ネットアップのハイブリッドクラウドデータソリューション - Spark と Hadoop はお客様のユースケースに基づいています

NetApp Solutions

NetApp
April 10, 2024

This PDF was generated from https://docs.netapp.com/ja-jp/netapp-solutions/data-analytics/hdcs-sh-solution-overview.html on April 10, 2024. Always check docs.netapp.com for the latest.

目次

ネットアップのハイブリッドクラウドデータソリューション - Spark と Hadoop はお客様のユースケースに基づいています . . . 1
TR-4657 ：ネットアップのハイブリッドクラウドデータソリューション - Spark と Hadoop はお客様のユースケースに基づいています . . . 1
ネットアップのデータファブリックを基盤としたビッグデータアーキテクチャ . . . 2
Hadoop データ保護機能とネットアップ . . . 6
Hadoop のデータ保護のユースケースの概要 . . . 6
使用事例 1 ： Hadoop データのバックアップ . . . 7
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ディザスタリカバリ . . . 12
ユースケース 3 ：既存の Hadoop データに対して DevTest を有効化 . . . 13
ユースケース 4 ：データ保護とマルチクラウド接続 . . . 15
ユースケース 5 ：分析ワークロードを高速化 . . . 16
まとめ . . . 18

ネットアップのハイブリッドクラウドデータソリューション - Spark と Hadoop はお客様のユースケースに基づいています

TR-4657 ：ネットアップのハイブリッドクラウドデータソリューション - Spark と Hadoop はお客様のユースケースに基づいています

ネットアップ、 Karthikeyan Nagalingam と Sathish Thyagarajan

本ドキュメントでは、 NetApp AFF および FAS ストレージシステム、 NetApp Cloud Volumes ONTAP 、ネットアップ接続ストレージ、 Spark および Hadoop 向けの NetApp FlexClone テクノロジ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。これらの解決策アーキテクチャを使用することで、お客様の環境に適したデータ保護解決策を選択できます。ネットアップは、お客様とのやり取りと、お客様の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 環境や Hadoop 環境、お客様の課題に対応するデータ保護が必要な理由
- ネットアップのビジョンと、そのビルディングブロックとサービスを基盤とす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
- 実際のお客様のユースケースに基づく、複数のアーキテクチャの長所と短所各ユースケースには、次のコンポーネントがあります。
  - お客様のシナリオ
  - 要件と課題
  - 解決策
  - ソリューションの概要

Hadoop のデータ保護を選ぶ理由

Hadoop 環境と Spark 環境では、次の点に注意する必要があります。

- * ソフトウェアや人為的なエラー。 * Hadoop データの処理中にソフトウェアを更新したときに人的エラーが発生すると、業務によって原因が予期せぬ結果を招く可能性がある動作不良になることがあります。このような場合は、障害や妥当でない結果が生じないように、データを保護する必要があります。たとえば、ソフトウェアアップデートの実行が不十分でトラフィック信号分析アプリケーションが実行されたため、トラフィック信号データをプレーンテキスト形式で適切に分析できない新機能があります。ソフトウェアは JSON やその他の非テキストファイル形式を分析して、リアルタイムトラフィック制御分析システムを生成し、データポイントが不足している予測結果を生成します。このような状況では、原因が出力不良の可能性があり、交通信号で事故につながるおそれがあります。データ保護機能を使用すると、以前の作業中のアプリケーションバージョンにすばやくロールバックできるため、この問題に対応できます。
- * サイズと拡張性。 * 分析データのサイズは日々増え続けています。その理由は、データソースとボリュームの数が増え続けることにあります。現在のビッグデータ市場では、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の各プラットフォームがデータの主要なソースとなっており、データは急速に増加しています。そのため、データを保護して、正確なデータ運用を確保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、このコマンドはバックアップ中のデータの整合性を提供しません。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 Snapshot 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 や **Spark** のお客様にとって、データ保護の課題が発生しています

Hadoop と Spark のお客様にとってよくある課題は、データ保護の際に本番クラスタ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向上させることです。

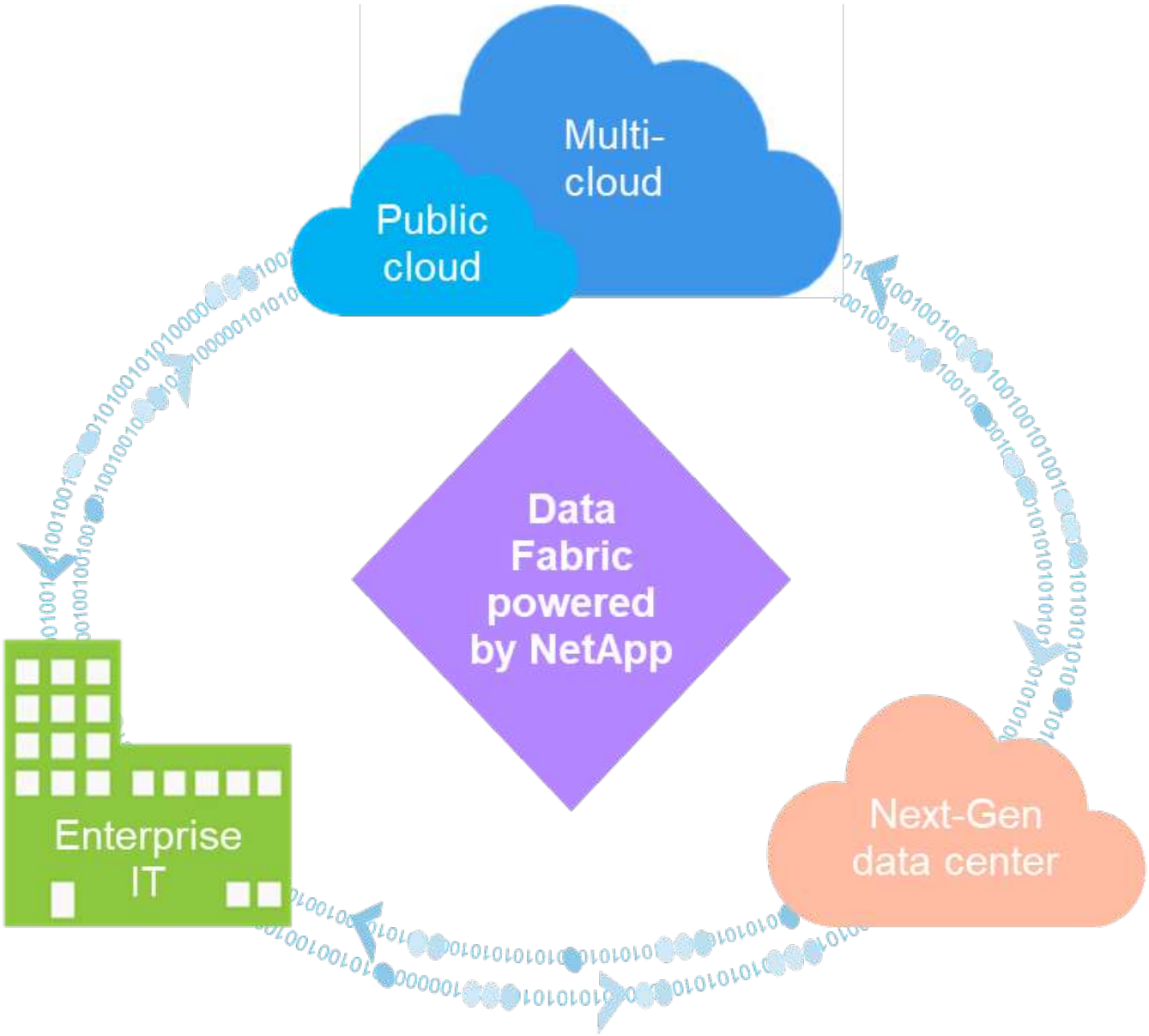
また、 RPO （目標復旧時点）と RTO （目標復旧時間）のダウンタイムを最小限に抑え、オンプレミスとクラウドベースのディザスタリカバリサイトを制御して、ビジネス継続性を最適化する必要もあります。この制御は、通常、エンタープライズレベルの管理ツールを使用して行われます。

