■ NetApp

NetAppとDremioの次世代ハイブリッド氷山レイクハウスソリューション

NetApp artificial intelligence solutions

NetApp August 18, 2025

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

目次

NetAppとDremioの次世代ハイブリッド氷山レイクハウスソリューション・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・	1
NetAppとDremioの次世代ハイブリッドIceberg Lakehouseソリューション · · · · · · · · · · · · · · · · · · ·	1
ソリューションの概要・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・	1
NetAppの概要 · · · · · · · · · · · · · · · · · · ·	1
Dremioの概要・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・	2
Dremio とNetApp のハイブリッド Iceberg Lakehouse	
ソリューションは顧客にどのような価値をもたらしますか?	3
テクノロジ要件	3
展開手順・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・	4
ソリューション検証の概要・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・	15
顧客ユースケース・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・	22
NetApp ActiveIQ のユースケース · · · · · · · · · · · · · · · · · · ·	22
自動車部品販売の顧客ユースケース。	23
まとめ	24
詳細情報の入手方法 詳細情報の入手方法	24

NetAppと**Dremio**の次世代ハイブリッド氷山レイクハウスソリューション

NetAppとDremioの次世代ハイブリッドIceberg Lakehouseソリューション

このドキュメントでは、ONTAP S3、NAS、 StorageGRIDなどのNetAppストレージ コントローラからのさまざまなデータ ソースを使用した Dremio の展開の詳細について説明します。展開中に、TPC-DS ベンチマーク ツールを使用して、さまざまなソースに対して 99 個の SQL クエリを実行しました。このドキュメントでは、 NetApp内の顧客ユースケースや、自動車部品販売の顧客に関連するユースケースについても説明します。

ソリューションの概要

Hybrid Iceberg Lakehouse ソリューションは、データ レイクの顧客が直面する顧客の課題に対処するための独自の利点を提供します。 Dremio Unified Lakehouse プラットフォームとNetApp ONTAP、 StorageGRID、 NetApp Cloud ソリューションを活用することで、企業はビジネス運営に大きな価値を付加できます。このソリューションは、 NetApp ソースを含む複数のデータ ソースへのアクセスを提供するだけでなく、全体的な分析パフォーマンスを向上させ、企業がビジネスの成長につながるビジネス インサイトを獲得できるよう支援します。

NetAppの概要

- ONTAPやStorageGRIDなどの NetApp の製品は、ストレージとコンピューティングの分離を可能にし、特定の要件に基づいて最適なリソース利用を実現します。この柔軟性により、お客様はNetAppストレージソリューションを使用してストレージを独自に拡張できるようになります。
- NetApp のストレージ コントローラを活用することで、顧客は NFS および S3 プロトコルを使用してベクター データベースにデータを効率的に提供できます。これらのプロトコルは、顧客データの保存を容易にし、ベクター データベース インデックスを管理するため、ファイルやオブジェクトの方法を通じてアクセスされるデータの複数のコピーが不要になります。
- * NetApp ONTAP は、AWS、Azure、Google Cloud などの主要なクラウド サービス プロバイダー全体で NAS およびオブジェクト ストレージのネイティブ サポートを提供します。この幅広い互換性により、シームレスな統合が保証され、顧客データのモビリティ、グローバルなアクセス性、災害復旧、動的なスケーラビリティ、高パフォーマンスが実現します。

StorageGRID

業界をリードする当社のオブジェクト ストレージ storageGRID は、自動データ配置、柔軟な導入オプション、階層化された消去コーディングによる比類のない耐久性を実現する強力なポリシー エンジンを提供します。単一の名前空間で数十億のオブジェクトとペタバイトのデータをサポートするスケーラブルなアーキテクチャを備えています。このソリューションはハイブリッド クラウド統合を可能にし、主要なクラウド プラットフォームへのデータ階層化を可能にします。 2019 年の IDC Marketscape Worldwide Object-Based Vendor Assessment でリーダーとして認められました。

さらに、storageGRID は、ソフトウェア定義のオブジェクト ストレージ、地理的冗長性、およびマルチサイト機能により、大規模な非構造化データの管理に優れています。ポリシーベースの情報ライフサイクル管理が組み込まれており、ミラーリングや検索などのクラウド統合機能を提供します。 Common Criteria、NF203 Digital Safe Component、ISO/IEC 25051、KPMG、Cohasset Compliance Assessment など、さまざまな認証を取得しています。

要約すると、 NetApp storageGRID は、大規模な非構造化データを効率的に管理するための強力な機能、スケーラビリティ、ハイブリッド クラウド統合、コンプライアンス認定を提供します。

NetApp ONTAP

NetApp ONTAPは、幅広いエンタープライズ機能を提供する堅牢なストレージ ソリューションです。これには、アプリケーションの一貫性と改ざん防止を備えた即時バックアップを提供するスナップショットが含まれています。 SnapRestore はオンデマンドでのバックアップのほぼ瞬時の復元を可能にし、 SnapMirror は統合されたリモート バックアップおよび災害復旧機能を提供します。このソリューションには、自律ランサムウェア保護 (ARP) も組み込まれており、複数の管理者による検証、FIPS 認定による保存データの暗号化、転送中のデータの暗号化、多要素認証 (MFA)、ロールベースのアクセス制御 (RBAC) などの機能によってデータのセキュリティを確保します。包括的なログ記録、監査、オンボードおよび外部キー管理、安全な消去、複数のテナントの安全な管理により、データのセキュリティとコンプライアンスがさらに強化されます。

NetApp ONTAP には、低い総所有コストで、高いレベルの整合性、パフォーマンス、保持力を備えた規制に 準拠したデータ保持を提供するSnapLockも搭載されています。 NetApp ONTAP 9 と完全に統合されており、 悪意のある行為、不正な管理者、ランサムウェアに対する保護を提供します。

