NetApp 과 Dremio의 차세대 하이브리드 아이스버그 레이크하우스 솔루션

NetApp artificial intelligence solutions

NetApp
August 18, 2025

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

목차

NetApp 과 Dremio의 차세대 하이브리드 아이스버그 레이크하우스 솔루션 . . . 1
NetApp 과 Dremio의 차세대 하이브리드 Iceberg Lakehouse 솔루션 . . . 1
솔루션 개요 . . . 1
NetApp 개요 . . . 1
드레미오 개요 . . . 2
Dremio와 NetApp Hybrid Iceberg Lakehouse 솔루션은 고객에게 어떤 가치를 제공합니까? . . . 2
기술 요구 사항 . . . 3
배포 절차 . . . 4
솔루션 검증 개요 . . . 14
고객 사용 사례 . . . 21
NetApp ActiveIQ 사용 사례 . . . 21
자동차 부품 판매 고객 활용 사례. . . . 22
결론 . . . 23
추가 정보를 찾을 수 있는 곳 . . . 23

NetApp 과 Dremio의 차세대 하이브리드 아이스버그 레이크하우스 솔루션

NetApp 과 Dremio의 차세대 하이브리드 Iceberg Lakehouse 솔루션

이 문서에서는 ONTAP S3, NAS, StorageGRID 포함한 NetApp 스토리지 컨트롤러의 다양한 데이터 소스를 활용한 Dremio의 배포 세부 정보에 대해 설명합니다. 배포하는 동안 TPC-DS 벤치마킹 도구를 사용하여 다양한 소스에서 99개의 SQL 쿼리를 실행했습니다. 이 문서에서는 NetApp 의 고객 사용 사례와 자동차 부품 판매 고객과 관련된 사용 사례도 살펴봅니다.

솔루션 개요

하이브리드 아이스버그 레이크하우스 솔루션은 데이터 레이크 고객이 직면한 과제를 해결하는 데 고유한 이점을 제공합니다. Dremio Unified Lakehouse 플랫폼과 NetApp ONTAP, StorageGRID, NetApp Cloud 솔루션을 활용하면 회사는 비즈니스 운영에 상당한 가치를 더할 수 있습니다. 이 솔루션은 NetApp 소스를 포함한 여러 데이터 소스에 대한 액세스를 제공할 뿐만 아니라 전반적인 분석 성능을 향상시키고 회사가 비즈니스 성장으로 이어지는 비즈니스 통찰력을 얻는 데 도움을 줍니다.

NetApp 개요

- ONTAP 및 StorageGRID 와 같은 NetApp의 제품은 스토리지와 컴퓨팅을 분리하여 특정 요구 사항에 따라 최적의 리소스 활용을 가능하게 합니다. 이러한 유연성 덕분에 고객은 NetApp 스토리지 솔루션을 사용하여 스토리지를 독립적으로 확장할 수 있습니다.
- NetApp의 스토리지 컨트롤러를 활용하면 고객은 NFS 및 S3 프로토콜을 사용하여 벡터 데이터베이스에 효율적으로 데이터를 제공할 수 있습니다. 이러한 프로토콜은 고객 데이터 저장을 용이하게 하고 벡터 데이터베이스 인덱스를 관리하며, 파일 및 객체 메서드를 통해 액세스하는 여러 데이터 사본의 필요성을 제거합니다.
- NetApp ONTAP AWS, Azure, Google Cloud 등 주요 클라우드 서비스 공급업체에서 NAS 및 개체 스토리지에 대한 기본 지원을 제공합니다. 이러한 광범위한 호환성은 원활한 통합을 보장하여 고객 데이터 이동성, 글로벌 접근성, 재해 복구, 동적 확장성 및 고성능을 구현합니다.

StorageGRID

업계를 선도하는 당사 객체 스토리지 storageGRID는 자동화된 데이터 배치를 위한 강력한 정책 엔진, 유연한 배포 옵션, 계층적 삭제 코딩을 통한 탁월한 내구성을 제공합니다. 단일 네임스페이스에서 수십억 개의 객체와 페타바이트 규모의 데이터를 지원하는 확장 가능한 아키텍처를 갖추고 있습니다. 이 솔루션은 하이브리드 클라우드 통합을 지원하여 주요 클라우드 플랫폼으로 데이터를 계층화할 수 있습니다. 2019 IDC Marketscape Worldwide Object-Based Vendor Assessment에서 리더로 인정받았습니다.

또한, storageGRID는 소프트웨어 정의 객체 스토리지, 지리적 중복성, 다중 사이트 기능을 통해 대규모 비정형 데이터를 관리하는 데 탁월합니다. 정책 기반 정보 수명 주기 관리를 통합하고 미러링 및 검색과 같은 클라우드 통합 기능을 제공합니다. Common Criteria, NF203 디지털 금고 부품, ISO/IEC 25051, KPMG, Cohasset 규정 준수 평가를 포함한 다양한 인증을 받았습니다.

요약하자면, NetApp storageGRID는 대규모 비정형 데이터를 효율적으로 관리하는 데 필요한 강력한 기능, 확장성,

하이브리드 클라우드 통합 및 규정 준수 인증을 제공합니다.

NetApp ONTAP

NetApp ONTAP 은 광범위한 엔터프라이즈 기능을 제공하는 강력한 스토리지 솔루션입니다. 여기에는 애플리케이션 일관적이고 변조 방지된 즉각 백업을 제공하는 스냅샷이 포함되어 있습니다. SnapRestore 필요에 따라 백업을 거의 즉시 복원할 수 있게 해주는 반면, SnapMirror 통합된 원격 백업 및 재해 복구 기능을 제공합니다. 이 솔루션은 다중 관리자 검증, FIPS 인증을 통한 저장 데이터 암호화, 전송 중 데이터 암호화, 다중 요소 인증(MFA), 역할 기반 액세스 제어(RBAC)와 같은 기능으로 데이터 보안을 보장하는 자율형 랜섬웨어 보호(ARP) 기능도 통합합니다. 포괄적인 로깅, 감사, 온보딩 및 외부 키 관리, 안전한 정리, 여러 테넌트의 안전한 관리를 통해 데이터 보안과 규정 준수가 더욱 강화됩니다.

NetApp ONTAP 또한 낮은 총소유비용으로 높은 수준의 무결성, 성능 및 보존을 제공하는 규정을 준수하는 데이터 보존 기능을 제공하는 SnapLock 기능을 갖추고 있습니다. NetApp ONTAP 9와 완벽하게 통합되어 악의적인 행위, 사기성 관리자 및 랜섬웨어로부터 보호 기능을 제공합니다.

이 솔루션은 전송 중 및 저장 데이터 암호화를 위한 NSE/NVE 암호화, 다중 관리자 액세스, 다중 관리자 검증을 포함합니다. Active IQ AI 기반 예측 분석과 시정 조치를 제공하고, QoS는 서비스 품질 워크로드 제어를 보장합니다. 관리 및 자동화 통합은 SysMgr/GUI/CLI/API를 통해 직관적으로 이루어집니다. FabricPool 자동 데이터 계층화를 지원하고, 해당 솔루션은 인라인 데이터 압축, 중복 제거 및 압축을 통해 효율성을 제공합니다. NetApp 고객에게 추가 비용 없이 워크로드 효율성 목표 달성을 보장합니다.

NetApp ONTAP NVMe/FC, FC, NVMe/TCP, iSCSI, NFS, SMB, S3 등 다양한 프로토콜을 지원하여 통합 스토리지 솔루션을 제공합니다. 전반적으로 NetApp ONTAP 다양한 스토리지 요구 사항을 충족하는 광범위한 엔터프라이즈 기능, 강력한 보안, 규정 준수, 효율성 및 다용성을 제공합니다.

드레미오 개요

Dremio는 셀프서비스 분석 및 AI를 위한 통합 레이크하우스 플랫폼입니다. Dremio 통합 분석 플랫폼은 기존 데이터웨어하우스 솔루션 비용의 일부만으로 레이크하우스의 유연성, 확장성 및 성능을 제공하여 사용자가 데이터에 더욱 가까이 접근할 수 있도록 합니다. Dremio는 복잡하고 비용이 많이 드는 데이터 통합 및 ETL을 없애는 "시프트-레프트" 분석을 지원하여 데이터 이동 없이 원활한 엔터프라이즈 규모 분석을 제공합니다. Dremio의 또 다른 특징은 다음과 같습니다.

- 범용 의미 계층과 긴밀하게 통합된 고성능 SQL 쿼리 엔진을 통해 사용하기 쉬운 셀프 서비스 분석이 가능하므로 클라우드와 온프레미스 모두에서 모든 데이터를 보다 쉽게 연결, 관리 및 분석할 수 있습니다.
- Dremio의 Apache Iceberg 기반 레이크하우스 관리 기능은 데이터 검색을 간소화하고, 데이터 최적화를 자동화하며, Git에서 영감을 받은 데이터 버전 관리를 통해 고성능 분석을 제공합니다.
- 오픈 소스와 개방형 표준을 기반으로 구축된 Dremio는 기업이 특정 분야에 얽매이지 않고 혁신을 위한 입지를 유지할 수 있도록 지원합니다. 대기업들은 Dremio를 모든 워크로드에 걸쳐 가장 가격 대비 성능이 뛰어나고 사용하기 가장 쉬운 레이크하우스 플랫폼으로 신뢰합니다.