データ量が膨大で増え続けているだけでなく、データの到着率も増加しているため、 Hadoop 環境と Spark 環境は複雑化しています。このようなシナリオでは、ソースデータから効率的で最新の DevTest 環境と QA 環境を迅速に構築することは困難です。ネットアップはこれらの課題を認識し、本ホワイトペーパーで紹介しているソリューションを提供しています。

ネットアップのデータファブリックを基盤としたビッグデータアーキテクチャ

ネットアップの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でデータ管理を簡易化、統合することで、デジタル変革を加速します。

ネットアップのデータファブリックは、一貫した統合的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）を提供し、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します。以下の図を参照してください。



実績のあるデータファブリックのユースケース

ネットアップのデータファブリックは、以下の 9 つのユースケースをお客様に提供します。

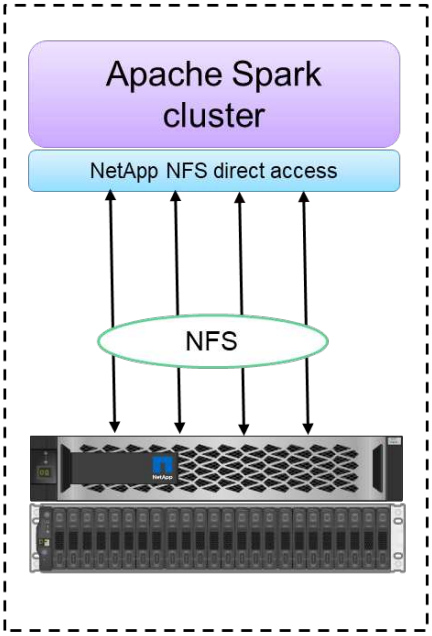
- 分析ワークロードを高速化
- DevOps 変革を加速
- クラウドとホスティングのインフラ構築
- クラウドデータサービスを統合
- データの保護とセキュリティ
- 非構造化データを最適化
- データセンターの効率化
- データの分析と管理を実現
- 簡易化と自動化

このドキュメントでは、 9 つのユースケースのうち 2 つを取り上げ、それぞれのソリューションを紹介します。

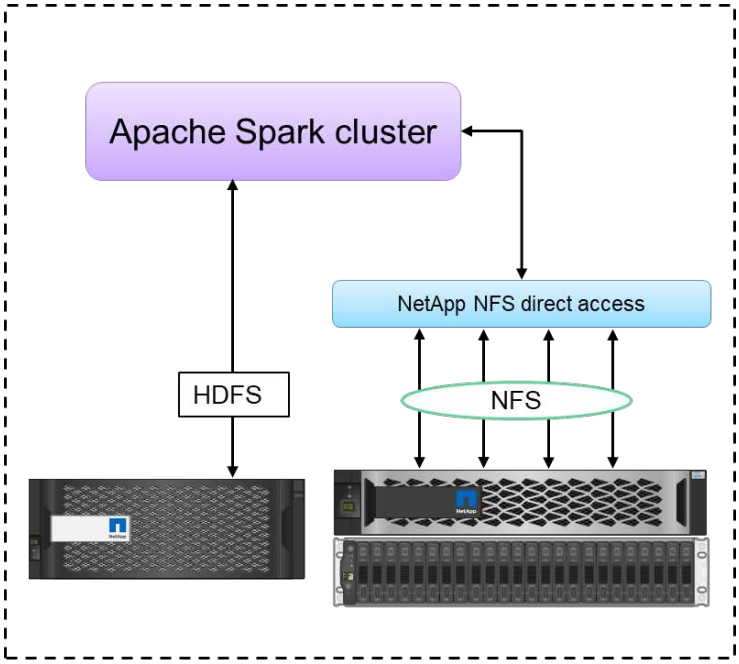
- 分析ワークロードを高速化
- データの保護とセキュリティ

NetApp NFS への直接アクセス

NetApp NFSを使用すると、既存または新規のNFSv3 / NFSv4データに対して、データを移動またはコピーすることなく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が作成されるため、ソースとデータを同期する必要がありません。たとえば、金融機関では、ある場所から別の場所へデータを移動する際に法的義務を果たす必要がありますが、これは容易な作業ではありません。このシナリオでは、 NetApp NFS の直接アクセスによって、元の場所から財務データが分析されます。もう 1 つの主な利点は、 NetApp NFS 直接アクセスを使用すると、ネイティブの Hadoop コマンドを使用して Hadoop データを保護しやすくなることと、ネットアップの充実したデータ管理ポートフォリオを活用してデータ保護ワークフローを実現できることです。



Configuration 1: NFS as primary storage



Configuration 2: HDFS and NFS in single Spark cluster

NetApp NFS 直接アクセスでは、 Hadoop クラスタと Spark クラスタに対して次の 2 種類の導入オプションを提供しています。

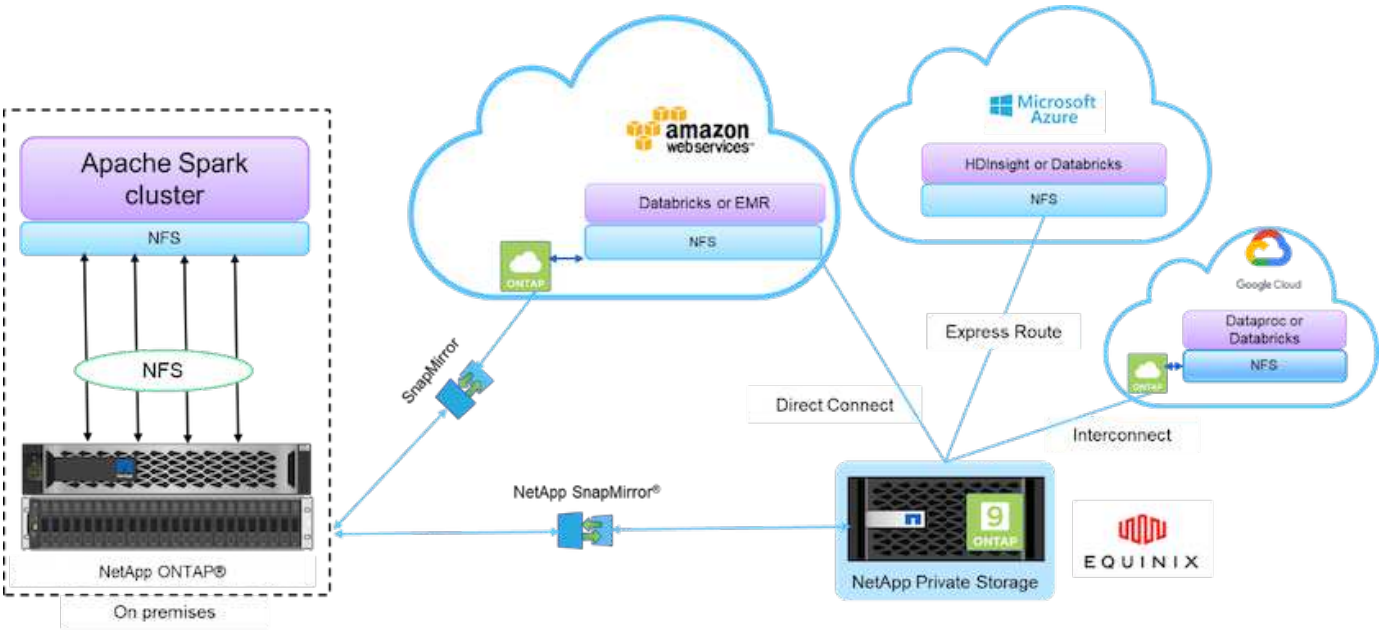
- デフォルトでは、 Hadoop / Spark クラスタは、データストレージとデフォルトのファイルシステムに Hadoop Distributed File System （ HDFS ； Hadoop 分散ファイルシステム）を使用しています。NetApp NFS の直接アクセスを使用すると、デフォルトの HDFS をデフォルトのファイルシステムとして NFS ストレージに置き換えることができるため、 NFS データに対する直接分析処理が可能になります。
- もう 1 つの導入オプションでは、 NetApp NFS 直接アクセスを使用して、 1 つの Hadoop / Spark クラスタ内に HDFS を追加のストレージとして構成することもできます。この場合、 NFS エクスポートを介してデータを共有し、 HDFS データと同じクラスタからデータにアクセスできます。

