NetAppハイブリッド クラウド データ ソリューション - 顧客のユースケースに基づいた Spark と Hadoop

NetApp artificial intelligence solutions

NetApp
August 18, 2025

This PDF was generated from https://docs.netapp.com/ja-jp/netapp-solutions-ai/data-analytics/hdcs-sh-solution-overview.html on August 18, 2025. Always check docs.netapp.com for the latest.

目次

NetAppハイブリッド クラウド データ ソリューション - 顧客のユースケースに基づいた Spark と Hadoop 1
TR-4657: NetAppハイブリッド クラウド データ ソリューション - 顧客のユースケースに基づく Spark と Hadoop 1
Hadoop データ保護の理由 1
Hadoop および Spark の顧客にとってのデータ保護の課題 2
ビッグデータ アーキテクチャ向けのNetApp搭載データ ファブリック 2
実証済みのデータファブリックの顧客ユースケース 3
Hadoopデータ保護とNetApp 6
Hadoop データ保護のユースケースの概要 6
ユースケース1: Hadoopデータのバックアップ 6
ユースケース 2: クラウドからオンプレミスへのバックアップと災害復旧 6
ユースケース3: 既存のHadoopデータでDevTestを有効にする 7
ユースケース4: データ保護とマルチクラウド接続 7
ユースケース5: 分析ワークロードの高速化 7
ユースケース1: Hadoopデータのバックアップ 7
要件と課題 7
顧客の既存のバックアップソリューションx 7
バックアップソリューション 8
ユースケース 2: クラウドからオンプレミスへのバックアップと災害復旧 12
シナリオ 12
要件と課題 13
解決策 13
ユースケース3: 既存のHadoopデータでDevTestを有効にする 13
シナリオ 14
要件と課題 14
解決策 14
ユースケース4: データ保護とマルチクラウド接続 15
シナリオ 15
要件と課題 15
解決策 15
ユースケース5: 分析ワークロードの高速化 16
シナリオ 16
要件と課題 16
解決策 17
まとめ 18
詳細情報の入手方法 18
謝辞 18
バージョン履歴 19

NetAppハイブリッド クラウド データ ソリューション - 顧客のユースケースに基づいた Spark と Hadoop

TR-4657: NetAppハイブリッド クラウド データ ソリューション - 顧客のユースケースに基づく Spark と Hadoop

NetApp 、Karthikeyan Nagalingam 氏と Sathish Thyagarajan 氏

このドキュメントでは、 NetApp AFFおよびFASストレージ システム、 NetApp Cloud Volumes ONTAP、 NetApp接続ストレージ、Spark および Hadoop 用のNetApp FlexCloneテクノロジーを使用したハイブリッド クラウド データ ソリューションについて説明します。これらのソリューション アーキテクチャにより、お客様は自社の環境に適したデータ保護ソリューションを選択できます。 NetApp は、顧客とのやり取りとビジネスユースケースに基づいてこれらのソリューションを設計しました。このドキュメントでは、次の詳細情報を提供します。

- Spark および Hadoop 環境でデータ保護が必要な理由と顧客の課題。
- NetApp のビジョンとその構成要素およびサービスによって実現されるデータ ファブリック。
- これらのビルディング ブロックを使用して、柔軟なデータ保護ワークフローを構築する方法を説明します。
- 実際の顧客の使用事例に基づいた、いくつかのアーキテクチャの長所と短所。各ユースケースでは、次のコンポーネントが提供されます。
  - 顧客シナリオ
  - 要件と課題
  - ソリューション
  - 解決策の要約

Hadoop データ保護の理由

Hadoop および Spark 環境では、次の懸念事項に対処する必要があります。

- ソフトウェアまたは人為的な障害。 Hadoop データ操作の実行中にソフトウェアを更新する際に人為的なエラーが発生すると、ジョブから予期しない結果が生じる可能性のある障害が発生する可能性があります。このような場合、障害や不合理な結果を避けるためにデータを保護する必要があります。たとえば、交通信号分析アプリケーションのソフトウェア更新が適切に実行されなかった結果、新しい機能がプレーンテキスト形式の交通信号データを適切に分析できないことがあります。ソフトウェアは依然として JSON やその他のテキスト以外のファイル形式を分析するため、リアルタイムの交通制御分析システムはデータ ポイントが欠落した予測結果を生成します。この状況により出力に不具合が生じ、交通信号での事故につながる可能性があります。データ保護では、以前の動作中のアプリケーション バージョンに迅速にロールバックする機能を提供することで、この問題に対処できます。
- *サイズとスケール。*データ ソースの数と量が増え続けているため、分析データのサイズは日々大きくなっています。ソーシャル メディア、モバイル アプリ、データ分析、クラウド コンピューティング プラッ

トフォームは、急速に拡大している現在のビッグ データ市場における主なデータ ソースであるため、正確なデータ操作を保証するためにデータを保護する必要があります。

- **Hadoop** のネイティブ データ保護。 Hadoop にはデータを保護するためのネイティブ コマンドがありますが、このコマンドではバックアップ中にデータの一貫性が確保されません。ディレクトリ レベルのバックアップのみをサポートします。 Hadoop によって作成されたスナップショットは読み取り専用であり、バックアップ データを直接再利用することはできません。

Hadoop および **Spark** の顧客にとってのデータ保護の課題

Hadoop および Spark の顧客にとっての共通の課題は、データ保護中に本番クラスターのパフォーマンスに悪影響を与えることなく、バックアップ時間を短縮し、バックアップの信頼性を高めることです。

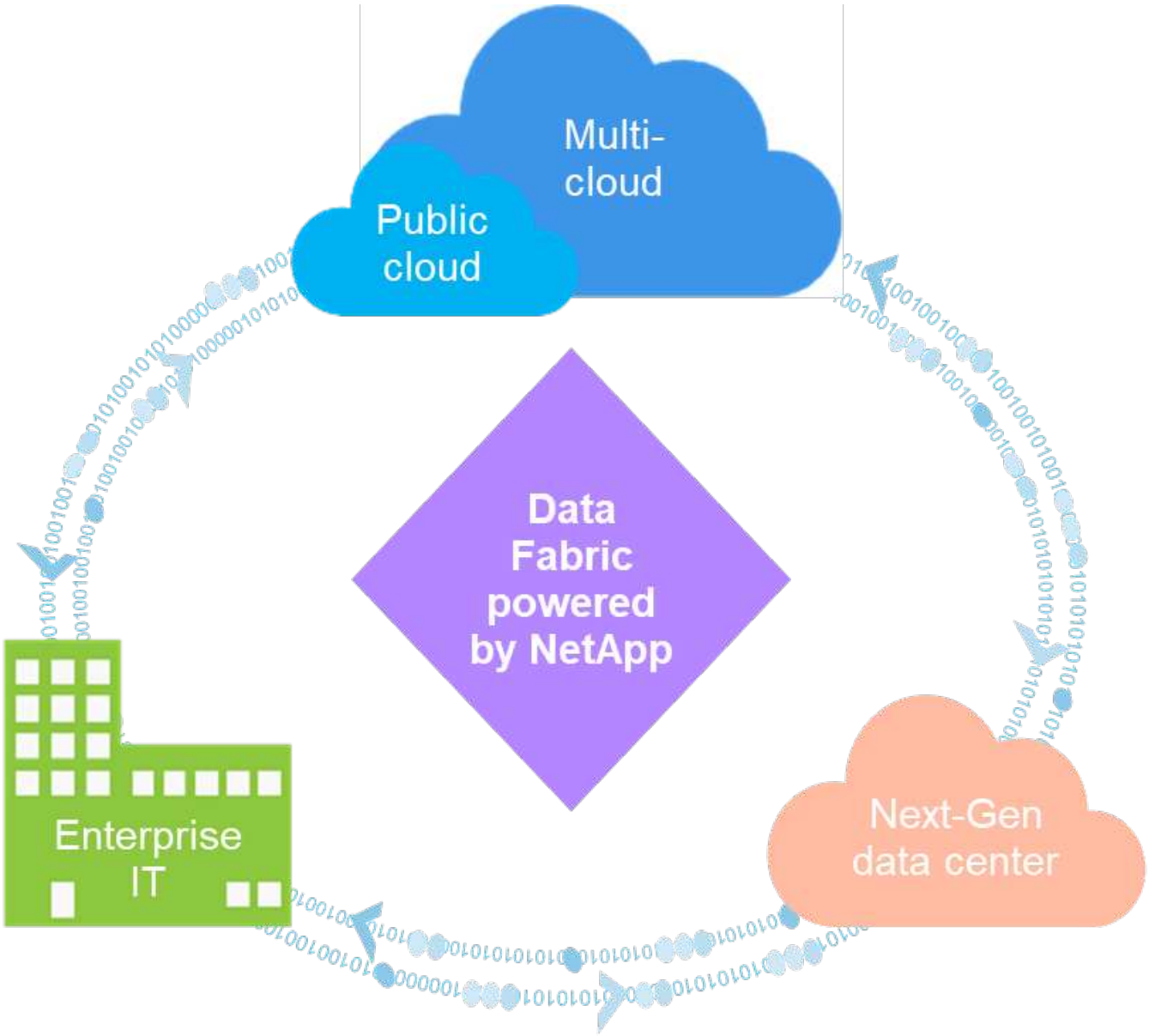
顧客は、最適なビジネス継続性を確保するために、復旧ポイント目標 (RPO) と復旧時間目標 (RTO) のダウンタイムを最小限に抑え、オンプレミスとクラウドベースの災害復旧サイトを制御する必要もあります。この制御は通常、エンタープライズ レベルの管理ツールによって実現されます。