このソリューションには、転送中および保存中のデータの暗号化のための NSE/NVE 暗号化、多要素管理者アクセス、および複数管理者検証が含まれます。 Active IQ はAI を活用した予測分析と修正アクションを提供し、QoS はサービス品質のワークロード制御を保証します。管理と自動化の統合は、SysMgr/GUI/CLI/API を通じて直感的に行えます。 FabricPool は自動データ階層化を可能にし、インライン データ圧縮、重複排除、および圧縮を通じて効率性を実現します。 NetApp は、顧客に費用をかけずにワークロード効率の目標を達成することを保証します。

NetApp ONTAP は、NVMe/FC、FC、NVMe/TCP、iSCSI、NFS、SMB、S3 などのさまざまなプロトコルをサポートし、統合ストレージ ソリューションを実現します。全体として、 NetApp ONTAP は、多様なストレージ ニーズを満たすための広範なエンタープライズ機能、強力なセキュリティ、コンプライアンス、効率性、汎用性を提供します。

Dremioの概要

Dremio は、セルフサービス分析と AI のための統合レイクハウス プラットフォームです。 Dremio 統合分析プラットフォームは、従来のデータ ウェアハウス ソリューションの数分の 1 のコストで、レイクハウスの柔軟性、拡張性、パフォーマンスを実現し、ユーザーをデータに近づけます。 Dremio は、「シフトレフト」分析を可能にして、複雑でコストのかかるデータ統合と ETL を排除し、データ移動のないシームレスなエンタープライズ規模の分析を実現します。 Dremio には以下の機能もあります:

- ユニバーサルなセマンティックレイヤーと緊密に統合された高性能なSQLクエリエンジンによって実現される使いやすいセルフサービス分析により、クラウドとオンプレミスの両方ですべてのデータの接続、管理、分析が容易になります。
- Dremio の Apache Iceberg ネイティブのレイクハウス管理機能は、データ検出を簡素化し、データ最適化を自動化して、Git にヒントを得たデータ バージョン管理による高性能分析を実現します。
- Dremio はオープンソースとオープン スタンダードに基づいて構築されており、企業がロックインを回避 し、イノベーションの態勢を維持できるようにします。エンタープライズ企業は、あらゆるワークロード で最高の価格性能比を備えた、最も使いやすいレイクハウス プラットフォームとして Dremio を信頼して

Dremio と**NetApp** のハイブリッド **Iceberg Lakehouse** ソリューションは顧客にどのような価値をもたらしますか**?**

- データ管理とアクセス性の向上: Dremio は、組織がデータ レイクから直接高速にデータをクエリできるようにするデータ レイクハウス プラットフォームでよく知られています。一方、 NetApp は、クラウド データ サービスとデータ ストレージ ソリューションの大手プロバイダーです。この共同オファーにより、お客様は企業のデータを効率的かつ効果的に保存、管理、アクセス、分析するための包括的なソリューションを利用できるようになります。
- パフォーマンスの最適化: NetApp のデータ ストレージに関する専門知識と Dremio のデータ処理およびデータ最適化の機能を組み合わせたパートナーシップにより、データ操作のパフォーマンスを向上させ、レイテンシを削減し、ビジネス インサイトを迅速に得るソリューションが提供されます。 Dremio は、NetApp 独自の社内 IT 分析インフラストラクチャにもパフォーマンス上のメリットをもたらしました。
- スケーラビリティ: Dremio とNetApp はどちらも、拡張性を考慮して設計されたソリューションを提供しています。この共同ソリューションは、顧客に高度にスケーラブルなデータ ストレージ、データ管理、および分析環境を提供します。ハイブリッド Iceberg Lakehouse 環境では、Dremio SQL クエリ エンジンとNetApp StorageGRIDを組み合わせることで、比類のないスケーラビリティ、同時実行性、クエリ パフォーマンスを実現し、あらゆるビジネスの分析ニーズに対応できます。
- データ セキュリティとガバナンス: 両社とも、データ セキュリティとガバナンスに重点を置いています。これらを組み合わせることで、強力なセキュリティとデータ ガバナンス機能が提供され、データが保護され、データ ガバナンスの要件が満たされることが保証されます。ロールベースのきめ細かなアクセス制御、包括的な監査、エンドツーエンドのデータ系統、統合 ID 管理、広範なコンプライアンスとセキュリティ フレームワークを備えた SSO などの機能により、企業の分析データ環境が安全に管理されます。
- コスト効率: Dremio のデータ レイク エンジンを NetApp のストレージ ソリューションと統合すること で、顧客はデータ管理とデータ移動に関連するコストを削減できます。組織は、従来のデータ レイク環境 から、 NetAppと Dremio で構成されたより最新のレイクハウス ソリューションに移行することもできます。このハイブリッド Iceberg Lakehouse ソリューションは、高速クエリ パフォーマンスと市場をリード するクエリ同時実行性を提供し、TCO を削減し、ビジネス洞察を得るまでの時間を短縮します。

テクノロジ要件

このドキュメントで実行された検証には、以下に概説するハードウェアおよびソフトウェア構成が使用されました。これらの構成は、環境を設定するためのガイドラインとして役立ちます。ただし、具体的なコンポーネントは個々の顧客の要件に応じて異なる場合があることに注意してください。

ハードウェア要件

ハードウェア	詳細
NetApp AFFストレージアレイ HA ペア	• A800
	• ONTAP 9.14.1
	• 48 x 3.49TB SSD-NVM
	• 2 つの S3 バケット: Dremio メタデータと顧客データ。