NetApp NFS 直接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ため、分析データを HDFS などの Hadoop インフラに移動する時間とパフォーマンスのかかるタスクは発生しません。
- レプリカの数を 3 つから 1 つに減らします。
- ユーザ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を保護します。
- Hortonworks データプラットフォームで認定されています。
- ハイブリッドデータ分析環境を実現
- 動的な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。

ビッグデータ向けのビルディングブロック

ネットアップのデータファブリックは、以下の図に示すように、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）を統合しています。



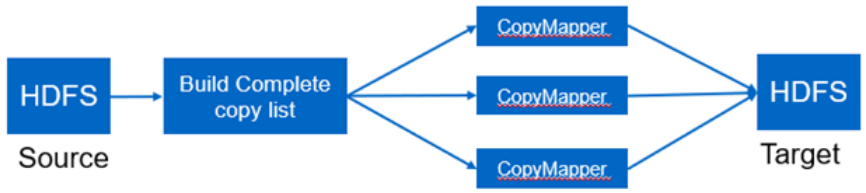
上の図の構成要素は次のとおりです。

- * NetApp NFS 直接アクセス。 * 最新の Hadoop クラスタと Spark クラスタを、ソフトウェアやドライバの追加の必要なしに NetApp NFS ボリュームに直接アクセスできます。
- * ネットアップの Cloud Volumes ONTAP とクラウドボリュームサービス。 * ソフトウェア定義型の接続ストレージ。 Amazon Web Services （ AWS ）で実行されている ONTAP または Microsoft Azure クラウドサービスで実行されている Azure NetApp Files （ ANF ）に基づいています。
- * NetApp SnapMirror テクノロジ * 。オンプレミスと ONTAP クラウドインスタンスまたは NPS インスタンス間で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 、 Azure Databricks などのクラウドベースの分析サービスを利用できます。

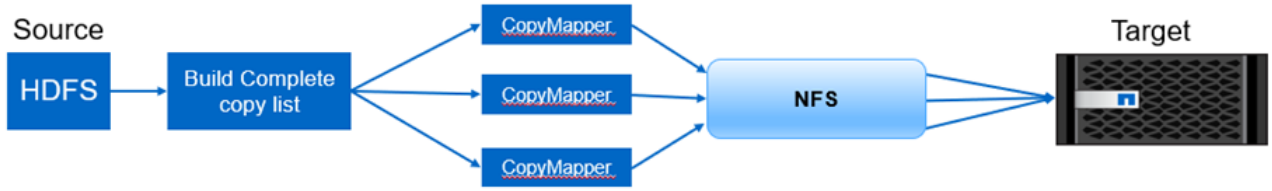
Hadoop データ保護機能とネットアップ

Hadoop ディストリビュータは、大規模なクラスタ間コピーとクラスタ内コピーに使用されるネイティブツールです。次の図に示す Hadoop ディストリビュータの基本的なプロセスは、 MapReduce などの Hadoop ネイティブツールを使用した一般的なバックアップワークフローで、 HDFS ソースから対応するターゲットに Hadoop データをコピーします。

NetApp NFS の直接アクセスを使用すると、 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、 HDFS ソースから MapReduce 経由で NFS 共有にデータをコピーできます。NetApp NFS への直接アクセスは、 DistCp ツールの NFS ドライバとして機能します。



Hadoop Distcp Basic Process



Hadoop Distcp and NetApp

Hadoop のデータ保護のユースケースの概要

このセクションでは、データ保護のユースケースの概要を概要で説明します。これが、本ドキュメントで重要となるのです。以降のセクションでは、それぞれのユースケースについて、お客様の問題（シナリオ）、要件と課題、ソリューションなどの詳細を説明します。

使用事例 1 ： Hadoop データのバックアップ

このユースケースでは、NetApp NFSボリュームを使用することで、大規模な金融機関では、長いバックアップウィンドウ時間を24時間以上からわずか数時間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ディザスタリカバリ

大規模な放送会社では、ネットアップのデータファブリックをビルディングブロックとして使用することで、オンデマンド、瞬時、データ転送などの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。 または、 Hadoop / Spark のクラスタの負荷に基づいて計算されました。

ユースケース 3 ：既存の Hadoop データに対して DevTest を有効化

ネットアップのソリューションは、オンラインの音楽配信企業が、スペース効率に優れた複数の Hadoop クラスタをさまざまなブランチオフィスに迅速に構築し、レポートを作成したり、定期的なポリシーを使用して日々の DevTest タスクを実行したりできるよう支援しました

ユースケース 4 ：データ保護とマルチクラウド接続

ある大手サービスプロバイダは、ネットアップの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お客様にマルチクラウド分析を提供していました。

ユースケース 5 ：分析ワークロードを高速化

最大規模の金融サービスおよび投資銀行の 1 つは、ネットアップのネットワーク接続型ストレージ解決策を使用して、 I/O 待ち時間を短縮し、定量的な金融分析プラットフォームを高速化しました。

使用事例 1 ： Hadoop データのバックアップ

シナリオ（ Scenario ）

このシナリオでは、大規模なオンプレミスの Hadoop リポジトリがあり、ディザスタリカバリのためにバックアップを作成したいと考えています。しかし、お客様の現在のバックアップ解決策はコストが高く、 24 時間以上のバックアップウィンドウに悩まされています。

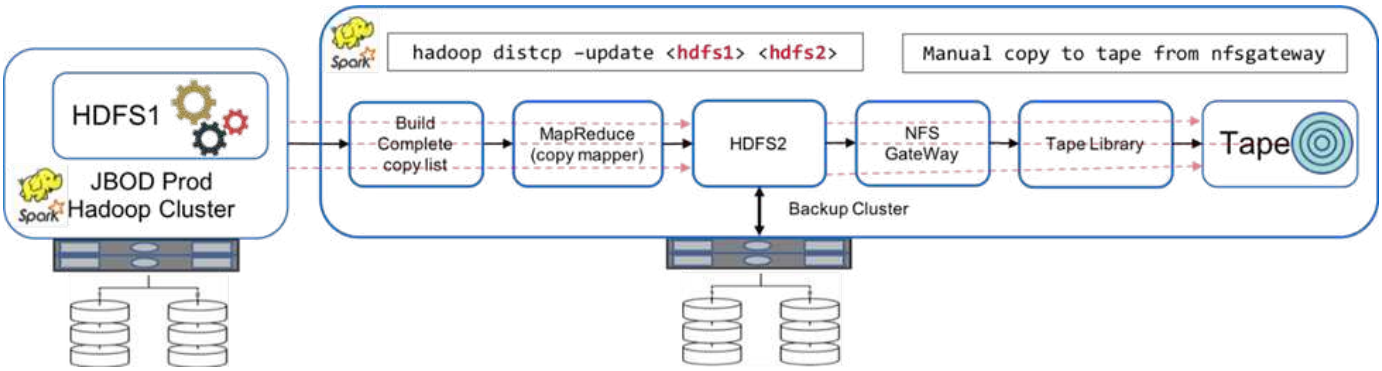
要件と課題

このユースケースの主な要件と課題は次のとおりです。

- ソフトウェアの下位互換性：
  - 提案する代替バックアップ解決策は、本番用 Hadoop クラスタで現在実行しているソフトウェアバージョンと互換性があることが必要です。
- コミットされた SLA を満たすためには、代替の解決策で非常に低い RPO と RTO を達成することを推奨します。
- ネットアップのバックアップ解決策で作成したバックアップは、データセンターのローカルに構築された Hadoop クラスタ、およびリモートサイトのディザスタリカバリロケーションで実行されている Hadoop クラスタで使用できます。
- 提案する解決策は対費用効果が高いものでなければなりません。
- 提案する解決策は、バックアップ処理中に実行中の本番環境の分析ジョブに与えるパフォーマンスへの影響を軽減する必要があります。

