Solução híbrida iceberg lakehouse de próxima geração da NetApp e Dremio

NetApp artificial intelligence solutions

NetApp
August 18, 2025

Índice

Solução híbrida iceberg lakehouse de próxima geração da NetApp e Dremio

A solução Iceberg Lakehouse híbrida de última geração da NetApp e Dremio

Neste documento, discutimos os detalhes de implantação do Dremio com diferentes fontes de dados de controladores de armazenamento NetApp , incluindo ONTAP S3, NAS e StorageGRID. Durante a implantação, usamos a ferramenta de benchmarking TPC-DS para executar 99 consultas SQL em várias fontes. O documento também explora casos de uso de clientes dentro da NetApp, bem como um caso de uso envolvendo um cliente de vendas de peças automotivas.

Visão geral da solução

A solução Hybrid Iceberg Lakehouse oferece benefícios exclusivos para abordar os desafios enfrentados pelos clientes do data lake. Ao aproveitar a plataforma Dremio Unified Lakehouse e as soluções NetApp ONTAP, StorageGRID e NetApp Cloud, as empresas podem agregar valor significativo às suas operações comerciais. A solução não apenas fornece acesso a várias fontes de dados, incluindo fontes da NetApp , mas também melhora o desempenho analítico geral e ajuda as empresas a gerar insights de negócios que levam ao crescimento dos negócios.

Visão geral da NetApp

- As ofertas da NetApp, como ONTAP e StorageGRID, permitem a separação de armazenamento e computação, possibilitando a utilização ideal de recursos com base em requisitos específicos. Essa flexibilidade permite que os clientes dimensionem seu armazenamento de forma independente usando soluções de armazenamento da NetApp
- Ao aproveitar os controladores de armazenamento da NetApp, os clientes podem fornecer dados de forma eficiente para seu banco de dados vetorial usando os protocolos NFS e S3. Esses protocolos facilitam o armazenamento de dados do cliente e gerenciam o índice do banco de dados vetorial, eliminando a necessidade de múltiplas cópias de dados acessadas por meio de métodos de arquivo e objeto.
- O NetApp ONTAP fornece suporte nativo para NAS e armazenamento de objetos nos principais provedores de serviços de nuvem, como AWS, Azure e Google Cloud. Essa ampla compatibilidade garante integração perfeita, permitindo mobilidade de dados do cliente, acessibilidade global, recuperação de desastres, escalabilidade dinâmica e alto desempenho.

StorageGRID

Nosso armazenamento de objetos líder do setor, o storageGRID, oferece um poderoso mecanismo de política para posicionamento automatizado de dados, opções flexíveis de implantação e durabilidade incomparável com codificação de eliminação em camadas. Ele tem uma arquitetura escalável que suporta bilhões de objetos e petabytes de dados em um único namespace. A solução permite a integração de nuvem híbrida, permitindo a hierarquização de dados nas principais plataformas de nuvem. Ela foi reconhecida como líder na Avaliação de Fornecedores Baseada em Objetos do IDC Marketscape Worldwide de 2019.

Além disso, o storageGRID se destaca no gerenciamento de dados não estruturados em escala com armazenamento de objetos definidos por software, redundância geográfica e recursos multisite. Ele incorpora gerenciamento de ciclo de vida de informações baseado em políticas e oferece recursos de integração em nuvem, como espelhamento e pesquisa. Possui diversas certificações, incluindo Common Criteria, NF203 Digital Safe Component, ISO/IEC 25051, KPMG e Cohasset Compliance Assessment.

Em resumo, o NetApp storageGRID oferece recursos poderosos, escalabilidade, integração de nuvem híbrida e certificações de conformidade para gerenciamento eficiente de dados não estruturados em escala.

NetApp ONTAP

O NetApp ONTAP é uma solução de armazenamento robusta que oferece uma ampla gama de recursos empresariais. Inclui o Snapshot, que fornece backups instantâneos consistentes com o aplicativo e à prova de violação. O SnapRestore permite restauração quase instantânea de backups sob demanda, enquanto o SnapMirror oferece recursos integrados de backup remoto e recuperação de desastres. A solução também incorpora a Proteção Autônoma contra Ransomware (ARP), garantindo a segurança dos dados com recursos como verificação de vários administradores, criptografia de dados em repouso com certificação FIPS, criptografia de dados em trânsito, autenticação multifator (MFA) e controle de acesso baseado em função (RBAC). Registro abrangente, auditoria, gerenciamento de chaves internas e externas, limpeza segura e gerenciamento seguro de múltiplos locatários aprimoram ainda mais a segurança e a conformidade dos dados.

O NetApp ONTAP também conta com o SnapLock, que fornece retenção de dados em conformidade com regulamentações, com altos níveis de integridade, desempenho e retenção a um baixo custo total de propriedade. Ele é totalmente integrado ao NetApp ONTAP 9 e oferece proteção contra atos maliciosos, administradores desonestos e ransomware.

A solução abrange criptografia NSE/NVE para criptografia de dados em voo e em repouso, acesso de administrador multifator e verificação de vários administradores. O Active IQ fornece análises preditivas e ações corretivas baseadas em IA, enquanto o QoS garante o controle da carga de trabalho de qualidade do serviço. A integração de gerenciamento e automação é intuitiva por meio de SysMgr/GUI/CLI/API. O FabricPool permite a hierarquização automática de dados, e a solução oferece eficiência por meio de compactação, desduplicação e compactação de dados em linha. A NetApp garante o cumprimento das metas de eficiência da carga de trabalho sem nenhum custo para o cliente.

O NetApp ONTAP oferece suporte a vários protocolos, incluindo NVMe/FC, FC, NVMe/TCP, iSCSI, NFS, SMB e S3, o que o torna uma solução de armazenamento unificada. No geral, o NetApp ONTAP oferece amplos recursos empresariais, segurança robusta, conformidade, eficiência e versatilidade para atender a diversas necessidades de armazenamento.

Visão geral do Dremio

Dremio é a plataforma Lakehouse unificada para análises de autoatendimento e IA. A plataforma Dremio Unified Analytics aproxima os usuários dos dados com flexibilidade, escalabilidade e desempenho do lakehouse por uma fração do custo das soluções de data warehouse legadas. O Dremio permite análises "shift-left" para eliminar a integração de dados complexa e dispendiosa e ETL, proporcionando análises contínuas em escala empresarial sem movimentação de dados. O Dremio também apresenta:

- Análises de autoatendimento fáceis de usar, possibilitadas por uma camada semântica universal e um mecanismo de consulta SQL altamente integrado e de alto desempenho, facilitando a conexão, o controle e a análise de todos os dados, tanto na nuvem quanto no local.
- Os recursos de gerenciamento de lakehouse nativos do Apache Iceberg do Dremio simplificam a descoberta de dados e automatizam a otimização de dados, oferecendo análises de alto desempenho com controle de versão de dados inspirado no Git.

- Baseado em código aberto e padrões abertos, o Dremio permite que as empresas evitem a dependência e permaneçam posicionadas para a inovação. Empresas corporativas confiam no Dremio como a plataforma lakehouse mais fácil de usar, com o melhor custo-benefício em todas as cargas de trabalho.

Que valor a solução Dremio e NetApp Hybrid Iceberg Lakehouse oferece aos clientes?