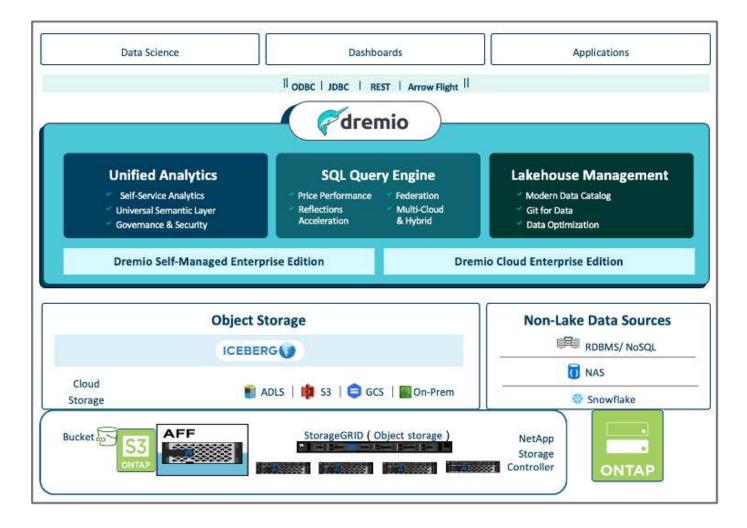
ハードウェア	詳細
富士通 PRIMERGY RX2540 M4 x 4	• CPU×64
	・インテル Xeon Gold 6142 CPU @ 2.60GHz
	• 256 GM 物理メモリ
	・1 x 100GbE ネットワークポート
ネットワーク	・100ギガビットイーサネット
StorageGRID	* 1 x SG100、3xSGF6024 * 3 x 24 x 7.68TB * 2 つの S3 バケット:Dremio メタデータと顧客データ。

ソフトウェア要件

ソフトウェア	詳細
ドレミオ	 バージョン - 25.0.3-202405170357270647- d2042e1b エンタープライズエディション
オンプレミス	5ノードのDremioクラスターマスターコーディネーター1名とエグゼキューター4名

展開手順

このリファレンスアーキテクチャの検証では、1つのコーディネータと4つのエグゼキュータで構成されるDremio構成を使用しました。



NetAppのセットアップ

- ・ストレージシステムの初期化
- *ストレージ仮想マシン(SVM)の作成
- ・論理ネットワークインターフェースの割り当て
- ・NFS、S3 の設定とライセンス

NFS (ネットワーク ファイル システム) の場合は、次の手順に従ってください。1. NFSv4 または NFSv3 用の Flex Group ボリュームを作成します。この検証のセットアップでは、48 台の SSD を使用しました。1 台の SSD はコントローラのルート ボリューム専用で、47 台の SSD は NFSv4 用に分散されています。 Flex Group ボリュームの NFS エクスポート ポリシーに、Dremio サーバー ネットワークに対する読み取り/書き込み権限があることを確認します。

1. すべての Dremio サーバーでフォルダーを作成し、各 Dremio サーバーの論理インターフェイス (LIF) を介してこのフォルダーに Flex Group ボリュームをマウントします。

S3 (Simple Storage Service) の場合は以下の手順に従ってください。

- 1. 「vserver object-store-server create」コマンドを使用して、HTTP が有効で管理ステータスが「up」に設定されたオブジェクト ストア サーバーをセットアップします。 HTTPS を有効にしてカスタム リスナーポートを設定するオプションがあります。
- 2. 「vserver object-store-server user create -user <username>」コマンドを使用して、オブジェクト ストア

サーバー ユーザーを作成します。

- 3. アクセス キーとシークレット キーを取得するには、次のコマンドを実行します: "set diag; vserver object-store-server user show -user <username>"。ただし、今後はこれらのキーはユーザー作成プロセス中に提供されるか、REST API 呼び出しを使用して取得できるようになります。
- 4. 手順 2 で作成したユーザーを使用して object-store-server グループを確立し、アクセスを許可します。この例では、「FullAccess」を指定しました。
- 5. タイプを「S3」に設定して 2 つの S3 バケットを作成します。 1 つは Dremio 構成用、もう 1 つは顧客データ用です。

飼育員のセットアップ

Dremio が提供する Zookeeper 構成を使用できます。この検証では、別のZookeeperを使用しました。このWebリンクに記載されている手順に従いました。 https://medium.com/@ahmetfurkandemir/distributed-hadoop-cluster-1-spark-with-all-dependincies-03c8ec616166

Dremioのセットアップ

この Web リンクに従って、tar ボール経由で Dremio をインストールしました。

1. Dremio グループを作成します。

```
sudo groupadd -r dremio
```

2. dremio ユーザーを作成します。

```
\verb|sudo| useradd -r -g dremio -d /var/lib/dremio -s /sbin/nologin dremio| \\
```

3. Dremio ディレクトリを作成します。

```
sudo mkdir /opt/dremio
sudo mkdir /var/run/dremio && sudo chown dremio:dremio /var/run/dremio
sudo mkdir /var/log/dremio && sudo chown dremio:dremio /var/log/dremio
sudo mkdir /var/lib/dremio && sudo chown dremio:dremio /var/lib/dremio
```

- 4. tarファイルをダウンロードしてください https://download.dremio.com/community-server/
- 5. Dremio を /opt/dremio ディレクトリに解凍します。

```
sudo tar xvf dremio-enterprise-25.0.3-202405170357270647-d2042e1b.tar.gz -C /opt/dremio --strip-components=1
```

6. 構成フォルダーへのシンボリック リンクを作成します。

```
sudo ln -s /opt/dremio/conf /etc/dremio
```

- 7. サービス構成を設定します (SystemD セットアップ)。
 - a. dremio デーモンのユニット ファイルを /opt/dremio/share/dremio.service から /etc/systemd/system/dremio.service にコピーします。
 - b. システムを再起動する

```
sudo systemctl daemon-reload
```