お客様の既存のバックアップソリューションx

次の図は、元の Hadoop ネイティブのバックアップ解決策を示しています。



本番環境のデータは、中間バックアップクラスタを通じてテープに保護されます。

- hadoop distcp-update <hdfs1 ><hdfs2>` コマンドを実行することにより、 HDFS1 データが HDFS2 にコピーされます。
- バックアップ・クラスタは NFS ゲートウェイとして機能し ' テープ・ライブラリを介して Linux'cp' コマンドを使用してデータを手動でテープにコピーします

元の Hadoop ネイティブバックアップ解決策には次のようなメリットがあります。

- 解決策は Hadoop ネイティブのコマンドをベースにしているため、新しい手順を習得する必要がなくなります。
- 解決策は、業界標準のアーキテクチャとハードウェアを活用しています。

元の Hadoop ネイティブバックアップ解決策には、次のような欠点があります。

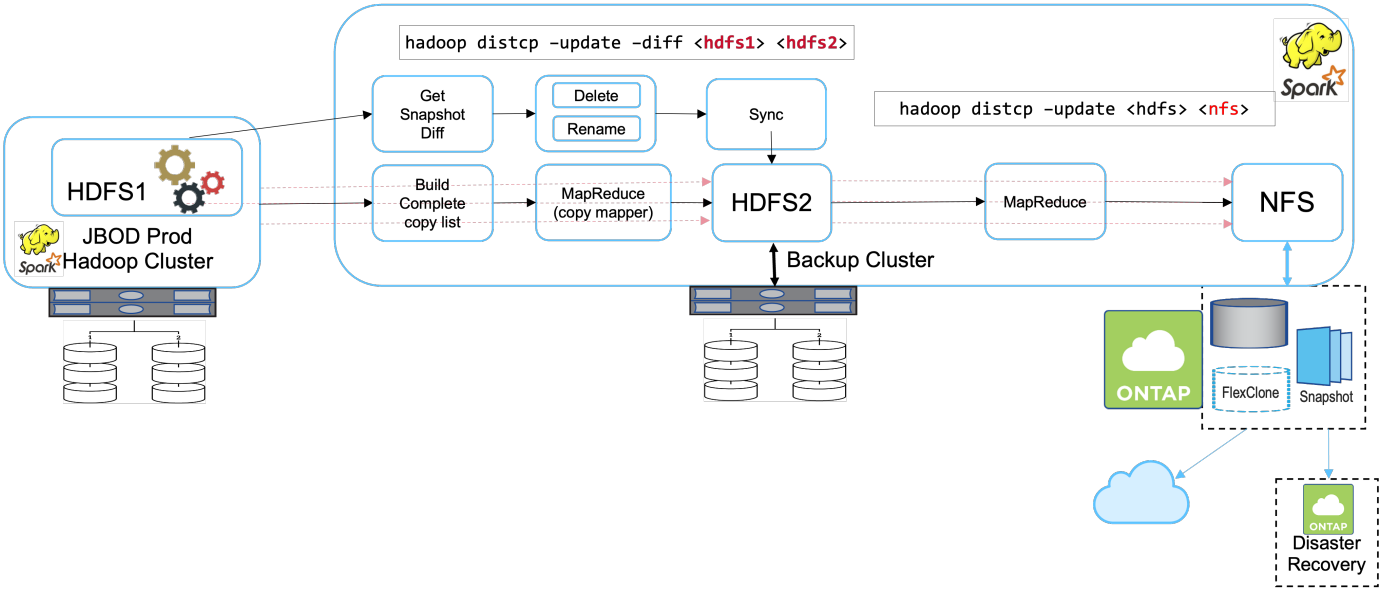
- バックアップ時間が長いと 24 時間を超えるため、本番環境のデータが脆弱になります。
- バックアップ時間中にクラスタのパフォーマンスが大幅に低下します。
- テープへのコピーは手動で行います。
- バックアップ解決策は、必要なハードウェアと、手動プロセスに必要な人的時間の点でコストが高くなります。

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いて、既存のバックアップシステムを検討し、 3 つのバックアップソリューションを提案しました。以降のサブセクションでは、解決策 A ～ 解決策 C というラベルの付いた 3 種類のバックアップソリューションについて説明します

解決策 **A** の略

解決策Aでは、次の図に示すように、バックアップHadoopクラスタからNetApp NFSストレージシステムにセカンダリバックアップが送信されるため、テープは必要ありません。



解決策 A の詳細なタスクは次のとおりです。

- 本番環境の Hadoop クラスタには、保護が必要な HDFS 内のお客様の分析データがあります。
- HDFS を使用するバックアップ Hadoop クラスタは、データの中間的な場所として機能します。Just a Bunch of Disks （ JBOD ）は、本番環境の Hadoop クラスタとバックアップの Hadoop クラスタの両方で HDFS にストレージを提供する。
- 「 hadoop distcp – update – diff<hdfs1 ><hdfs2>` コマンド」を実行することで、 Hadoop 本番クラスタの HDFS からバックアップクラスタの HDFS へと Hadoop 本番データを保護します。

> Hadoop スナップショットは、本番環境からバックアップ Hadoop クラスタへ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。
- を実行します `Hadoop distcp` コマンドMapReduceと複数のマッパーを活用して、分析データをバックアップHadoopクラスタからNFSに保護します。

  ネットアップストレージシステム上の NFS にデータを格納したあと、必要に応じて、ネットアップの Snapshot 、 SnapRestore 、および FlexClone テクノロジを使用して Hadoop データをバックアップ、リストア、および複製します。

> Hadoop データは、 SnapMirror テクノロジを使用してクラウドやディザスタリカバリロケーションに保護できます。

解決策 A には、次のような利点があります。

- Hadoop の本番データはバックアップクラスタから保護されます。
- HDFS データは NFS を通じて保護されるため、クラウドやディザスタリカバリの場所を保護できます。
- バックアップ処理をバックアップクラスタにオフロードすることでパフォーマンスを向上します。
- 手動でのテープ操作が不要になります