- **Gerenciamento de dados e acessibilidade aprimorados**: A Dremio é conhecida por sua plataforma de data lakehouse que permite que organizações consultem dados diretamente de seus data lakes em alta velocidade. A NetApp, por outro lado, é uma provedora líder de serviços de dados em nuvem e soluções de armazenamento de dados. A oferta conjunta fornece aos clientes uma solução abrangente para armazenar, gerenciar, acessar e analisar os dados de sua empresa de forma eficiente e eficiente.
- **Otimização de desempenho**: Com a experiência da NetApp em armazenamento de dados e os recursos da Dremio em processamento e otimização de dados, a parceria oferece uma solução que melhora o desempenho das operações de dados, reduz a latência e aumenta a velocidade do insight de negócios. O Dremio até mesmo trouxe benefícios de desempenho para a infraestrutura analítica de TI interna da NetApp.
- **Escalabilidade**: Tanto o Dremio quanto o NetApp oferecem uma solução projetada para escalar. A solução conjunta fornece aos clientes ambientes de armazenamento, gerenciamento de dados e análise de dados altamente escaláveis. Em um ambiente Hybrid Iceberg Lakehouse, o mecanismo de consulta Dremio SQL emparelhado com o NetApp StorageGRID oferece escalabilidade, simultaneidade e desempenho de consulta incomparáveis, capaz de lidar com as necessidades analíticas de qualquer negócio.
- **Segurança e governança de dados**: Ambas as empresas têm um forte foco em segurança e governança de dados. Juntos, eles oferecem recursos robustos de segurança e governança de dados, garantindo que os dados sejam protegidos e que os requisitos de governança de dados sejam atendidos. Recursos como controles de acesso detalhados e baseados em funções, auditoria abrangente, linhagem de dados de ponta a ponta, gerenciamento unificado de identidade e SSO com uma ampla estrutura de conformidade e segurança garantem que os ambientes de dados analíticos das empresas sejam seguros e governados.
- **Eficiência de custos**: Ao integrar o mecanismo de data lake da Dremio com as soluções de armazenamento da NetApp, os clientes podem reduzir os custos associados ao gerenciamento e à movimentação de dados. As organizações também podem migrar de ambientes de data lake legados para uma solução de lakehouse mais moderna, composta por NetApp e Dremio. Esta solução Hybrid Iceberg Lakehouse oferece desempenho de consulta de alta velocidade e simultaneidade de consulta líder de mercado, o que reduz o TCO e o tempo para obter insights de negócios.

Requisitos de tecnologia

As configurações de hardware e software descritas abaixo foram utilizadas para validações realizadas neste documento. Essas configurações servem como um guia para ajudar você a configurar seu ambiente. No entanto, observe que os componentes específicos podem variar dependendo dos requisitos individuais do cliente.

Requisitos de hardware

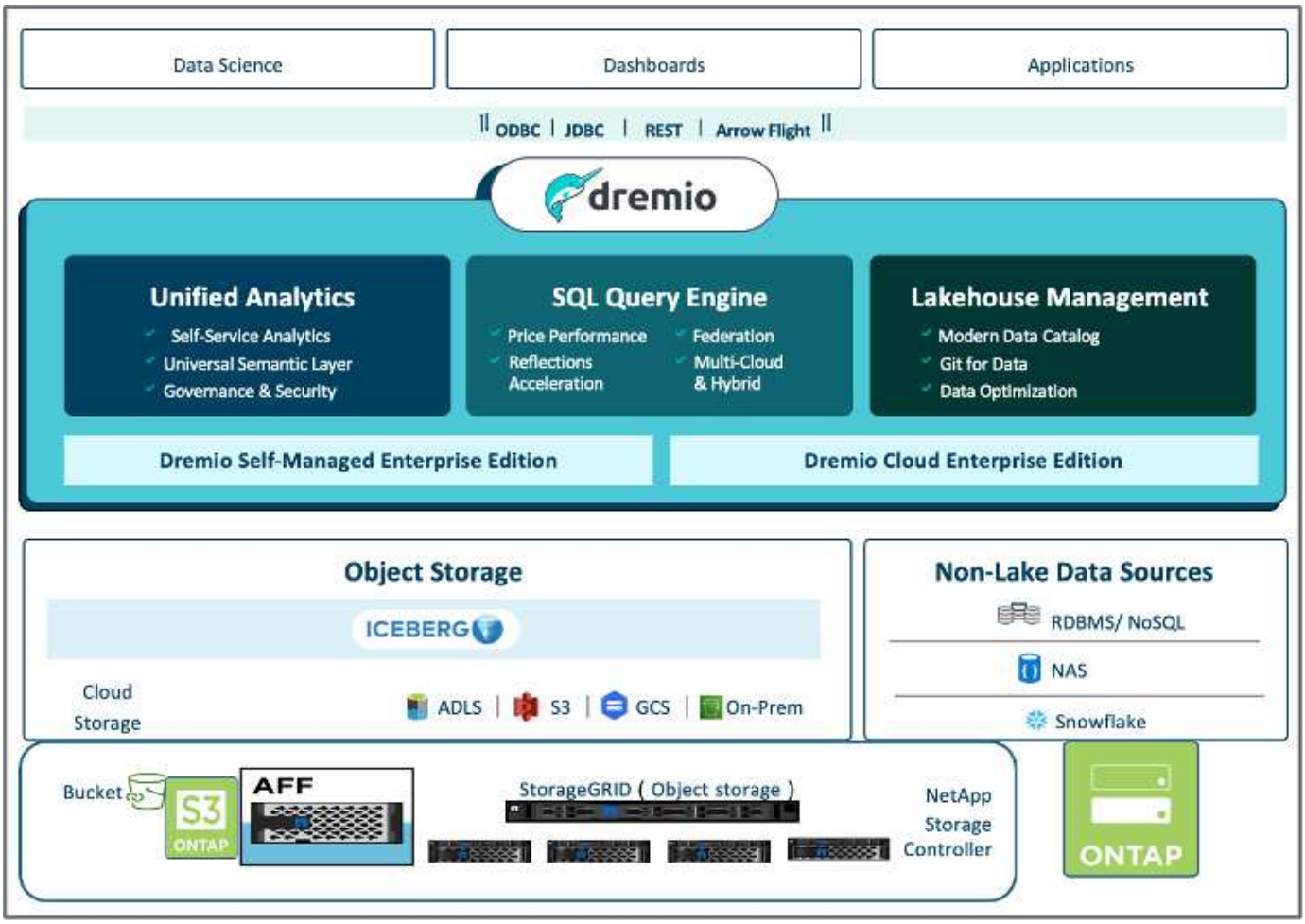
Hardware	Detalhes
Par de matriz de armazenamento AFF da NetApp HA	• A800 • ONTAP 9.14.1 • 48 SSDs NVM de 3,49 TB • Dois buckets S3: metadados do Dremio e dados do cliente.
4 x FUJITSU PRIMERGY RX2540 M4	• 64 CPUs • CPU Intel Xeon Gold 6142 a 2,60 GHz • 256 GM de memória física • 1 porta de rede 100GbE
Rede	• 100 GbE
StorageGRID	* 1 x SG100, 3xSGF6024 * 3 x 24 x 7,68 TB * Dois buckets S3: metadados Dremio e dados do cliente.

Requisitos de software

Software	Detalhes
Dremio	• versão - 25.0.3-202405170357270647-d2042e1b • Edição Enterprise
No local	• Cluster Dremio de 5 nós • 1 coordenador mestre e 4 executores

Procedimento de Implantação

Nesta validação da arquitetura de referência, utilizamos uma configuração Dremio composta por um coordenador e quatro executores



Configuração do NetApp

- Inicialização do sistema de armazenamento
- Criação de máquina virtual de armazenamento (SVM)
- Atribuição de interfaces de rede lógicas
- NFS, configuração e licenciamento S3

Siga os passos abaixo para NFS (Network File System): 1. Crie um volume Flex Group para NFSv4 ou NFSv3. Em nossa configuração para esta validação, usamos 48 SSDs, 1 SSD dedicado ao volume raiz do controlador e 47 SSDs distribuídos para NFSv4. Verifique se a política de exportação NFS para o volume do Flex Group tem permissões de leitura/gravação para a rede de servidores Dremio.

1. Em todos os servidores Dremio, crie uma pasta e monte o volume do Flex Group nessa pasta por meio de uma Interface Lógica (LIF) em cada servidor Dremio.

Siga as etapas abaixo para o S3 (Serviço de Armazenamento Simples):

1. Configure um servidor de armazenamento de objetos com HTTP habilitado e o status do administrador definido como 'ativo' usando o comando "vserver object-store-server create". Você tem a opção de habilitar HTTPS e definir uma porta de escuta personalizada.
2. Crie um usuário object-store-server usando o comando "vserver object-store-server user create -user <nome de usuário>".
3. Para obter a chave de acesso e a chave secreta, você pode executar o seguinte comando: "set diag;

vserver object-store-server user show -user <nome de usuário>". No entanto, a partir de agora, essas chaves serão fornecidas durante o processo de criação do usuário ou poderão ser recuperadas usando chamadas de API REST.

4. Estabeleça um grupo de objetos-armazenamento-servidor usando o usuário criado na etapa 2 e conceda acesso. Neste exemplo, fornecemos "FullAccess".
5. Crie dois buckets S3 definindo seu tipo como "S3". Um para configuração do Dremio e um para dados do cliente.