Hadoop および Spark 環境は、データ量が膨大かつ増加しているだけでなく、データの到着速度も増加しているため、複雑になっています。このシナリオでは、ソース データから効率的で最新の DevTest および QA 環境を迅速に作成することが困難になります。 NetApp はこれらの課題を認識しており、このホワイト ペーパーで紹介するソリューションを提供しています。

ビッグデータ アーキテクチャ向けの**NetApp**搭載データ ファブリック

NetAppが提供するデータ ファブリックは、クラウド環境とオンプレミス環境全体のデータ管理を簡素化および統合し、デジタル変革を加速します。

NetApp搭載のデータ ファブリックは、次の図に示すように、データの可視性と分析、データのアクセスと制御、データの保護とセキュリティを実現する、一貫性のある統合データ管理サービスとアプリケーション (構成要素) を提供します。



実証済みのデータファブリックの顧客ユースケース

NetAppを搭載したデータ ファブリックは、顧客に次の 9 つの実証済みユース ケースを提供します。

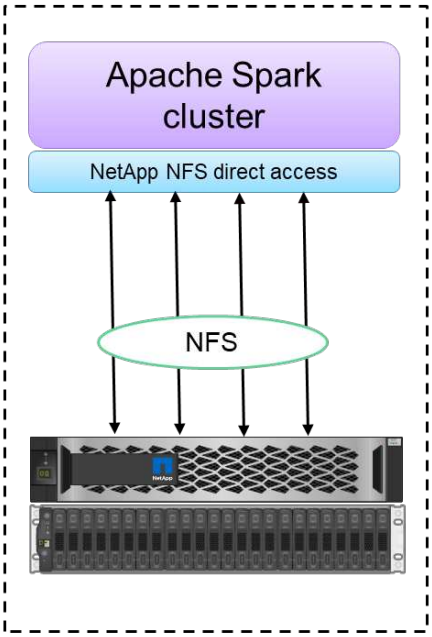
- 分析ワークロードを加速
- DevOps変革を加速
- クラウドホスティングインフラストラクチャの構築
- クラウドデータサービスを統合する
- データの保護とセキュリティ確保
- 非構造化データを最適化する
- データセンターの効率化
- データの洞察と制御を提供
- 簡素化と自動化

このドキュメントでは、9 つのユースケースのうち 2 つ (およびそのソリューション) について説明します。

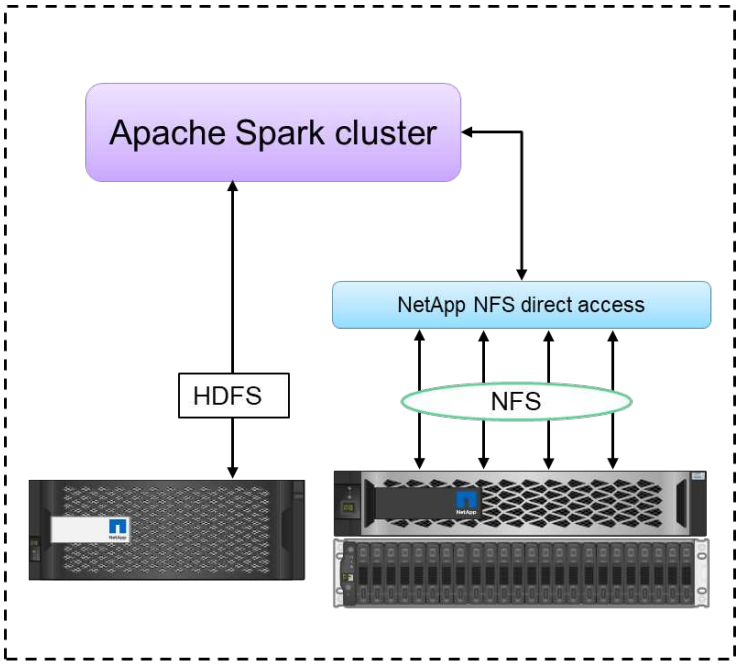
- 分析ワークロードを加速
- データの保護とセキュリティ確保

NetApp NFS ダイレクトアクセス

NetApp NFS を使用すると、データを移動またはコピーすることなく、既存または新規の NFSv3 または NFSv4 データに対してビッグ データ分析ジョブを実行できます。データの複数のコピーを防ぎ、ソースとデータを同期する必要がなくなります。たとえば、金融分野では、ある場所から別の場所へのデータの移動は法的義務を満たす必要があり、これは簡単な作業ではありません。このシナリオでは、 NetApp NFS ダイレクト アクセスが元の場所から財務データを分析します。もう 1 つの重要な利点は、 NetApp NFS ダイレクト アクセスを使用すると、ネイティブ Hadoop コマンドを使用して Hadoop データの保護が簡素化され、NetApp の豊富なデータ管理ポートフォリオを活用したデータ保護ワークフローが実現されることです。



Configuration 1: NFS as primary storage



Configuration 2: HDFS and NFS in single Spark cluster

NetApp NFS ダイレクト アクセスは、Hadoop/Spark クラスターに対して 2 種類の導入オプションを提供します。

- デフォルトでは、Hadoop/Spark クラスターは、データ ストレージとデフォルトのファイル システムとして Hadoop Distributed File System (HDFS) を使用します。 NetApp NFS ダイレクト アクセスでは、デフォルトの HDFS を NFS ストレージに置き換えてデフォルトのファイル システムとして使用できるため、NFS データに対する直接的な分析操作が可能になります。
- 別の導入オプションとして、 NetApp NFS ダイレクト アクセスでは、単一の Hadoop/Spark クラスター内の HDFS とともに NFS を追加ストレージとして構成することがサポートされています。この場合、顧客は NFS エクスポートを通じてデータを共有し、HDFS データと同じクラスターからデータにアクセスできます。

NetApp NFS ダイレクト アクセスを使用する主な利点は次のとおりです。

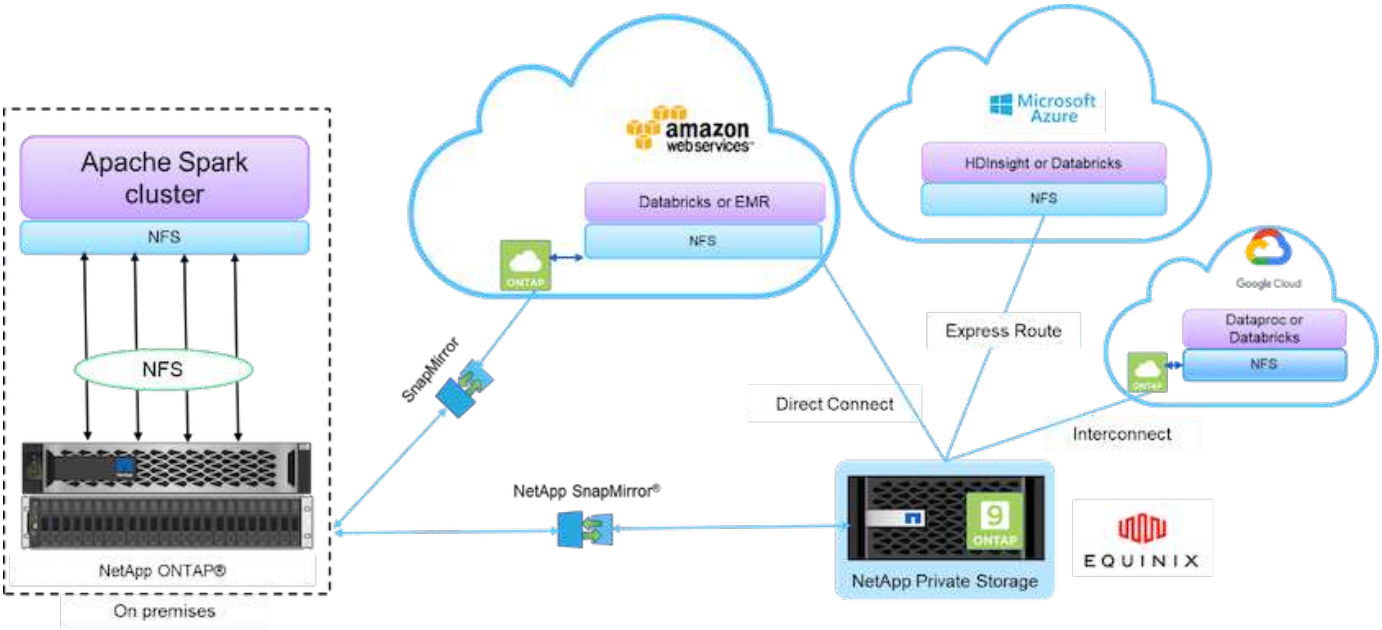
- 現在の場所からデータを分析することで、分析データを HDFS などの Hadoop インフラストラクチャに移