Dremio와 **NetApp Hybrid Iceberg Lakehouse** 솔루션은 고객에게 어떤 가치를 제공합니까**?**

- 향상된 데이터 관리 및 접근성: Dremio는 조직이 데이터 레이크에서 고속으로 직접 데이터를 쿼리할 수 있도록 하는 데이터 레이크하우스 플랫폼으로 잘 알려져 있습니다. 반면, NetApp 클라우드 데이터 서비스와 데이터 저장 솔루션 분야의 선두주자입니다. 이 공동 제안은 고객에게 기업의 데이터를 효율적이고 효과적으로 저장, 관리, 액세스 및 분석할 수 있는 포괄적인 솔루션을 제공합니다.
- 성능 최적화: NetApp의 데이터 스토리지 전문성과 Dremio의 데이터 처리 및 데이터 최적화 역량을 바탕으로, 이 파트너십은 데이터 운영의 성능을 개선하고, 대기 시간을 줄이고, 비즈니스 통찰력을 얻는 속도를 높여주는 솔루션을 제공합니다. Dremio는 NetApp의 내부 IT 분석 인프라에도 성능상의 이점을 제공했습니다.

- 확장성: Dremio와 NetApp 모두 확장 가능하도록 설계된 솔루션을 제공합니다. 이 공동 솔루션은 고객에게 확장성이 뛰어난 데이터 저장, 데이터 관리 및 분석 환경을 제공합니다. 하이브리드 Iceberg Lakehouse 환경에서 Dremio SQL 쿼리 엔진은 NetApp StorageGRID 와 결합되어 모든 비즈니스의 분석 요구 사항을 처리할 수 있는 탁월한 확장성, 동시성 및 쿼리 성능을 제공합니다.
- 데이터 보안 및 거버넌스: 두 회사 모두 데이터 보안과 거버넌스에 중점을 두고 있습니다. 이 두 가지 솔루션을 함께 사용하면 강력한 보안 및 데이터 거버넌스 기능을 제공하여 데이터가 보호되고 데이터 거버넌스 요구 사항이 충족되도록 할 수 있습니다. 역할 기반 및 세분화된 액세스 제어, 포괄적인 감사, 종단 간 데이터 계보, 통합 ID 관리, 광범위한 규정 준수 및 보안 프레임워크를 갖춘 SSO와 같은 기능을 통해 회사의 분석 데이터 환경이 안전하고 관리되도록 보장합니다.
- 비용 효율성: Dremio의 데이터 레이크 엔진을 NetApp의 스토리지 솔루션과 통합함으로써 고객은 데이터 관리 및 데이터 이동과 관련된 비용을 줄일 수 있습니다. 조직은 기존 데이터 레이크 환경에서 NetApp 과 Dremio로 구성된 보다 현대적인 레이크하우스 솔루션으로 전환할 수도 있습니다. 이 하이브리드 Iceberg Lakehouse 솔루션은 고속 쿼리 성능과 시장 선도적인 쿼리 동시성을 제공하여 TCO를 낮추고 비즈니스 통찰력을 얻는 시간을 단축합니다.

기술 요구 사항

아래에 설명된 하드웨어 및 소프트웨어 구성은 이 문서에서 수행된 검증에 활용되었습니다. 이러한 구성은 환경을 설정하는 데 도움이 되는 지침으로 사용되지만, 특정 구성 요소는 개별 고객 요구 사항에 따라 달라질 수 있습니다.

하드웨어 요구 사항

하드웨어	세부
NetApp AFF 스토리지 어레이 HA 쌍	• A800 • ONTAP 9.14.1 • 48 x 3.49TB SSD-NVM • 두 개의 S3 버킷: Dremio 메타데이터와 고객 데이터.
후지쯔 프라이머지 RX2540 M4 4개	• 64개의 CPU • 인텔 제온 골드 6142 CPU @ 2.60GHz • 256 GM 물리적 메모리 • 1 x 100GbE 네트워크 포트
네트워킹	• 100GbE
StorageGRID	* 1 x SG100, 3xSGF6024 * 3 x 24 x 7.68TB * 2개의 S3 버킷: Dremio 메타데이터 및 고객 데이터.

소프트웨어 요구 사항

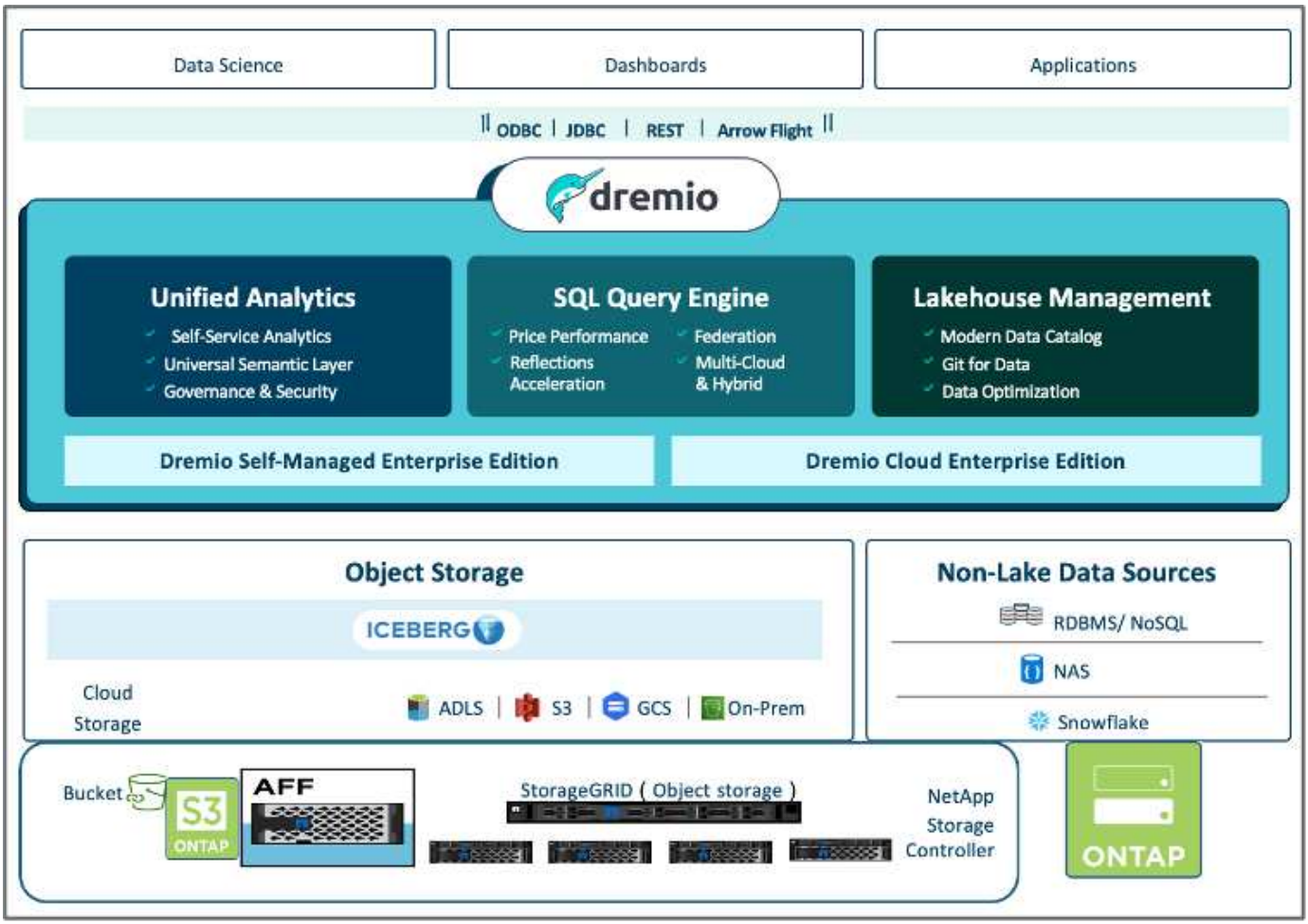
소프트웨어	세부
드레미오	• 버전 - 25.0.3-202405170357270647-d2042e1b • 엔터프라이즈 에디션

소프트웨어	세부
온프레미스	• 5개 노드 Dremio 클러스터 • 1명의 마스터 코디네이터와 4명의 실행자

배포 절차

이 참조 아키텍처 검증에서 우리는 하나의 코디네이터와 네 개의 실행자로 구성된 Dremio 구성을 사용했습니다

.



NetApp 설정

- 스토리지 시스템 초기화
- 스토리지 가상 머신(SVM) 생성
- 논리적 네트워크 인터페이스 할당
- NFS, S3 구성 및 라이센싱

NFS(네트워크 파일 시스템)의 경우 아래 단계를 따르세요. 1. NFSv4 또는 NFSv3에 대한 Flex Group 볼륨을 만듭니다. 이 검증을 위한 설정에서 우리는 48개의 SSD를 사용했는데, 그 중 1개는 컨트롤러의 루트 볼륨에 전용으로 사용되었고, 47개는 NFSv4에 분산되어 있었습니다. Flex Group 볼륨에 대한 NFS 내보내기 정책에 Dremio 서버

네트워크에 대한 읽기/쓰기 권한이 있는지 확인합니다.

1. 모든 Dremio 서버에서 폴더를 만들고 각 Dremio 서버의 논리 인터페이스(LIF)를 통해 Flex Group 볼륨을 이 폴더에 마운트합니다.

S3(Simple Storage Service)의 경우 아래 단계를 따르세요.

1. "vserver object-store-server create" 명령을 사용하여 HTTP를 활성화하고 관리 상태를 'up'으로 설정한 객체 저장소 서버를 설정합니다. HTTPS를 활성화하고 사용자 정의 리스너 포트를 설정할 수 있습니다.
2. "vserver object-store-server user create -user <username>" 명령을 사용하여 object-store-server 사용자를 생성합니다.
3. 액세스 키와 비밀 키를 얻으려면 다음 명령을 실행하세요: "set diag; vserver object-store-server user show -user <username>". 하지만 앞으로는 이러한 키가 사용자 생성 프로세스 중에 제공되거나 REST API 호출을 사용하여 검색할 수 있습니다.
4. 2단계에서 생성한 사용자를 사용하여 객체-저장소-서버 그룹을 설정하고 액세스 권한을 부여합니다. 이 예에서는 "FullAccess"를 제공했습니다.
5. 유형을 "S3"로 설정하여 두 개의 S3 버킷을 만듭니다. 하나는 Dremio 구성용이고 다른 하나는 고객 데이터용입니다.

