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

NetApp artificial intelligence solutions

NetApp
August 18, 2025

Inhalt

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

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

Karthikeyan Nagalingam und Sathish Thyagarajan, NetApp

Dieses Dokument beschreibt Hybrid-Cloud-Datenlösungen mit NetApp AFF und FAS Speichersystemen, NetApp Cloud Volumes ONTAP, NetApp Connected Storage und NetApp FlexClone -Technologie für Spark und Hadoop. Diese Lösungsarchitekturen ermöglichen es Kunden, eine geeignete Datenschutzlösung für ihre Umgebung auszuwählen. NetApp hat diese Lösungen auf der Grundlage der Interaktion mit Kunden und ihren geschäftlichen Anwendungsfällen entwickelt. Dieses Dokument enthält die folgenden detaillierten Informationen:

- Warum wir Datenschutz für Spark- und Hadoop-Umgebungen und Kundenherausforderungen benötigen.
- Das Data Fabric basiert auf NetApp Vision und seinen Bausteinen und Services.
- Wie diese Bausteine zum Entwerfen flexibler Datenschutz-Workflows verwendet werden können.
- Die Vor- und Nachteile mehrerer Architekturen basierend auf realen Anwendungsfällen von Kunden. Jeder Anwendungsfall stellt die folgenden Komponenten bereit:
  - Kundenszenarien
  - Anforderungen und Herausforderungen
  - Lösungen
  - Zusammenfassung der Lösungen

Warum Hadoop-Datenschutz?

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

- **Software- oder menschliches Versagen.** Menschliche Fehler bei Software-Updates während der Durchführung von Hadoop-Datenoperationen können zu fehlerhaftem Verhalten führen, das unerwartete Ergebnisse des Auftrags zur Folge haben kann. In einem solchen Fall müssen wir die Daten schützen, um Fehler oder unangemessene Ergebnisse zu vermeiden. Beispielsweise kann aufgrund eines schlecht ausgeführten Software-Updates einer Anwendung zur Verkehrssignalanalyse eine neue Funktion die im Klartext vorliegenden Verkehrssignaldaten nicht richtig analysieren. Die Software analysiert weiterhin JSON und andere Nicht-Text-Dateiformate, was dazu führt, dass das Echtzeit-Verkehrskontrollanalysesystem Vorhersageergebnisse erzeugt, bei denen Datenpunkte fehlen. Diese Situation kann zu fehlerhaften Ausgaben führen, die zu Unfällen an der Ampel führen können. Der Datenschutz kann dieses Problem lösen, indem er die Möglichkeit bietet, schnell zur vorherigen funktionierenden Anwendungsversion zurückzukehren.
- **Größe und Maßstab.** Aufgrund der ständig zunehmenden Anzahl von Datenquellen und Datenvolumina wächst der Umfang der Analysedaten täglich. Soziale Medien, mobile Apps, Datenanalysen und Cloud-Computing-Plattformen sind die wichtigsten Datenquellen im aktuellen Big-Data-Markt, der sehr schnell

wächst. Daher müssen die Daten geschützt werden, um einen genauen Datenbetrieb zu gewährleisten.

- **Hadoops nativer Datenschutz.** Hadoop verfügt über einen nativen Befehl zum Schutz der Daten, dieser Befehl gewährleistet jedoch keine Datenkonsistenz während der Sicherung. Es unterstützt nur die Sicherung auf Verzeichnisebene. Die von Hadoop erstellten Snapshots sind schreibgeschützt und können nicht zur direkten Wiederverwendung der Sicherungsdaten verwendet werden.

Datenschutzherausforderungen für Hadoop- und Spark-Kunden

Eine häufige Herausforderung für Hadoop- und Spark-Kunden besteht darin, die Sicherungszeit zu verkürzen und die Zuverlässigkeit der Sicherung zu erhöhen, ohne die Leistung des Produktionsclusters während der Datensicherung negativ zu beeinflussen.

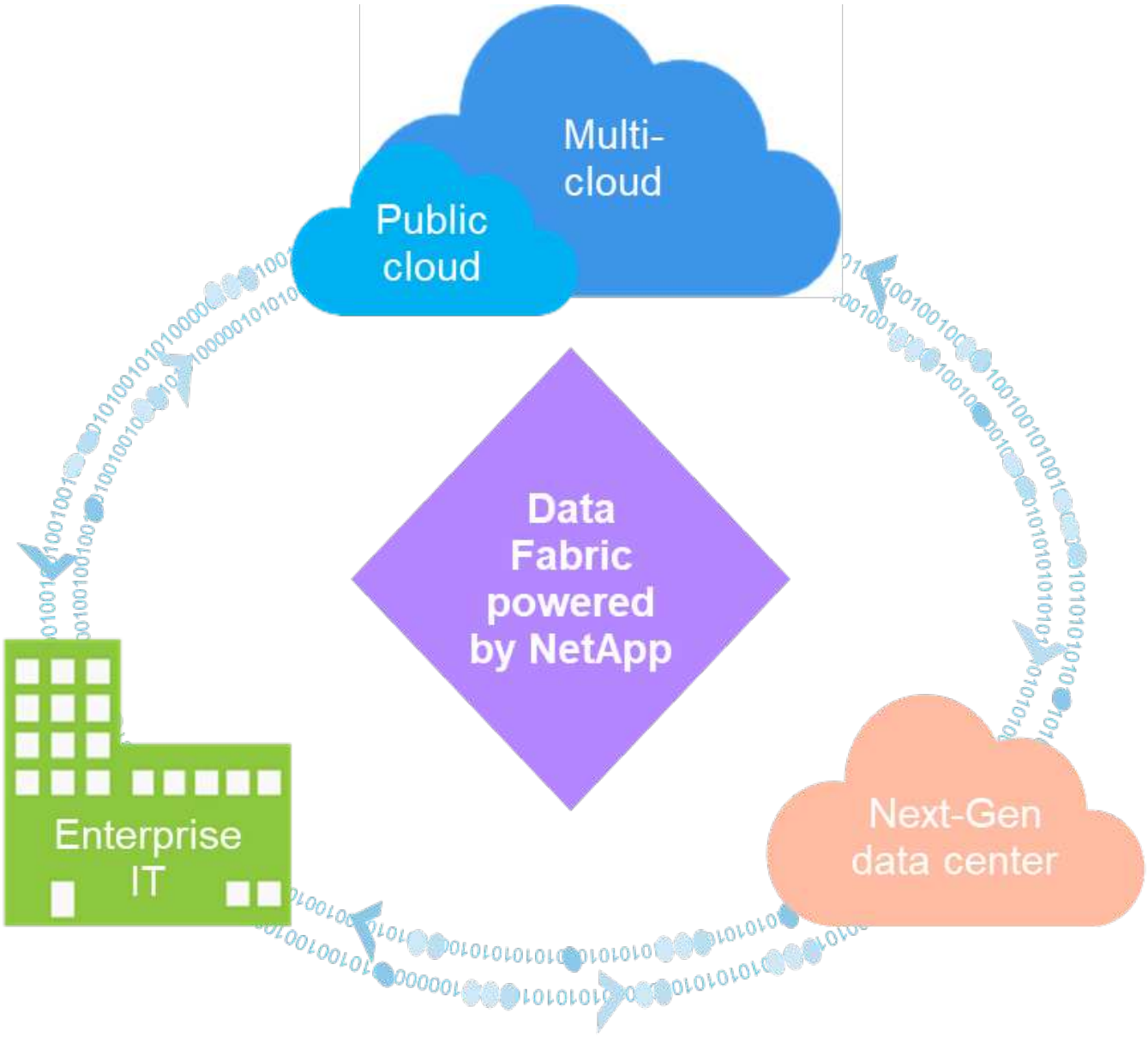
Kunden müssen außerdem die Ausfallzeiten im Hinblick auf Recovery Point Objective (RPO) und Recovery Time Objective (RTO) minimieren und ihre lokalen und Cloud-basierten Disaster-Recovery-Sites für optimale Geschäftskontinuität kontrollieren. Diese Kontrolle wird in der Regel durch den Einsatz von Managementtools auf Unternehmensebene erreicht.

Die Hadoop- und Spark-Umgebungen sind kompliziert, da nicht nur das Datenvolumen riesig ist und wächst, sondern auch die Geschwindigkeit, mit der diese Daten eintreffen, zunimmt. Dieses Szenario erschwert die schnelle Erstellung effizienter und aktueller DevTest- und QA-Umgebungen aus den Quelldaten. NetApp ist sich dieser Herausforderungen bewusst und bietet die in diesem Dokument vorgestellten Lösungen an.

Data Fabric powered by NetApp für Big Data-Architektur

Das von NetApp unterstützte Datengewebe vereinfacht und integriert das Datenmanagement in Cloud- und lokalen Umgebungen, um die digitale Transformation zu beschleunigen.

Die von NetApp betriebene Datenstruktur bietet konsistente und integrierte Datenverwaltungsdienste und -anwendungen (Bausteine) für Datentransparenz und -einblicke, Datenzugriff und -kontrolle sowie Datenschutz und -sicherheit, wie in der folgenden Abbildung dargestellt.