動する、時間とパフォーマンスを消費するタスクを回避します。

- レプリカの数を 3 から 1 に減らします。
- ユーザーはコンピューティングとストレージを切り離して、個別に拡張できます。
- ONTAPの豊富なデータ管理機能を活用して、エンタープライズ データ保護を提供します。
- Hortonworks データ プラットフォームの認定を受けています。
- ハイブリッド データ分析の展開を可能にします。
- 動的マルチスレッド機能を活用してバックアップ時間を短縮します。

ビッグデータの構成要素

NetAppが提供するデータ ファブリックは、次の図に示すように、データ アクセス、制御、保護、セキュリティのためのデータ管理サービスとアプリケーション (ビルディング ブロック) を統合します。



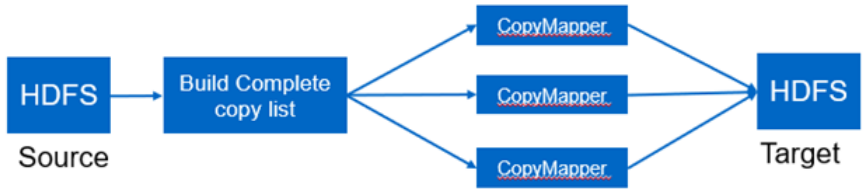
上の図の構成要素は次のとおりです。

- * NetApp NFS ダイレクト アクセス。*追加のソフトウェアやドライバーを必要とせずに、最新の Hadoop および Spark クラスターにNetApp NFS ボリュームへの直接アクセスを提供します。
- * NetApp Cloud Volumes ONTAPとGoogle Cloud NetApp Volumes 。* Amazon Web Services (AWS) または Microsoft Azure クラウド サービスのAzure NetApp Files (ANF) で実行されるONTAPに基づくソフトウェア定義の接続ストレージ。
- * NetApp SnapMirrorテクノロジー*。オンプレミスとONTAP Cloud または NPS インスタンス間のデータ保護機能を提供します。
- *クラウド サービス プロバイダー*これらのプロバイダーには、AWS、Microsoft Azure、Google Cloud 、IBM Cloud が含まれます。
- **PaaS** AWS の Amazon Elastic MapReduce (EMR) や Databricks、Microsoft Azure HDInsight や Azure Databricks などのクラウドベースの分析サービス。

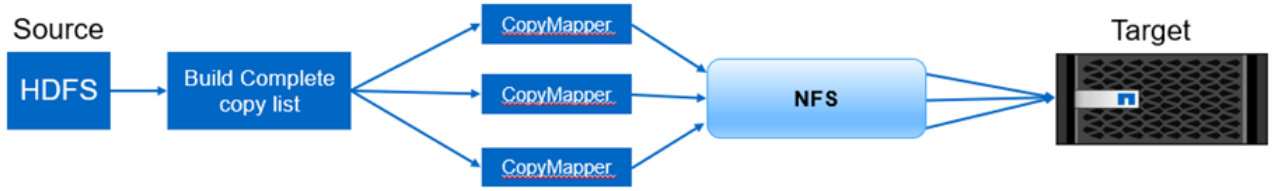
Hadoopデータ保護とNetApp

Hadoop DistCp は、大規模なクラスター間およびクラスター内のコピーに使用されるネイティブ ツールです。下の図に示す Hadoop DistCp の基本プロセスは、MapReduce などの Hadoop ネイティブ ツールを使用して Hadoop データを HDFS ソースから対応するターゲットにコピーする一般的なバックアップ ワークフローです。

NetApp NFS ダイレクト アクセスを使用すると、顧客は Hadoop DistCp ツールのターゲット デスティネーションとして NFS を設定し、MapReduce を介して HDFS ソースから NFS 共有にデータをコピーできます。NetApp NFS ダイレクト アクセスは、DistCp ツールの NFS ドライバーとして機能します。



Hadoop Distcp Basic Process



Hadoop Distcp and NetApp

Hadoop データ保護のユースケースの概要

このセクションでは、このホワイト ペーパーの焦点となるデータ保護のユース ケースの概要を説明します。残りのセクションでは、顧客の問題 (シナリオ)、要件と課題、解決策など、各ユースケースの詳細について説明します。

ユースケース1: Hadoopデータのバックアップ

このユースケースでは、 NetApp NFS ボリュームにより、大手金融機関は長いバックアップ ウィンドウ時間を 24 時間以上から数時間未満に短縮することができました。

ユースケース 2: クラウドからオンプレミスへのバックアップと災害復旧

大手放送会社は、 NetAppを搭載したデータ ファブリックを構成要素として使用することで、オンデマンド、瞬時、Hadoop/Spark クラスターの負荷に基づくなど、さまざまなデータ転送モードに応じてクラウド データをオンプレミスのデータ センターにバックアップするという要件を満たすことができました。

ユースケース3: 既存のHadoopデータでDevTestを有効にする

NetAppソリューションは、オンライン音楽配信会社がさまざまなブランチに複数のスペース効率の高い Hadoop クラスターを迅速に構築し、レポートを作成し、スケジュールされたポリシーを使用して毎日の DevTest タスクを実行するのに役立ちました。

ユースケース4: データ保護とマルチクラウド接続

大手サービス プロバイダーは、 NetAppが提供するデータ ファブリックを使用して、さまざまなクラウド インスタンスから顧客にマルチクラウド分析を提供しました。

ユースケース5: 分析ワークロードの高速化

最大規模の金融サービスおよび投資銀行の 1 社は、 NetAppネットワーク接続ストレージ ソリューションを使用して、I/O 待機時間を短縮し、定量的金融分析プラットフォームを高速化しました。

ユースケース1: Hadoopデータのバックアップ

このシナリオでは、顧客はオンプレミスの大規模な Hadoop リポジトリを所有しており、災害復旧のためにそれをバックアップしたいと考えています。しかし、顧客の現在のバックアップ ソリューションはコストが高く、バックアップ ウィンドウが 24 時間を超える長い時間を要するという問題がありました。

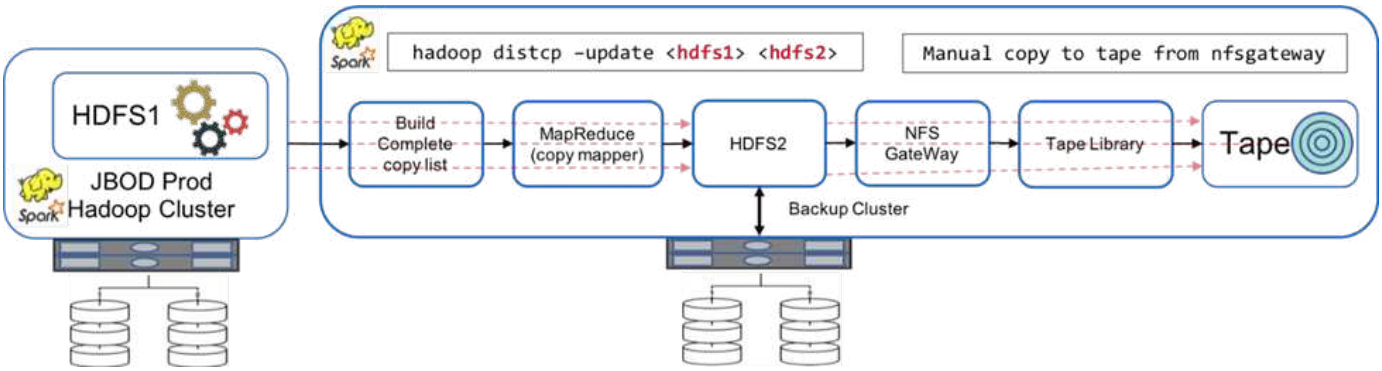
要件と課題

このユースケースの主な要件と課題は次のとおりです。

- ソフトウェアの下位互換性:
  - 提案される代替バックアップ ソリューションは、実稼働 Hadoop クラスターで使用されている現在実行中のソフトウェア バージョンと互換性がある必要があります。
- 約束された SLA を満たすには、提案された代替ソリューションでは非常に低い RPO と RTO を実現する必要があります。
- NetAppバックアップ ソリューションによって作成されたバックアップは、データセンターにローカルに構築された Hadoop クラスターだけでなく、リモート サイトの災害復旧場所で稼働している Hadoop クラスターでも使用できます。
- 提案されるソリューションはコスト効率がよいものでなければなりません。
- 提案されたソリューションは、バックアップ時間中に、現在実行中の本番環境の分析ジョブに対するパフォーマンスの影響を軽減する必要があります。

顧客の既存のバックアップソリューションx

下の図は、元の Hadoop ネイティブ バックアップ ソリューションを示しています。



実稼働データは中間バックアップ クラスターを通じてテープに対して保護されます。

- HDFS1のデータは、以下のコマンドを実行することでHDFS2にコピーされます。 `hadoop distcp -update <hdfs1> <hdfs2>`指示。
- バックアップクラスタはNFSゲートウェイとして機能し、データはLinux経由で手動でテープにコピーされます。 `cp`テープライブラリを介してコマンドを実行します。