Configuração do tratador de zoológico

Você pode usar a configuração do zookeeper fornecida pelo Dremio. Nesta validação, usamos um zookeeper separado. Seguimos os passos mencionados neste link da web https://medium.com/@ahmetfurkandemir/distributed-hadoop-cluster-1-spark-with-all-dependincies-03c8ec616166

Configuração do Dremio

Seguimos este link para instalar o Dremio via tar ball.

1. Crie um grupo Dremio.

```
sudo groupadd -r dremio
```

2. Crie um usuário dremio.

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

3. Crie diretórios Dremio.

```
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. Baixe o arquivo tar de https://download.dremio.com/community-server/
5. Descompacte o Dremio no diretório /opt/dremio.

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

6. Crie um link simbólico para a pasta de configuração.

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

7. Configure sua configuração de serviço (configuração do SystemD).
   a. Copie o arquivo de unidade do daemon dremio de /opt/dremio/share/dremio.service para /etc/systemd/system/dremio.service.
   b. Reiniciar o sistema

```
sudo systemctl daemon-reload
```

   c. Habilitar o dremio para iniciar na inicialização.

```
sudo systemctl enable dremio
```

8. Configurar o Dremio no coordenador. Veja a configuração do Dremio para mais informações
   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#
```

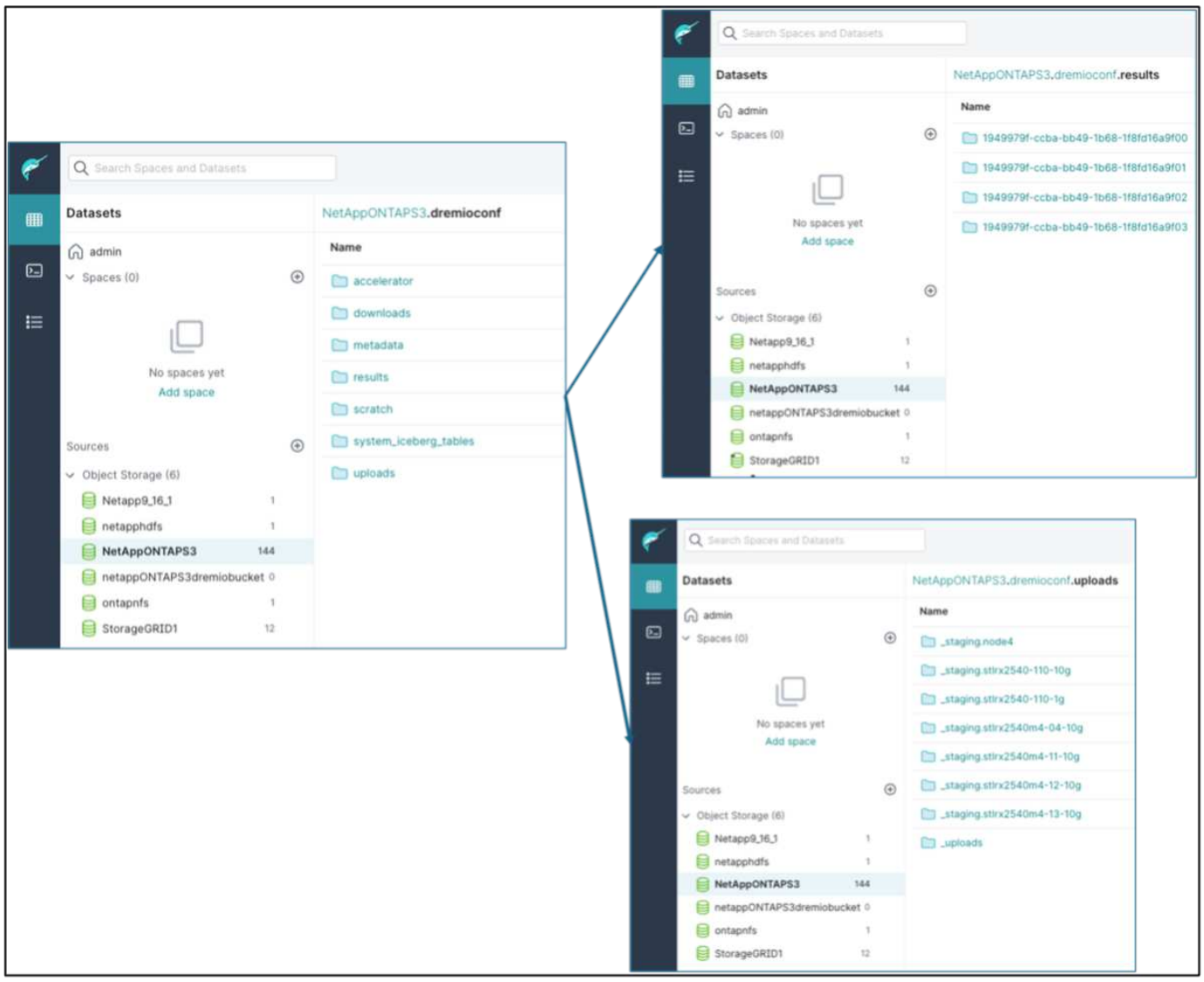
   b. Core-site.xml

```
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. A configuração do Dremio é armazenada no armazenamento de objetos do NetApp . Em nossa validação, o bucket "dremioconf" reside em um bucket ontap S3. A imagem abaixo mostra alguns detalhes das pastas "scratch" e "uploads" do bucket S3 "dremioconf".



1. Configurar o Dremio nos executores. Em nossa configuração, temos 3 executores.