- ネットアップのツールを使用してエンタープライズ管理機能を利用できます。
- 既存の環境への変更は最小限で済みます。
- 対費用効果の高い解決策です。

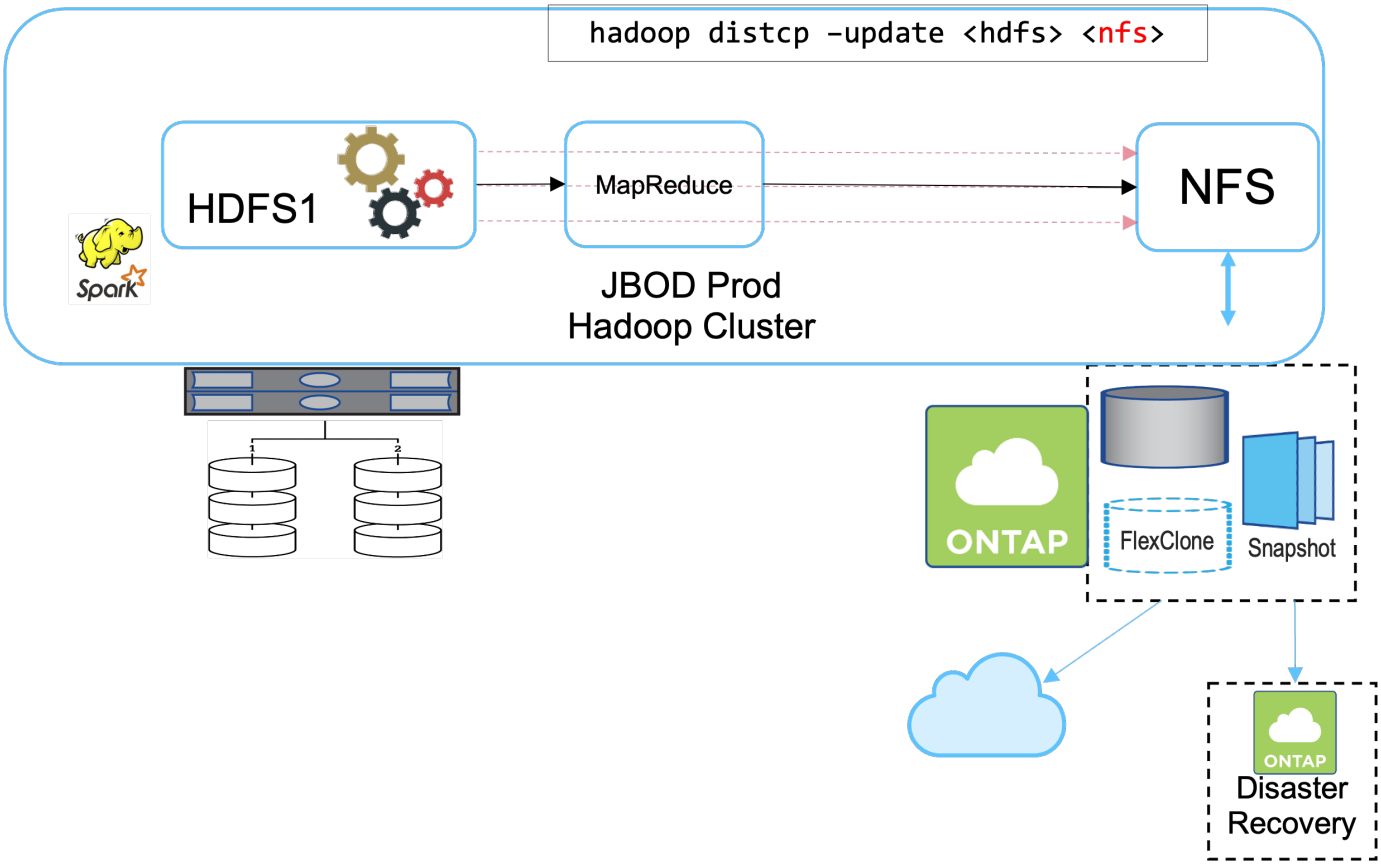
この解決策の欠点は、パフォーマンスを向上させるためにバックアップクラスタと追加のマッパが必要であることです。

お客様は最近、解決策 A を導入しました。シンプルさ、コスト、全体的なパフォーマンスが理由です。

この解決策では、 JBOD の代わりに ONTAP の SAN ディスクを使用できます。このオプションを選択すると、バックアップクラスタのストレージ負荷が ONTAP にオフロードされますが、問題となるのは SAN ファブリックスイッチが必要な場合です。

解決策 **B**

解決策BはNFSボリュームを本番用Hadoopクラスタに追加するため、次の図に示すように、バックアップHadoopクラスタは必要ありません。



解決策 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、本番用 Hadoop クラスタに対して 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、Hadoopデータを本番用クラスタのHDFSからNFSに保護します。

- ネットアップストレージシステム上の NFS にデータを格納したあと、 Snapshot 、 SnapRestore 、および FlexClone テクノロジを使用して、必要に応じて Hadoop データをバックアップ、リストア、および複製します。

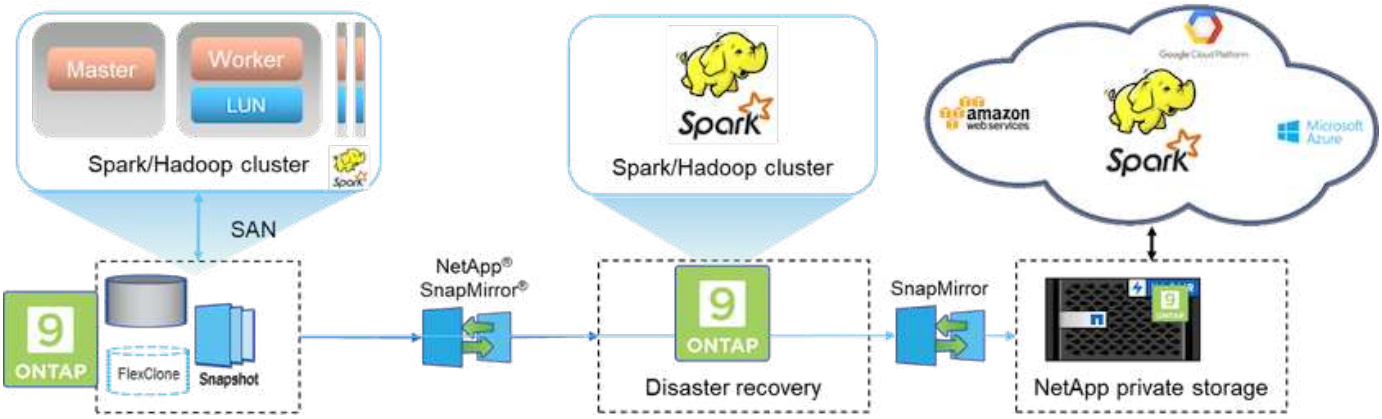
解決策 B には次のような利点があります。

- 本番環境クラスタは、バックアップ解決策用に若干変更されるため、実装が簡単になり、インフラコストを削減できます。
- バックアップ処理のためのバックアップクラスタは必要ありません。
- HDFS の本番環境のデータは、 NFS データへの変換によって保護されます。
- 解決策では、ネットアップのツールを使用してエンタープライズ管理機能を実行できます。

この解決策の欠点は、本番クラスタに実装されており、本番クラスタに管理者タスクを追加できることです。

解決策 **C**

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



解決策 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 データストレージの本番用 Hadoop クラスタでプロビジョニングされます。
- NetApp Snapshot テクノロジと SnapMirror テクノロジを使用して、本番用 Hadoop クラスタの HDFS データをバックアップします。
- バックアップはストレージレイヤにあるため、 Snapshot コピーのバックアッププロセス中は Hadoop / Spark クラスタの本番環境でパフォーマンスが低下することはありません。