C. 起動時に dremio が起動するようにします。

```
sudo systemctl enable dremio
```

- 8. コーディネーター上で Dremio を構成します。詳細については、Dremio 構成を参照してください。
 - a. Dremio.conf

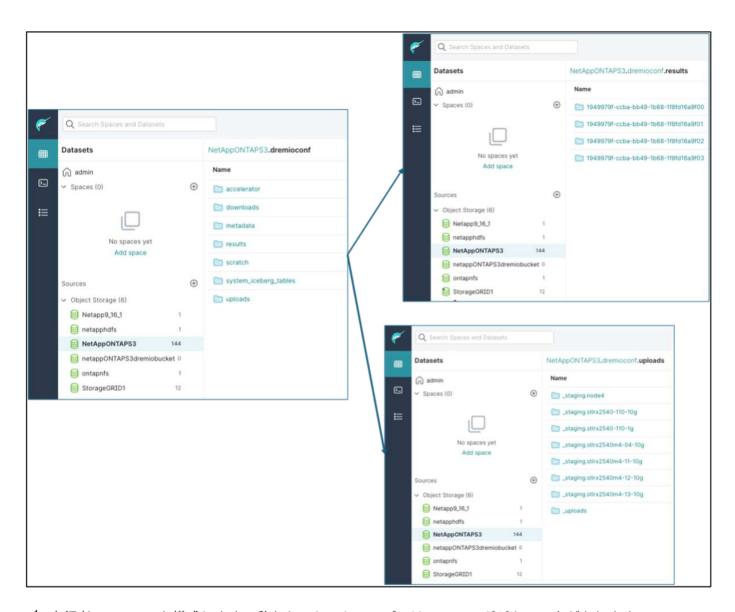
```
root@hadoopmaster:/usr/src/tpcds# cat /opt/dremio/conf/dremio.conf
paths: {
  # the local path for dremio to store data.
  local: ${DREMIO HOME}"/dremiocache"
  # the distributed path Dremio data including job results,
downloads, uploads, etc
  #dist: "hdfs://hadoopmaster:9000/dremiocache"
  dist: "dremioS3://dremioconf"
services: {
  coordinator.enabled: true,
  coordinator.master.enabled: true,
  executor.enabled: false,
 flight.use session service: false
}
zookeeper: "10.63.150.130:2181,10.63.150.153:2181,10.63.150.151:2181"
services.coordinator.master.embedded-zookeeper.enabled: false
root@hadoopmaster:/usr/src/tpcds#
```

b. コアサイト.xml

```
root@hadoopmaster:/usr/src/tpcds# cat /opt/dremio/conf/core-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
 Licensed under the Apache License, Version 2.0 (the "License");
 you may not use this file except in compliance with the License.
 You may obtain a copy of the License at
   http://www.apache.org/licenses/LICENSE-2.0
 Unless required by applicable law or agreed to in writing, software
 distributed under the License is distributed on an "AS IS" BASIS,
 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or
implied.
 See the License for the specific language governing permissions and
 limitations under the License. See accompanying LICENSE file.
-->
<!-- Put site-specific property overrides in this file. -->
<configuration>
   property>
       <name>fs.dremioS3.impl</name>
       <value>com.dremio.plugins.s3.store.S3FileSystem</value>
   </property>
   cproperty>
               <name>fs.s3a.access.key</name>
               <value>24G4C1316APP2BIPDE5S
   </property>
   property>
               <name>fs.s3a.endpoint
               <value>10.63.150.69:80
       </property>
    cproperty>
           <name>fs.s3a.secret.key</name>
           <value>Zd28p43rgZaU44PX ftT279z9nt4jBSro97j87Bx</value>
   </property>
   cproperty>
           <name>fs.s3a.aws.credentials.provider</name>
           <description>The credential provider type.</description>
<value>org.apache.hadoop.fs.s3a.SimpleAWSCredentialsProvider</value>
    </property>
   cproperty>
               <name>fs.s3a.path.style.access
               <value>false</value>
```

```
</property>
   cproperty>
           <name>hadoop.proxyuser.dremio.hosts
           <value>*</value>
   </property>
   cproperty>
           <name>hadoop.proxyuser.dremio.groups</name>
           <value>*</value>
   </property>
   cproperty>
           <name>hadoop.proxyuser.dremio.users
           <value>*</value>
   </property>
   cproperty>
       <name>dremio.s3.compat
       <description>Value has to be set to true.</description>
       <value>true</value>
   </property>
   cproperty>
       <name>fs.s3a.connection.ssl.enabled
       <description>Value can either be true or false, set to true
to use SSL with a secure Minio server.</description>
       <value>false</value>
   </property>
</configuration>
root@hadoopmaster:/usr/src/tpcds#
```

9. Dremio 構成はNetAppオブジェクト ストレージに保存されます。私たちの検証では、「dremioconf」バケットは ontap S3 バケットに存在します。下の図は、「dremioconf」S3 バケットの「scratch」および「uploads」フォルダの詳細を示しています。



- 1. 実行者で Dremio を構成します。私たちのセットアップには3つのエグゼキュータがあります。
 - a. dremio.conf

```
paths: {
  # the local path for dremio to store data.
  local: ${DREMIO_HOME}"/dremiocache"

  # the distributed path Dremio data including job results,
downloads, uploads, etc
  #dist: "hdfs://hadoopmaster:9000/dremiocache"
  dist: "dremioS3:///dremioconf"
}

services: {
  coordinator.enabled: false,
  coordinator.master.enabled: false,
  executor.enabled: true,
  flight.use_session_service: true
}

zookeeper: "10.63.150.130:2181,10.63.150.153:2181,10.63.150.151:2181"
services.coordinator.master.embedded-zookeeper.enabled: false
```

b. Core-site.xml – コーディネーターの構成と同じです。



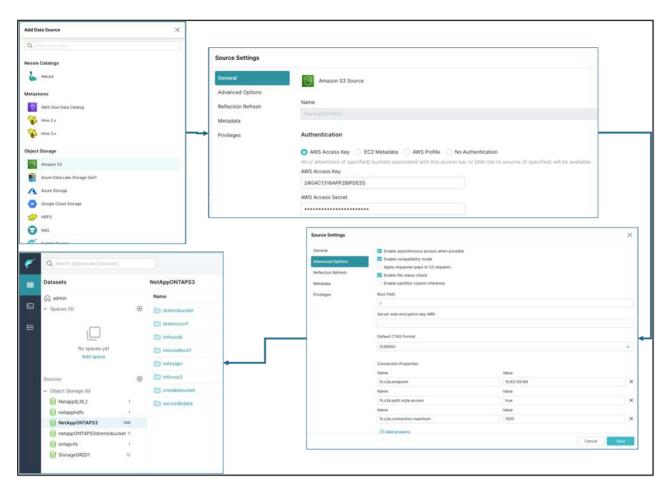
NetApp は、Datalake および Lakehouse 環境の主要なオブジェクト ストレージ ソリューションとしてStorageGRID を推奨しています。さらに、ファイル/オブジェクトの二重性のためにNetApp ONTAPが採用されています。このドキュメントでは、お客様のご要望に応じてONTAP S3 のテストを実施し、データ ソースとして正常に機能していることが確認されています。