オリジナルの Hadoop ネイティブ バックアップ ソリューションの利点は次のとおりです。

- このソリューションは Hadoop ネイティブ コマンドに基づいているため、ユーザーは新しい手順を学習する必要がありません。
- このソリューションは、業界標準のアーキテクチャとハードウェアを活用します。

オリジナルの Hadoop ネイティブ バックアップ ソリューションの欠点は次のとおりです。

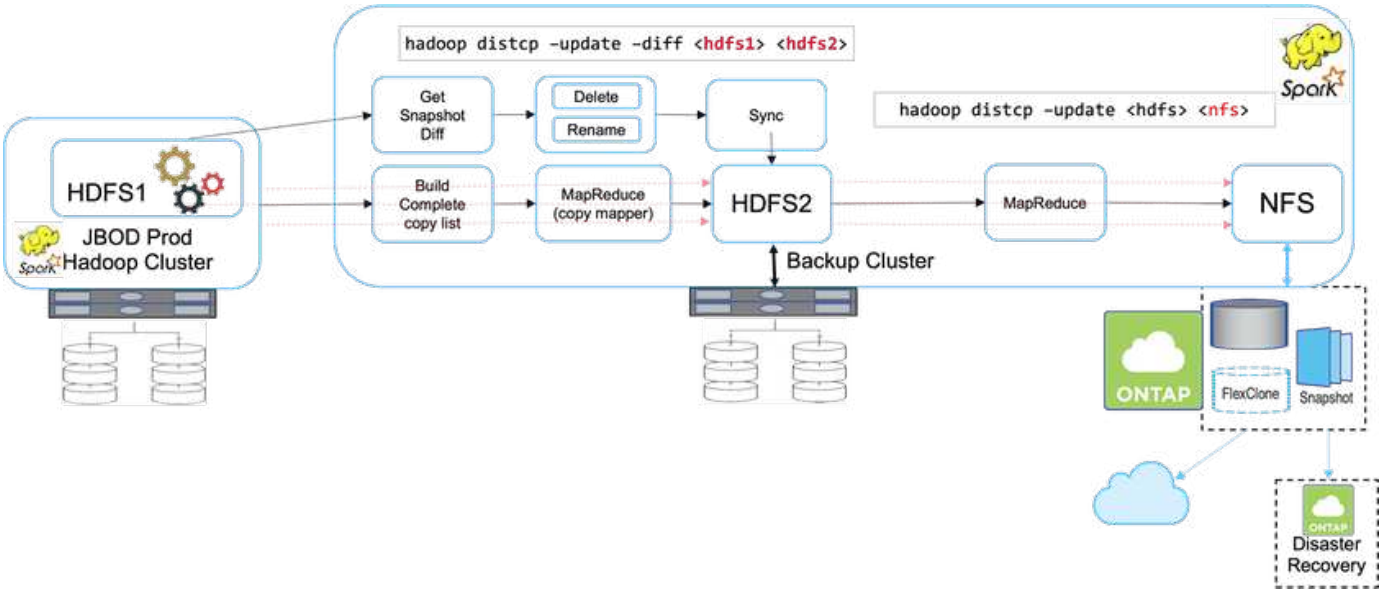
- 長いバックアップウィンドウの時間が 24 時間を超えると、運用データが脆弱になります。
- バックアップ時間中にクラスターのパフォーマンスが大幅に低下します。
- テープへのコピーは手動で行います。
- バックアップ ソリューションは、必要なハードウェアと手動プロセスに必要な人的時間の点で高価です。

バックアップソリューション

これらの課題と要件に基づき、既存のバックアップ システムを考慮して、3 つのバックアップ ソリューションが提案されました。次のサブセクションでは、ソリューション A からソリューション C までの 3 つの異なるバックアップ ソリューションのそれぞれについて説明します。

解決策**A**

ソリューション A では、バックアップ Hadoop クラスターがセカンダリ バックアップをNetApp NFS ストレージ システムに送信するため、下の図に示すように、テープは不要になります。



ソリューション A の詳細なタスクは次のとおりです。

- 実稼働 Hadoop クラスターには、保護が必要な顧客の分析データが HDFS 内にあります。
- HDFS を使用したバックアップ Hadoop クラスターは、データの中間場所として機能します。 JBOD (Just a Bunch of Disks) は、本番環境とバックアップ環境の両方の Hadoop クラスターで HDFS のストレージを提供します。
- Hadoopの本番データは、本番クラスタのHDFSからバックアップクラスタのHDFSまで、以下のコマンドを実行することで保護されます。 `Hadoop distcp –update –diff <hdfs1> <hdfs2>`指示。

> Hadoop スナップショットは、本番環境からバックアップ Hadoop クラスターまでのデータを保護するために使用されます。

- NetApp ONTAPストレージ コントローラは、バックアップ Hadoop クラスターにプロビジョニングされる NFS エクスポート ボリュームを提供します。
- 実行することで `Hadoop distcp`MapReduce と複数のマッパーを活用したコマンドにより、分析データはバックアップ Hadoop クラスターから NFS に保護されます。

  データがNetAppストレージ システム上の NFS に保存された後、必要に応じてNetApp Snapshot、SnapRestore、 FlexCloneテクノロジを使用して Hadoop データのバックアップ、復元、複製が行われます。

> SnapMirrorテクノロジーを使用すると、Hadoop データをクラウドや災害復旧場所まで保護できます。

ソリューション A の利点は次のとおりです。

- Hadoop 実稼働データはバックアップ クラスターから保護されます。
- HDFS データは NFS を通じて保護され、クラウドおよび災害復旧場所への保護が可能になります。
- バックアップ操作をバックアップ クラスターにオフロードすることでパフォーマンスが向上します。
- 手動のテープ操作を排除

- NetAppツールを通じてエンタープライズ管理機能を実現します。
- 既存の環境への変更は最小限で済みます。
- コスト効率の高いソリューションです。

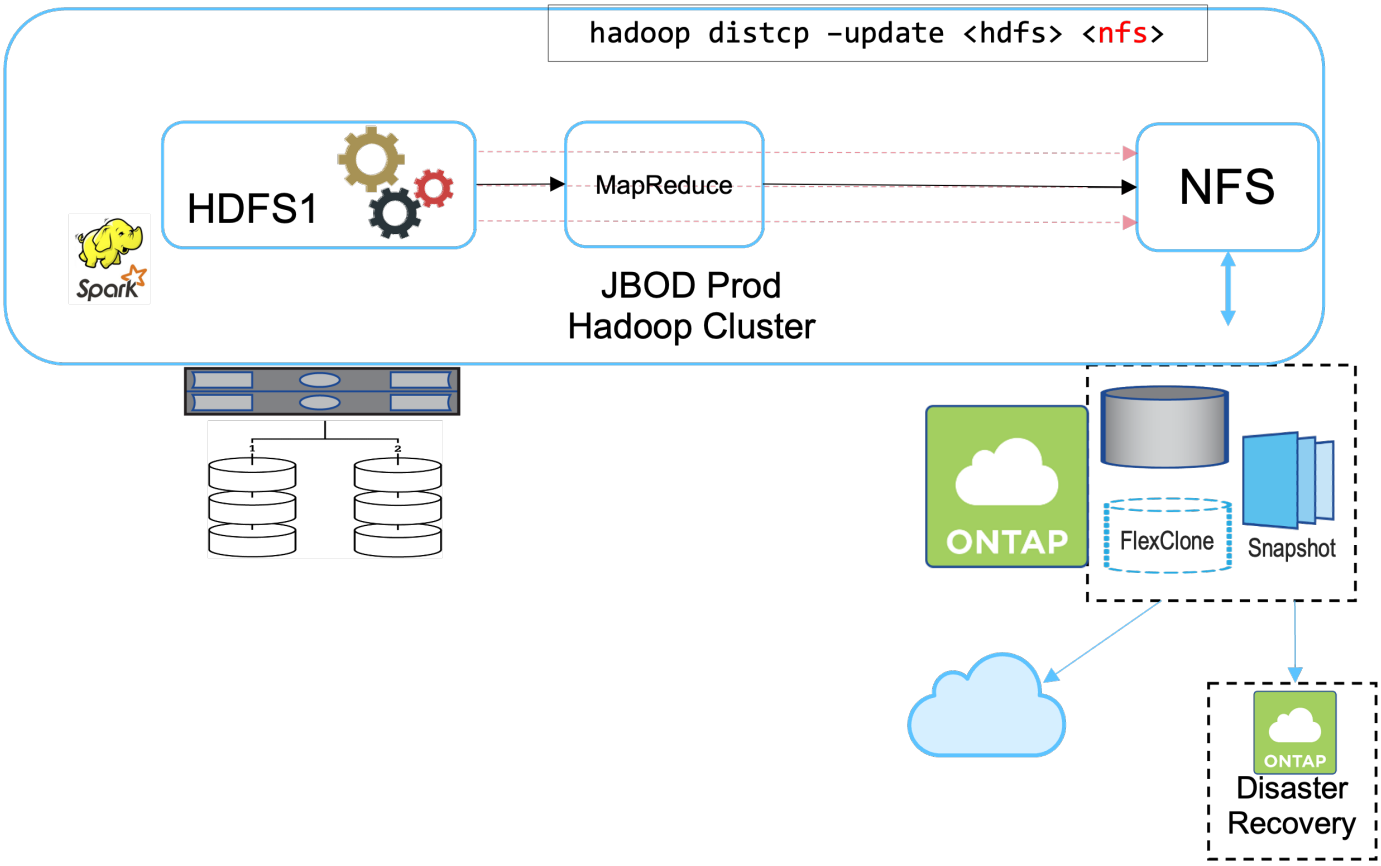
このソリューションの欠点は、パフォーマンスを向上させるためにバックアップ クラスターと追加のマッパーが必要になることです。