Zookeeper 설정

Dremio가 제공하는 Zookeeper 구성을 사용할 수 있습니다. 이 검증에서 우리는 별도의 Zookeeper를 사용했습니다. 우리는 이 웹 링크에 언급된 단계를 따랐습니다. https://medium.com/@ahmetfurkandemir/distributed-hadoop-cluster-1-spark-with-all-dependincies-03c8ec616166

Dremio 설정

우리는 이 웹링크를 따라 타르볼을 통해 Dremio를 설치했습니다.

1. Dremio 그룹을 만드세요.

```
sudo groupadd -r dremio
```

2. dremio 사용자를 생성합니다.

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

3. Dremio 디렉토리를 만듭니다.

```
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 파일을 다운로드하세요 https://download.dremio.com/community-server/

5. Dremio를 /opt/dremio 디렉터리에 압축을 풉니다.

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

6. 구성 폴더에 대한 심볼릭 링크를 만듭니다.

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

7. 서비스 구성(SystemD 설정)을 설정합니다.
   a. dremio 데몬의 단위 파일을 /opt/dremio/share/dremio.service에서 /etc/systemd/system/dremio.service로 복사합니다.
   b. 시스템 재시작

```
sudo systemctl daemon-reload
```

   c. 부팅 시 dremio를 시작하도록 설정합니다.

```
sudo systemctl enable dremio
```

8. 코디네이터에서 Dremio를 구성합니다. 자세한 내용은 Dremio 구성을 참조하세요.
   a. 드레미오.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#
```