Bewährte Anwendungsfälle für Data Fabric-Kunden

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

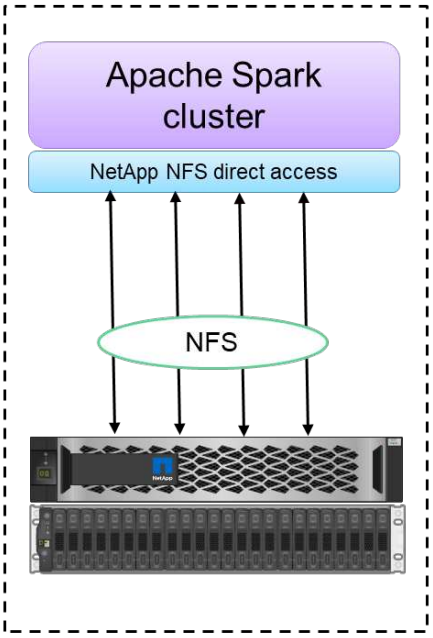
- Beschleunigen Sie Analyse-Workloads
- Beschleunigen Sie die DevOps-Transformation
- Erstellen Sie eine Cloud-Hosting-Infrastruktur
- Integrieren Sie Cloud-Datendienste
- Schützen und sichern Sie Daten
- Optimieren Sie unstrukturierte Daten
- Steigern Sie die Effizienz Ihres Rechenzentrums
- Liefern Sie Dateneinblicke und Kontrolle
- Vereinfachen und automatisieren

Dieses Dokument behandelt zwei der neun Anwendungsfälle (zusammen mit ihren Lösungen):

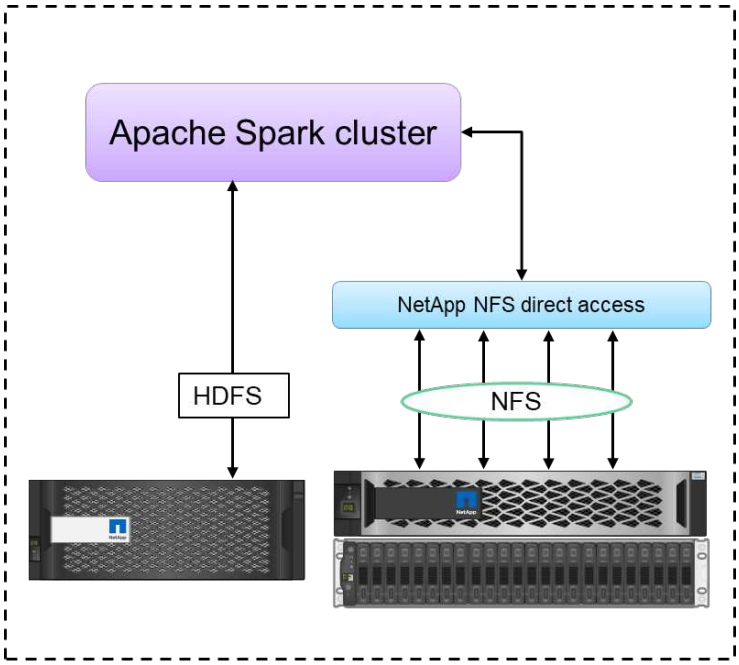
- Beschleunigen Sie Analyse-Workloads
- Schützen und sichern Sie Daten

NetApp NFS-Direktzugriff

Mit NetApp NFS können Kunden Big Data-Analysejobs auf ihren vorhandenen oder neuen NFSv3- oder NFSv4-Daten ausführen, ohne die Daten zu verschieben oder zu kopieren. Es verhindert mehrere Kopien der Daten und macht die Synchronisierung der Daten mit einer Quelle überflüssig. Im Finanzsektor beispielsweise muss die Bewegung von Daten von einem Ort zu einem anderen gesetzlichen Verpflichtungen genügen, was keine leichte Aufgabe ist. In diesem Szenario analysiert der NetApp NFS-Direktzugriff die Finanzdaten von ihrem ursprünglichen Speicherort. Ein weiterer wichtiger Vorteil besteht darin, dass die Verwendung des NetApp NFS-Direktzugriffs den Schutz von Hadoop-Daten durch die Verwendung nativer Hadoop-Befehle vereinfacht und Datenschutz-Workflows ermöglicht, die das umfangreiche Datenmanagement-Portfolio von NetApp nutzen.



Configuration 1: NFS as primary storage



Configuration 2: HDFS and NFS in single Spark cluster

Der NetApp NFS-Direktzugriff bietet zwei Arten von Bereitstellungsoptionen für Hadoop/Spark-Cluster:

- Standardmäßig verwenden die Hadoop/Spark-Cluster das Hadoop Distributed File System (HDFS) zur Datenspeicherung und als Standarddateisystem. Der NetApp NFS-Direktzugriff kann das Standard-HDFS durch NFS-Speicher als Standarddateisystem ersetzen und ermöglicht so direkte Analysevorgänge auf NFS-Daten.
- In einer anderen Bereitstellungsoption unterstützt der NetApp NFS-Direktzugriff die Konfiguration von NFS als zusätzlichen Speicher zusammen mit HDFS in einem einzelnen Hadoop/Spark-Cluster. In diesem Fall kann der Kunde Daten über NFS-Exporte freigeben und zusammen mit HDFS-Daten vom selben Cluster aus darauf zugreifen.

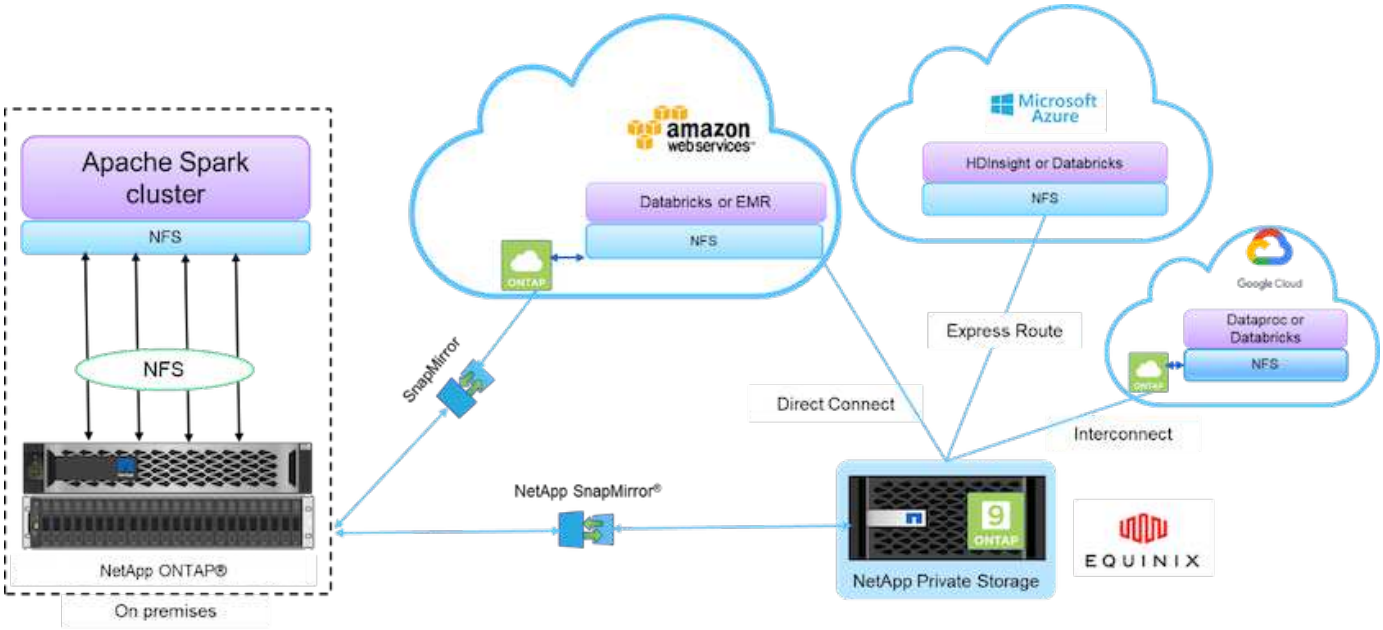
Zu den wichtigsten Vorteilen der Verwendung des NetApp NFS-Direktzugriffs gehören:

- Analysiert die Daten von ihrem aktuellen Standort aus, wodurch die zeit- und leistungsintensive Aufgabe des Verschiebens von Analysedaten in eine Hadoop-Infrastruktur wie HDFS vermieden wird.

- Reduziert die Anzahl der Replikate von drei auf eins.
- Ermöglicht Benutzern, Rechenleistung und Speicher zu entkoppeln, um sie unabhängig voneinander zu skalieren.
- Bietet Unternehmensdatenschutz durch Nutzung der umfassenden Datenverwaltungsfunktionen von ONTAP.
- Ist mit der Hortonworks-Datenplattform zertifiziert.
- Ermöglicht die Bereitstellung hybrider Datenanalysen.
- Reduziert die Sicherungszeit durch Nutzung der dynamischen Multithread-Funktion.