    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 – o mesmo que a configuração do coordenador.

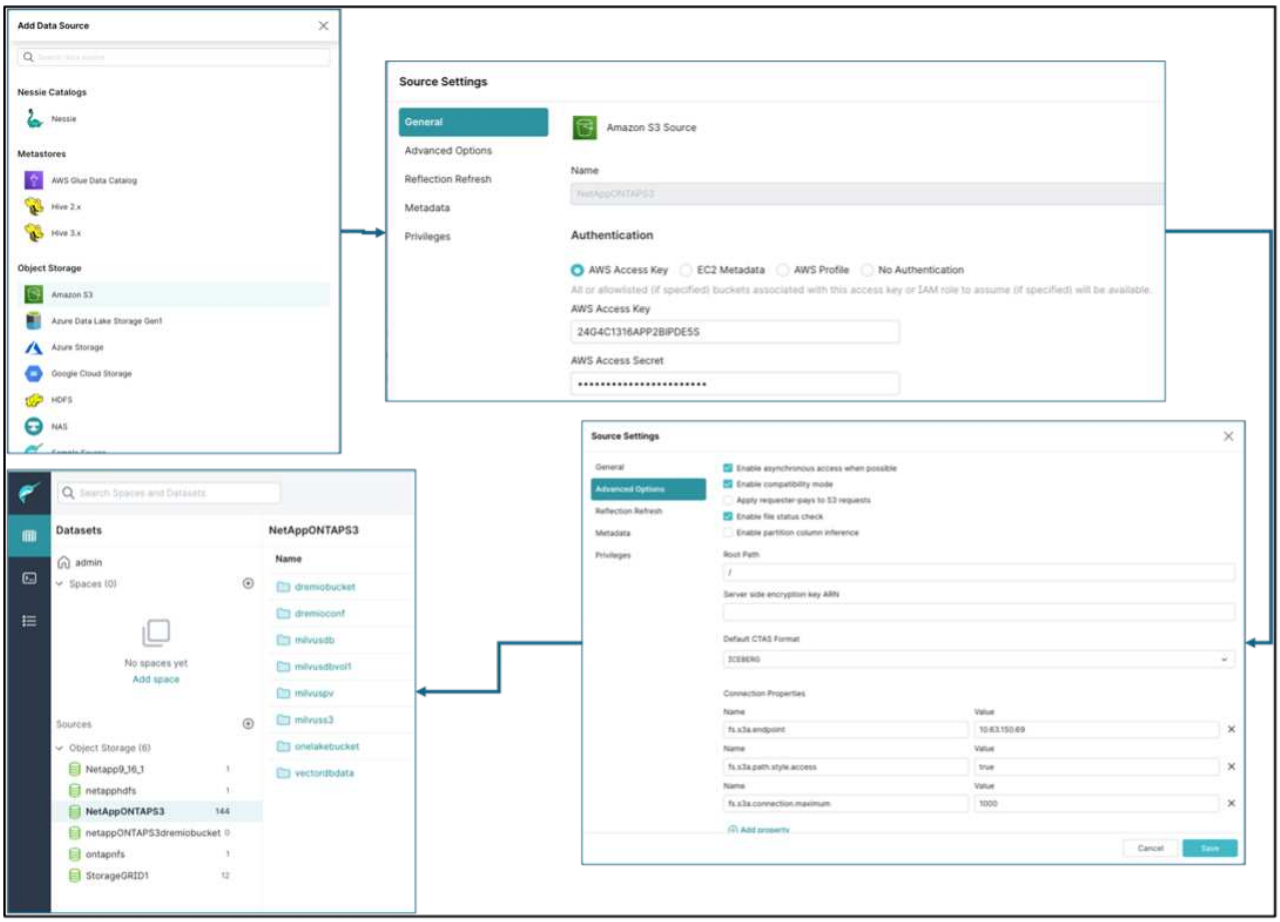
> A NetApp recomenda o StorageGRID como sua principal solução de armazenamento de objetos para ambientes Datalake e Lakehouse. Além disso, o NetApp ONTAP é empregado para dualidade arquivo/objeto. No contexto deste documento, conduzimos testes no ONTAP S3 em resposta a uma solicitação do cliente, e ele funciona com sucesso como uma fonte de dados.

Configuração de múltiplas fontes

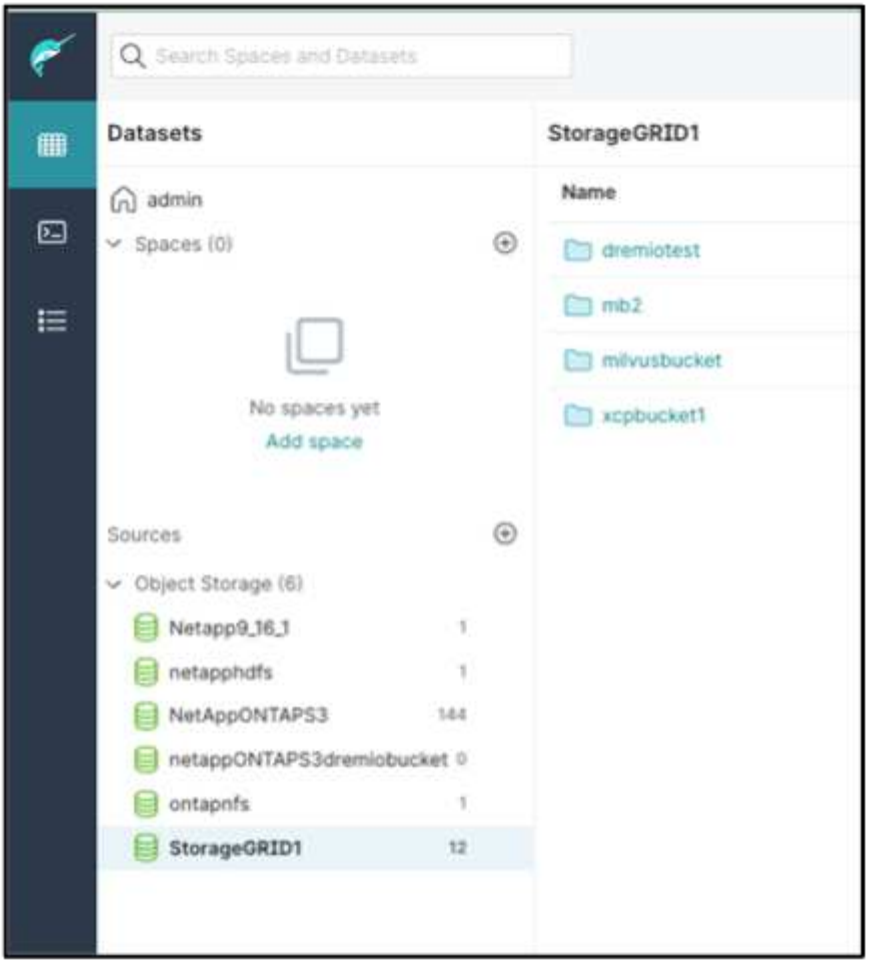
1. Configure o ONTAP S3 e o storageGRID como uma fonte s3 no Dremio.
   a. Painel do Dremio → conjuntos de dados → fontes → adicionar fonte.
   b. Na seção geral, atualize o acesso e a chave secreta da AWS
   c. Na opção avançada, ative o modo de compatibilidade e atualize as propriedades de conexão com os detalhes abaixo. O IP/nome do ponto de extremidade do controlador de armazenamento NetApp do ontap S3 ou do storageGRID.

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

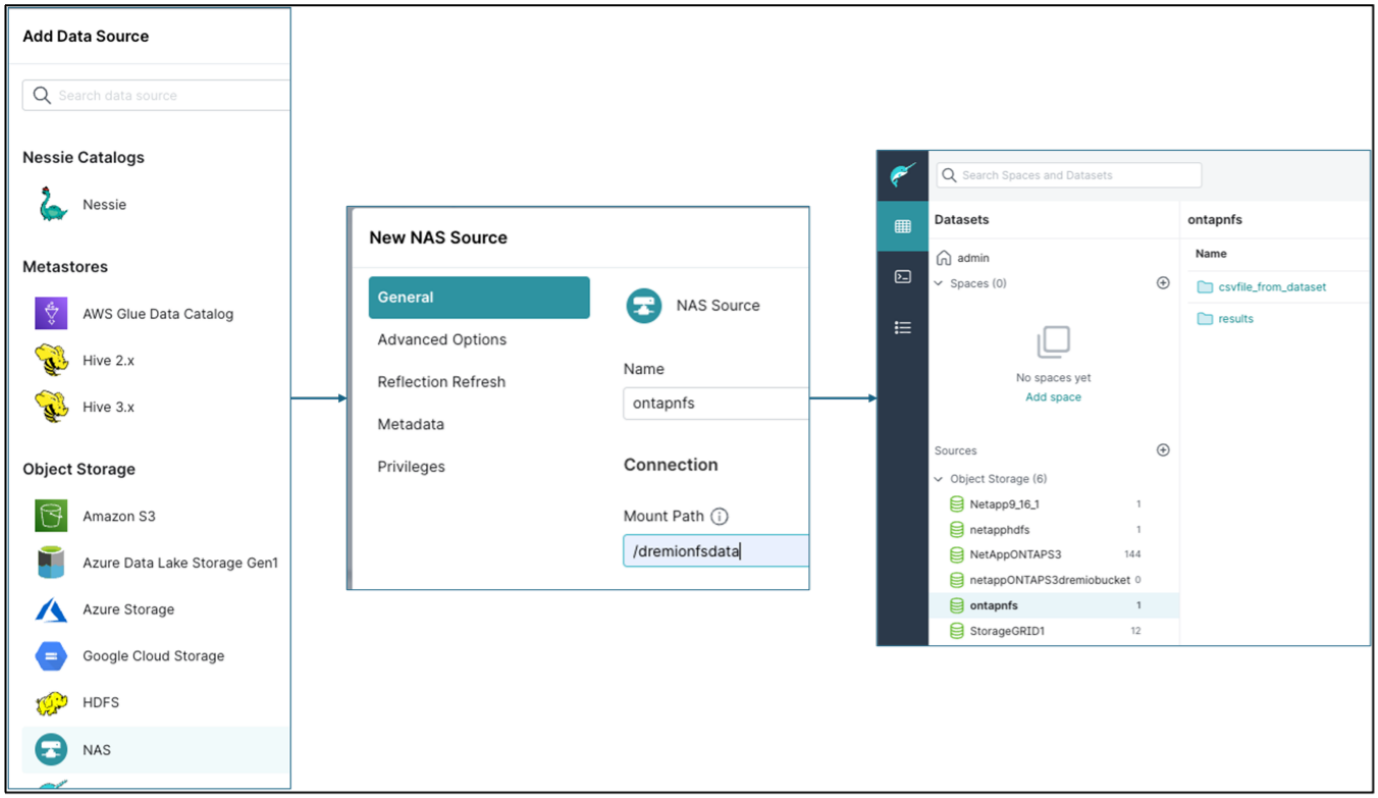
   d. Habilitar o cache local quando possível, Percentual máximo do cache total disponível para uso quando possível = 100
   e. Em seguida, visualize a lista de buckets do armazenamento de objetos do NetApp
   .



f. Exibição de exemplo dos detalhes do bucket

storageGRID 

2. Configure o NAS (especificamente o NFS) como uma fonte no Dremio.
   a. Painel do Dremio → conjuntos de dados → fontes → adicionar fonte.
   b. Na seção geral, insira o nome e o caminho de montagem do NFS. Certifique-se de que o caminho de montagem do NFS esteja montado na mesma pasta em todos os nós do cluster Dremio.


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:~#
```

Visão geral da verificação da solução

Nesta seção, executamos consultas de teste SQL de várias fontes para verificar a funcionalidade, testar e verificar o transbordamento para o armazenamento NetApp .