顧客は最近、シンプルさ、コスト、全体的なパフォーマンスを理由にソリューション A を導入しました。

このソリューションでは、JBOD の代わりにONTAPの SAN ディスクを使用できます。このオプションは、バックアップ クラスタ ストレージの負荷をONTAPにオフロードしますが、SAN ファブリック スイッチが必要になるという欠点があります。

解決策**B**

ソリューション B は、本番 Hadoop クラスターに NFS ボリュームを追加し、下の図に示すように、バックアップ Hadoop クラスターの必要性を排除します。



ソリューション B の詳細なタスクは次のとおりです。

- NetApp ONTAPストレージ コントローラは、本番 Hadoop クラスターに NFS エクスポートをプロビジョニングします。

  Hadoopネイティブ `hadoop distcp`コマンドは、本番クラスターの HDFS から NFS への Hadoop データを保護します。

- データがNetAppストレージ システム上の NFS に保存された後、必要に応じて Snapshot、SnapRestore、 FlexCloneテクノロジを使用して Hadoop データのバックアップ、復元、複製が行われます。

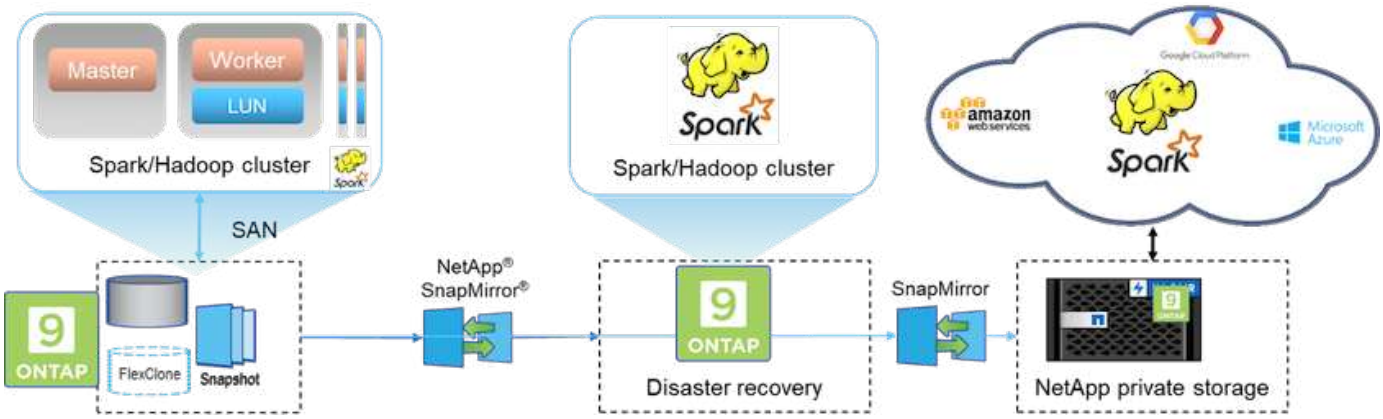
ソリューション B の利点は次のとおりです。

- 運用クラスターはバックアップ ソリューション用にわずかに変更されており、実装が簡素化され、追加のインフラストラクチャ コストが削減されます。
- バックアップ操作用のバックアップ クラスターは必要ありません。
- HDFS 実稼働データは、NFS データへの変換中に保護されます。
- このソリューションは、 NetAppツールを通じてエンタープライズ管理機能を実現します。

このソリューションの欠点は、本番クラスターに実装されるため、本番クラスターに追加の管理者タスクが追加される可能性があることです。

解決策**C**

ソリューション C では、下の図に示すように、 NetApp SAN ボリュームが HDFS ストレージの Hadoop 本番クラスターに直接プロビジョニングされます。



ソリューション C の詳細な手順は次のとおりです。

- NetApp ONTAP SAN ストレージは、HDFS データ ストレージ用に本番 Hadoop クラスターにプロビジョニングされます。
- NetApp Snapshot およびSnapMirrorテクノロジーは、実稼働 Hadoop クラスターから HDFS データをバックアップするために使用されます。
- バックアップはストレージ レイヤーで行われるため、スナップショット コピーのバックアップ プロセス中に Hadoop/Spark クラスターの運用パフォーマンスに影響はありません。

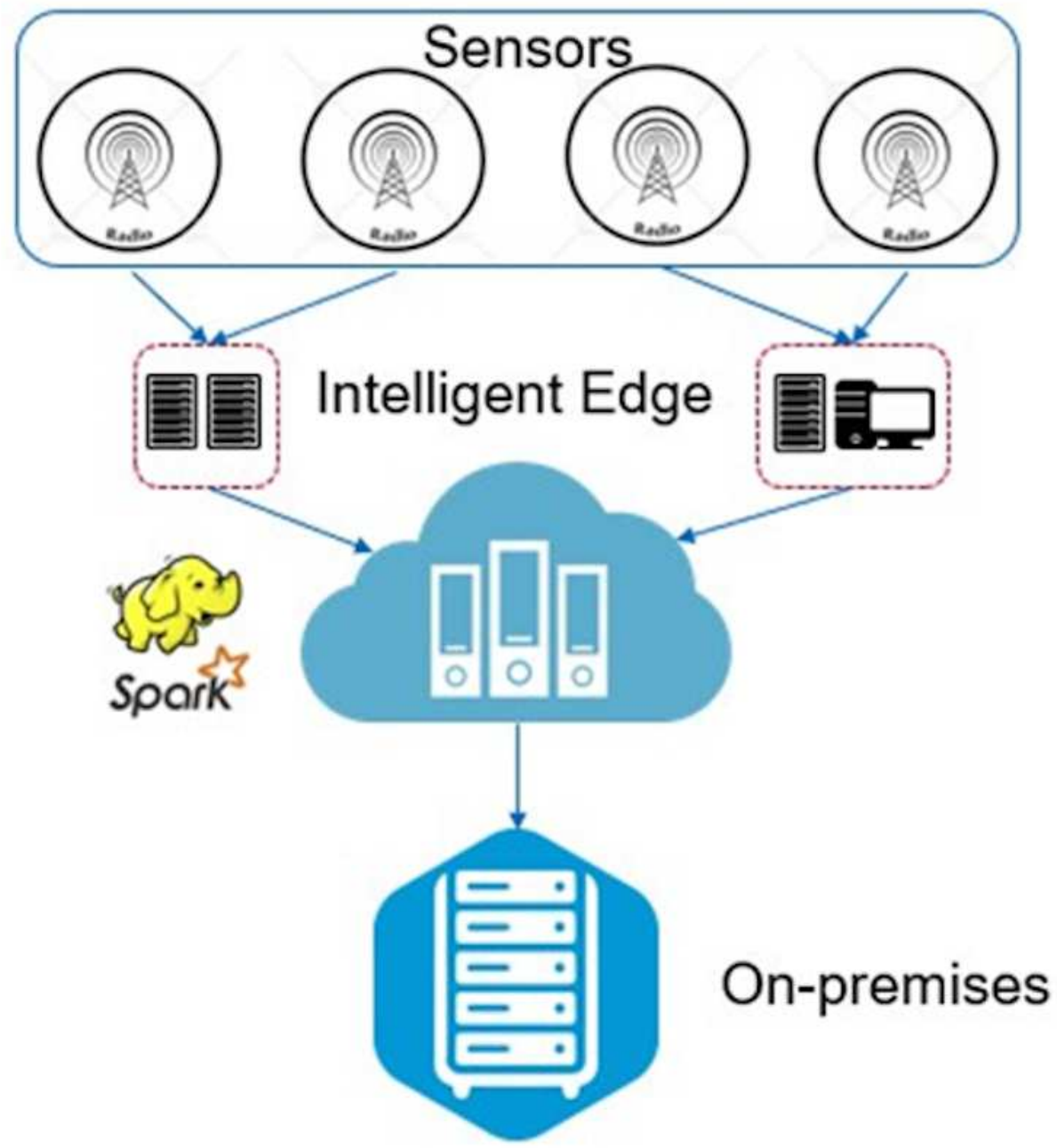
> スナップショット テクノロジーは、データのサイズに関係なく、数秒で完了するバックアップを提供します。