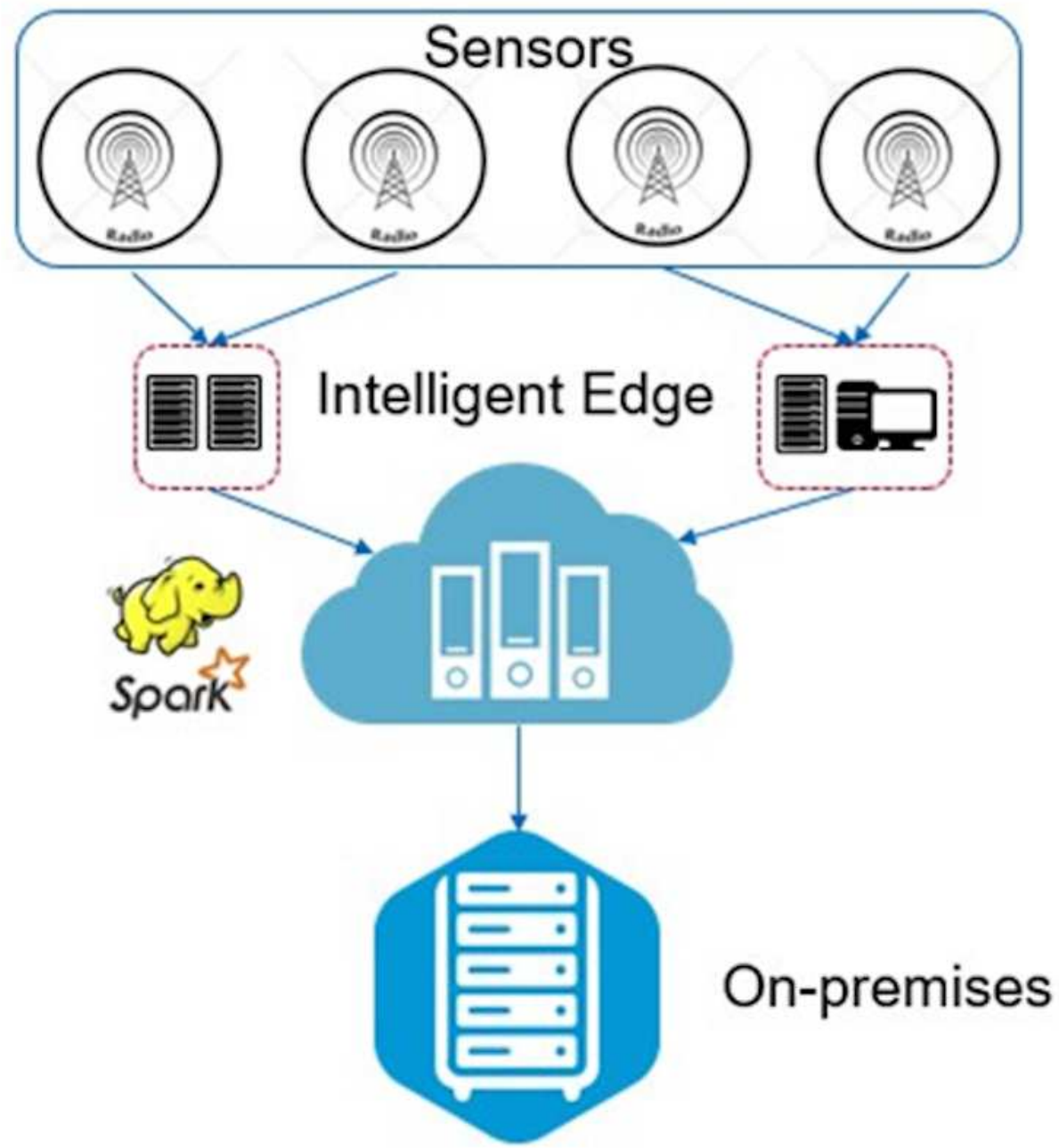
> Snapshot テクノロジを使用すると、データのサイズに関係なく数秒で完了するバックアップを作成できます。

解決策 C には次のような利点があります。

- スペース効率に優れたバックアップは、 Snapshot テクノロジを使用して作成できます。
- ネットアップのツールを使用してエンタープライズ管理機能を利用できます。

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ディザスタリカバリ

このユースケースは、放送局のお客様が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場合を基準にしています。以下の図を参照してください。



シナリオ（ **Scenario** ）

このシナリオでは、 IoT センサーのデータがクラウドに取り込まれ、 AWS 内のオープンソースの Apache Spark クラスタを使用して分析されます。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が要件です。

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。

- データ保護を有効原因にしても、本番環境の Spark / Hadoop クラスタのパフォーマンスへの影響は一切ありません。
- 効率的かつ安全な方法で、クラウドセンサーデータをオンプレミスに移動して保護する必要があります。
- オンデマンド、瞬時、クラスタの低負荷時など、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柔軟に転送できます。

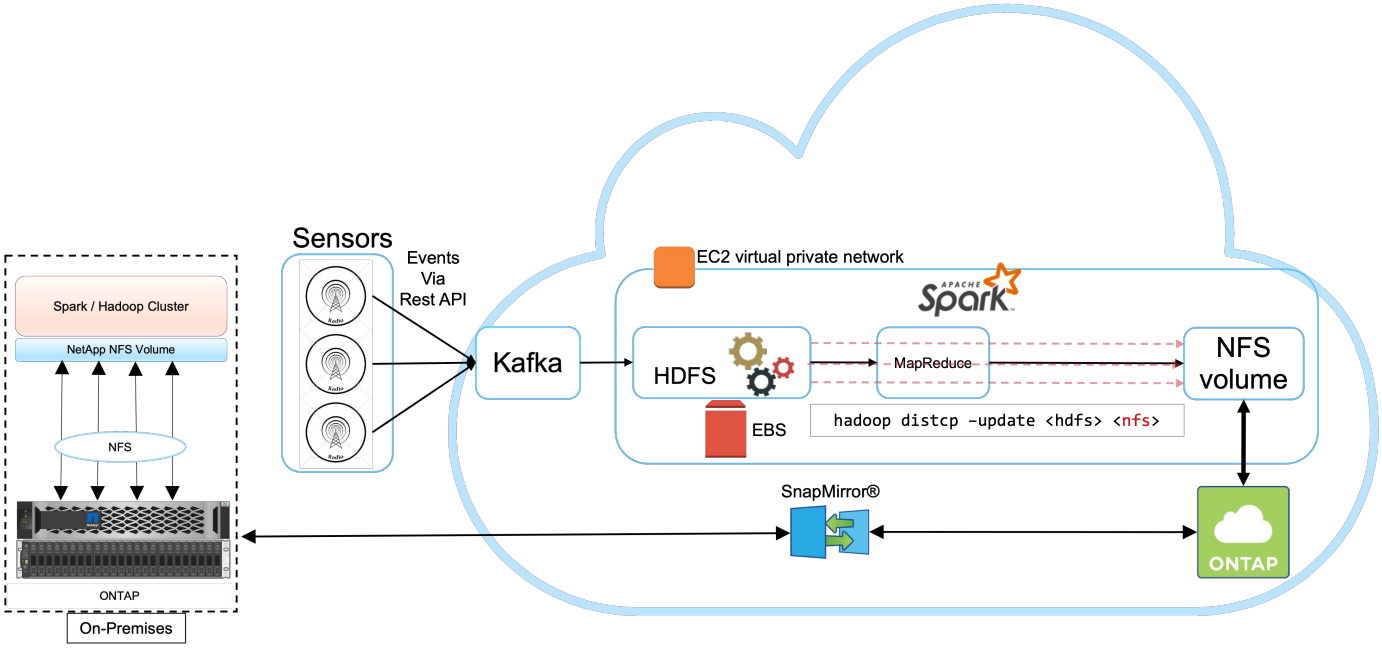
解決策

お客様は、 Spark クラスタの HDFS ストレージとして AWS Elastic Block Store （ EBS ）を使用して、 Kafka 経由でリモートセンサーからデータを受け取り、取り込むことになりました。そのため、 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ストレージのデータをONTAP NFS共有にコピーします。データが ONTAP クラウド上の NFS に配置されると、 SnapMirror テクノロジを使用して、必要に応じてクラウドからオンプレミスのストレージにデータを安全かつ効率的にミラーリングできます。

この図は、クラウドからオンプレミスの解決策へのバックアップとディザスタリカバリを示しています。



ユースケース 3 ：既存の Hadoop データに対して DevTest を有効化

このユースケースでは、 DevTest と Reporting の目的で同じデータセンターとリモートサイトに大量の分析データを格納した既存の Hadoop クラスタをベースに、新しい

Hadoop / Spark クラスタを迅速かつ効率的に構築することがお客様の要件となります。

シナリオ（ **Scenario** ）

このシナリオでは、大規模な Hadoop データレイク実装をオンプレミスとディザスタリカバリサイトで使用して、複数の Spark / Hadoop クラスタを構築しています。

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 用など、同じ本番環境のデータへのアクセスを必要とする用途に複数の Hadoop クラスタを作成この課題は、非常に大規模な Hadoop クラスタを、スペース効率に優れた方法で何度も瞬時にクローニングすることです。
- 運用効率を高めるために、 Hadoop データを DevTest チームとレポートチームに同期します。
- 業務用クラスタと新規クラスタに同じクレデンシャルを使用して 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、本番クラスタに影響を与えずに効率的に QA クラスタを作成