Bausteine für Big Data

Das von NetApp betriebene Datengewebe integriert Datenverwaltungsdienste und Anwendungen (Bausteine) für Datenzugriff, -kontrolle, -schutz und -sicherheit, wie in der folgenden Abbildung dargestellt.



Zu den Bausteinen in der obigen Abbildung gehören:

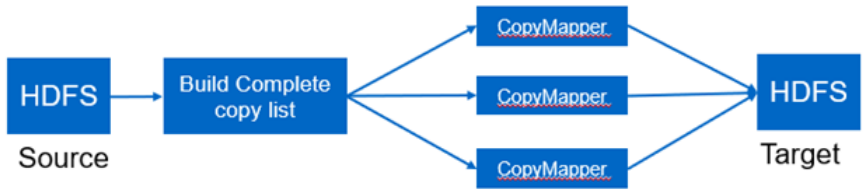
- * NetApp NFS-Direktzugriff.* Bietet den neuesten Hadoop- und Spark-Clustern direkten Zugriff auf NetApp NFS-Volumes ohne zusätzliche Software- oder Treiberanforderungen.
- * NetApp Cloud Volumes ONTAP und Google Cloud NetApp Volumes.* Softwaredefinierter verbundener Speicher basierend auf ONTAP , der in Amazon Web Services (AWS) oder Azure NetApp Files (ANF) in Microsoft Azure-Clouddiensten ausgeführt wird.
- * NetApp SnapMirror -Technologie*. Bietet Datenschutzfunktionen zwischen lokalen und ONTAP Cloud- oder NPS-Instanzen.
- **Cloud-Dienstanbieter.** Zu diesen Anbietern gehören AWS, Microsoft Azure, Google Cloud und IBM Cloud.
- **PaaS.** Cloudbasierte Analysedienste wie Amazon Elastic MapReduce (EMR) und Databricks in AWS sowie Microsoft Azure HDInsight und Azure Databricks.

Hadoop-Datenschutz und NetApp

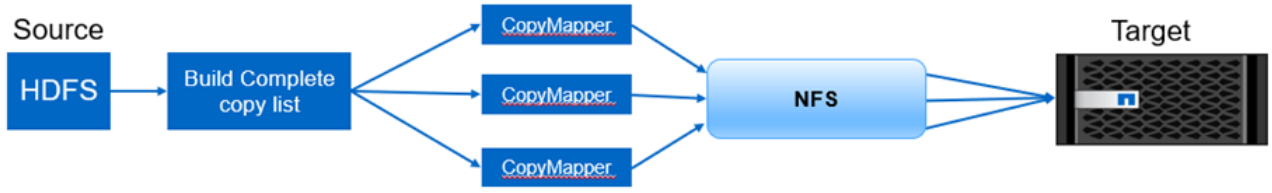
Hadoop DistCp ist ein natives Tool zum Kopieren großer Datenmengen zwischen und

innerhalb von Clustern. Der in der folgenden Abbildung dargestellte grundlegende Hadoop DistCp-Prozess ist ein typischer Sicherungsworkflow, bei dem native Hadoop-Tools wie MapReduce verwendet werden, um Hadoop-Daten von einer HDFS-Quelle auf ein entsprechendes Ziel zu kopieren.

Der NetApp NFS-Direktzugriff ermöglicht es Kunden, NFS als Ziel für das Hadoop DistCp-Tool festzulegen, um die Daten über MapReduce von der HDFS-Quelle in eine NFS-Freigabe zu kopieren. Der NetApp NFS-Direktzugriff fungiert als NFS-Treiber für das DistCp-Tool.



Hadoop Distcp Basic Process



Hadoop Distcp and NetApp

Übersicht über Anwendungsfälle für den Hadoop-Datenschutz

Dieser Abschnitt bietet eine allgemeine Beschreibung der Anwendungsfälle zum Datenschutz, die den Schwerpunkt dieses Dokuments bilden. Die verbleibenden Abschnitte enthalten weitere Details zu jedem Anwendungsfall, beispielsweise zum Kundenproblem (Szenario), zu Anforderungen und Herausforderungen sowie zu Lösungen.

Anwendungsfall 1: Sichern von Hadoop-Daten

In diesem Anwendungsfall konnte ein großes Finanzinstitut mithilfe des NetApp NFS-Volumes die lange Backup-Zeit von über 24 Stunden auf knapp ein paar Stunden reduzieren.

Anwendungsfall 2: Backup und Disaster Recovery von der Cloud in die lokale Umgebung

Durch die Verwendung der von NetApp bereitgestellten Datenstruktur als Bausteine konnte ein großes Rundfunkunternehmen seine Anforderung erfüllen, Cloud-Daten in seinem Rechenzentrum vor Ort zu sichern, und zwar abhängig von den verschiedenen Datenübertragungsmodi, beispielsweise auf Abruf, sofort oder basierend auf der Hadoop/Spark-Clusterlast.

Anwendungsfall 3: DevTest für vorhandene Hadoop-Daten aktivieren

Mithilfe von NetApp -Lösungen konnte ein Online-Musikvertrieb schnell mehrere platzsparende Hadoop-Cluster in verschiedenen Zweigstellen erstellen, um mithilfe geplanter Richtlinien Berichte zu erstellen und tägliche DevTest-Aufgaben auszuführen.

Anwendungsfall 4: Datenschutz und Multicloud-Konnektivität

Ein großer Dienstanbieter nutzte das von NetApp betriebene Daten-Fabric, um seinen Kunden Multicloud-Analysen aus verschiedenen Cloud-Instanzen bereitzustellen.

Anwendungsfall 5: Beschleunigen analytischer Workloads

Eine der größten Finanzdienstleistungs- und Investmentbanken nutzte die Network Attached Storage-Lösung von NetApp , um die I/O-Wartezeit zu reduzieren und ihre Plattform für quantitative Finanzanalysen zu beschleunigen.

Anwendungsfall 1: Sichern von Hadoop-Daten

In diesem Szenario verfügt der Kunde über ein großes lokales Hadoop-Repository und möchte es für die Notfallwiederherstellung sichern. Die aktuelle Backup-Lösung des Kunden ist jedoch kostspielig und leidet unter einem langen Backup-Fenster von mehr als 24 Stunden.

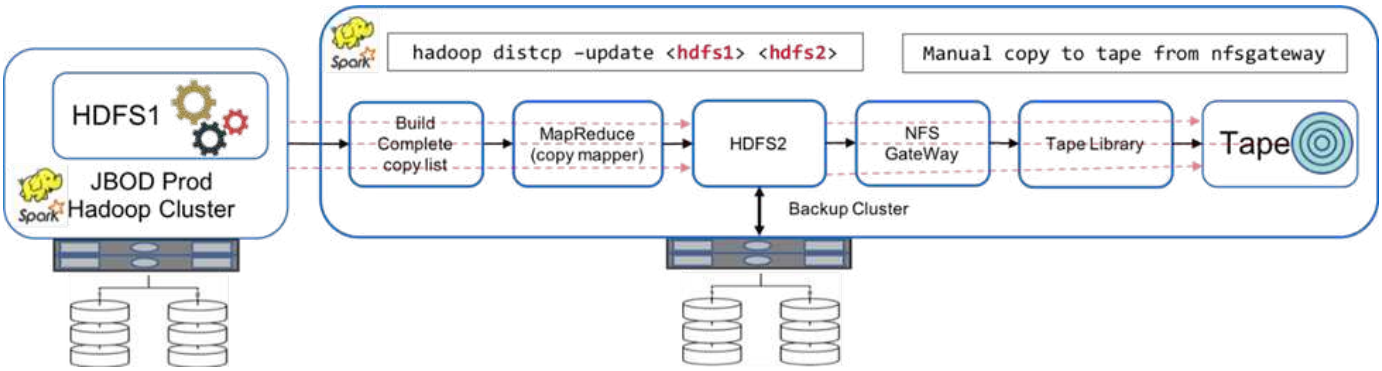
Anforderungen und Herausforderungen

Zu den wichtigsten Anforderungen und Herausforderungen für diesen Anwendungsfall gehören:

- Software-Abwärtskompatibilität:
  - Die vorgeschlagene alternative Sicherungslösung sollte mit den aktuell im Produktions-Hadoop-Cluster verwendeten Softwareversionen kompatibel sein.
- Um die vereinbarten SLAs einzuhalten, sollte die vorgeschlagene Alternativlösung sehr niedrige RPOs und RTOs erreichen.
- Das von der NetApp Backup-Lösung erstellte Backup kann sowohl im lokal im Rechenzentrum erstellten Hadoop-Cluster als auch im Hadoop-Cluster verwendet werden, der am Disaster-Recovery-Standort am Remote-Standort ausgeführt wird.
- Die vorgeschlagene Lösung muss kosteneffizient sein.
- Die vorgeschlagene Lösung muss die Leistungseinbußen bei den aktuell laufenden, produktiven Analysejobs während der Sicherungszeiten reduzieren.

