSCHWEIGENDE GEWÄHRLEISTUNG DER MARKTGÄNGIGKEIT UND EIGNUNG FÜR EINEN BESTIMMTEN ZWECK, DIE HIERMIT AUSGESCHLOSSEN WERDEN. NETAPP ÜBERNIMMT KEINERLEI HAFTUNG FÜR DIREKTE, INDIREKTE, ZUFÄLLIGE, BESONDERE, BEISPIELHAFTE SCHÄDEN ODER FOLGESCHÄDEN (EINSCHLIESSLICH, JEDOCH NICHT BESCHRÄNKT AUF DIE BESCHAFFUNG VON ERSATZWAREN ODER -DIENSTLEISTUNGEN, NUTZUNGS-, DATEN- ODER GEWINNVERLUSTE ODER UNTERBRECHUNG DES GESCHÄFTSBETRIEBS), UNABHÄNGIG DAVON, WIE SIE VERURSACHT WURDEN UND AUF WELCHER HAFTUNGSTHEORIE SIE BERUHEN, OB AUS VERTRAGLICH FESTGELEGTER HAFTUNG, VERSCHULDENSUNABHÄNGIGER HAFTUNG ODER DELIKTSHAFTUNG (EINSCHLIESSLICH FAHRLÄSSIGKEIT ODER AUF ANDEREM WEGE), DIE IN IRGENDEINER WEISE AUS DER NUTZUNG DIESER SOFTWARE RESULTIEREN, SELBST WENN AUF DIE MÖGLICHKEIT DERARTIGER SCHÄDEN HINGEWIESEN WURDE.

NetApp behält sich das Recht vor, die hierin beschriebenen Produkte jederzeit und ohne Vorankündigung zu ändern. NetApp übernimmt keine Verantwortung oder Haftung, die sich aus der Verwendung der hier beschriebenen Produkte ergibt, es sei denn, NetApp hat dem ausdrücklich in schriftlicher Form zugestimmt. Die Verwendung oder der Erwerb dieses Produkts stellt keine Lizenzierung im Rahmen eines Patentrechts, Markenrechts oder eines anderen Rechts an geistigem Eigentum von NetApp dar.

Das in diesem Dokument beschriebene Produkt kann durch ein oder mehrere US-amerikanische Patente, ausländische Patente oder anhängige Patentanmeldungen geschützt sein.

ERLÄUTERUNG ZU „RESTRICTED RIGHTS“: Nutzung, Vervielfältigung oder Offenlegung durch die US-Regierung unterliegt den Einschränkungen gemäß Unterabschnitt (b)(3) der Klausel „Rights in Technical Data – Noncommercial Items“ in DFARS 252.227-7013 (Februar 2014) und FAR 52.227-19 (Dezember 2007).

Die hierin enthaltenen Daten beziehen sich auf ein kommerzielles Produkt und/oder einen kommerziellen Service (wie in FAR 2.101 definiert) und sind Eigentum von NetApp, Inc. Alle technischen Daten und die Computersoftware von NetApp, die unter diesem Vertrag bereitgestellt werden, sind gewerblicher Natur und wurden ausschließlich unter Verwendung privater Mittel entwickelt. Die US-Regierung besitzt eine nicht ausschließliche, nicht übertragbare, nicht unterlizenzierbare, weltweite, limitierte unwiderrufliche Lizenz zur Nutzung der Daten nur in Verbindung mit und zur Unterstützung des Vertrags der US-Regierung, unter dem die Daten bereitgestellt wurden. Sofern in den vorliegenden Bedingungen nicht anders angegeben, dürfen die Daten ohne vorherige schriftliche Genehmigung von NetApp, Inc. nicht verwendet, offengelegt, vervielfältigt, geändert, aufgeführt oder angezeigt werden. Die Lizenzrechte der US-Regierung für das US-Verteidigungsministerium sind auf die in DFARS-Klausel 252.227-7015(b) (Februar 2014) genannten Rechte beschränkt.

Markeninformationen

NETAPP, das NETAPP Logo und die unter http://www.netapp.com/TM aufgeführten Marken sind Marken von NetApp, Inc. Andere Firmen und Produktnamen können Marken der jeweiligen Eigentümer sein.