ソリューション C の利点は次のとおりです。

- スナップショット テクノロジーを使用することで、スペース効率の高いバックアップを作成できます。
- NetAppツールを通じてエンタープライズ管理機能を実現します。

ユースケース 2: クラウドからオンプレミスへのバックアップと災害復旧

このユースケースは、下の図に示すように、クラウドベースの分析データをオンプレミスのデータセンターにバックアップする必要がある放送業界の顧客に基づいています。



シナリオ

このシナリオでは、IoT センサー データがクラウドに取り込まれ、AWS 内のオープン ソース Apache Spark クラスターを使用して分析されます。要件は、処理されたデータをクラウドからオンプレミスにバックアップすることです。

要件と課題

このユースケースの主な要件と課題は次のとおりです。

- データ保護を有効にしても、クラウド内の本番環境の Spark/Hadoop クラスターのパフォーマンスには影響しません。
- クラウド センサー データは、効率的かつ安全な方法でオンプレミスに移動され、保護される必要があります。
- オンデマンド、瞬時、クラスター負荷が低い時間など、さまざまな条件下でクラウドからオンプレミスにデータを転送する柔軟性。

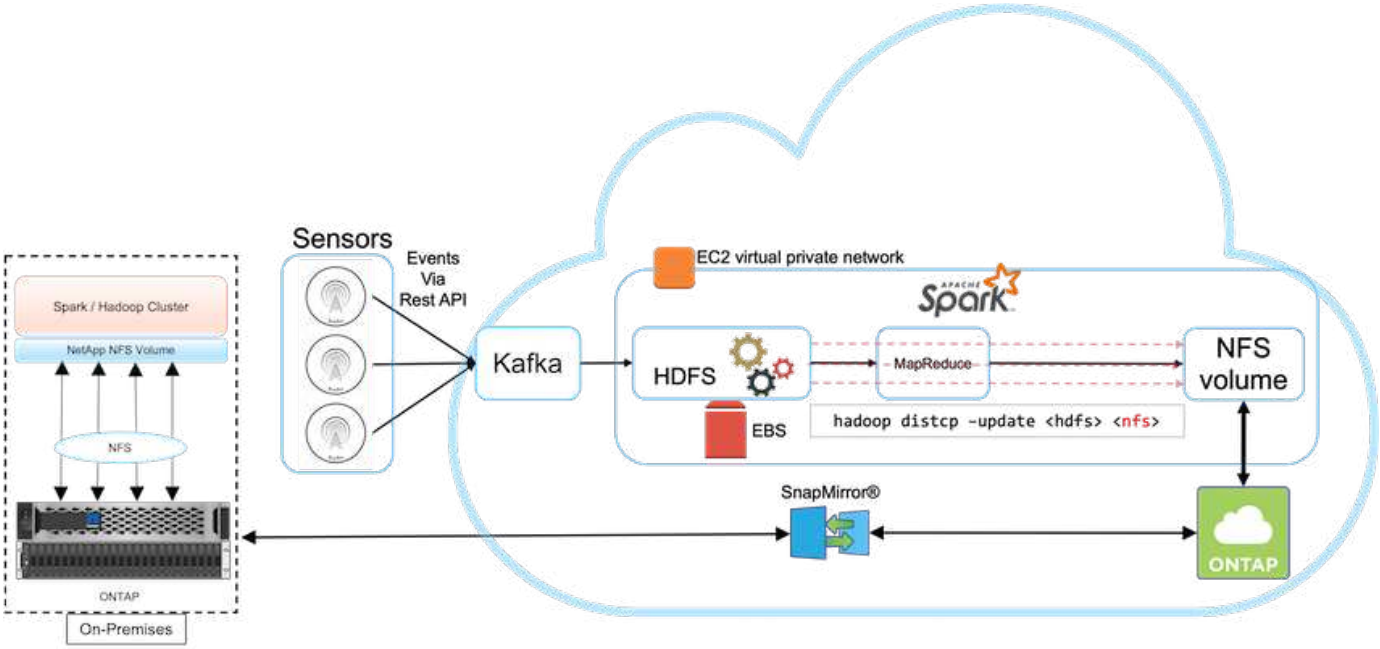
解決策

顧客は、Spark クラスターの HDFS ストレージに AWS Elastic Block Store (EBS) を使用して、Kafka を介してリモート センサーからデータを受信して取り込みます。したがって、HDFS ストレージはバックアップ データのソースとして機能します。

これらの要件を満たすために、 NetApp ONTAP Cloud が AWS に導入され、Spark/Hadoop クラスターのバックアップ ターゲットとして機能する NFS 共有が作成されます。

NFS 共有が作成されたら、HDFS EBS ストレージからONTAP NFS 共有にデータをコピーします。データがONTAP Cloud の NFS に保存された後、 SnapMirrorテクノロジーを使用して、必要に応じてクラウドのデータをオンプレミスのストレージに安全かつ効率的にミラーリングできます。

この画像は、クラウドからオンプレミス ソリューションへのバックアップと災害復旧を示しています。



ユースケース3: 既存のHadoopデータでDevTestを有効にする

このユースケースでは、顧客の要件は、同じデータセンターとリモート ロケーションで、DevTest とレポート作成の目的で大量の分析データを含む既存の Hadoop クラスターに基づいて、新しい Hadoop/Spark クラスターを迅速かつ効率的に構築することで

す。

シナリオ

このシナリオでは、オンプレミスおよび災害復旧場所の大規模な Hadoop データ レイク実装から複数の Spark/Hadoop クラスターが構築されます。

要件と課題

このユースケースの主な要件と課題は次のとおりです。

- DevTest、QA、または同じ本番データへのアクセスが必要なその他の目的のために、複数の Hadoop クラスターを作成します。ここでの課題は、非常に大規模な Hadoop クラスターを、瞬時に、かつ非常にスペース効率の高い方法で複数回複製することです。
- 運用効率を高めるために、Hadoop データを DevTest およびレポート チームと同期します。
- 同じ資格情報を使用して、本番環境と新しいクラスター全体で Hadoop データを分散します。
- スケジュールされたポリシーを使用して、本番クラスターに影響を与えずに QA クラスターを効率的に作成します。

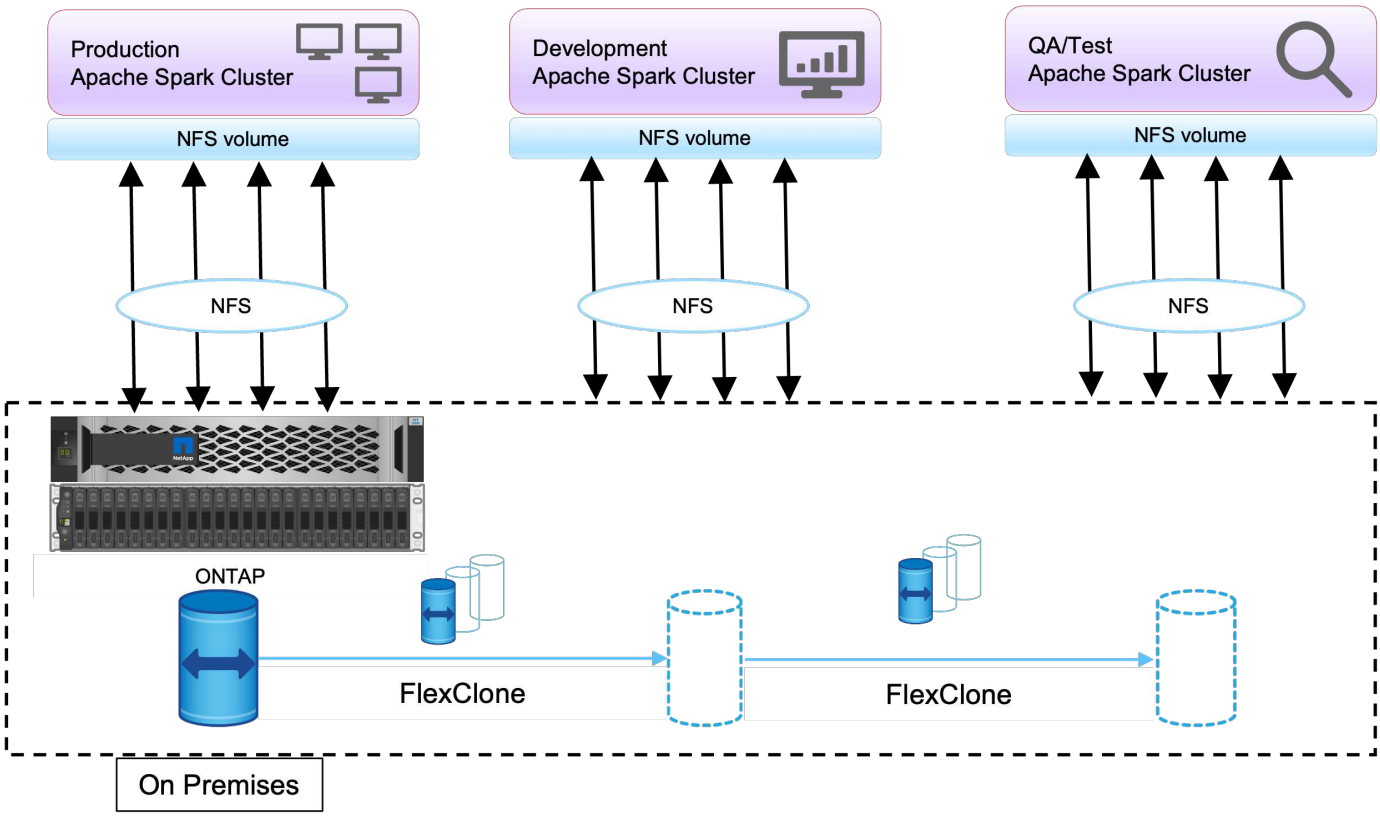
解決策

FlexCloneテクノロジーは、前述の要件に応えるために使用されます。 FlexCloneテクノロジーは、スナップショット コピーの読み取り/書き込みコピーです。親スナップショット コピー データからデータを読み取り、新しいブロックまたは変更されたブロックに対してのみ追加のスペースを消費します。高速かつスペース効率に優れています。