Vorhandene Backup-Lösung des Kundenx

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



Die Produktionsdaten werden über den Zwischensicherungscluster auf Band gesichert:

- HDFS1-Daten werden nach HDFS2 kopiert, indem der `hadoop distcp -update <hdfs1> <hdfs2>` Befehl.
- Der Backup-Cluster fungiert als NFS-Gateway und die Daten werden manuell über das Linux-System auf Band kopiert. `cp` Befehl durch die Bandbibliothek.

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

- Die Lösung basiert auf nativen Hadoop-Befehlen, sodass der Benutzer keine neuen Verfahren erlernen muss.
- Die Lösung nutzt eine dem Industriestandard entsprechende Architektur und Hardware.

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

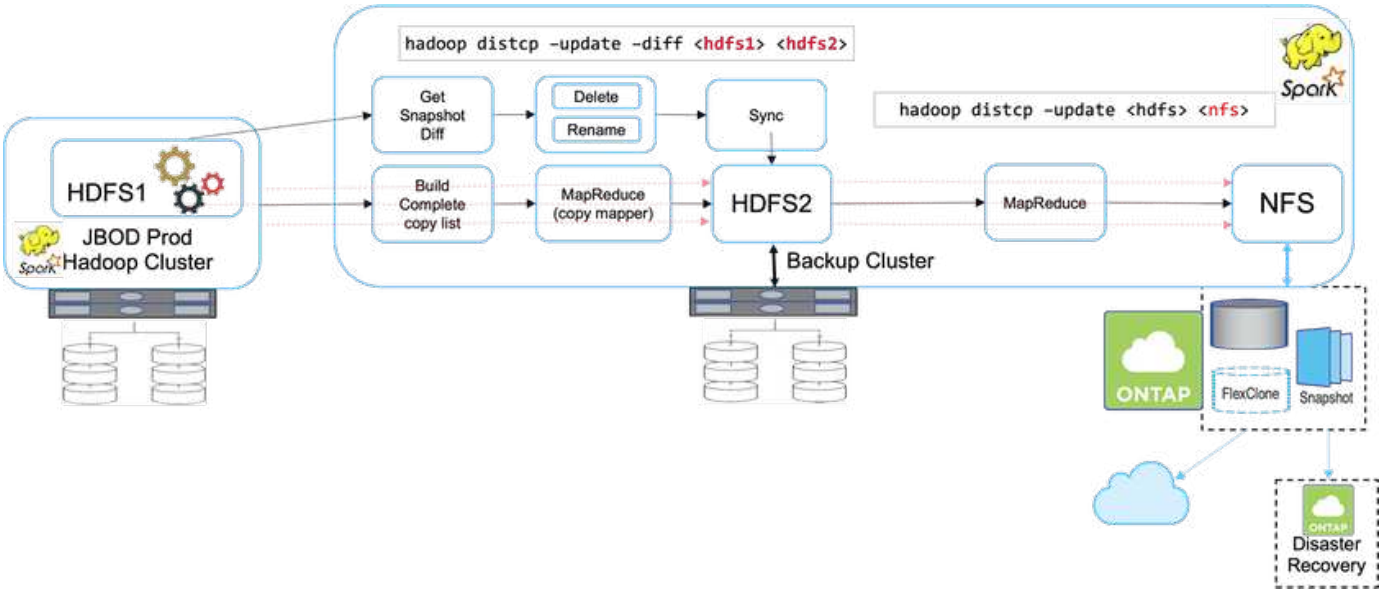
- Das lange Backup-Fenster beträgt mehr als 24 Stunden, wodurch die Produktionsdaten gefährdet sind.
- Erhebliche Leistungseinbußen des Clusters während der Sicherungszeiten.
- Das Kopieren auf Band ist ein manueller Vorgang.
- Die Backup-Lösung ist hinsichtlich der erforderlichen Hardware und der für manuelle Prozesse erforderlichen Arbeitsstunden teuer.

Backup-Lösungen

Basierend auf diesen Herausforderungen und Anforderungen und unter Berücksichtigung des vorhandenen Backup-Systems wurden drei mögliche Backup-Lösungen vorgeschlagen. In den folgenden Unterabschnitten werden die drei verschiedenen Sicherungslösungen mit den Bezeichnungen Lösung A bis Lösung C beschrieben.

Lösung A

In Lösung A sendet der Hadoop-Backup-Cluster die sekundären Backups an NetApp NFS-Speichersysteme, wodurch die Bandanforderung entfällt, wie in der folgenden Abbildung dargestellt.



Die detaillierten Aufgaben für Lösung A umfassen:

- Der Produktions-Hadoop-Cluster verfügt über die Analysedaten des Kunden im HDFS, die geschützt werden müssen.
- Der Backup-Hadoop-Cluster mit HDFS fungiert als Zwischenspeicherort für die Daten. Nur eine Reihe von Festplatten (JBOD) stellt den Speicher für HDFS sowohl in den Produktions- als auch in den Backup-Hadoop-Clustern bereit.
- Schützen Sie die Hadoop-Produktionsdaten vom Produktionscluster HDFS zum Backup-Cluster HDFS, indem Sie Folgendes ausführen: `Hadoop distcp –update –diff <hdfs1> <hdfs2>` Befehl.

> Der Hadoop-Snapshot wird verwendet, um die Daten vor der Produktion im Hadoop-Backup-Cluster zu schützen.

- Der NetApp ONTAP Speichercontroller stellt ein per NFS exportiertes Volume bereit, das für den Backup-Hadoop-Cluster bereitgestellt wird.
- Durch Ausführen des `Hadoop distcp` Durch den Einsatz von MapReduce und mehreren Mappern werden die Analysedaten vom Backup-Hadoop-Cluster auf NFS geschützt.

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

> Hadoop-Daten können mithilfe der SnapMirror -Technologie sowohl in der Cloud als auch an Disaster-Recovery-Standorten geschützt werden.

Zu den Vorteilen von Lösung A gehören:

- Hadoop-Produktionsdaten werden vor dem Backup-Cluster geschützt.
- HDFS-Daten werden durch NFS geschützt und ermöglichen so den Schutz von Cloud- und Notfallwiederherstellungsstandorten.
- Verbessert die Leistung durch Auslagerung von Sicherungsvorgängen auf den Sicherungscluster.
- Eliminiert manuelle Bandvorgänge

- Ermöglicht Unternehmensverwaltungsfunktionen über NetApp -Tools.
- Erfordert nur minimale Änderungen an der vorhandenen Umgebung.
- Ist eine kostengünstige Lösung.

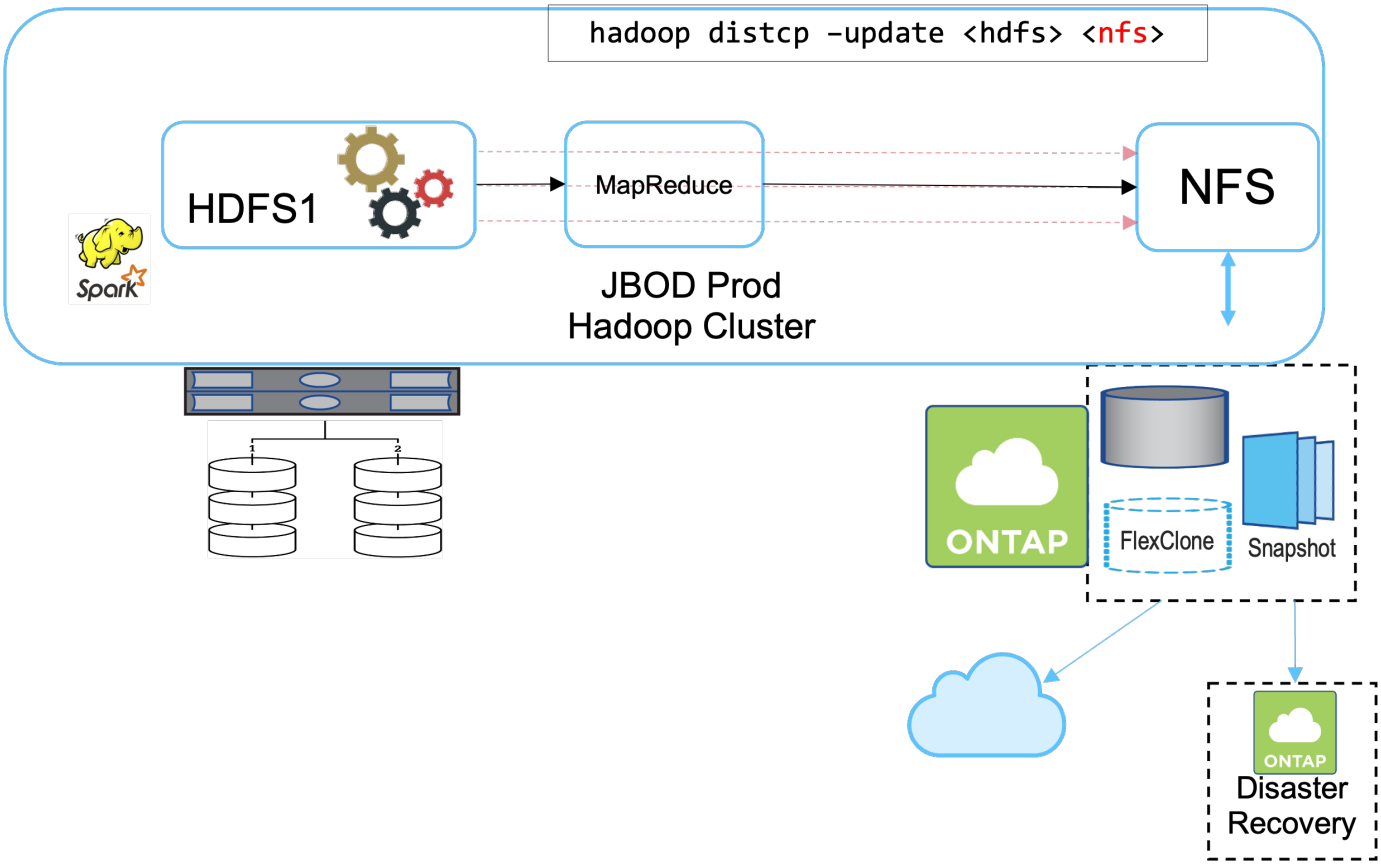
Der Nachteil dieser Lösung besteht darin, dass sie einen Backup-Cluster und zusätzliche Mapper zur Leistungsverbesserung erfordert.