Consulta SQL no armazenamento de objetos

1. Defina a memória para 250 GB por servidor em dremio.env

```
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:~#
```

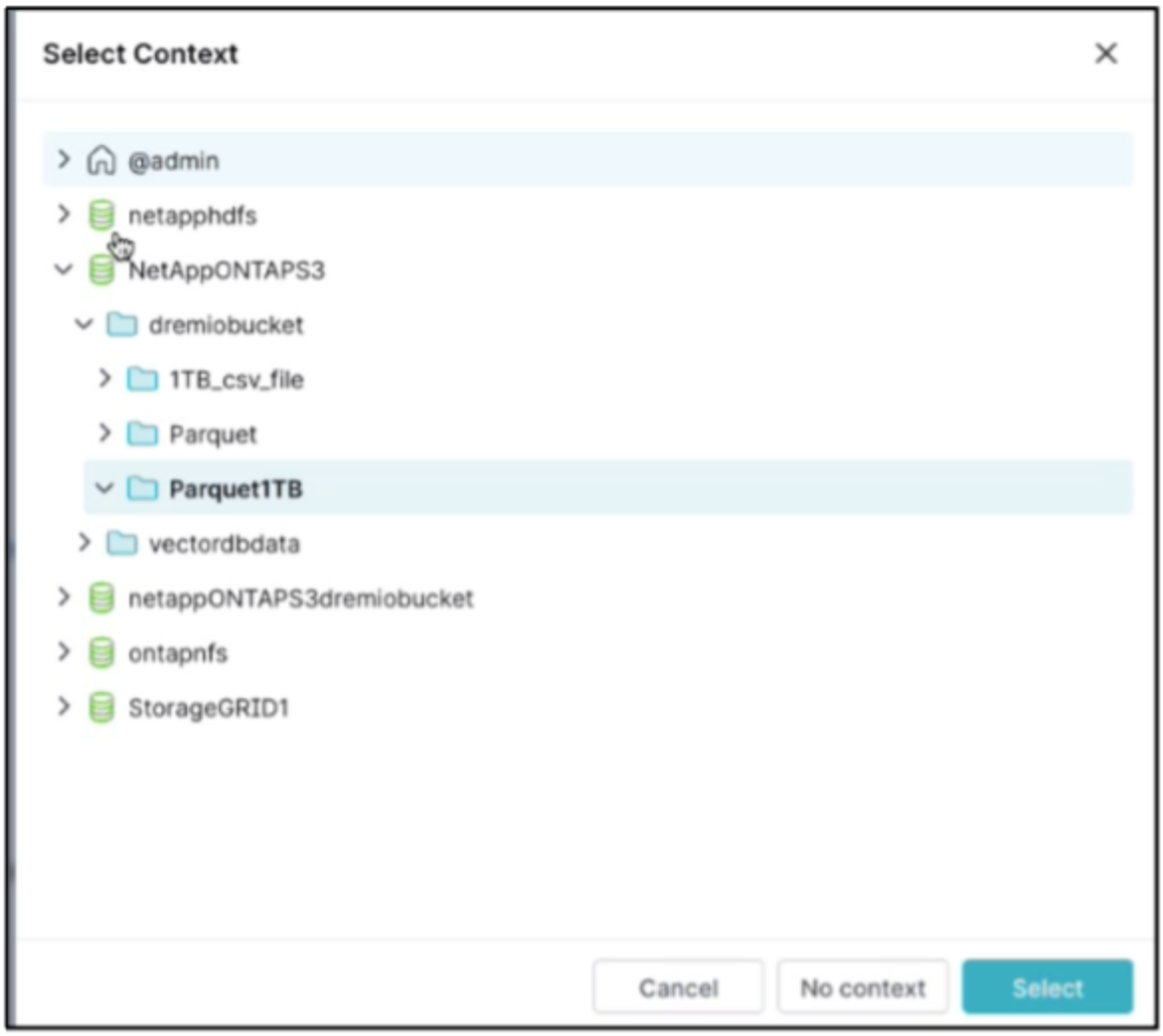
2. Verifique o local do spillover (${DREMIO_HOME}"/dremiocache) no arquivo dremio.conf e os detalhes de armazenamento.

```
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
```

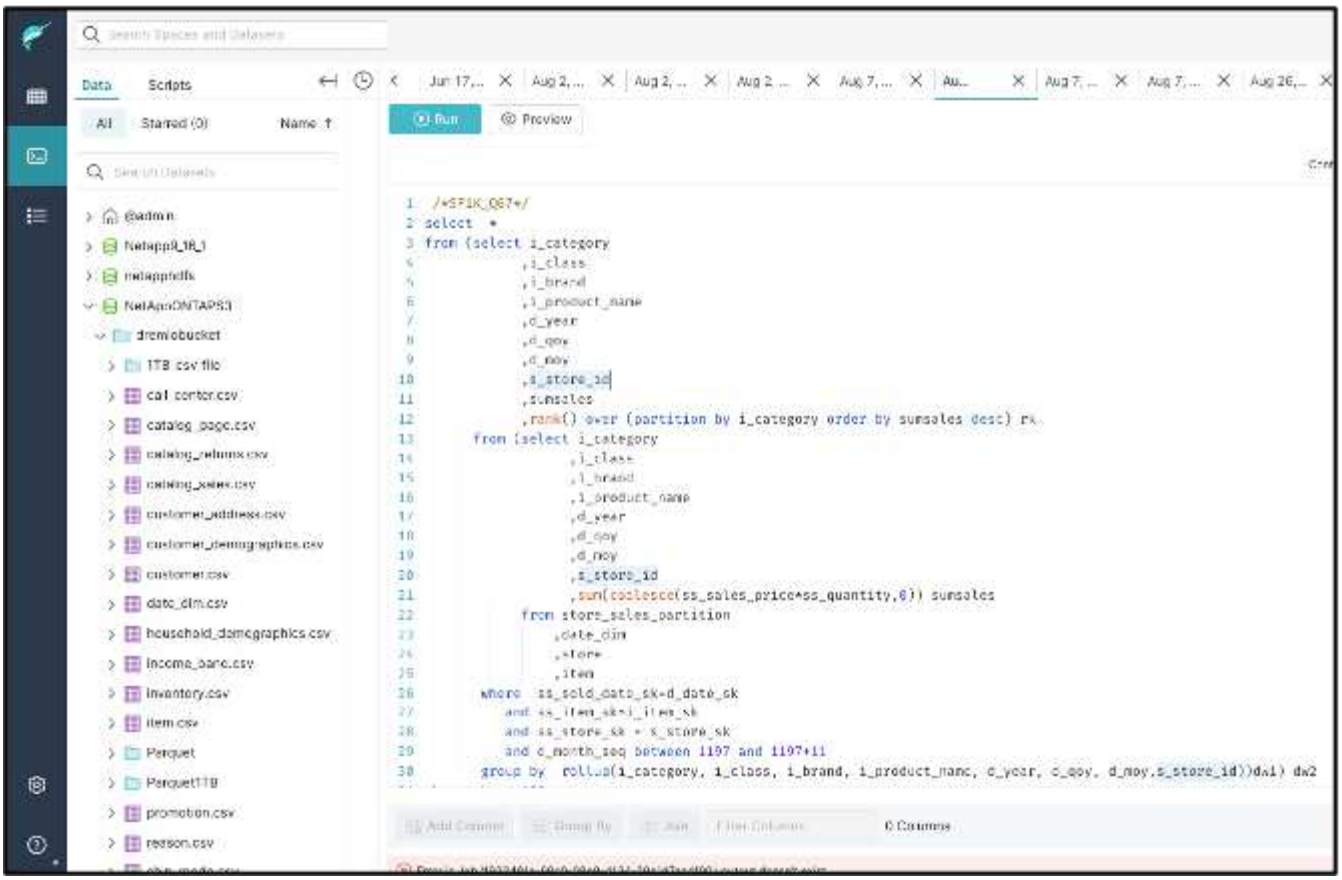
3. Aponte o local de vazamento do Dremio para o armazenamento NetApp NFS

```
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:~#
```