複数のソースの設定

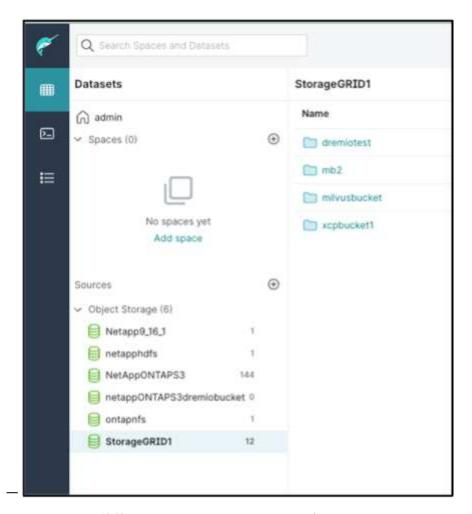
- 1. Dremio でONTAP S3 と storageGRID を s3 ソースとして設定します。
 - a. Dremio ダッシュボード \rightarrow データセット \rightarrow ソース \rightarrow ソースの追加。
 - b. 一般セクションでAWSアクセスキーとシークレットキーを更新してください
 - c. 詳細オプションで互換モードを有効にし、以下の詳細で接続プロパティを更新します。 ontap S3 または storageGRID のいずれかのNetAppストレージ コントローラからのエンドポイント IP/名前。

```
fs.s3a.endoint = 10.63.150.69
fs.s3a.path.style.access = true
fs.s3a.connection.maximum=1000
```

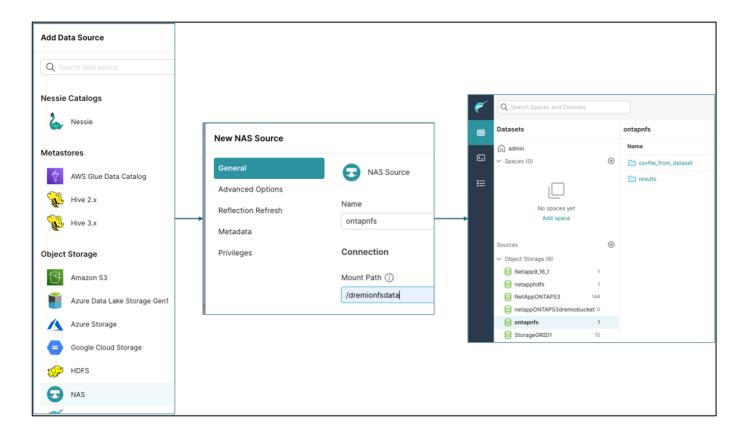
- d. 可能な場合はローカル キャッシュを有効にする、可能な場合に使用する利用可能なキャッシュの合計 の最大パーセント = 100
- e. 次に、 NetAppオブジェクト ストレージからバケットのリストを表示します。



f. storageGRID バケットの詳細のサンプルビュ



- 2. Dremio で NAS (具体的には NFS) をソースとして設定します。
 - a. Dremio ダッシュボード \rightarrow データセット \rightarrow ソース \rightarrow ソースの追加。
 - b. 一般セクションで、名前と NFS マウント パスを入力します。 NFS マウント パスが Dremio クラスター内のすべてのノード上の同じフォルダーにマウントされていることを確認してください。



+

```
root@hadoopmaster:~# for i in hadoopmaster hadoopnode1 hadoopnode2
hadoopnode3 hadoopnode4; do ssh $i "date; hostname; du -hs
/opt/dremio/data/spill/; df -h //dremionfsdata "; done
Fri Sep 13 04:13:19 PM UTC 2024
hadoopmaster
du: cannot access '/opt/dremio/data/spill/': No such file or directory
Filesystem
                            Size Used Avail Use% Mounted on
10.63.150.69:/dremionfsdata 2.1T 921M 2.0T 1% /dremionfsdata
Fri Sep 13 04:13:19 PM UTC 2024
hadoopnode1
12K /opt/dremio/data/spill/
Filesystem
                            Size Used Avail Use% Mounted on
10.63.150.69:/dremionfsdata 2.1T 921M 2.0T 1% /dremionfsdata
Fri Sep 13 04:13:19 PM UTC 2024
hadoopnode2
12K /opt/dremio/data/spill/
Filesystem
                            Size Used Avail Use% Mounted on
10.63.150.69:/dremionfsdata 2.1T 921M 2.0T 1% /dremionfsdata
Fri Sep 13 16:13:20 UTC 2024
hadoopnode3
16K /opt/dremio/data/spill/
Filesystem
                            Size Used Avail Use% Mounted on
10.63.150.69:/dremionfsdata 2.1T 921M 2.0T 1% /dremionfsdata
Fri Sep 13 04:13:21 PM UTC 2024
node4
12K /opt/dremio/data/spill/
Filesystem
                            Size Used Avail Use% Mounted on
10.63.150.69:/dremionfsdata 2.1T 921M 2.0T 1% /dremionfsdata
root@hadoopmaster:~#
```