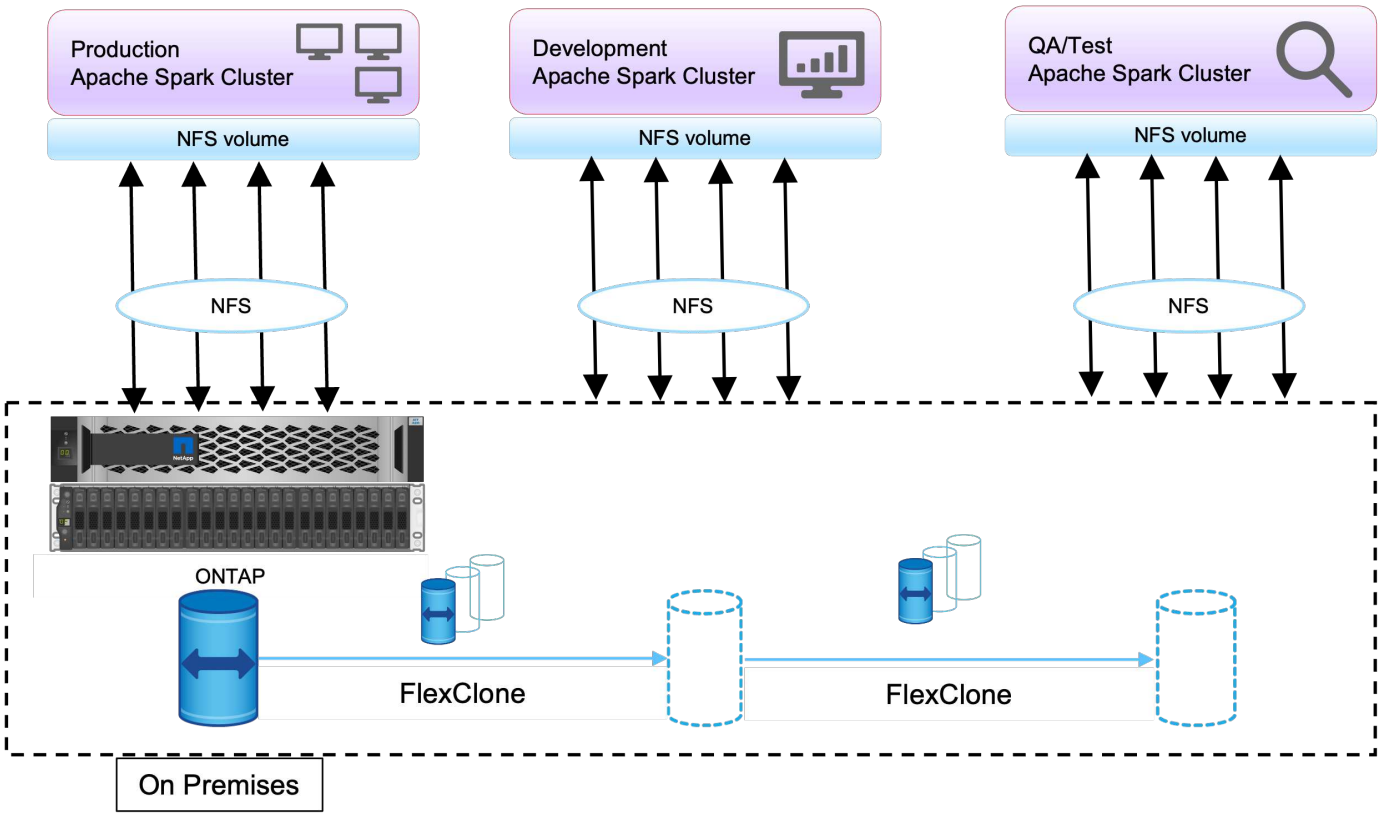
解決策

FlexClone テクノロジは、直前に説明した要件を回答に適用するために使用されます。FlexClone テクノロジは、 Snapshot コピーの読み取り / 書き込みコピーです。親 Snapshot コピーのデータを読み取り、新規または変更されたブロック用に追加のスペースのみを消費します。高速でスペース効率に優れています。

まず、ネットアップの整合グループを使用して既存のクラスタの Snapshot コピーを作成し、

NetApp System Manager またはストレージ管理プロンプト内の Snapshot コピー。整合グループ Snapshot コピーはアプリケーションと整合性のあるグループ Snapshot コピーであり、整合グループ Snapshot コピーに基づいて FlexClone ボリュームが作成されます。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留意する必要があります。Snapshot コピーの作成後、次の図に示すように、DevTest および Reporting 用に新しい Hadoop クラスタをインストールする必要があります。新しいHadoopクラスタのクローンNFSボリュームがNFSデータにアクセスします。

この図は、 DevTest 用の Hadoop クラスタを示しています。



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、お客様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という課題を抱えるクラウドサービスパートナーに適しています。

シナリオ（ Scenario ）

このシナリオでは、さまざまなソースから AWS で受信した 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 と Azure 上にある Spark / Hadoop クラスタに接続されています。これにより、同じデータにアクセスする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を実現できます。

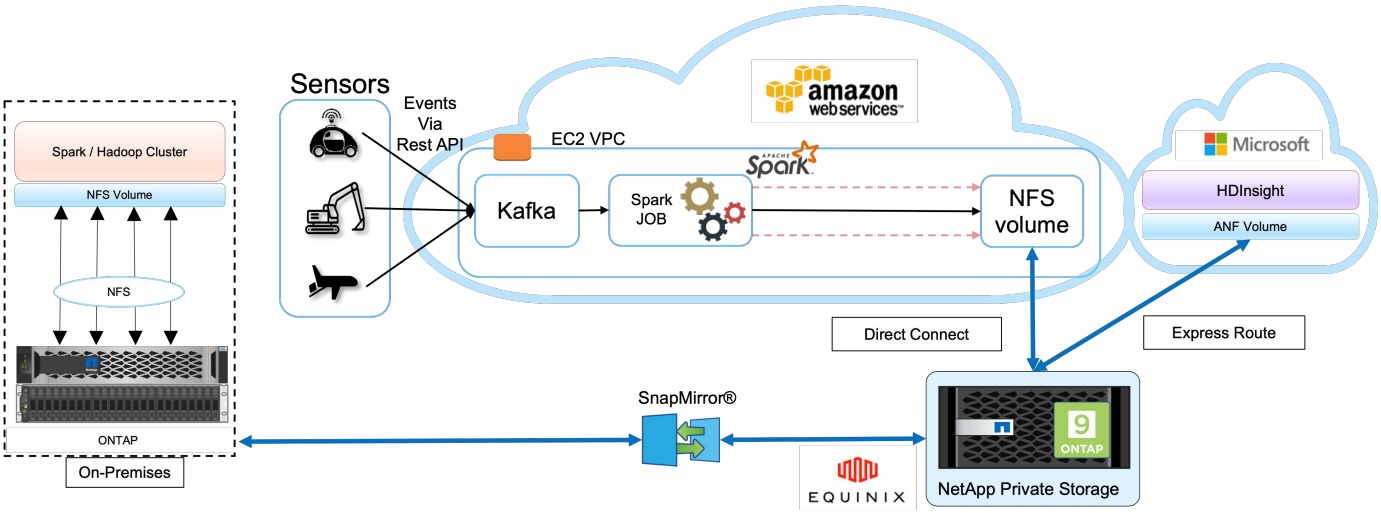
要件と課題

このユースケースの主な要件と課題は次のとおりです。

- お客様は、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。
- オンプレミスやクラウドなどのさまざまなソースから、さまざまなセンサーやハブを介してデータを受信する必要があります。
- 解決策は、効率性とコスト効率に優れている必要があります。
- 主な課題は、オンプレミスと異なるクラウドの間でハイブリッド分析サービスを提供する、対費用効果の高い効率的な解決策を構築することです。

解決策

この図は、データ保護とマルチクラウド接続解決策を示しています。



上の図に示すように、センサーからのデータはストリーミングされ、 Kafka を介して AWS Spark クラスタに取り込まれます。データは NPS 内の NFS 共有に格納されます。 NPS は、 Equinix データセンター内のクラウドプロバイダの外部にあります。NetApp NPSはDirect Connect接続とExpress Route接続を介してAmazon AWSとMicrosoft Azureに接続されるため、お客様はAmazonとAWSの両方の分析クラスタから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にわたるクラウド分析の実現を解決します。