Der Kunde hat vor Kurzem Lösung A aufgrund ihrer Einfachheit, Kosten und Gesamtleistung implementiert.

Bei dieser Lösung können SAN-Festplatten von ONTAP anstelle von JBOD verwendet werden. Diese Option verlagert die Speicherlast des Backup-Clusters auf ONTAP. Der Nachteil besteht jedoch darin, dass SAN-Fabric-Switches erforderlich sind.

Lösung B

Lösung B fügt dem Produktions-Hadoop-Cluster ein NFS-Volume hinzu, wodurch der Backup-Hadoop-Cluster überflüssig wird, wie in der folgenden Abbildung gezeigt.



Die detaillierten Aufgaben für Lösung B umfassen:

- Der NetApp ONTAP Speichercontroller stellt den NFS-Export zum Produktions-Hadoop-Cluster bereit.

  Der Hadoop-Native `hadoop distcp` Der Befehl schützt die Hadoop-Daten vom Produktionscluster HDFS auf NFS.

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

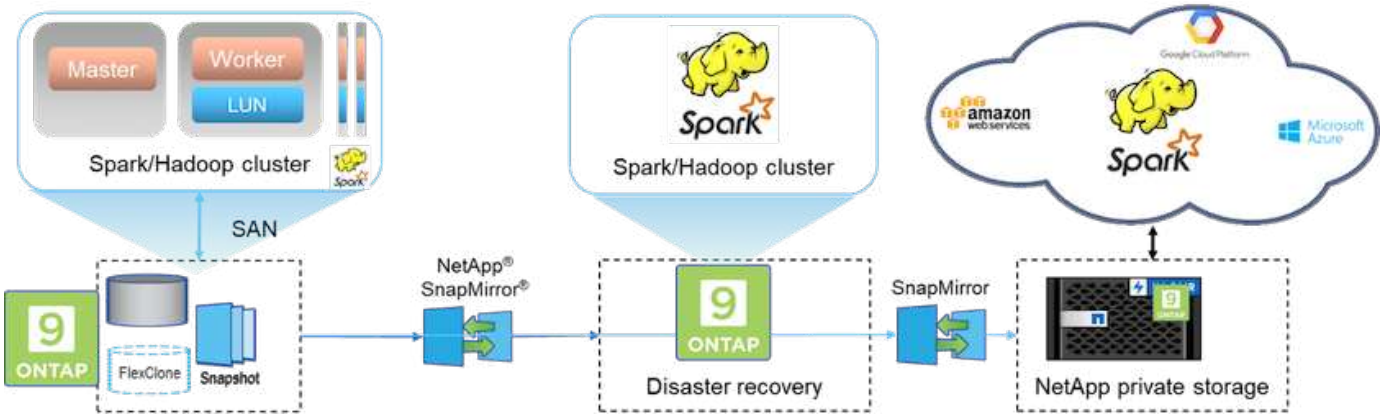
Zu den Vorteilen der Lösung B gehören:

- Der Produktionscluster wird für die Backup-Lösung leicht modifiziert, was die Implementierung vereinfacht und zusätzliche Infrastrukturkosten reduziert.
- Ein Backup-Cluster für den Backup-Vorgang ist nicht erforderlich.
- HDFS-Produktionsdaten werden bei der Konvertierung in NFS-Daten geschützt.
- Die Lösung ermöglicht Unternehmensverwaltungsfunktionen über NetApp -Tools.

Der Nachteil dieser Lösung besteht darin, dass sie im Produktionscluster implementiert wird, was zu zusätzlichen Administratoraufgaben im Produktionscluster führen kann.

Lösung C

In Lösung C werden die NetApp SAN-Volumes direkt für den Hadoop-Produktionscluster zur HDFS-Speicherung bereitgestellt, wie in der folgenden Abbildung dargestellt.



Die detaillierten Schritte für Lösung C umfassen:

- NetApp ONTAP SAN-Speicher wird im Produktions-Hadoop-Cluster für die HDFS-Datenspeicherung bereitgestellt.
- Zum Sichern der HDFS-Daten aus dem Produktions-Hadoop-Cluster werden die Technologien NetApp Snapshot und SnapMirror verwendet.
- Während des Snapshot-Kopiersicherungsprozesses kommt es zu keinen Leistungseinbußen bei der Produktion des Hadoop/Spark-Clusters, da die Sicherung auf der Speicherebene erfolgt.

> Die Snapshot-Technologie ermöglicht Backups, die unabhängig von der Datengröße in Sekundenschnelle abgeschlossen sind.

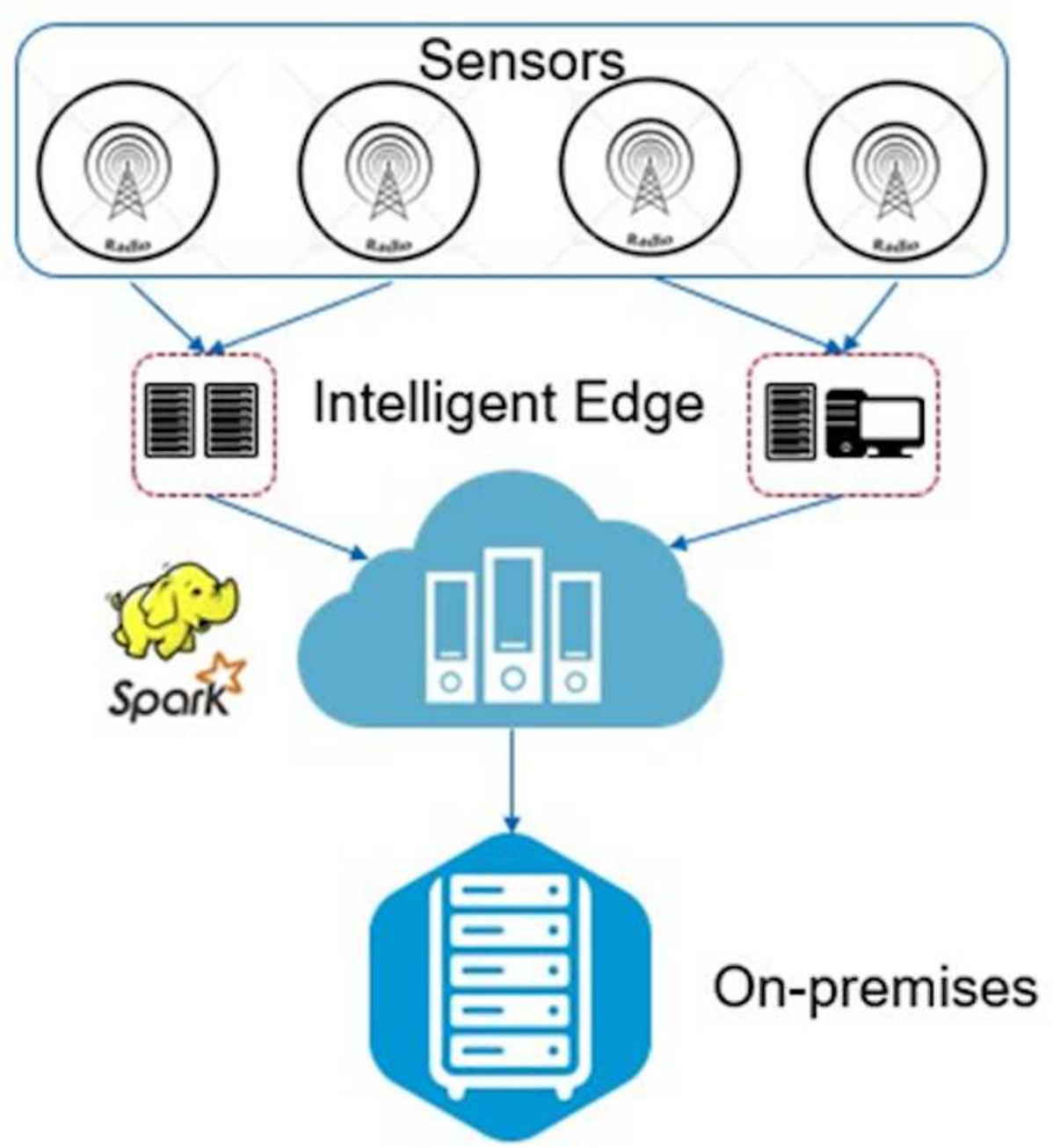
Zu den Vorteilen der Lösung C gehören:

- Mithilfe der Snapshot-Technologie können platzsparende Backups erstellt werden.
- Ermöglicht Unternehmensverwaltungsfunktionen über NetApp -Tools.