ソリューション検証の概要

このセクションでは、複数のソースから SQL テスト クエリを実行して機能を検証し、 NetAppストレージへのスピルオーバーをテストおよび検証しました。

オブジェクトストレージのSQLクエリ

1. dremio.envでサーバーあたりのメモリを250GBに設定する

root@hadoopmaster:~# for i in hadoopmaster hadoopnode1 hadoopnode2 hadoopnode3 hadoopnode4; do ssh \$i "hostname; grep -i DREMIO MAX MEMORY SIZE MB /opt/dremio/conf/dremio-env; cat /proc/meminfo | grep -i memtotal"; done hadoopmaster #DREMIO MAX MEMORY SIZE MB=120000 DREMIO MAX MEMORY SIZE MB=250000 MemTotal: 263515760 kB hadoopnode1 #DREMIO_MAX_MEMORY_SIZE MB=120000 DREMIO MAX MEMORY SIZE MB=250000 MemTotal: 263515860 kB hadoopnode2 #DREMIO MAX MEMORY SIZE MB=120000 DREMIO MAX MEMORY SIZE MB=250000 263515864 kB MemTotal: hadoopnode3 #DREMIO MAX MEMORY SIZE MB=120000 DREMIO MAX MEMORY SIZE MB=250000 MemTotal: 264004556 kB node4 #DREMIO MAX MEMORY SIZE MB=120000 DREMIO MAX MEMORY SIZE MB=250000 MemTotal: 263515484 kB root@hadoopmaster:~#

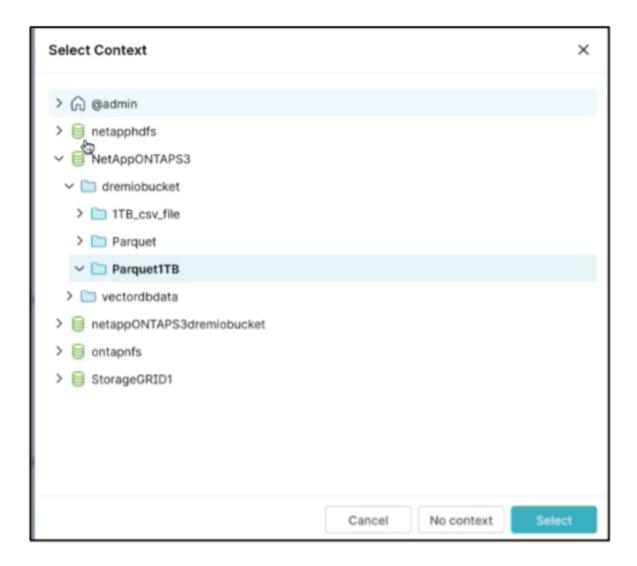
2. dremio.conf ファイルとストレージの詳細で、スピルオーバーの場所 (\${DREMIO_HOME}"/dremiocache)を確認します。

```
paths: {
  # the local path for dremio to store data.
  local: ${DREMIO HOME}"/dremiocache"
  # the distributed path Dremio data including job results, downloads,
uploads, etc
  #dist: "hdfs://hadoopmaster:9000/dremiocache"
  dist: "dremioS3://dremioconf"
}
services: {
 coordinator.enabled: true,
 coordinator.master.enabled: true,
 executor.enabled: false,
 flight.use_session service: false
}
zookeeper: "10.63.150.130:2181,10.63.150.153:2181,10.63.150.151:2181"
services.coordinator.master.embedded-zookeeper.enabled: false
```

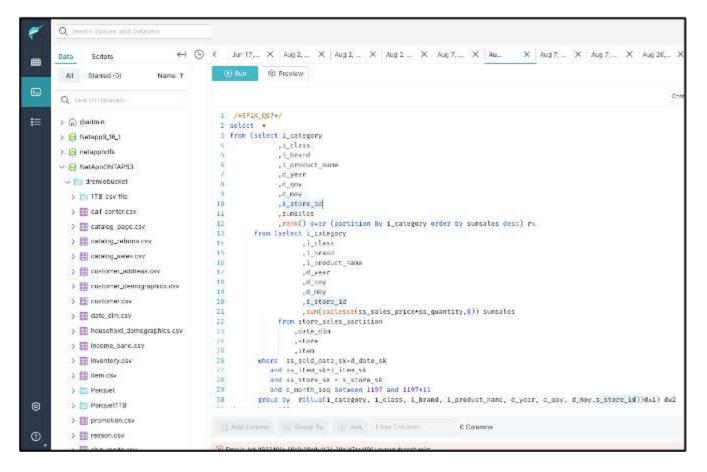
3. Dremioのスピルオーバー場所をNetApp NFSストレージに向ける

```
root@hadoopnode1:~# ls -ltrh /dremiocache
total 4.0K
drwx----- 3 nobody nogroup 4.0K Sep 13 16:00 spilling stlrx2540m4-12-
10q 45678
root@hadoopnode1:~# ls -ltrh /opt/dremio/dremiocache/
drwxr-xr-x 3 dremio dremio 4.0K Aug 22 18:19 spill old
drwxr-xr-x 4 dremio dremio 4.0K Aug 22 18:19 cm
lrwxrwxrwx 1 root root 12 Aug 22 19:03 spill -> /dremiocache
root@hadoopnode1:~# ls -ltrh /dremiocache
total 4.0K
drwx----- 3 nobody nogroup 4.0K Sep 13 16:00 spilling_stlrx2540m4-12-
10g 45678
root@hadoopnode1:~# df -h /dremiocache
                                       Size Used Avail Use% Mounted on
Filesystem
10.63.150.159:/dremiocache hadoopnode1 2.1T 209M 2.0T 1%
/dremiocache
root@hadoopnode1:~#
```

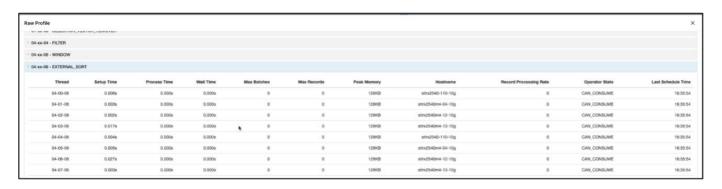
4. コンテキストを選択します。私たちのテストでは、 ONTAP S3 にある TPCDS によって生成された parquet ファイルに対してテストを実行しました。 Dremioダッシュボード \rightarrow SQLランナー \rightarrow コンテキスト \rightarrow NetAppONTAPS3 \rightarrow Parquet1TB