そのため、オンプレミスと NPS ストレージはどちらも ONTAP ソフトウェアを実行するため、 SnapMirror を使用して NPS データをオンプレミスクラスタにミラーリングし、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実現できます。

パフォーマンスを最大限に高めるために、通常は複数のネットワークインターフェイスと直接接続 / エクスプレス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推奨します。

ユースケース 5 ：分析ワークロードを高速化

このシナリオでは、 NetApp NFS ストレージ解決策を使用して大規模な金融サービスおよび投資銀行の分析プラットフォームを最新化し、資産管理および定量的ビジネスユニットの投資リスクおよび派生物の分析を大幅に改善しました。

シナリオ（ **Scenario** ）

お客様の既存の環境では、分析プラットフォームに使用される Hadoop インフラストラクチャは、 Hadoop サーバの内部ストレージを活用しています。JBOD 環境の専有特性により、組織内の多くの社内顧客は、リアルタイムデータの繰り返し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利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のに最適な能力は、量的資産管理事業部門にとって好ましくないものとなっていました。

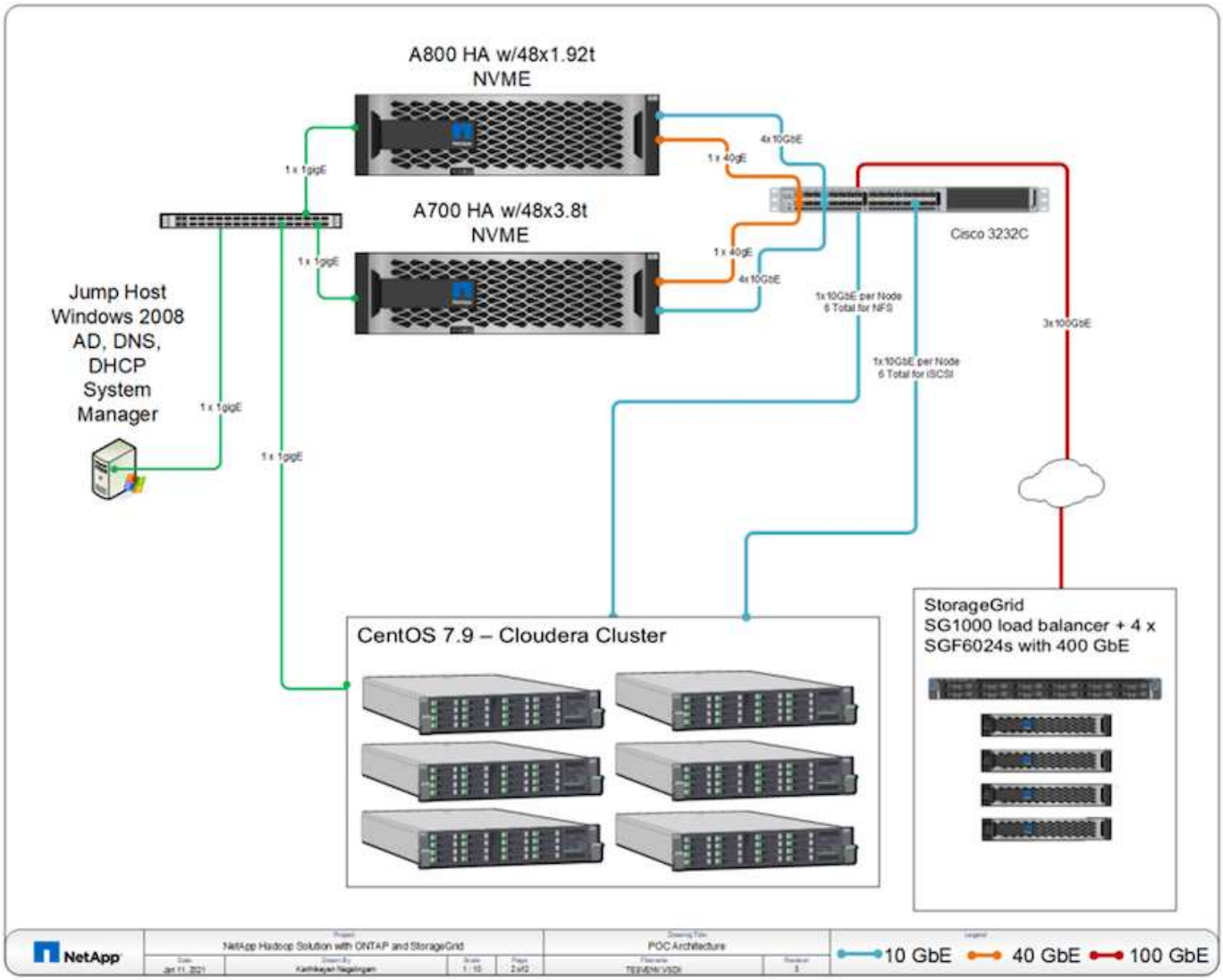
要件と課題

銀行の定量事業部門は、正確でタイムリーな予測を実現するための効率的な予測方法を求めていました。そのためには、インフラを刷新し、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て、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、リスクを分析する必要性を認識しました。

解決策

お客様は、既存の Spark 解決策の JBOD を使用していました。その後、 NetApp ONTAP 、 NetApp StorageGRID 、 MinIO Gateway to NFS を活用して、銀行の定量的財務グループの I/O 待機時間を短縮し、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しました。この図は、 Spark の解決策とネットアップストレージを示しています。



上の図に示すように、 Spark 搭載の 6 ノード Hadoop クラスタで NFS プロトコルと S3 プロトコルを使用して寄木細工のファイルにアクセスするために AFF A800 、 A700 システム、 StorageGRID を導入し、データ分析処理用に糸と Hive のメタデータサービスを用意しました。

お客様の古い環境にある DAS （直接接続型ストレージ）解決策には、コンピューティングとストレージを個別に拡張するという欠点がありました。NetApp ONTAP 解決策 for Spark を使用することで、銀行の財務分析事業部門は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リソースをより効率的に提供することができました。

NFS で ONTAP を使用することで、 Spark の SQL ジョブにはコンピュートサーバの CPU がほぼフルに活用され、 I/O 待機時間が 70% 近く削減されました。その結果、 Spark のワークロードの処理能力とパフォーマ

ンスが向上しました。また、 CPU 利用率の向上により、お客様は GPUDirect などの GPU を活用してプラットフォームをさらに最新化できるようになりました。さらに、 StorageGRID は Spark の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 プロトコル経由で 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は、 Cloud Volumes ONTAP 、 Azure NetApp Files 、 NetApp Cloud Volumes Service を推奨します。

まとめ

このセクションでは、ネットアップが提供するさまざまな Hadoop データ保護要件を満たすユースケースとソリューションの概要を説明します。ネットアップのデータファブリックを使用することで、お客様は次のことが可能になります。

- ネットアップの充実したデータ管理機能と Hadoop ネイティブワークフローとの統合により、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 クラスタのバックアップ時間を約 70% 短縮します。
- Hadoop クラスタのバックアップによるパフォーマンスへの影響を排除します。
- 異なるクラウドプロバイダからのマルチクラウドデータ保護とデータアクセスを、単一の分析データソースに同時に提供できます。
- FlexClone テクノロジを使用すると、スペース効率に優れた高速な Hadoop クラスタコピーを作成できます。

追加情報の参照先

このドキュメントに記載されている情報の詳細については、以下のドキュメントや Web サイトを参照してください。

- ネットアップの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"

- ネットアップ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

- ネットアップの Apache Spark 向けストレージソリューション

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

- ネットアップが有効にしたデータファブリック上の Apache Hadoop

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

謝辞

- ネットアップ、 ANZ 地域セールス担当、 Paul Burland 氏
- ネットアップ、ビジネス開発マネージャー、 Hoseb Dermanilian 氏
- ネットアップ、 MPSG ディレクター、 Lee Dorrior 氏

- ネットアップ、 ANZ ビクトリア地区 SE 、システムエンジニア David Thiessen 氏

バージョン履歴

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

著作権に関する情報

Copyright © 2024 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。