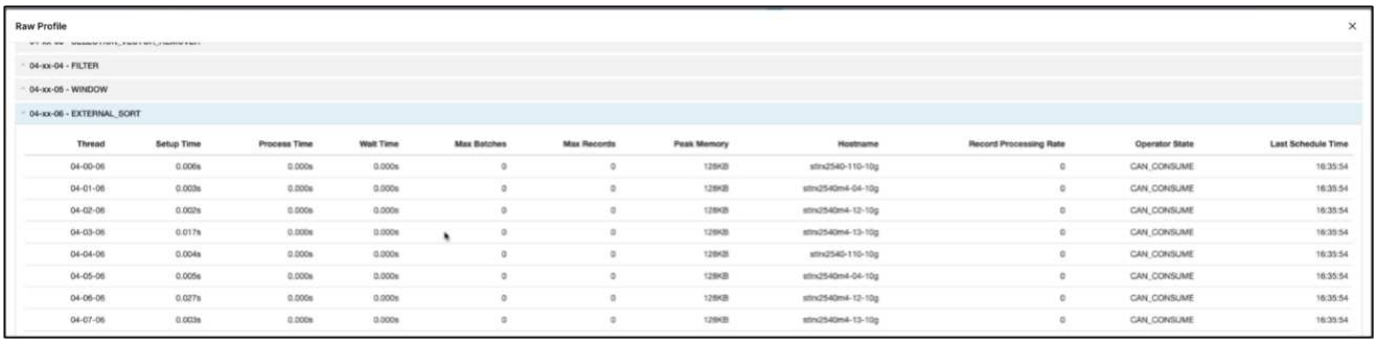
4. Selecione o contexto. Em nosso teste, executamos o teste em arquivos parquet gerados pelo TPCDS residentes no ONTAP S3. Painel Dremio → SQL runner → contexto → NetAppONTAPS3→Parquet1TB



1. Execute a consulta TPC-DS67 no painel do Dremio



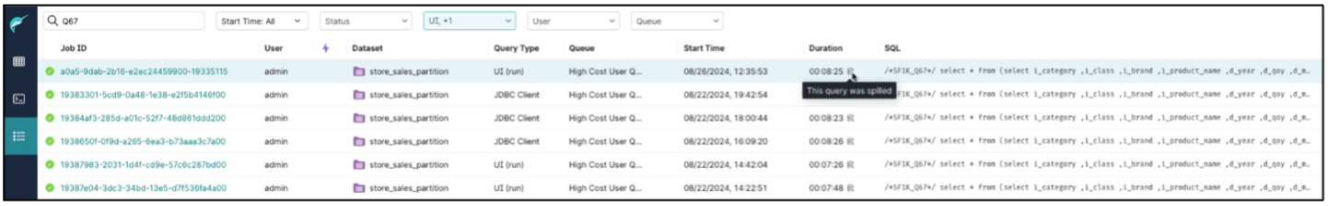
1. Verifique se o trabalho está sendo executado em todos os executores. Painel do Dremio → jobs → <jobid> → perfil bruto → selecione EXTERNAL_SORT → Nome do host



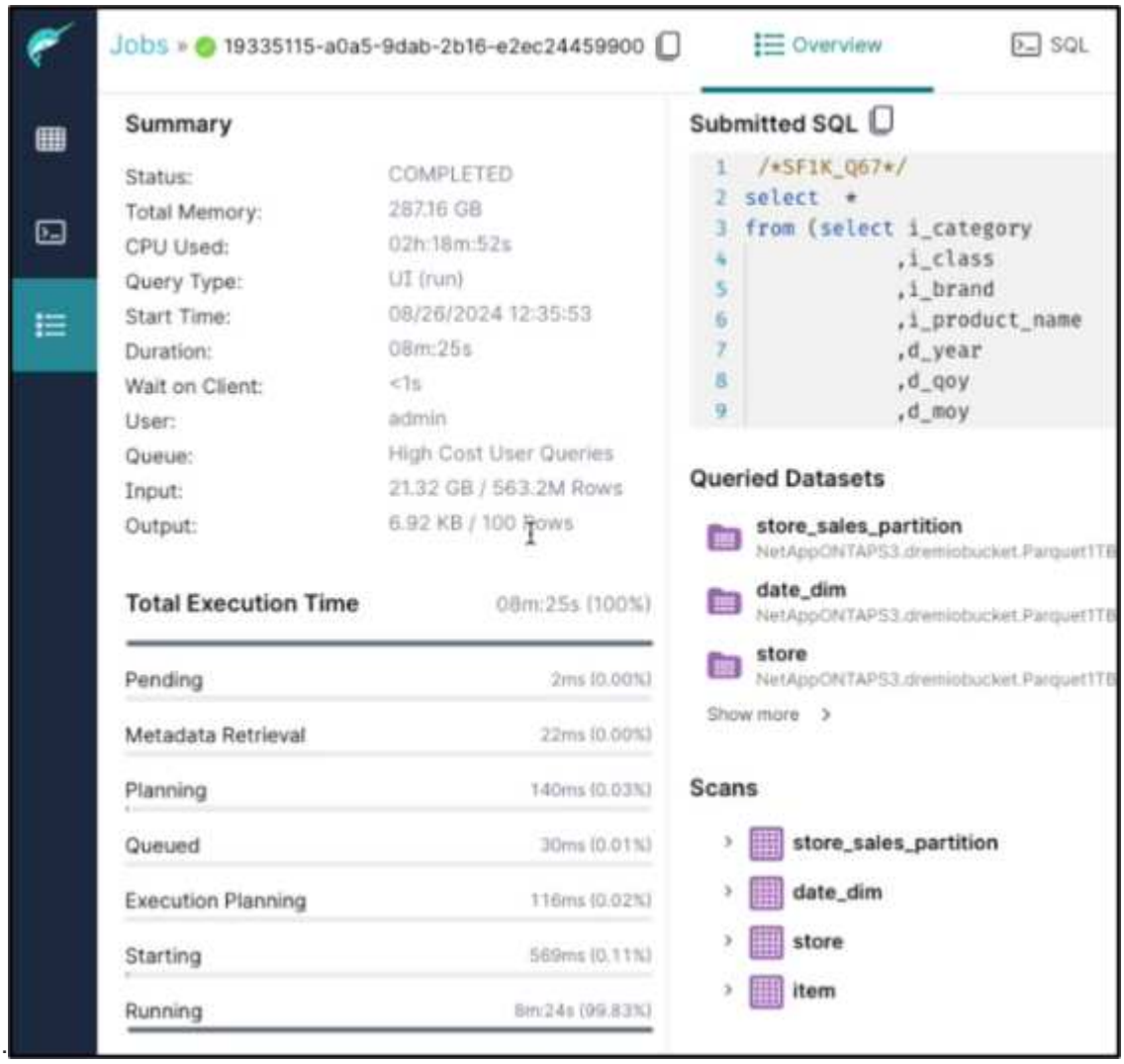
1. Quando a consulta SQL estiver em execução, você poderá verificar a pasta dividida para armazenamento em cache de dados no controlador de armazenamento NetApp .

```
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
```