1. DremioダッシュボードからTPC-DSクエリ67を実行する



1. すべての実行プログラムでジョブが実行されていることを確認します。 Dremioダッシュボード \rightarrow ジョブ \rightarrow <jobid> \rightarrow 生のプロファイル \rightarrow EXTERNAL_SORTを選択 \rightarrow ホスト名



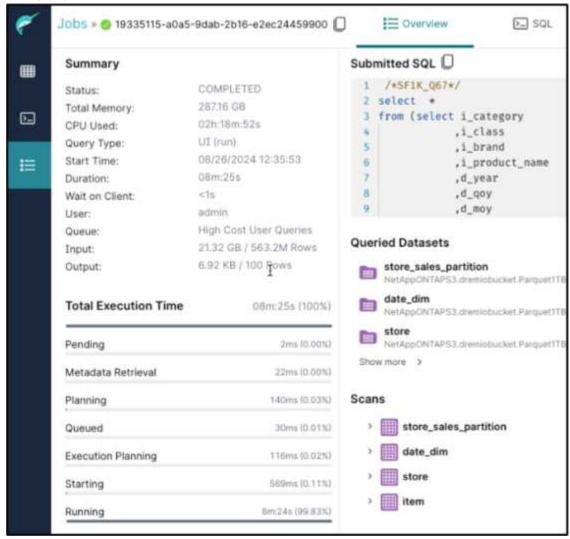
SQL クエリが実行されている場合、 NetAppストレージ コントローラ内のデータ キャッシュの分割フォルダーを確認できます。

```
root@hadoopnode1:~# 1s -ltrh /dremiocache
total 4.0K
drwx----- 3 nobody nogroup 4.0K Sep 13 16:00 spilling_stlrx2540m4-12-
10g_45678
root@hadoopnode1:~# 1s -ltrh /dremiocache/spilling_stlrx2540m4-12-
10g_45678/
total 4.0K
drwxr-xr-x 2 root daemon 4.0K Sep 13 16:23 1726243167416
```

2. SQLクエリはスピルオーバーで完了しまし



3. ジョブ完了の概



4. こぼれたデータのサイズを確認す

要。

EXTERNAL_SORT 04-06

1->

Runtime 1.68m (100%)

Startup 49.09ms (0.05%)

Processing 39.62s (39.36%)

IO Wait 1.02m (60.6%)

Overview/Main

Batches Processed: 104333

Records Processed: 387.6M

Peak Memory: 199 MB

Bytes Sent: 44 GB

Number of Threads: 180

Operator Statistics

Merge Time Nanos: Ons

Spill Count: 360

Spill Time Nanos: 37.68m

Total Spilled Data Size: 20,339,702,765

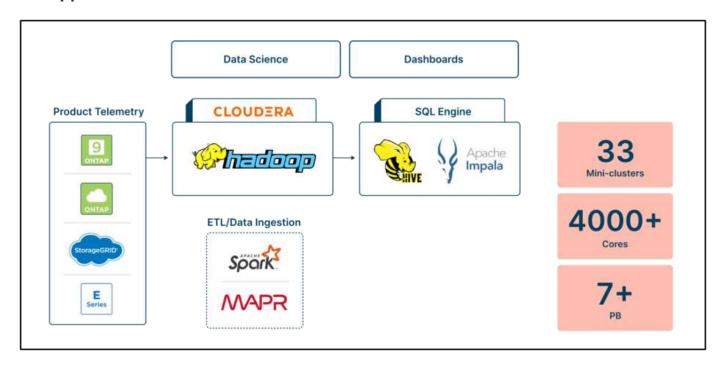
Batches Spilled: 97,854

る

NAS およびStorageGRIDオブジェクト ストレージにも同じ手順が適用されます。

顧客ユースケース

NetApp ActivelQ のユースケース

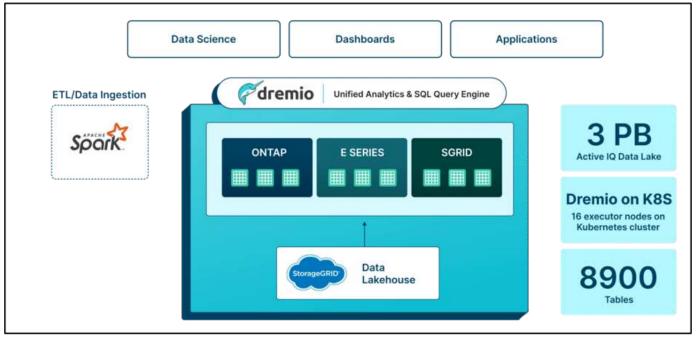


課題: 当初は多数のユースケースをサポートするために設計された NetApp 独自の社内Active IQソリューションは、社内ユーザーと顧客の両方に包括的なサービスを提供するソリューションへと進化しました。しかし、基盤となる Hadoop/MapR ベースのバックエンド インフラストラクチャでは、データの急速な増加と効率的なデータ アクセスの必要性により、コストとパフォーマンスに関する課題が生じていました。ストレージを拡張すると、不要なコンピューティング リソースが追加され、コストが増加します。

さらに、Hadoop クラスターの管理には時間がかかり、専門知識が必要でした。データのパフォーマンスと管理の問題により状況はさらに複雑になり、クエリには平均 45 分かかり、誤った構成によりリソースが不足するようになりました。これらの課題に対処するため、NetApp は既存のレガシー Hadoop 環境に代わる手段を模索し、Dremio 上に構築された新しい最新ソリューションによってコストが削減され、ストレージとコンピューティングが分離され、パフォーマンスが向上し、データ管理が簡素化され、きめ細かな制御が可能になり、災害復旧機能も提供されると判断しました。