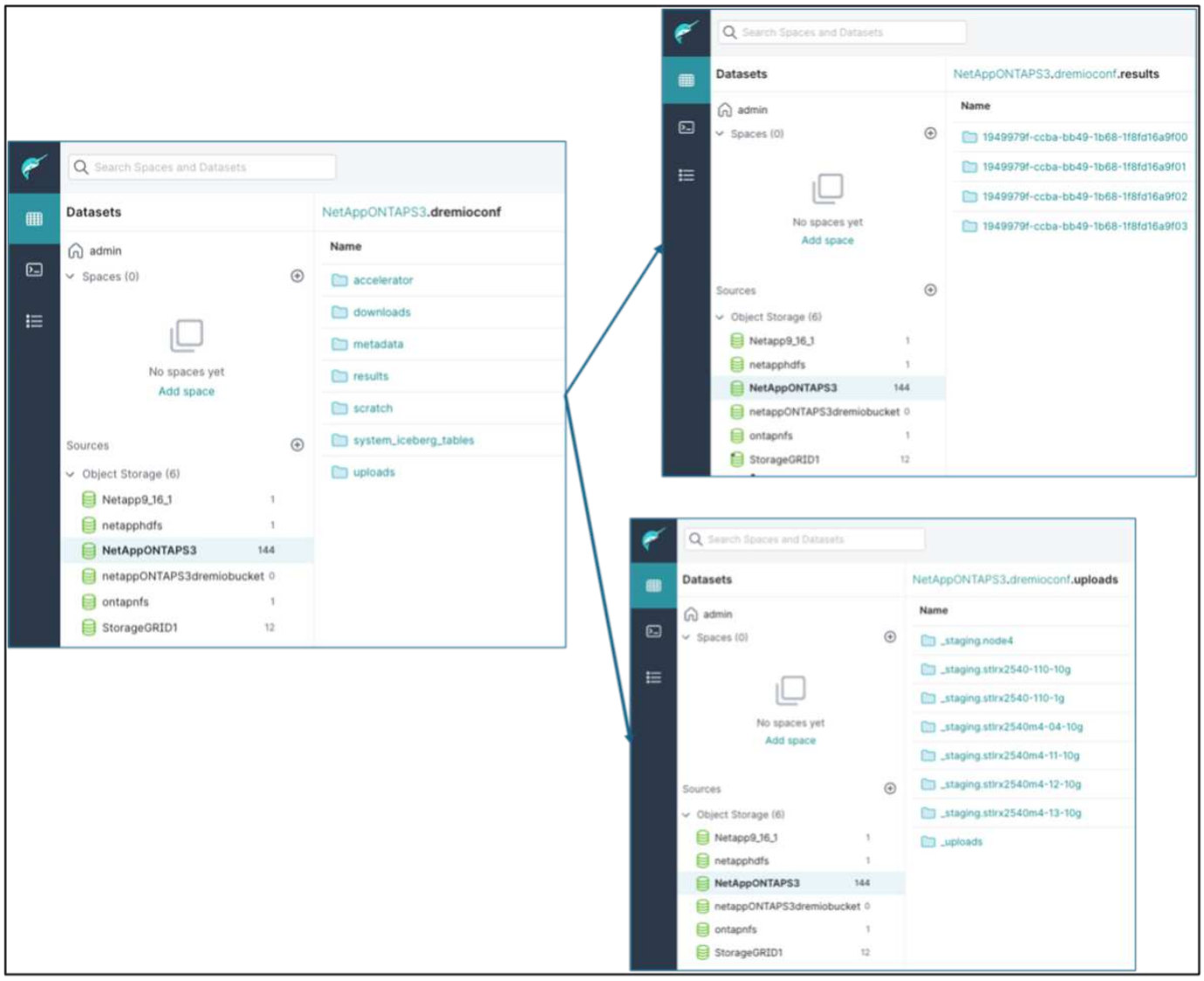
b. 코어-사이트.xml

```
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>
    <property>
                <name>fs.s3a.access.key</name>
                <value>24G4C1316APP2BIPDE5S</value>
    </property>
    <property>
                <name>fs.s3a.endpoint</name>
                <value>10.63.150.69:80</value>
        </property>
    <property>
            <name>fs.s3a.secret.key</name>
            <value>Zd28p43rgZaU44PX_ftT279z9nt4jBSro97j87Bx</value>
    </property>
    <property>
            <name>fs.s3a.aws.credentials.provider</name>
            <description>The credential provider type.</description>

<value>org.apache.hadoop.fs.s3a.SimpleAWSCredentialsProvider</value>
    </property>
    <property>
                <name>fs.s3a.path.style.access</name>
                <value>false</value>
        </property>
    <property>
            <name>hadoop.proxyuser.dremio.hosts</name>
            <value>*</value>
    </property>
    <property>
            <name>hadoop.proxyuser.dremio.groups</name>
            <value>*</value>
    </property>
    <property>
            <name>hadoop.proxyuser.dremio.users</name>
            <value>*</value>
    </property>
    <property>
        <name>dremio.s3.compat</name>
        <description>Value has to be set to true.</description>
        <value>true</value>
    </property>
    <property>
        <name>fs.s3a.connection.ssl.enabled</name>
```

```
        <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 구성은 NetApp 개체 스토리지에 저장됩니다. 검증 결과, "dremioconf" 버킷은 ontap S3 버킷에 있습니다. 아래 그림은 "dremioconf" S3 버킷의 "scratch" 및 "uploads" 폴더의 일부 세부 정보를 보여줍니다.



1. 실행자에서 Dremio를 구성합니다. 우리의 설정에는 3명의 실행자가 있습니다.
   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 – 코디네이터 구성과 동일합니다.

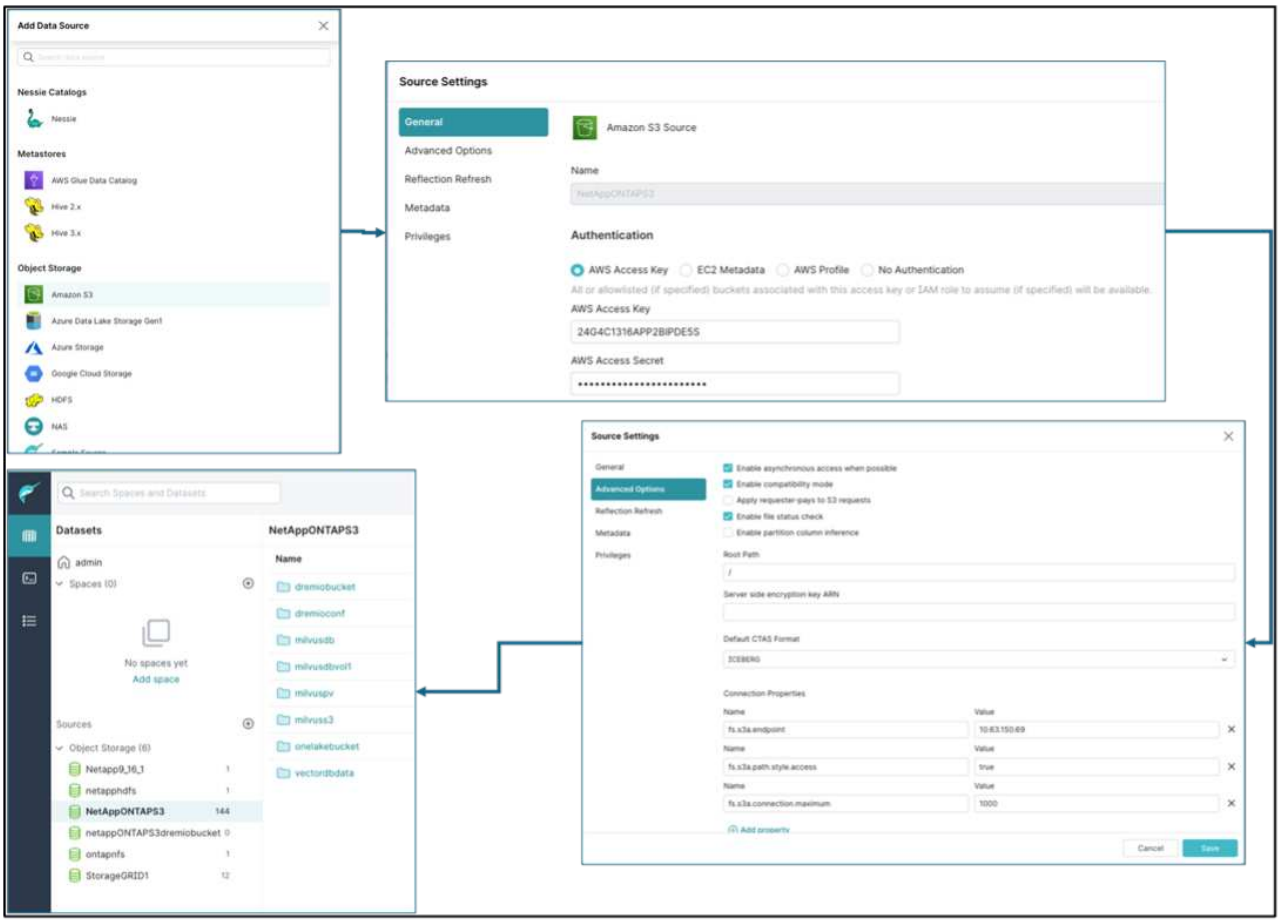
NetApp Datalake 및 Lakehouse 환경을 위한 기본 개체 스토리지 솔루션으로 StorageGRID 권장합니다. 또한 NetApp ONTAP 파일/객체 이중성을 위해 사용됩니다. 이 문서의 맥락에서 우리는 고객 요청에 따라 ONTAP S3에 대한 테스트를 수행했으며, 이는 데이터 소스로서 성공적으로 기능했습니다.

다중 소스 설정

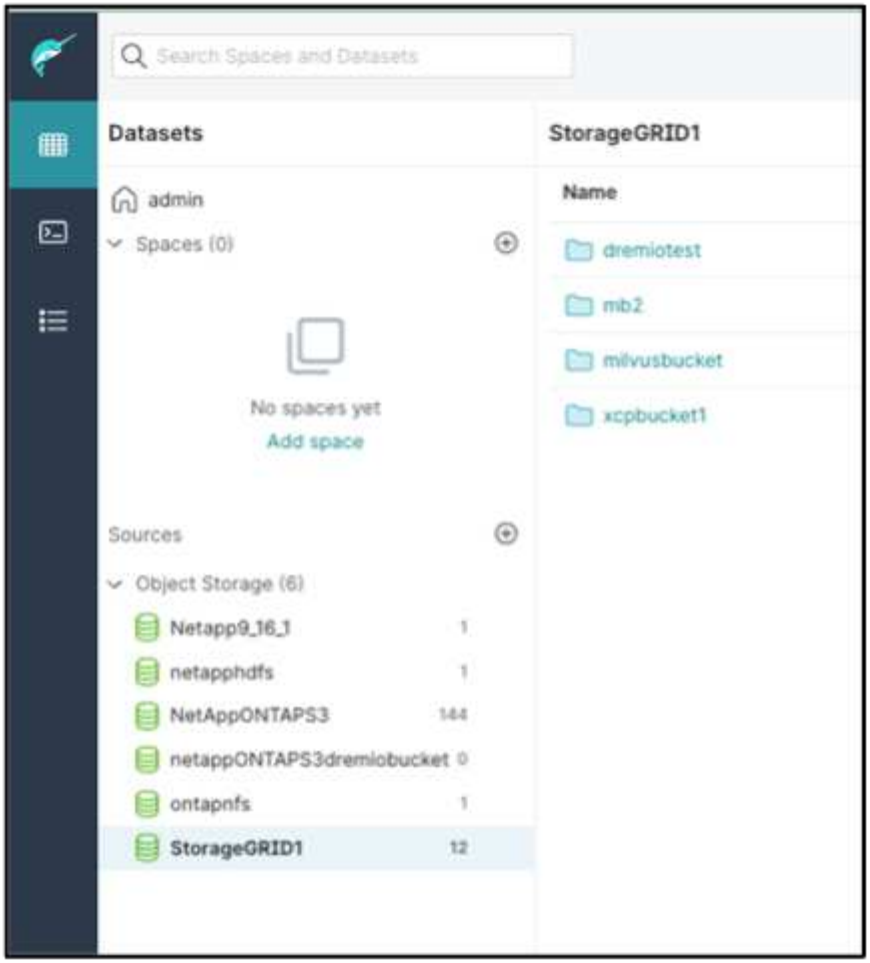
1. Dremio에서 ONTAP S3 및 storageGRID를 S3 소스로 구성합니다.
   a. Dremio 대시보드 → 데이터 세트 → 소스 → 소스 추가.
   b. 일반 섹션에서 AWS 액세스 및 비밀 키를 업데이트하세요.
   c. 고급 옵션에서 호환 모드를 활성화하고 아래 세부 정보로 연결 속성을 업데이트합니다. NetApp 스토리지 컨트롤러의 엔드포인트 IP/이름은 ontap S3 또는 storageGRID에서 가져옵니다.

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

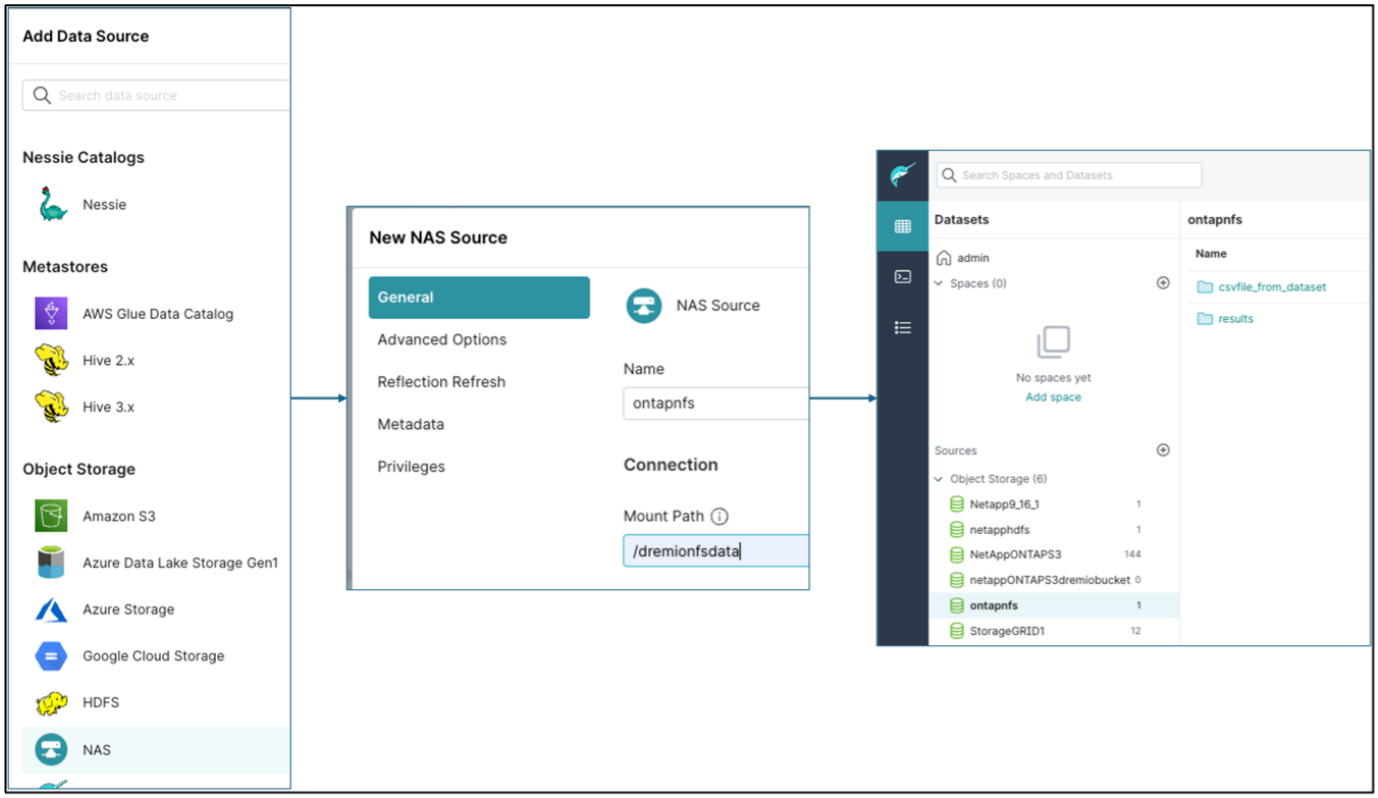
   d. 가능하면 로컬 캐싱을 활성화하고, 가능한 경우 사용할 수 있는 총 캐시의 최대 백분율은 100입니다.
   e. 그런 다음 NetApp 개체 스토리지의 버킷 목록을 봅니다
   .



f. storageGRID 버킷 세부 정보의 샘플

보기 

2. Dremio에서 NAS(특히 NFS)를 소스로 구성합니다.
   a. Dremio 대시보드 → 데이터 세트 → 소스 → 소스 추가.
   b. 일반 섹션에 이름과 NFS 마운트 경로를 입력합니다. Dremio 클러스터의 모든 노드에서 NFS 마운트 경로가 동일한 폴더에 마운트되었는지 확인하세요.


Add Data Source
Search data source
Nessie Catalogs
Nessie
Metastores
AWS Glue Data Catalog
Hive 2.x
Hive 3.x
Object Storage
Amazon S3
Azure Data Lake Storage Gen1
Azure Storage
Google Cloud Storage
HDFS
NAS
New NAS Source
General
Advanced Options
Reflection Refresh
Metadata
Privileges
NAS Source
Name
ontapnfs
Connection
Mount Path
/dremionfsdata
Search Spaces and Datasets
Datasets
admin
Spaces (0)
No spaces yet
Add space
Sources
Object Storage (6)
Netapp9_16_1 1
netapphdfs 1
NetAppONTAPS3 144
netappONTAPS3dremiobucket 0
ontapnfs 1
StorageGRID1 12
ontapnfs
Name
csvfile_from_dataset
results

+

```
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:~#
```

솔루션 검증 개요

이 섹션에서는 여러 소스에서 SQL 테스트 쿼리를 실행하여 기능을 검증하고 NetApp 스토리지로의 스필오버를 테스트하고 확인했습니다.

객체 스토리지에 대한 **SQL** 쿼리

1. dremio.env에서 서버당 메모리를 250GB로 설정합니다.