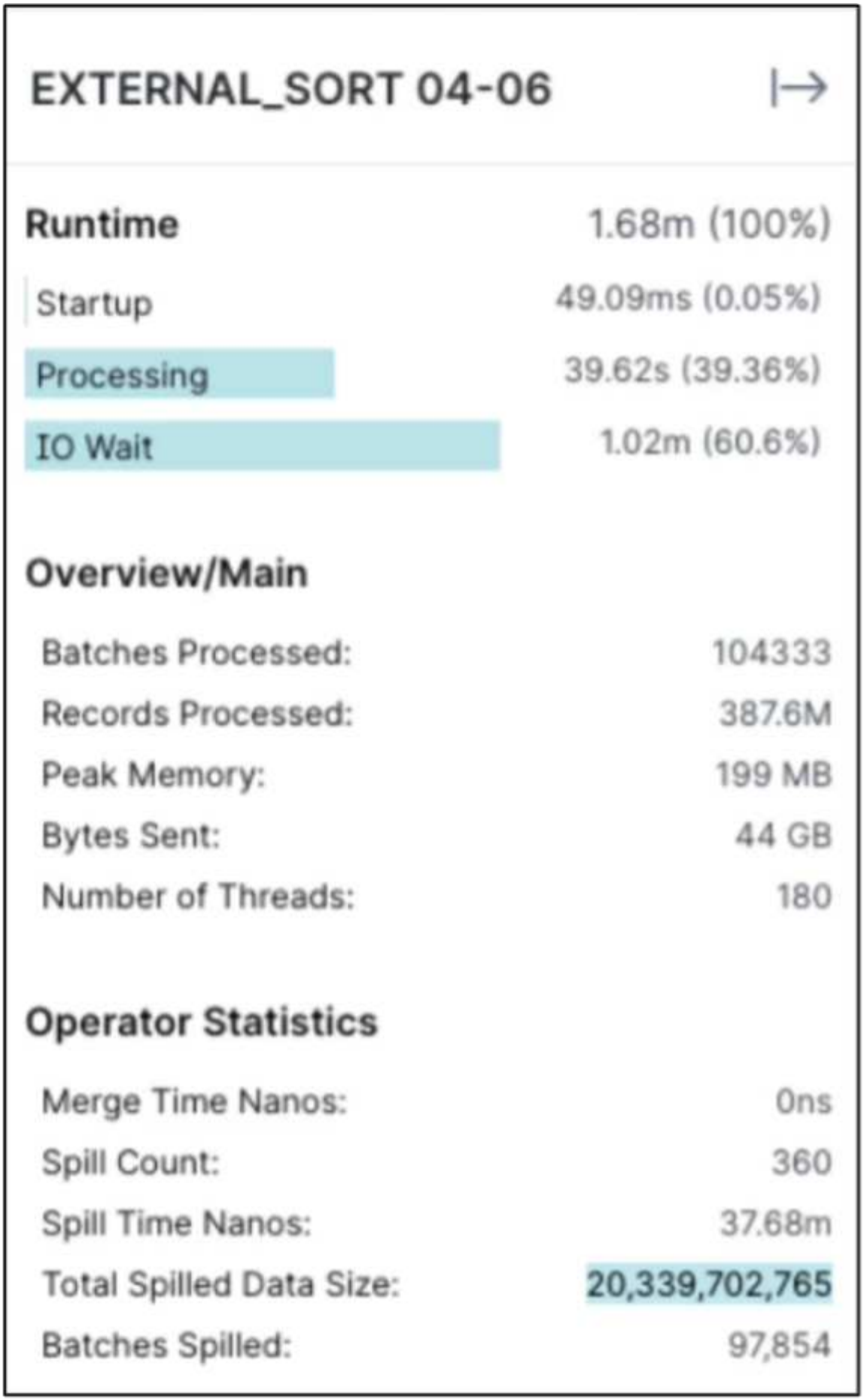
2. A consulta SQL foi concluída com transbordamento



3. Resumo da conclusão do trabalho.



4. Verifique o tamanho dos dados

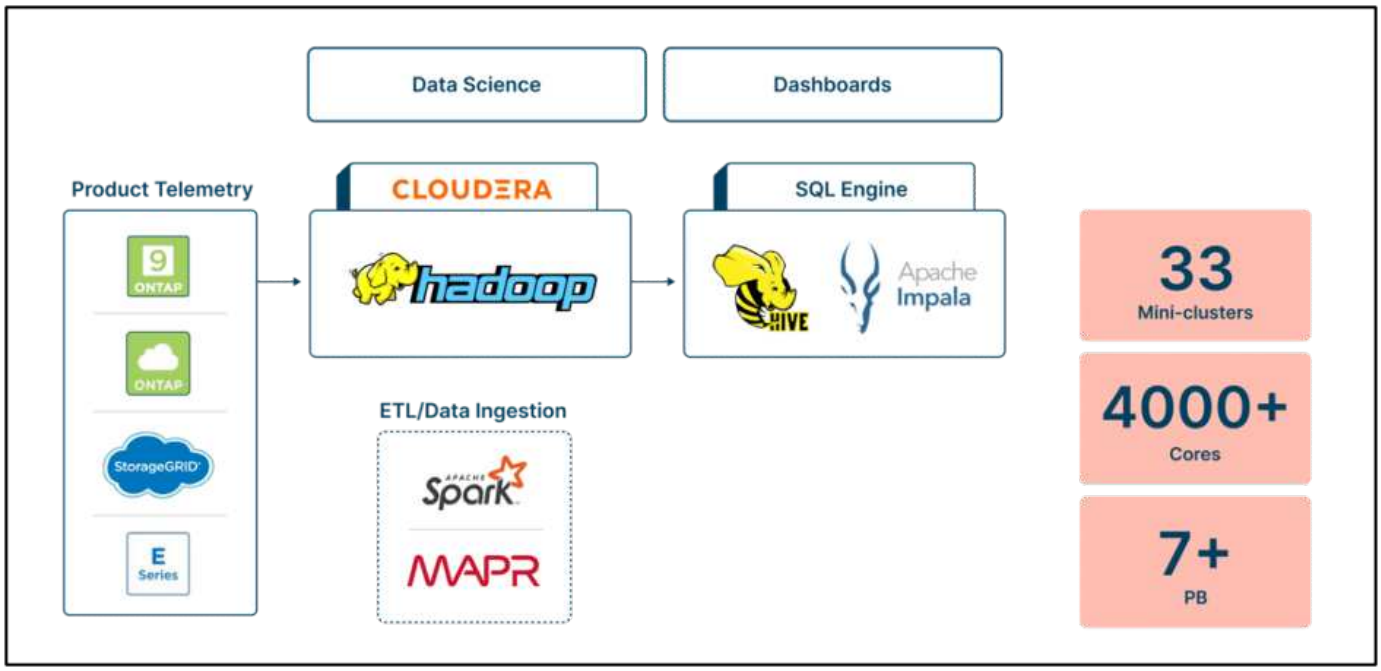

EXTERNAL_SORT 04-06
Runtime 1.68m (100%)
Startup 49.09ms (0.05%)
Processing 39.62s (39.36%)
IO Wait 1.02m (60.6%)
Overview/Main
Batches Processed: 104333
Records Processed: 387.6M
Peak Memory: 199 MB
Bytes Sent: 44 GB
Number of Threads: 180
Operator Statistics
Merge Time Nanos: 0ns
Spill Count: 360
Spill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

vazados

O mesmo procedimento se aplica ao armazenamento de objetos NAS e StorageGRID .