Anwendungsfall 2: Backup und Disaster Recovery von der Cloud in die lokale Umgebung

Dieser Anwendungsfall basiert auf einem Rundfunkkunden, der Cloud-basierte

Analysedaten in seinem lokalen Rechenzentrum sichern muss, wie in der folgenden Abbildung dargestellt.



Szenario

In diesem Szenario werden die IoT-Sensordaten in die Cloud eingespeist und mithilfe eines Open-Source-Apache-Spark-Clusters innerhalb von AWS analysiert. Voraussetzung ist, dass die verarbeiteten Daten aus der Cloud vor Ort gesichert werden.

Anforderungen und Herausforderungen

Zu den wichtigsten Anforderungen und Herausforderungen für diesen Anwendungsfall gehören:

- Das Aktivieren des Datenschutzes sollte keine Auswirkungen auf die Leistung des Spark/Hadoop-

Produktionsclusters in der Cloud haben.

- Cloud-Sensordaten müssen effizient und sicher vor Ort verschoben und geschützt werden.
- Flexibilität bei der Datenübertragung von der Cloud zu lokalen Standorten unter verschiedenen Bedingungen, z. B. bei Bedarf, sofort und bei geringer Cluster-Auslastung.

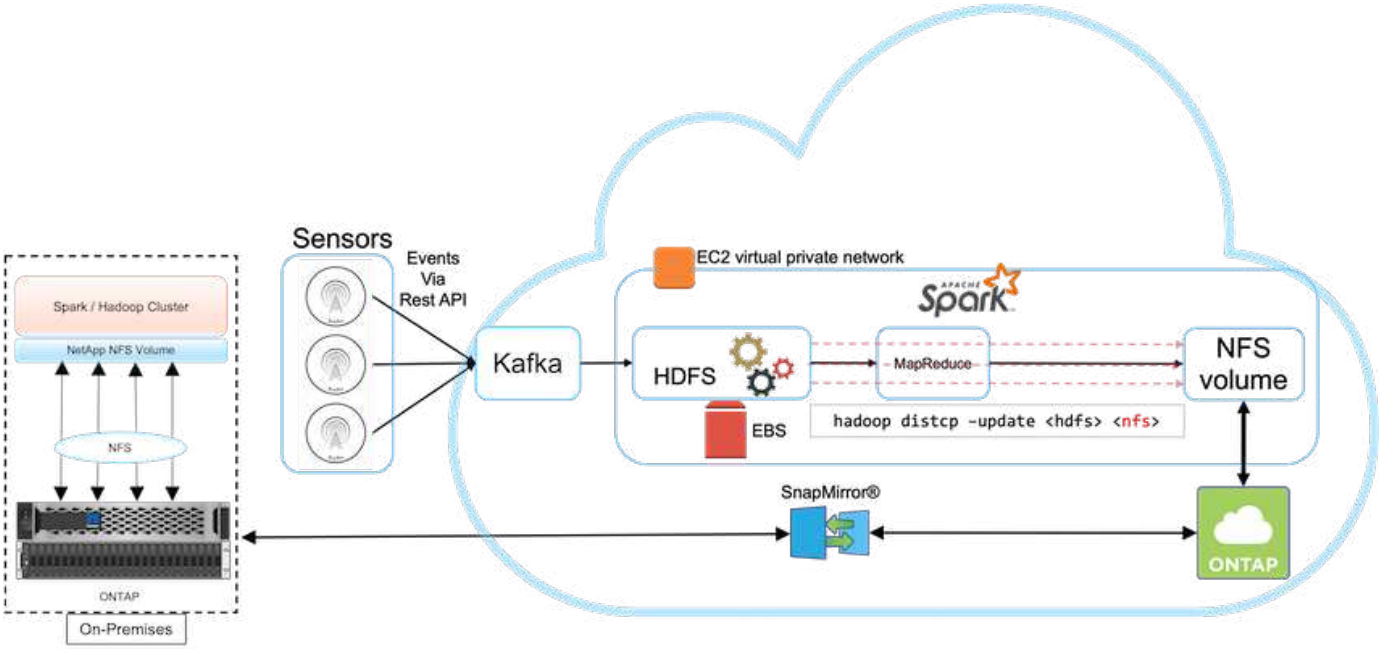
Lösung

Der Kunde verwendet AWS Elastic Block Store (EBS) für seinen Spark-Cluster-HDFS-Speicher, um Daten von Remote-Sensoren über Kafka zu empfangen und aufzunehmen. Folglich fungiert der HDFS-Speicher als Quelle für die Sicherungsdaten.

Um diese Anforderungen zu erfüllen, wird NetApp ONTAP Cloud in AWS bereitgestellt und eine NFS-Freigabe erstellt, die als Sicherungsziel für den Spark/Hadoop-Cluster fungiert.

Nachdem die NFS-Freigabe erstellt wurde, kopieren Sie die Daten aus dem HDFS-EBS-Speicher in die ONTAP -NFS-Freigabe. Nachdem die Daten in NFS in ONTAP Cloud gespeichert wurden, können die Daten mithilfe der SnapMirror -Technologie bei Bedarf sicher und effizient aus der Cloud in den lokalen Speicher gespiegelt werden.

Dieses Bild zeigt die Backup- und Disaster-Recovery-Lösung von der Cloud zur lokalen Lösung.



Anwendungsfall 3: DevTest für vorhandene Hadoop-Daten aktivieren

In diesem Anwendungsfall besteht die Anforderung des Kunden darin, schnell und effizient neue Hadoop/Spark-Cluster auf der Grundlage eines vorhandenen Hadoop-Clusters zu erstellen, der eine große Menge an Analysedaten für DevTest- und Berichtszwecke im selben Rechenzentrum sowie an Remote-Standorten enthält.

Szenario

In diesem Szenario werden mehrere Spark/Hadoop-Cluster aus einer großen Hadoop-Data-Lake-Implementierung vor Ort sowie an Disaster-Recovery-Standorten erstellt.

Anforderungen und Herausforderungen

Zu den wichtigsten Anforderungen und Herausforderungen für diesen Anwendungsfall gehören:

- Erstellen Sie mehrere Hadoop-Cluster für DevTest, Qualitätssicherung oder andere Zwecke, die Zugriff auf dieselben Produktionsdaten erfordern. Die Herausforderung besteht darin, einen sehr großen Hadoop-Cluster mehrere Male gleichzeitig und auf sehr platzsparende Weise zu klonen.
- Synchronisieren Sie die Hadoop-Daten mit DevTest- und Berichtsteams, um die betriebliche Effizienz zu steigern.
- Verteilen Sie die Hadoop-Daten, indem Sie dieselben Anmeldeinformationen in der Produktion und in neuen Clustern verwenden.
- Verwenden Sie geplante Richtlinien, um QA-Cluster effizient zu erstellen, ohne den Produktionscluster zu beeinträchtigen.