まず、 NetApp整合性グループを使用して、既存のクラスターのスナップショット コピーを作成しました。

NetApp System Manager またはストレージ管理プロンプト内のスナップショット コピー。整合性グループのスナップショット コピーは、アプリケーション整合性グループのスナップショット コピーであり、FlexCloneボリュームは整合性グループのスナップショット コピーに基づいて作成されます。 FlexCloneボリュームは親ボリュームの NFS エクスポート ポリシーを継承することに注意してください。スナップショット コピーが作成された後、次の図に示すように、DevTest とレポート作成の目的で新しい Hadoop クラスターをインストールする必要があります。新しい Hadoop クラスターからクローンされた NFS ボリュームは NFS データにアクセスします。

この画像は、DevTest の Hadoop クラスターを示しています。



ユースケース4: データ保護とマルチクラウド接続

このユースケースは、顧客のビッグ データ分析データにマルチクラウド接続を提供する役割を担うクラウド サービス パートナーに関係します。

シナリオ

このシナリオでは、さまざまなソースから AWS に受信された IoT データが NPS の中央の場所に保存されます。 NPS ストレージは、AWS および Azure にある Spark/Hadoop クラスターに接続され、複数のクラウドで実行されるビッグ データ分析アプリケーションが同じデータにアクセスできるようになります。

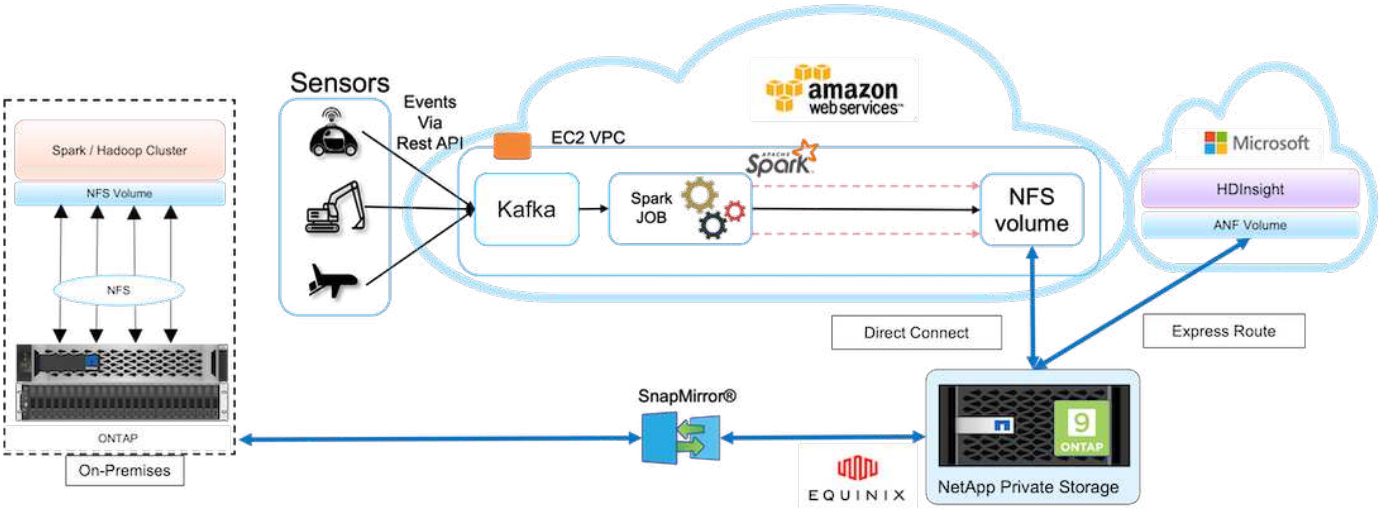
要件と課題

このユースケースの主な要件と課題は次のとおりです。

- 顧客は複数のクラウドを使用して同じデータに対して分析ジョブを実行したいと考えています。
- データは、さまざまなセンサーやハブを介して、オンプレミスやクラウドなどのさまざまなソースから受信する必要があります。
- ソリューションは効率的かつ費用対効果の高いものでなければなりません。
- 主な課題は、オンプレミスとさまざまなクラウド間でハイブリッド分析サービスを提供する、コスト効率が高く効率的なソリューションを構築することです。

解決策

この画像は、データ保護とマルチクラウド接続ソリューションを示しています。



上の図に示すように、センサーからのデータは Kafka を介してストリーミングされ、AWS Spark クラスターに取り込まれます。データは、Equinix データセンター内のクラウド プロバイダーの外部にある NPS にある NFS 共有に保存されます。 NetApp NPS は、それぞれ Direct Connect 接続と Express Route 接続を介して Amazon AWS と Microsoft Azure に接続されているため、顧客は Amazon と AWS の両方の分析クラスターから NFS データにアクセスできます。このアプローチは、複数のハイパースケーラーにわたるクラウド分析の問題を解決します。

その結果、オンプレミスと NPS ストレージの両方でONTAPソフトウェアが実行されるため、 SnapMirror はNPS データをオンプレミス クラスターにミラーリングし、オンプレミスと複数のクラウドにわたるハイブリッド クラウド分析を提供できます。

最高のパフォーマンスを得るために、 NetAppでは通常、複数のネットワーク インターフェイスと直接接続/高速ルートを使用してクラウド インスタンスからデータにアクセスすることをお勧めします。

ユースケース5: 分析ワークロードの高速化

このシナリオでは、大手金融サービスおよび投資銀行の分析プラットフォームがNetApp NFS ストレージ ソリューションを使用して最新化され、資産管理および定量ビジネス ユニットの投資リスクとデリバティブの分析が大幅に改善されました。

シナリオ

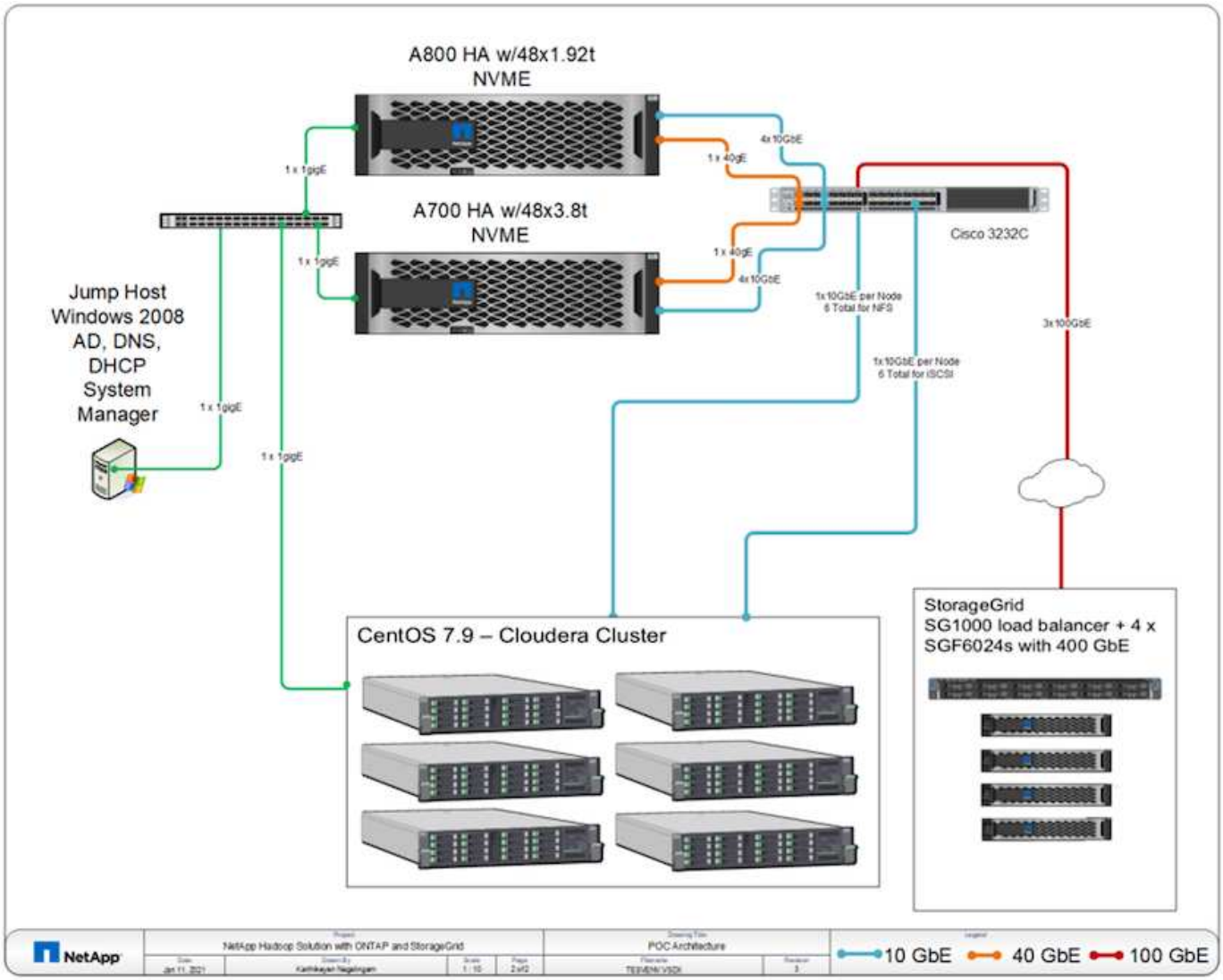
顧客の既存の環境では、分析プラットフォームに使用されている Hadoop インフラストラクチャは、Hadoop サーバーの内部ストレージを活用していました。 JBOD 環境の独自仕様のため、組織内の多くの内部顧客は、リアルタイム データの定期的なサンプルに依存するシミュレーションであるモンテ カルロ定量モデルを活用できませんでした。市場動向の不確実性の影響を理解する能力が最適ではなかったため、定量資産運用事業部門にとって不利な状況となっていました。