Casos de uso do cliente

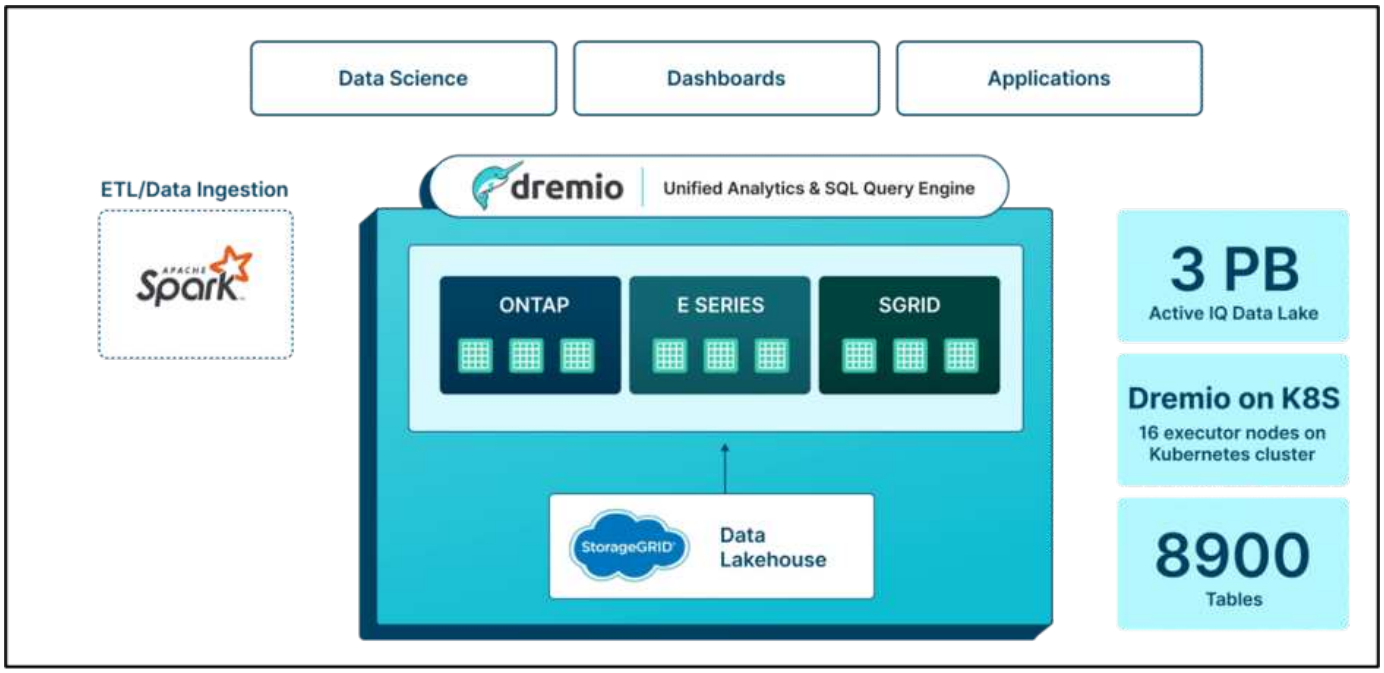
Caso de uso do NetApp ActiveIQ



Desafio: A solução interna Active IQ da NetApp, projetada inicialmente para dar suporte a diversos casos de uso, evoluiu para uma oferta abrangente para usuários internos e clientes. No entanto, a infraestrutura de backend subjacente baseada em Hadoop/MapR apresentou desafios em termos de custo e desempenho, devido ao rápido crescimento de dados e à necessidade de acesso eficiente aos dados. Aumentar a escala do armazenamento significava adicionar recursos de computação desnecessários, resultando em aumento de custos.

Além disso, gerenciar o cluster Hadoop consumia tempo e exigia conhecimento especializado. Problemas de desempenho e gerenciamento de dados complicaram ainda mais a situação, com consultas levando em média 45 minutos e escassez de recursos devido a configurações incorretas. Para enfrentar esses desafios, a NetApp buscou uma alternativa ao ambiente Hadoop legado existente e determinou que uma nova solução moderna construída no Dremio reduziria custos, separaria o armazenamento e a computação, melhoraria o desempenho, simplificaria o gerenciamento de dados, ofereceria controles detalhados e forneceria recursos de recuperação de desastres.

Solução:



A Dremio permitiu que a NetApp modernizasse sua infraestrutura de dados baseada em Hadoop em uma abordagem em fases, fornecendo um roteiro para análises unificadas. Ao contrário de outros fornecedores que exigiram mudanças significativas no processamento de dados, a Dremio se integrou perfeitamente aos pipelines existentes, economizando tempo e despesas durante a migração. Ao fazer a transição para um ambiente totalmente conteinerizado, a NetApp reduziu a sobrecarga de gerenciamento, melhorou a segurança e aumentou a resiliência. A adoção de ecossistemas abertos como Apache Iceberg e Arrow pela Dremio garantiu proteção para o futuro, transparência e extensibilidade.

Como substituição para a infraestrutura Hadoop/Hive, o Dremio ofereceu funcionalidade para casos de uso secundários por meio da camada semântica. Embora os mecanismos existentes de ETL e ingestão de dados baseados em Spark tenham permanecido, o Dremio forneceu uma camada de acesso unificada para facilitar a descoberta e a exploração de dados sem duplicação. Essa abordagem reduziu significativamente os fatores de replicação de dados e desvinculou o armazenamento e a computação.

Benefícios: Com o Dremio, a NetApp obteve reduções de custos significativas ao minimizar o consumo de computação e os requisitos de espaço em disco em seus ambientes de dados. O novo Active IQ Data Lake é composto por 8.900 tabelas que armazenam 3 petabytes de dados, em comparação com a infraestrutura anterior, com mais de 7 petabytes. A migração para o Dremio também envolveu a transição de 33 miniclusters e 4.000 núcleos para 16 nós executores em clusters do Kubernetes. Mesmo com reduções significativas nos recursos de computação, a NetApp experimentou melhorias notáveis de desempenho. Ao acessar os dados diretamente pelo Dremio, o tempo de execução da consulta diminuiu de 45 minutos para 2 minutos, resultando em um tempo 95% mais rápido para obter insights para manutenção preditiva e otimização. A migração também resultou em uma redução de mais de 60% nos custos de computação, consultas mais de 20 vezes mais rápidas e uma economia de mais de 30% no custo total de propriedade (TCO).

Caso de uso do cliente de vendas de peças automotivas.

Desafios: Nesta empresa global de vendas de peças automotivas, os grupos executivos e corporativos de planejamento financeiro e análise não conseguiram obter uma visão consolidada dos relatórios de vendas e foram forçados a ler os relatórios de métricas de vendas de cada linha de negócios e tentar consolidá-los. Isso fez com que os clientes tomassem decisões com dados que tinham pelo menos um dia. O tempo de espera para obter novos insights analíticos normalmente levaria mais de quatro semanas. A solução de problemas em pipelines de dados exigiria ainda mais tempo, acrescentando três dias ou mais ao cronograma já longo. O lento processo de desenvolvimento de relatórios, bem como o desempenho dos relatórios, forçava a comunidade de analistas a esperar continuamente que os dados fossem processados ou carregados, em vez

de permitir que eles encontrassem novos insights de negócios e impulsionassem novos comportamentos empresariais. Esses ambientes problemáticos eram compostos por vários bancos de dados diferentes para diferentes linhas de negócios, resultando em vários silos de dados. O ambiente lento e fragmentado complicou a governança de dados, pois havia muitas maneiras de os analistas chegarem à sua própria versão da verdade em vez de uma única fonte de verdade. A abordagem custou mais de US$ 1,9 milhão em custos com plataforma de dados e pessoas. Manter a plataforma legada e atender às solicitações de dados exigia sete engenheiros técnicos de campo (FTEs) por ano. Com o aumento das solicitações de dados, a equipe de inteligência de dados não conseguiu dimensionar o ambiente legado para atender às necessidades futuras

Solução: Armazene e gerencie de forma econômica grandes tabelas Iceberg no NetApp Object Store. Crie domínios de dados usando a camada semântica do Dremio, permitindo que usuários empresariais criem, pesquisem e compartilhem produtos de dados facilmente.

Benefícios para o cliente: • Arquitetura de dados existente melhorada e otimizada e tempo reduzido para insights de quatro semanas para apenas algumas horas • Tempo de solução de problemas reduzido de três dias para apenas algumas horas • Custos de plataforma e gerenciamento de dados reduzidos em mais de US$ 380.000 • (2) FTEs de esforço de inteligência de dados economizados por ano

Conclusão

Concluindo, este relatório técnico forneceu detalhes abrangentes de implantação do q Hybrid Iceberg Lakehouse com Dremio em conjunto com várias fontes de dados de controladores de armazenamento NetApp , incluindo ONTAP S3, NAS e StorageGRID. O processo de implantação foi executado com sucesso, e a ferramenta de benchmarking TPC-DS foi utilizada para executar 99 consultas SQL nas diferentes fontes de dados. O relatório também explorou casos de uso de clientes dentro do NetApp, demonstrando a versatilidade e eficácia do Dremio em atender a diversos requisitos de negócios. Além disso, um caso de uso específico envolvendo um cliente de vendas de peças automotivas foi examinado, destacando a aplicação prática e os benefícios de aproveitar o Dremio para análises de dados e insights.

No geral, este documento serve como um recurso valioso para entender a implantação e o uso do Dremio com controladores de armazenamento NetApp , mostrando seus recursos e potencial para impulsionar a tomada de decisões e a otimização baseadas em dados em vários setores.

Onde encontrar informações adicionais

Para saber mais sobre as informações descritas neste documento, revise os seguintes documentos e/ou sites:

- Instalação do tratador de zoológico

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

- Dremio

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

- Configurando Dremio com storageGRID

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

configure-dremio-data-source

- Caso de uso do NetApp

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

Informações sobre direitos autorais

Informações sobre marcas comerciais