```
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
MemTotal:       263515864 kB
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:~#
```

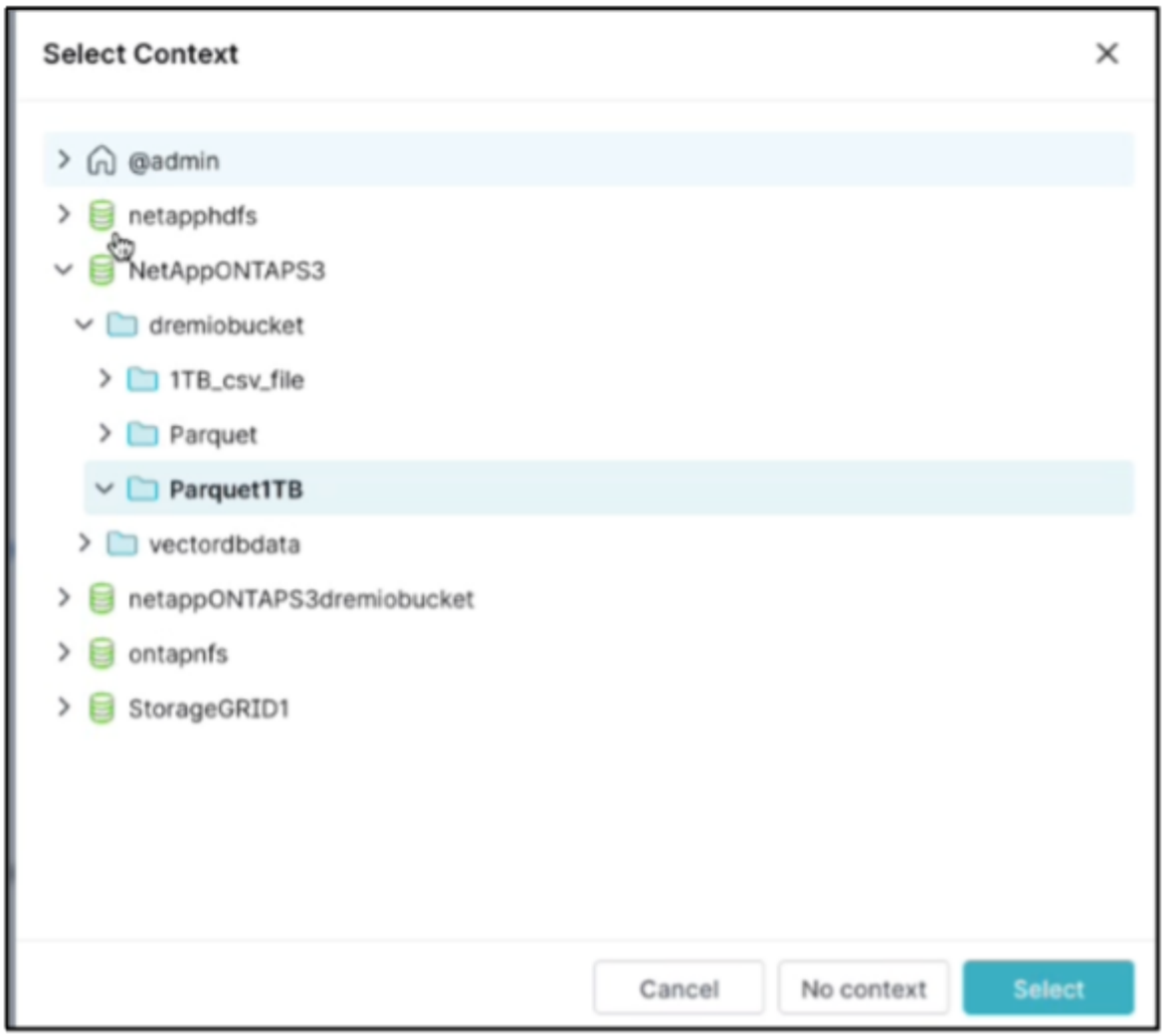
2. dremio.conf 파일에서 스필오버 위치(${DREMIO_HOME}"/dremiocache)와 저장소 세부 정보를 확인하세요.

```
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
```

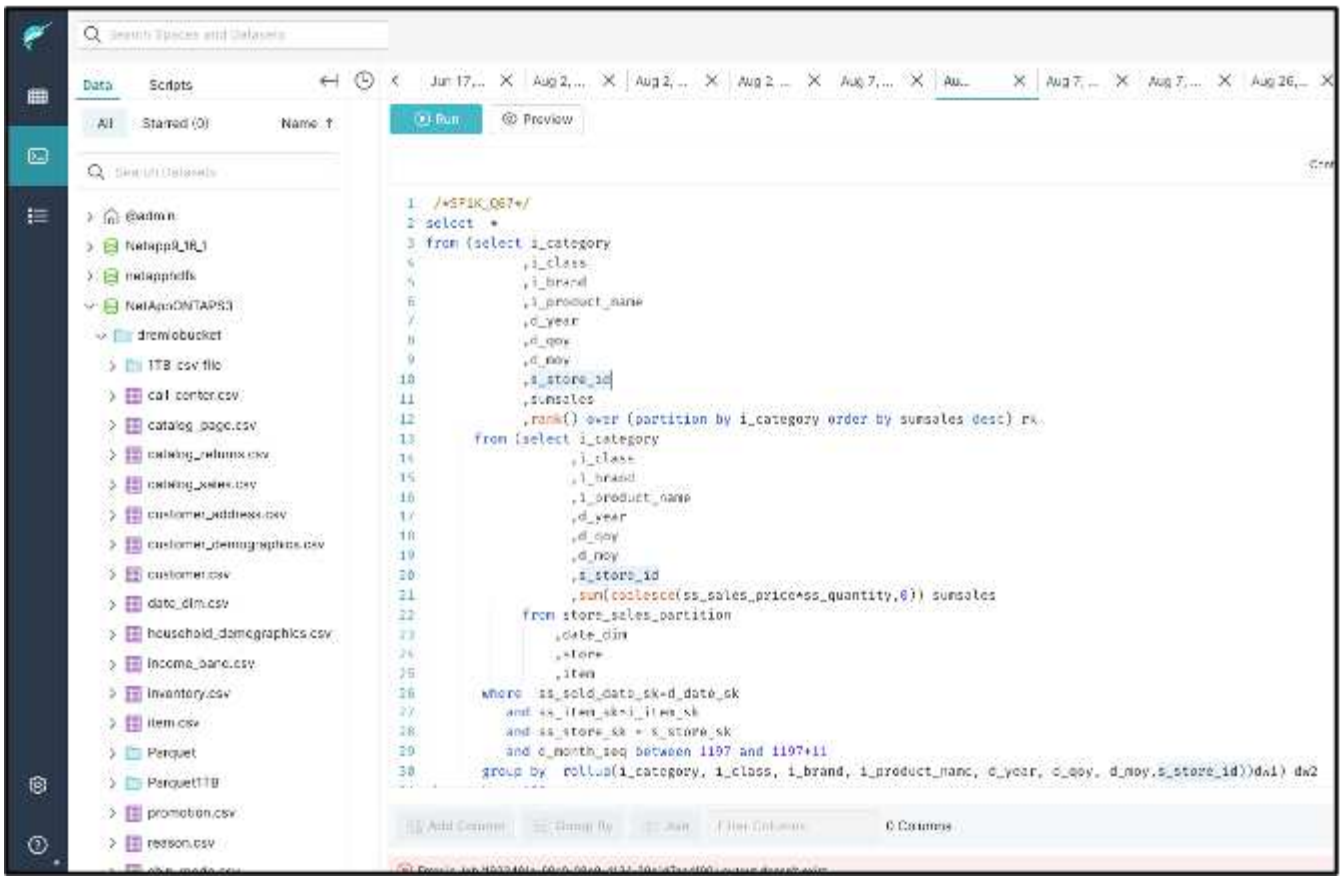
3. Dremio 스필오버 위치를 NetApp NFS 스토리지로 지정합니다.