解決

22



Dremio により、NetApp は段階的なアプローチで Hadoop ベースのデータ インフラストラクチャを最新化し、統合分析のロードマップを提供できるようになりました。データ処理に大幅な変更を必要とする他のベンダーとは異なり、Dremio は既存のパイプラインとシームレスに統合し、移行時の時間と費用を節約しました。 NetApp は、完全にコンテナ化された環境に移行することで、管理オーバーヘッドを削減し、セキュリティを強化し、耐障害性を強化しました。 Dremio は、Apache Iceberg や Arrow などのオープン エコシステムを採用することで、将来性、透明性、拡張性を確保しました。

Dremio は、Hadoop/Hive インフラストラクチャの代替として、セマンティック レイヤーを通じて二次的なユース ケース向けの機能を提供しました。既存の Spark ベースの ETL およびデータ取り込みメカニズムはそのままに、Dremio は、重複のないデータの検出と探索を容易にする統合アクセス レイヤーを提供しました。このアプローチにより、データ複製係数が大幅に削減され、ストレージとコンピューティングが分離されました。

メリット: Dremio を使用することで、NetApp はデータ環境におけるコンピューティング消費とディスク容量要件を最小限に抑え、大幅なコスト削減を実現しました。新しいActive IQデータ レイクは、3 ペタバイトのデータを保持する 8,900 個のテーブルで構成されています。以前のインフラストラクチャでは、7 ペタバイト以上でした。 Dremio への移行には、33 個のミニ クラスターと 4,000 個のコアから Kubernetes クラスター上の 16 個のエグゼキューター ノードへの移行も含まれていました。コンピューティング リソースが大幅に減少したにもかかわらず、 NetApp はパフォーマンスの顕著な向上を実現しました。 Dremio を介してデータに直接アクセスすることで、クエリの実行時間が 45 分から 2 分に短縮され、予測メンテナンスと最適化のための洞察を得るまでの時間が 95% 高速化されました。この移行により、コンピューティング コストが 60%以上削減され、クエリが 20 倍以上高速化され、総所有コスト (TCO) が 30% 以上節約されました。

自動車部品販売の顧客ユースケース。

課題: この世界的な自動車部品販売会社では、経営陣と企業財務計画および分析グループが販売レポートの統合ビューを取得できず、個々の事業ラインの販売指標レポートを読み取って統合するしかありませんでした。その結果、顧客は少なくとも 1 日前のデータに基づいて意思決定を行うようになりました。新しい分析情報を取得するには、通常 4 週間以上かかります。データ パイプラインのトラブルシューティングにはさらに多くの時間が必要となり、すでに長いタイムラインにさらに 3 日以上が追加されます。レポート開発プロセスとレポートのパフォーマンスが遅いため、アナリスト コミュニティは、新しいビジネス インサイトを見つけたり、新しいビジネス行動を推進したりするのではなく、データの処理または読み込みが完了するまで継続的に待機する必要がありました。これらの問題のある環境は、さまざまな事業部門の多数の異なるデータベースで構成されており、多数のデータ サイロが発生していました。遅くて断片化された環境では、アナリストが単一の真実のソースではなく独自の真実を導き出す方法が多すぎるため、データ ガバナンスが複雑になりま

した。このアプローチでは、データ プラットフォームと人件費で 190 万ドル以上かかりました。レガシープラットフォームを維持し、データ要求に応えるには、年間 7 人のフィールド テクニカル エンジニア (FTE) が必要でした。データ要求の増加に伴い、データ インテリジェンス チームは、将来のニーズに合わせてレガシー環境を拡張することができませんでした。

ソリューション: NetApp Object Store で大規模な Iceberg テーブルをコスト効率よく保存および管理します。 Dremio のセマンティック レイヤーを使用してデータ ドメインを構築し、ビジネス ユーザーがデータ製品を簡単に作成、検索、共有できるようにします。

顧客へのメリット: • 既存のデータアーキテクチャを改善および最適化し、洞察を得るまでの時間を4週間から 数時間に短縮 • トラブルシューティング時間を3日から数時間に短縮 • データプラットフォームと管理コスト を38万ドル以上削減 • 年間2 FTEのデータインテリジェンス作業を削減

まとめ

結論として、この技術レポートでは、ONTAP S3、NAS、 Storage GRIDなどのNet App ストレージ コントローラーのさまざまなデータ ソースと組み合わせた、Dremio を使用した q Hybrid Iceberg Lakehouse の包括的な展開の詳細について説明しました。展開プロセスは正常に実行され、TPC-DS ベンチマーク ツールを使用して、さまざまなデータソースにわたって 99 個の SQL クエリが実行されました。このレポートでは、Net App 内の顧客の使用事例も調査し、多様なビジネス要件を満たす Dremio の汎用性と有効性を実証しました。さらに、自動車部品販売の顧客を対象とする特定のユースケースが検討され、データ分析と洞察のために Dremio を活用する実際的なアプリケーションと利点が強調されました。

全体として、このドキュメントは、 NetAppストレージ コントローラを使用した Dremio の導入と使用方法を 理解するための貴重なリソースとして機能し、さまざまな業界でデータ主導の意思決定と最適化を推進する機能と可能性を示しています。

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については、次のドキュメントや Web サイトを参照してください。

Zookeeperのインストール

https://medium.com/@ahmetfurkandemir/distributed-hadoop-cluster-1-spark-with-all-dependincies-03c8ec616166

・ドレミオ

https://docs.dremio.com/current/get-started/cluster-deployments/deployment-models/standalone/standalone-tarball/

* storageGRIDを使用したDremioの設定

https://docs.netapp.com/us-en/storagegrid-enable/tools-apps-guides/configure-dremio-storagegrid.html#configure-dremio-data-source

NetAppのユースケース

https://www.dremio.com/customers/netapp/

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。