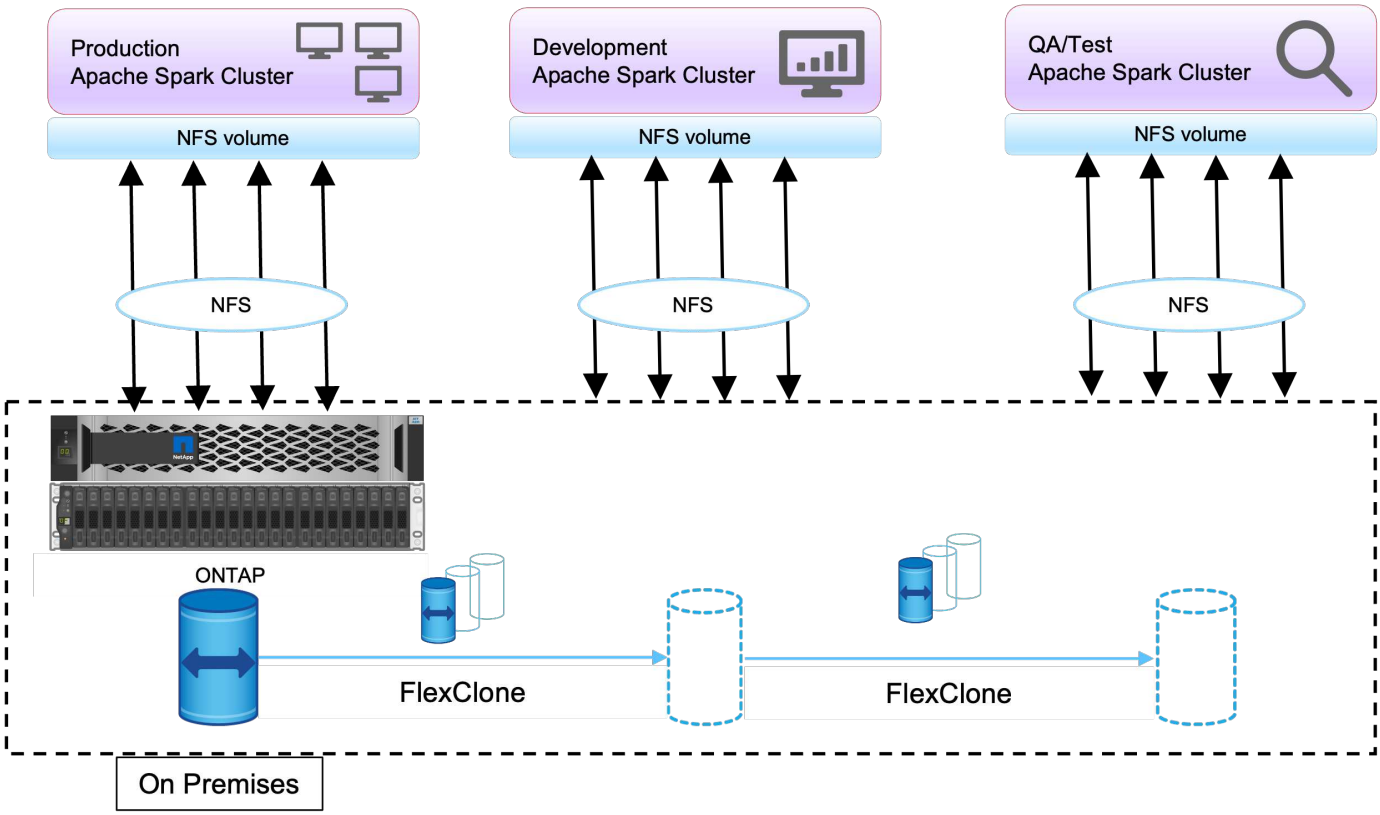
Lösung

Um die gerade beschriebenen Anforderungen zu erfüllen, wird die FlexClone -Technologie verwendet. Die FlexClone -Technologie ist die Lese-/Schreibkopie einer Snapshot-Kopie. Es liest die Daten aus den übergeordneten Snapshot-Kopiedaten und verbraucht nur zusätzlichen Speicherplatz für neue/geänderte Blöcke. Es ist schnell und platzsparend.

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

Snapshot-Kopien im NetApp System Manager oder in der Storage-Admin-Eingabeaufforderung. Bei den Snapshot-Kopien der Konsistenzgruppe handelt es sich um anwendungskonsistente Snapshot-Kopien der Gruppe, und das FlexClone -Volume wird basierend auf Snapshot-Kopien der Konsistenzgruppe erstellt. Es ist erwähnenswert, dass ein FlexClone -Volume die NFS-Exportrichtlinie des übergeordneten Volumes erbt. Nachdem die Snapshot-Kopie erstellt wurde, muss ein neuer Hadoop-Cluster für DevTest- und Berichtszwecke installiert werden, wie in der folgenden Abbildung 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: Datenschutz und Multicloud-Konnektivität

Dieser Anwendungsfall ist für einen Cloud-Service-Partner relevant, der die Aufgabe hat, Multicloud-Konnektivität für die Big-Data-Analysedaten von Kunden bereitzustellen.

Szenario

In diesem Szenario werden IoT-Daten, die in AWS aus verschiedenen Quellen empfangen werden, an einem zentralen Ort in NPS gespeichert. Der NPS-Speicher ist mit Spark/Hadoop-Clustern in AWS und Azure verbunden, sodass Big Data-Analyseanwendungen in mehreren Clouds ausgeführt werden können und auf dieselben Daten zugreifen.

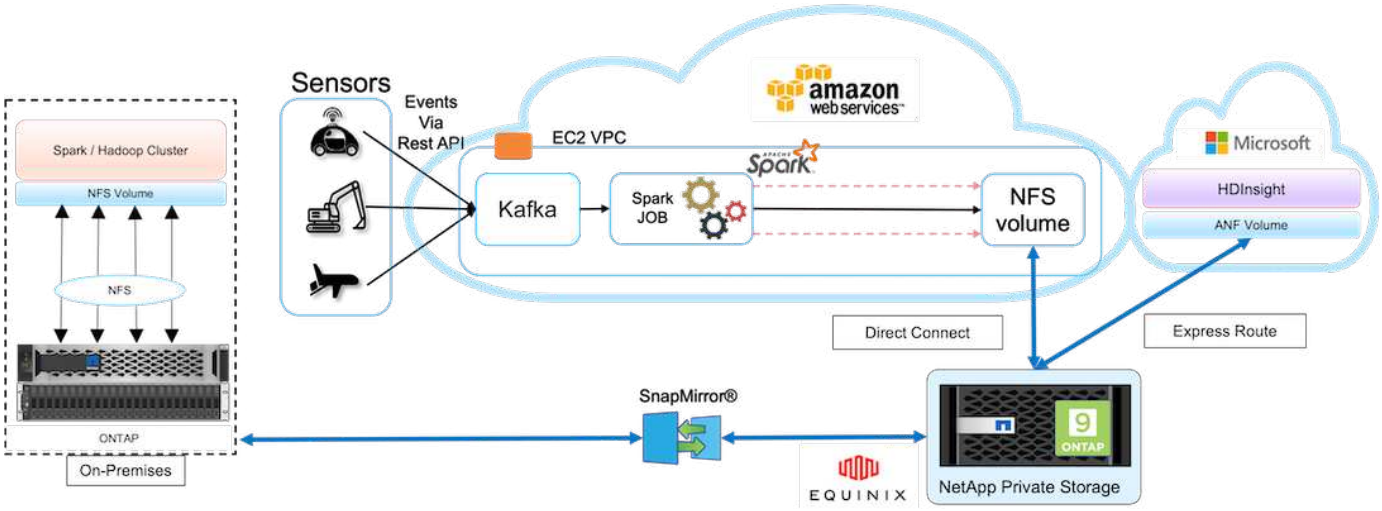
Anforderungen und Herausforderungen

Zu den wichtigsten Anforderungen und Herausforderungen für diesen Anwendungsfall gehören:

- Kunden möchten Analyseaufträge mit denselben Daten über mehrere Clouds ausführen.
- Daten müssen aus verschiedenen Quellen, beispielsweise vor Ort und aus der Cloud, über verschiedene Sensoren und Hubs empfangen werden.
- Die Lösung muss effizient und kostengünstig sein.
- Die größte Herausforderung besteht darin, eine kostengünstige und effiziente Lösung zu entwickeln, die hybride Analysedienste zwischen lokalen Standorten und verschiedenen Clouds bereitstellt.

Lösung

Dieses Bild veranschaulicht die Lösung für Datenschutz und Multicloud-Konnektivität.



Wie in der obigen Abbildung gezeigt, werden Daten von Sensoren gestreamt und über Kafka in den AWS Spark-Cluster aufgenommen. Die Daten werden in einer NFS-Freigabe gespeichert, die sich in NPS befindet, das sich außerhalb des Cloud-Anbieters in einem Equinix-Rechenzentrum befindet. Da NetApp NPS über Direct Connect- bzw. Express Route-Verbindungen mit Amazon AWS und Microsoft Azure verbunden ist, können Kunden auf die NFS-Daten sowohl von Amazon- als auch von AWS-Analyseclustern zugreifen. Dieser Ansatz löst das Problem, Cloud-Analysen über mehrere Hyperscaler hinweg durchzuführen.

Da sowohl der lokale als auch der NPS-Speicher mit ONTAP -Software laufen, kann SnapMirror die NPS-Daten in den lokalen Cluster spiegeln und so Hybrid-Cloud-Analysen über lokale und mehrere Clouds hinweg bereitstellen.

Für eine optimale Leistung empfiehlt NetApp normalerweise die Verwendung mehrerer Netzwerkschnittstellen und Direktverbindungen/Expressrouten für den Zugriff auf die Daten von Cloud-Instanzen.

Anwendungsfall 5: Beschleunigen analytischer Workloads

In diesem Szenario wurde die Analyseplattform einer großen Finanzdienstleistungs- und Investmentbank mithilfe der NetApp NFS-Speicherlösung modernisiert, um eine deutliche Verbesserung bei der Analyse von Anlagerisiken und Derivaten für ihre Vermögensverwaltungs- und quantitative Geschäftseinheit zu erreichen.