```
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:~# ls -ltrh /opt/dremio/dremiocache/
total 8.0K
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
Filesystem                                Size  Used Avail Use% Mounted on
10.63.150.159:/dremiocache_hadoopnode1  2.1T  209M  2.0T   1%
/dremiocache
root@hadoopnode1:~#
```

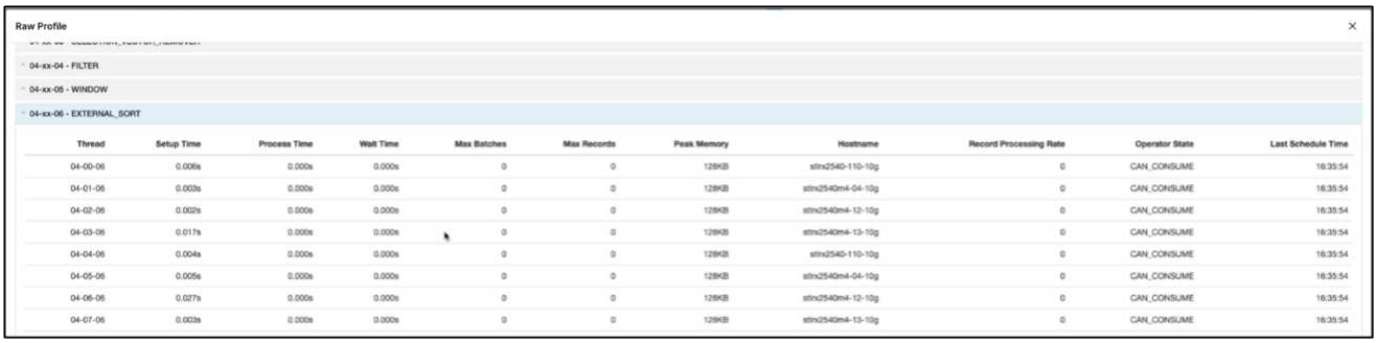
4. 맥락을 선택하세요. 우리 테스트에서는 ONTAP S3에 있는 TPCDS에서 생성된 Parquet 파일을 대상으로 테스트를 실행했습니다. Dremio 대시보드 → SQL 러너 → 컨텍스트 → NetAppONTAPS3→Parquet1TB



1. Dremio 대시보드에서 TPC-DS 쿼리67을 실행합니다.



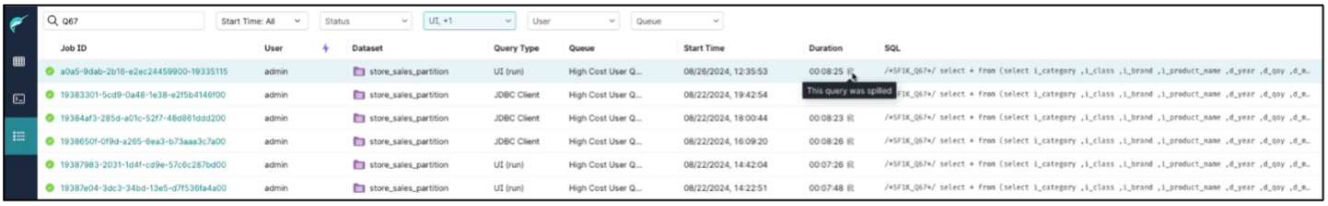
1. 모든 실행자에서 작업이 실행 중인지 확인하세요. Dremio 대시보드 → 작업 → <작업 ID> → 원시 프로필 → EXTERNAL_SORT 선택 → 호스트 이름



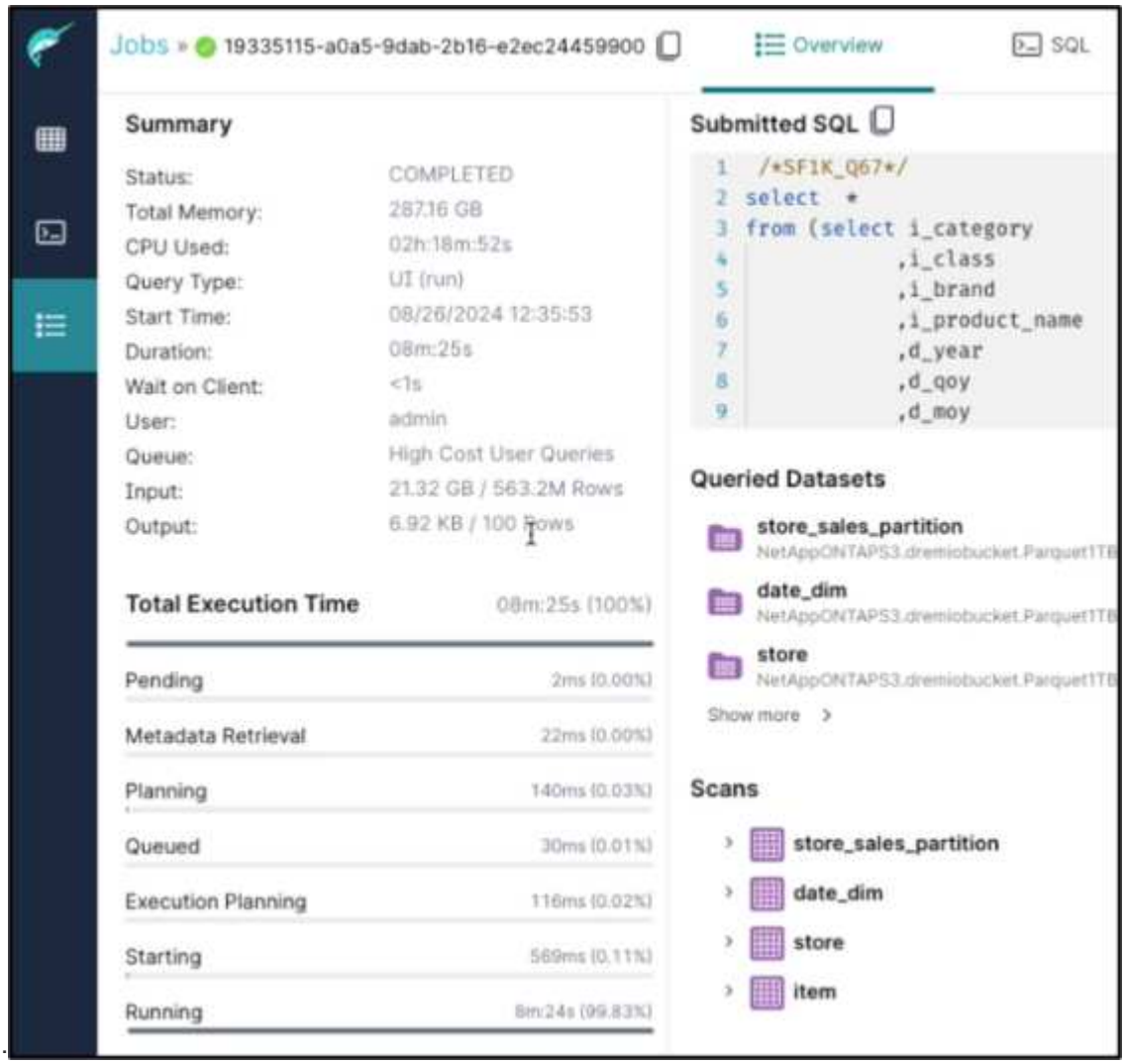
1. SQL 쿼리가 실행 중일 때 NetApp 스토리지 컨트롤러에서 데이터 캐싱을 위한 분할 폴더를 확인할 수 있습니다.