要件と課題

銀行の定量ビジネス部門は、正確でタイムリーな予測を達成するための効率的な予測方法を求めていました。これを実現するために、チームは、投資モデルを効率的にシミュレートし、潜在的な利益を測定し、リスクを分析するために、インフラストラクチャを最新化し、既存の I/O 待機時間を短縮し、Hadoop や Spark などの分析アプリケーションのパフォーマンスを向上させる必要があることを認識しました。

解決策

顧客は既存の Spark ソリューションに JBOD を導入していました。その後、 NetApp ONTAP、 NetApp StorageGRID、および MinIO Gateway to NFS を活用して、潜在的な利益とリスクを評価する投資モデルのシミュレーションと分析を実行する銀行の定量財務グループの I/O 待機時間を短縮しました。この画像は、NetAppストレージを使用した Spark ソリューションを示しています。



上の図に示すように、 AFF A800、A700 システム、およびStorageGRIDが導入され、データ分析操作用の Spark、YARN、および Hive メタデータ サービスを備えた 6 ノードの Hadoop クラスターで NFS および S3 プロトコルを介して parquet ファイルにアクセスしました。

顧客の古い環境における直接接続ストレージ (DAS) ソリューションには、コンピューティングとストレージを個別に拡張できないという欠点がありました。同銀行の金融分析事業部門は、Spark 向けNetApp ONTAPソリューションを導入することで、ストレージをコンピューティングから切り離し、必要に応じてインフラストラクチャ リソースをより効率的にシームレスに提供できるようになりました。

ONTAP をNFS と併用することで、コンピューティング サーバーの CPU が Spark SQL ジョブにほぼ完全に利用され、I/O 待機時間が 70% 近く短縮され、Spark ワークロードのコンピューティング能力とパフォーマンスが向上しました。その後、CPU 使用率の向上により、顧客は GPUDirect などの GPU を活用してプラットフォームをさらに近代化することも可能になりました。さらに、 StorageGRID はSpark ワークロードに低コストのストレージ オプションを提供し、MinIO Gateway は S3 プロトコルを介して NFS データへの安全なア

クセスを提供します。クラウド内のデータについては、 NetApp はCloud Volumes ONTAP、 Azure NetApp Files、 Google Cloud NetApp Volumes を推奨しています。

まとめ

このセクションでは、さまざまな Hadoop データ保護要件を満たすためにNetAppが提供するユースケースとソリューションの概要を示します。 NetAppが提供するデータ ファブリックを使用すると、お客様は次のことが可能になります。

- NetApp の豊富なデータ管理機能と Hadoop ネイティブ ワークフローとの統合を活用することで、適切なデータ保護ソリューションを柔軟に選択できます。
- Hadoop クラスターのバックアップ ウィンドウ時間を約 70% 削減します。
- Hadoop クラスターのバックアップによって生じるパフォーマンスへの影響を排除します。
- マルチクラウド データ保護と、さまざまなクラウド プロバイダーから単一の分析データ ソースへのデータ アクセスを同時に提供します。
- FlexCloneテクノロジーを使用して、高速かつスペース効率の高い Hadoop クラスターのコピーを作成します。

詳細情報の入手方法

このドキュメントに記載されている情報の詳細は、以下のドキュメントやウェブサイトをご参照ください。

- NetAppビッグデータ分析ソリューション

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

- NetAppストレージを使用した Apache Spark ワークロード

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

- Apache Spark向けNetAppストレージソリューション

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

- NetAppによって実現されるデータファブリック上の Apache Hadoop

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

謝辞

- ポール・バーランド、 NetApp 、ANZビクトリア地区営業担当
- NetAppビジネス開発マネージャー、Hoseb Dermanilian
- Lee Dorrier、 NetApp MPSG ディレクター
- David Thiessen、システム エンジニア、ANZ Victoria District SE、 NetApp

バージョン履歴

version	日付	ドキュメントのバージョン履歴
バージョン1.0	2018年1月	初版リリース
バージョン2.0	2021年10月	ユースケース#5を更新: 分析ワークロードの高速化
バージョン3.0	2023年11月	NIPAM の詳細を削除しました

著作権に関する情報

Copyright © 2025 NetApp, Inc. All Rights Reserved. Printed in the U.S.このドキュメントは著作権によって保護されています。著作権所有者の書面による事前承諾がある場合を除き、画像媒体、電子媒体、および写真複写、記録媒体、テープ媒体、電子検索システムへの組み込みを含む機械媒体など、いかなる形式および方法による複製も禁止します。

ネットアップの著作物から派生したソフトウェアは、次に示す使用許諾条項および免責条項の対象となります。

このソフトウェアは、ネットアップによって「現状のまま」提供されています。ネットアップは明示的な保証、または商品性および特定目的に対する適合性の暗示的保証を含み、かつこれに限定されないいかなる暗示的な保証も行いません。ネットアップは、代替品または代替サービスの調達、使用不能、データ損失、利益損失、業務中断を含み、かつこれに限定されない、このソフトウェアの使用により生じたすべての直接的損害、間接的損害、偶発的損害、特別損害、懲罰的損害、必然的損害の発生に対して、損失の発生の可能性が通知されていたとしても、その発生理由、根拠とする責任論、契約の有無、厳格責任、不法行為（過失またはそうでない場合を含む）にかかわらず、一切の責任を負いません。

ネットアップは、ここに記載されているすべての製品に対する変更を随時、予告なく行う権利を保有します。ネットアップによる明示的な書面による合意がある場合を除き、ここに記載されている製品の使用により生じる責任および義務に対して、ネットアップは責任を負いません。この製品の使用または購入は、ネットアップの特許権、商標権、または他の知的所有権に基づくライセンスの供与とはみなされません。

このマニュアルに記載されている製品は、1つ以上の米国特許、その他の国の特許、および出願中の特許によって保護されている場合があります。

権利の制限について：政府による使用、複製、開示は、DFARS 252.227-7013（2014年2月）およびFAR 5252.227-19（2007年12月）のRights in Technical Data -Noncommercial Items（技術データ - 非商用品目に関する諸権利）条項の(b)(3)項、に規定された制限が適用されます。

本書に含まれるデータは商用製品および / または商用サービス（FAR 2.101の定義に基づく）に関係し、データの所有権はNetApp, Inc.にあります。本契約に基づき提供されるすべてのネットアップの技術データおよびコンピュータ ソフトウェアは、商用目的であり、私費のみで開発されたものです。米国政府は本データに対し、非独占的かつ移転およびサブライセンス不可で、全世界を対象とする取り消し不能の制限付き使用権を有し、本データの提供の根拠となった米国政府契約に関連し、当該契約の裏付けとする場合にのみ本データを使用できます。前述の場合を除き、NetApp, Inc.の書面による許可を事前に得ることなく、本データを使用、開示、転載、改変するほか、上演または展示することはできません。国防総省にかかる米国政府のデータ使用権については、DFARS 252.227-7015(b)項（2014年2月）で定められた権利のみが認められます。

商標に関する情報

NetApp、NetAppのロゴ、http://www.netapp.com/TMに記載されているマークは、NetApp, Inc.の商標です。その他の会社名と製品名は、それを所有する各社の商標である場合があります。