Szenario

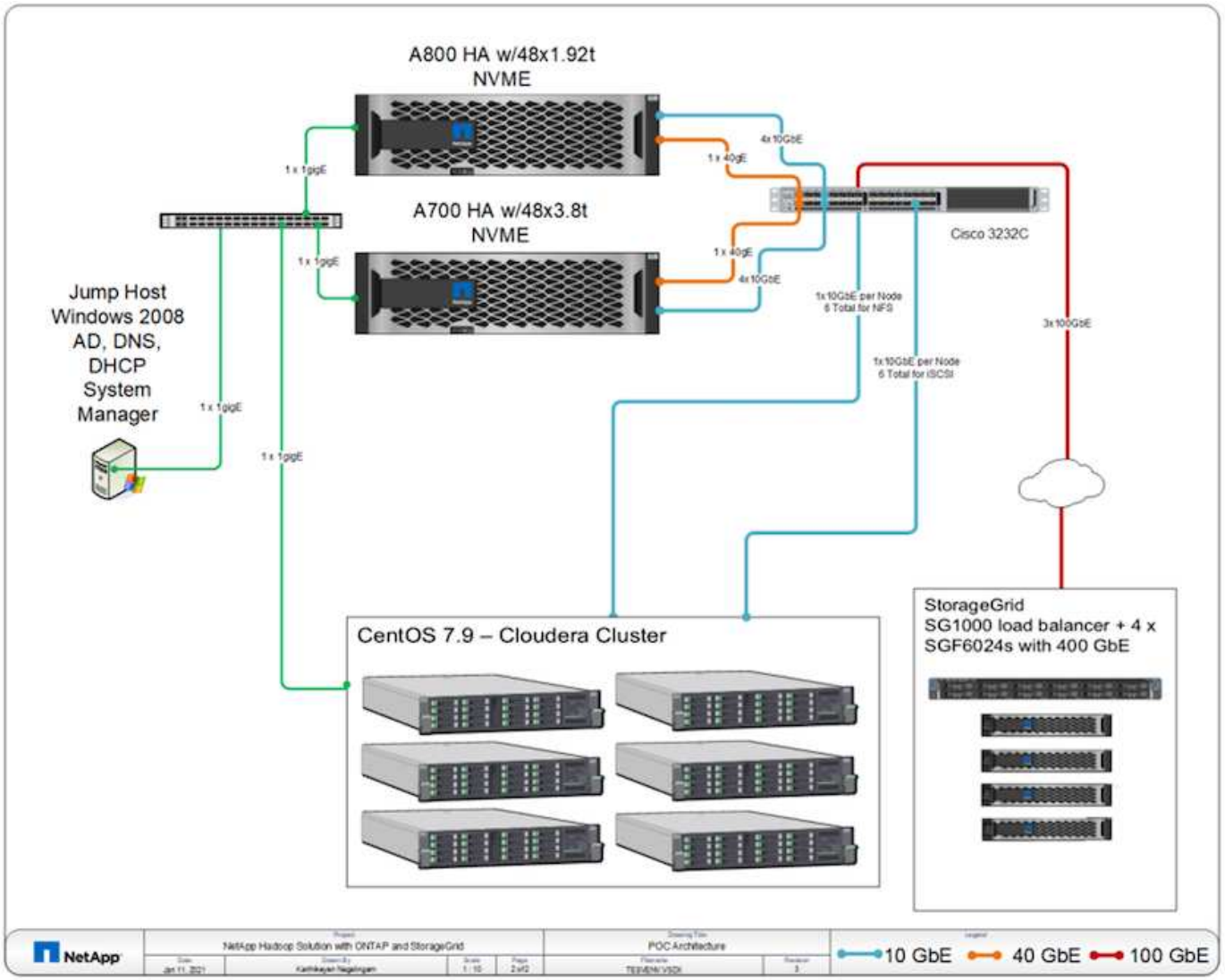
In der bestehenden Umgebung des Kunden nutzte die für die Analyseplattform verwendete Hadoop-Infrastruktur den internen Speicher der Hadoop-Server. Aufgrund der proprietären Natur der JBOD-Umgebung konnten viele interne Kunden innerhalb der Organisation ihr quantitatives Monte-Carlo-Modell nicht nutzen, eine Simulation, die auf wiederkehrenden Stichproben von Echtzeitdaten basiert. Die unzureichende Fähigkeit, die Auswirkungen der Unsicherheit bei Marktbewegungen zu verstehen, wirkte sich nachteilig auf die Geschäftseinheit für quantitatives Asset Management aus.

Anforderungen und Herausforderungen

Die quantitative Geschäftseinheit der Bank wollte eine effiziente Prognosemethode, um genaue und zeitnahe Vorhersagen zu erzielen. Um dies zu erreichen, erkannte das Team die Notwendigkeit, die Infrastruktur zu modernisieren, die bestehende E/A-Wartezeit zu reduzieren und die Leistung der Analyseanwendungen wie Hadoop und Spark zu verbessern, um Investitionsmodelle effizient zu simulieren, potenzielle Gewinne zu messen und Risiken zu analysieren.

Lösung

Der Kunde hatte JBOD für seine bestehende Spark-Lösung. NetApp ONTAP, NetApp StorageGRID und MinIO Gateway to NFS wurden dann genutzt, um die E/A-Wartezeit für die quantitative Finanzgruppe der Bank zu reduzieren, die Simulationen und Analysen von Investitionsmodellen durchführt, um potenzielle Gewinne und Risiken zu bewerten. Dieses Bild zeigt die Spark-Lösung mit NetApp -Speicher.



Wie in der Abbildung oben gezeigt, wurden AFF A800 und A700-Systeme sowie StorageGRID bereitgestellt, um über NFS- und S3-Protokolle in einem Hadoop-Cluster mit sechs Knoten mit Spark sowie YARN- und Hive-Metadatendiensten für Datenanalysevorgänge auf Parquet-Dateien zuzugreifen.

Eine Direct-Attached-Storage-Lösung (DAS) in der alten Umgebung des Kunden hatte den Nachteil, dass Rechenleistung und Speicher unabhängig voneinander skaliert werden mussten. Mit der NetApp ONTAP

-Lösung für Spark konnte die Geschäftseinheit für Finanzanalysen der Bank Speicher und Rechenleistung entkoppeln und Infrastrukturressourcen bei Bedarf nahtlos und effektiver bereitstellen.

Durch die Verwendung von ONTAP mit NFS wurden die CPUs des Compute-Servers für Spark SQL-Jobs fast vollständig genutzt und die E/A-Wartezeit um fast 70 % reduziert, was zu einer besseren Rechenleistung und Leistungssteigerung für Spark-Workloads führte. Durch die anschließende Erhöhung der CPU-Auslastung konnte der Kunde auch GPUs wie GPUDirect für eine weitere Modernisierung der Plattform nutzen. Darüber hinaus bietet StorageGRID eine kostengünstige Speicheroption für Spark-Workloads und MinIO Gateway bietet sicheren Zugriff auf NFS-Daten über das S3-Protokoll. Für Daten in der Cloud empfiehlt NetApp Cloud Volumes ONTAP, Azure NetApp Files und Google Cloud NetApp Volumes.

Abschluss

Dieser Abschnitt bietet eine Zusammenfassung der Anwendungsfälle und Lösungen von NetApp zur Erfüllung verschiedener Hadoop-Datenschutzanforderungen. Durch die Verwendung der von NetApp bereitgestellten Datenstruktur können Kunden:

- Profitieren Sie von der Flexibilität, die richtigen Datenschutzlösungen auszuwählen, indem Sie die umfassenden Datenverwaltungsfunktionen von NetApp und die Integration mit nativen Hadoop-Workflows nutzen.
- Reduzieren Sie die Zeit des Backup-Fensters Ihres Hadoop-Clusters um fast 70 %.
- Beseitigen Sie alle Leistungseinbußen, die durch Hadoop-Cluster-Backups entstehen.
- Bieten Sie Multicloud-Datenschutz und Datenzugriff von verschiedenen Cloud-Anbietern gleichzeitig auf eine einzige Quelle von Analysedaten.
- Erstellen Sie mithilfe der FlexClone -Technologie schnelle und platzsparende Hadoop-Clusterkopien.

Wo Sie weitere Informationen finden

Weitere Informationen zu den in diesem Dokument beschriebenen Informationen finden Sie in den folgenden Dokumenten und/oder auf den folgenden Websites:

- NetApp Big Data Analytics-Lösungen

  "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 auf Data Fabric, unterstützt von NetApp

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

Danksagung

- Paul Burland, Vertriebsmitarbeiter, ANZ Victoria District Sales, NetApp

- Hoseb Dermanilian, Business Development Manager, NetApp
- Lee Dorrier, Direktor MPSG, NetApp
- David Thiessen, Systemingenieur, ANZ Victoria District SE, NetApp

Versionsverlauf

Version	Datum	Dokumentversionsverlauf
Version 1.0	Januar 2018	Erstveröffentlichung
Version 2.0	Oktober 2021	Aktualisiert mit Anwendungsfall Nr. 5: Beschleunigen Sie die analytische Arbeitslast
Version 3.0	November 2023	NIPAM-Details entfernt