```
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:~# ls -ltrh /dremiocache/spilling_stlrx2540m4-12-
10g_45678/
total 4.0K
drwxr-xr-x 2 root daemon 4.0K Sep 13 16:23 1726243167416
```

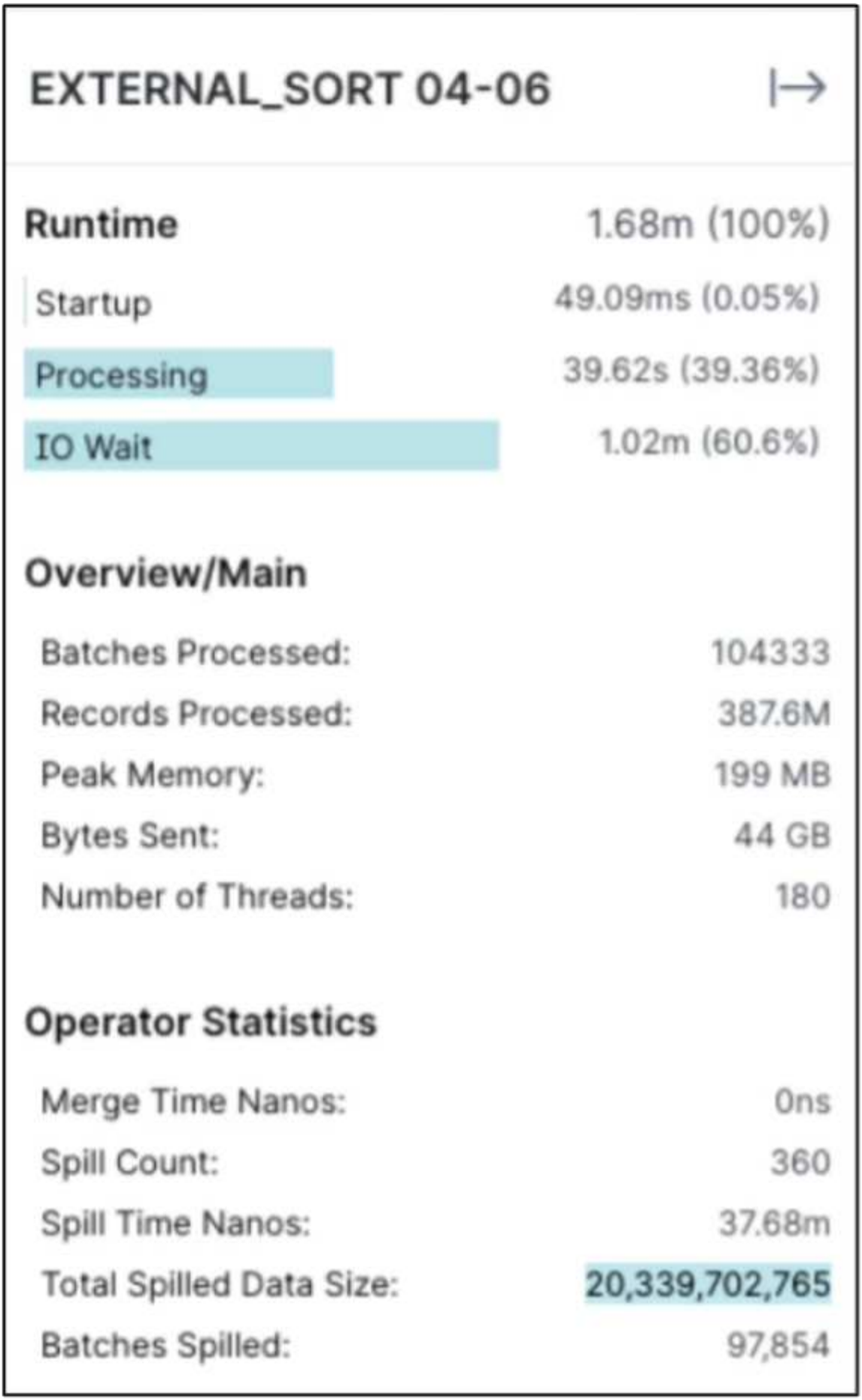
2. 스필오버로 완료된 SQL
쿼리



3. 작업 완료 요약.



4. 유출된 데이터 크기를

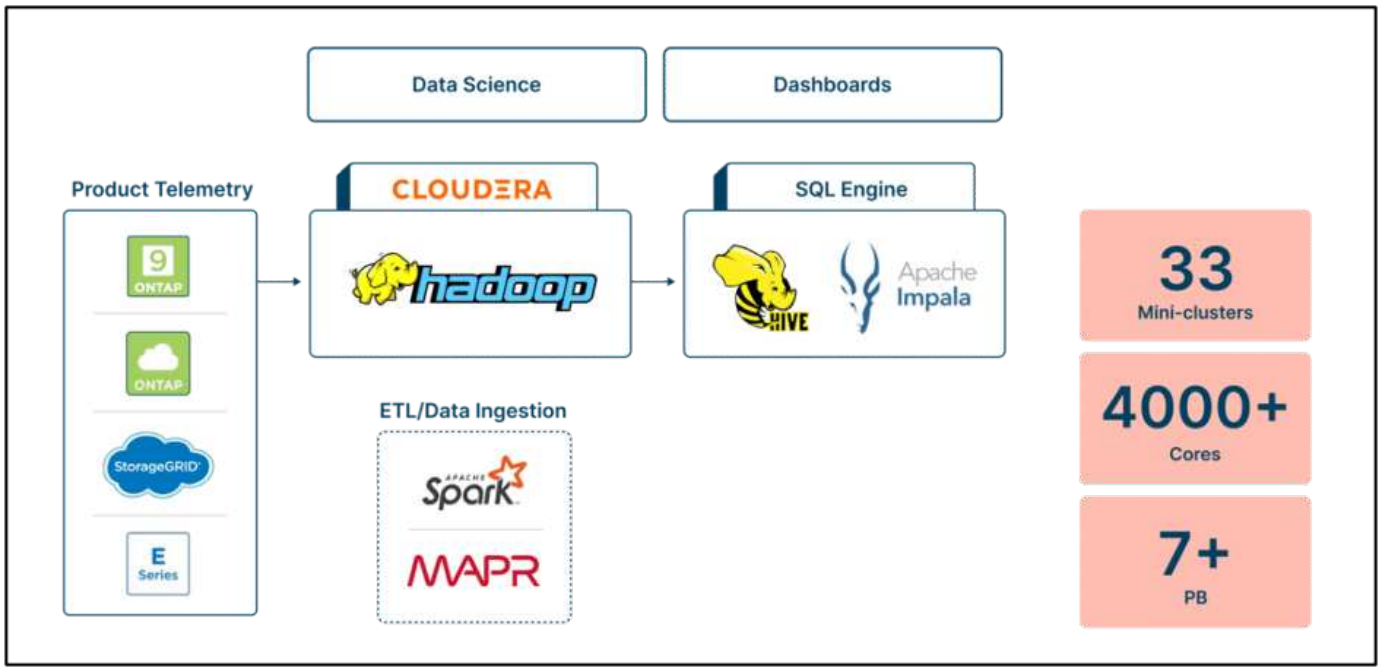

EXTERNAL_SORT 04-06
Runtime
1.68m (100%)
Startup
49.09ms (0.05%)
Processing
39.62s (39.36%)
IO Wait
1.02m (60.6%)
Overview/Main
Batches Processed:
104333
Records Processed:
387.6M
Peak Memory:
199 MB
Bytes Sent:
44 GB
Number of Threads:
180
Operator Statistics
Merge Time Nanos:
0ns
Spill Count:
360
Spill Time Nanos:
37.68m
Total Spilled Data Size:
20,339,702,765
Batches Spilled:
97,854

확인하세요

NAS 및 StorageGRID Object Storage에도 동일한 절차가 적용됩니다.

고객 사용 사례

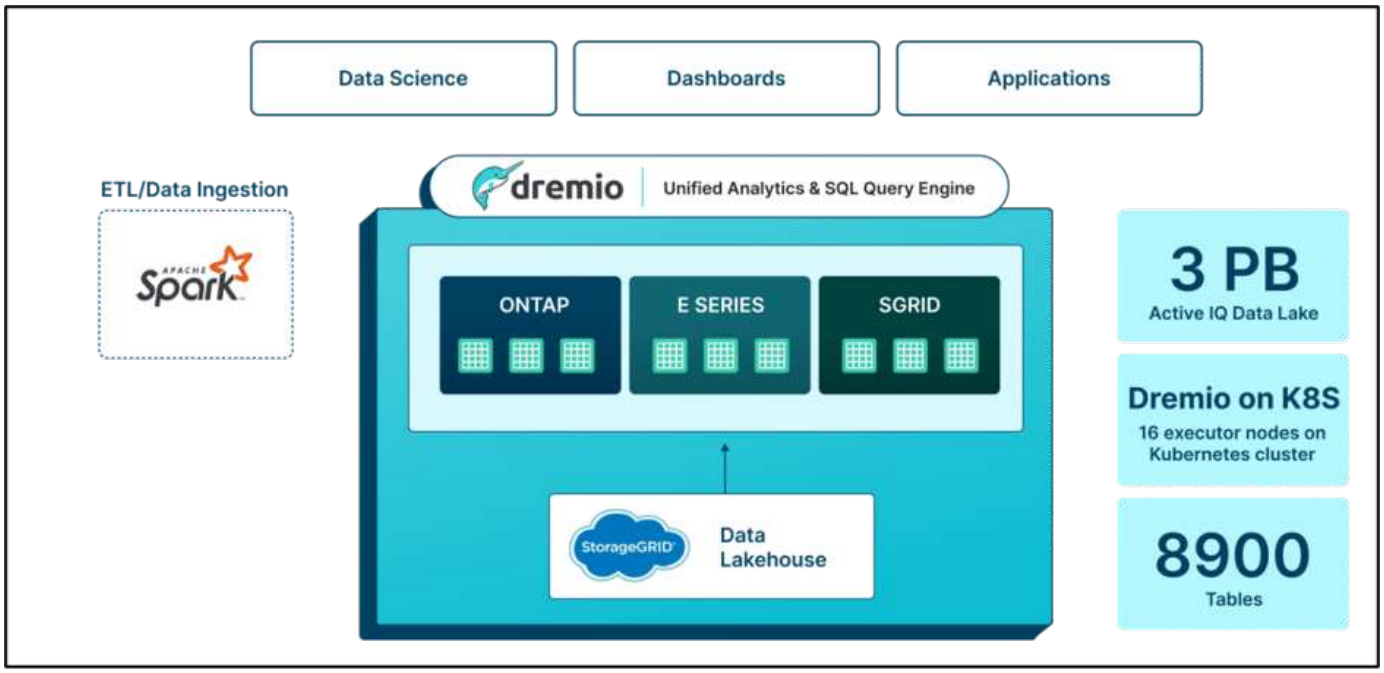
NetApp ActiveIQ 사용 사례



과제: NetApp의 자체 Active IQ 솔루션은 원래 다양한 사용 사례를 지원하도록 설계되었지만, 이제는 내부 사용자와 고객 모두를 위한 포괄적인 솔루션으로 발전했습니다. 그러나 데이터의 급속한 성장과 효율적인 데이터 액세스의 필요성으로 인해 Hadoop/MapR 기반 백엔드 인프라는 비용과 성능 측면에서 과제를 안고 있었습니다. 저장소를 확장하면 불필요한 컴퓨팅 리소스가 추가되어 비용이 증가하게 됩니다.

게다가 Hadoop 클러스터를 관리하는 일은 시간이 많이 걸리고 전문적인 지식이 필요했습니다. 데이터 성능 및 관리 문제로 인해 상황이 더욱 복잡해졌으며, 쿼리를 실행하는 데 평균 45분이 걸리고 구성 오류로 인해 리소스가 부족해졌습니다. 이러한 과제를 해결하기 위해 NetApp 기존의 레거시 Hadoop 환경에 대한 대안을 모색했고 Dremio를 기반으로 하는 새로운 현대적 솔루션이 비용을 절감하고, 스토리지와 컴퓨팅을 분리하고, 성능을 개선하고, 데이터 관리를 간소화하고, 세분화된 제어 기능을 제공하고, 재해 복구 기능을 제공할 것이라고 판단했습니다.

해결책:



Dremio는 NetApp 단계적 접근 방식으로 Hadoop 기반 데이터 인프라를 현대화할 수 있도록 지원하여 통합 분석을 위한 로드맵을 제공했습니다. 데이터 처리를 크게 변경해야 했던 다른 공급업체와 달리 Dremio는 기존 파이프라인과 완벽하게 통합되어 마이그레이션하는 동안 시간과 비용을 절감했습니다. NetApp 완전히 컨테이너화된 환경으로 전환함으로써 관리 오버헤드를 줄이고, 보안을 강화하고, 복원력을 향상시켰습니다. Dremio는 Apache Iceberg와 Arrow와 같은 개방형 생태계를 도입하여 미래 지향적인 운영, 투명성, 확장성을 확보했습니다.

Dremio는 Hadoop/Hive 인프라를 대체하여 의미 계층을 통해 2차 사용 사례에 대한 기능을 제공했습니다. 기존의 Spark 기반 ETL 및 데이터 수집 메커니즘은 그대로 유지하면서 Dremio는 중복 없이 데이터를 보다 쉽게 검색하고 탐색할 수 있는 통합 액세스 계층을 제공했습니다. 이러한 접근 방식은 데이터 복제 요소를 크게 줄이고 스토리지와 컴퓨팅을 분리했습니다.

이점: NetApp Dremio를 통해 데이터 환경에서 컴퓨팅 소비와 디스크 공간 요구 사항을 최소화하여 상당한 비용 절감을 달성했습니다. 새로운 Active IQ Data Lake는 7페타바이트가 넘는 기존 인프라에 비해 3페타바이트의 데이터를 보유한 8,900개의 테이블로 구성됩니다. Dremio로의 마이그레이션에는 Kubernetes 클러스터에서 33개의 미니 클러스터와 4,000개의 코어를 16개의 실행자 노드로 전환하는 작업도 포함되었습니다. 컴퓨팅 리소스가 크게 감소했음에도 불구하고 NetApp 놀라운 성능 향상을 경험했습니다. Dremio를 통해 데이터에 직접 액세스함으로써 쿼리 런타임이 45분에서 2분으로 단축되었고, 그 결과 예측 유지 관리 및 최적화를 위한 통찰력을 얻는 시간이 95% 빨라졌습니다. 마이그레이션을 통해 컴퓨팅 비용이 60% 이상 절감되고, 쿼리 속도가 20배 이상 빨라졌으며, 총 소유 비용(TCO)도 30% 이상 절감되었습니다.

자동차 부품 판매 고객 활용 사례.

과제: 이 글로벌 자동차 부품 판매 회사 내에서 임원 및 기업 재무 계획 및 분석 그룹은 판매 보고서에 대한 통합된 보기를 얻지 못하고 각 사업부 판매 지표 보고서를 읽고 이를 통합하려고 시도해야 했습니다. 이로 인해 고객은 적어도 하루 이상 지난 데이터를 바탕으로 결정을 내리는 경우가 많아졌습니다. 새로운 분석 통찰력을 얻는 데 걸리는 리드타임은 일반적으로 4주 이상 걸립니다. 데이터 파이프라인의 문제를 해결하려면 더 많은 시간이 필요하며, 이미 긴 일정에 3일 이상이 추가될 것입니다. 보고서 개발 프로세스가 느리고 보고서 성능도 좋지 않아 분석가 커뮤니티는 데이터가 처리되거나 로드될 때까지 계속 기다려야 했고, 그 결과 새로운 비즈니스 통찰력을 찾고 새로운 비즈니스 행동을 촉진하는 데 어려움을 겪었습니다. 이러한 문제가 있는 환경은 다양한 사업 부문을 위한 수많은 데이터베이스로 구성되어 있었고, 이로 인해 수많은 데이터 사일로가 생겨났습니다. 느리고 단편화된 환경은 분석가가 단일 진실 소스에 의존하는 대신, 각자의 진실 버전을 도출해낼 방법이 너무 많았기 때문에 데이터 거버넌스를 복잡하게 만들었습니다. 이 접근 방식에는 데이터 플랫폼 비용과 인건비로 190만 달러가 넘게 들었습니다. 기존 플랫폼을 유지 관리하고 데이터 요청을 충족하려면 연간 7명의 현장 기술 엔지니어(FTE)가 필요했습니다. 데이터 요청이 증가함에 따라 데이터 인텔리전스 팀은 향후 요구 사항을 충족하기 위해 레거시 환경을 확장할 수 없었습니다.

해결책: NetApp Object Store에서 대규모 Iceberg 테이블을 비용 효율적으로 저장하고 관리합니다. Dremio의 의미 계층을 사용하여 데이터 도메인을 구축하면 비즈니스 사용자가 쉽게 데이터 제품을 만들고, 검색하고, 공유할 수 있습니다.

고객 혜택: • 기존 데이터 아키텍처 개선 및 최적화, 통찰력 확보 시간 4주에서 단 몇 시간으로 단축 • 문제 해결 시간 3일에서 단 몇 시간으로 단축 • 데이터 플랫폼 및 관리 비용 38만 달러 이상 절감 • 연간 데이터 인텔리전스 작업 FTE 2명 절감

결론

결론적으로 이 기술 보고서는 ONTAP S3, NAS, StorageGRID 포함한 NetApp 스토리지 컨트롤러의 다양한 데이터 소스와 함께 Dremio를 사용한 q Hybrid Iceberg Lakehouse의 포괄적인 배포 세부 정보를 제공했습니다. 배포 프로세스가 성공적으로 실행되었으며, TPC-DS 벤치마킹 도구를 활용하여 다양한 데이터 소스에서 99개의 SQL 쿼리를 수행했습니다. 이 보고서에서는 NetApp 의 고객 사용 사례도 살펴보았고, 이를 통해 다양한 비즈니스 요구 사항을 충족하는 데 있어 Dremio의 다재다능함과 효과성을 입증했습니다. 또한 자동차 부품 판매 고객을 포함한 특정 사용 사례를 조사하여 데이터 분석 및 통찰력을 위해 Dremio를 활용하는 실제적 적용과 이점을 강조했습니다.

전반적으로 이 문서는 NetApp 스토리지 컨트롤러와 함께 Dremio를 배포하고 사용하는 방법을 이해하는 데 귀중한 자료가 되며, 다양한 산업에서 데이터 중심 의사 결정과 최적화를 추진하는 기능과 잠재력을 보여줍니다.

추가 정보를 찾을 수 있는 곳

이 문서에 설명된 정보에 대해 자세히 알아보려면 다음 문서 및/또는 웹사이트를 검토하세요.

- Zookeeper 설치

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

- 드레미오

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

- storageGRID를 사용하여 Dremio 구성

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

- NetApp 사용 사례

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

저작권 정보

Copyright © 2025 NetApp, Inc. All Rights Reserved. 미국에서 인쇄됨 본 문서의 어떠한 부분도 저작권 소유자의 사전 서면 승인 없이는 어떠한 형식이나 수단(복사, 녹음, 녹화 또는 전자 검색 시스템에 저장하는 것을 비롯한 그래픽, 전자적 또는 기계적 방법)으로도 복제될 수 없습니다.

NetApp이 저작권을 가진 자료에 있는 소프트웨어에는 아래의 라이센스와 고지사항이 적용됩니다.

본 소프트웨어는 NetApp에 의해 '있는 그대로' 제공되며 상품성 및 특정 목적에의 적합성에 대한 명시적 또는 묵시적 보증을 포함하여(이에 제한되지 않음) 어떠한 보증도 하지 않습니다. NetApp은 대체품 또는 대체 서비스의 조달, 사용 불능, 데이터 손실, 이익 손실, 영업 중단을 포함하여(이에 국한되지 않음), 이 소프트웨어의 사용으로 인해 발생하는 모든 직접 및 간접 손해, 우발적 손해, 특별 손해, 징벌적 손해, 결과적 손해의 발생에 대하여 그 발생 이유, 책임론, 계약 여부, 엄격한 책임, 불법 행위(과실 또는 그렇지 않은 경우)와 관계없이 어떠한 책임도 지지 않으며, 이와 같은 손실의 발생 가능성이 통지되었다 하더라도 마찬가지입니다.

NetApp은 본 문서에 설명된 제품을 언제든지 예고 없이 변경할 권리를 보유합니다. NetApp은 NetApp의 명시적인 서면 동의를 받은 경우를 제외하고 본 문서에 설명된 제품을 사용하여 발생하는 어떠한 문제에도 책임을 지지 않습니다. 본 제품의 사용 또는 구매의 경우 NetApp에서는 어떠한 특허권, 상표권 또는 기타 지적 재산권이 적용되는 라이센스도 제공하지 않습니다.

본 설명서에 설명된 제품은 하나 이상의 미국 특허, 해외 특허 또는 출원 중인 특허로 보호됩니다.

제한적 권리 표시: 정부에 의한 사용, 복제 또는 공개에는 DFARS 252.227-7013(2014년 2월) 및 FAR 52.227-19(2007년 12월)의 기술 데이터-비상업적 품목에 대한 권리(Rights in Technical Data -Noncommercial Items) 조항의 하위 조항 (b)(3)에 설명된 제한사항이 적용됩니다.

여기에 포함된 데이터는 상업용 제품 및/또는 상업용 서비스(FAR 2.101에 정의)에 해당하며 NetApp, Inc.의 독점 자산입니다. 본 계약에 따라 제공되는 모든 NetApp 기술 데이터 및 컴퓨터 소프트웨어는 본질적으로 상업용이며 개인 비용만으로 개발되었습니다. 미국 정부는 데이터가 제공된 미국 계약과 관련하여 해당 계약을 지원하는 데에만 데이터에 대한 전 세계적으로 비독점적이고 양도할 수 없으며 재사용이 불가능하며 취소 불가능한 라이센스를 제한적으로 가집니다. 여기에 제공된 경우를 제외하고 NetApp, Inc.의 사전 서면 승인 없이는 이 데이터를 사용, 공개, 재생산, 수정, 수행 또는 표시할 수 없습니다. 미국 국방부에 대한 정부 라이센스는 DFARS 조항 252.227-7015(b)(2014년 2월)에 명시된 권한으로 제한됩니다.

상표 정보

NETAPP, NETAPP 로고 및 http://www.netapp.com/TM에 나열된 마크는 NetApp, Inc.의 상표입니다. 기타 회사 및 제품 이름은 해당 소유자의 상표일 수 있습니다.