Solución híbrida iceberg lakehouse de próxima generación de NetApp y Dremio

NetApp artificial intelligence solutions

NetApp
August 18, 2025

Tabla de contenidos

Solución híbrida iceberg lakehouse de próxima generación de NetApp y Dremio

La solución Iceberg Lakehouse híbrida de próxima generación de NetApp y Dremio

En este documento, analizamos los detalles de implementación de Dremio con diferentes fuentes de datos de los controladores de almacenamiento de NetApp , incluidos ONTAP S3, NAS y StorageGRID. Durante la implementación, utilizamos la herramienta de evaluación comparativa TPC-DS para ejecutar 99 consultas SQL en varias fuentes. El documento también explora casos de uso de clientes dentro de NetApp, así como un caso de uso que involucra a un cliente de ventas de repuestos de automóviles.

Descripción general de la solución

La solución Hybrid Iceberg Lakehouse ofrece beneficios únicos para abordar los desafíos que enfrentan los clientes del lago de datos. Al aprovechar la plataforma Dremio Unified Lakehouse y las soluciones NetApp ONTAP, StorageGRID y NetApp Cloud, las empresas pueden agregar valor significativo a sus operaciones comerciales. La solución no solo brinda acceso a múltiples fuentes de datos, incluidas las fuentes de NetApp , sino que también mejora el rendimiento analítico general y ayuda a las empresas a generar conocimientos comerciales que conducen al crecimiento empresarial.

Descripción general de NetApp

- Las ofertas de NetApp, como ONTAP y StorageGRID, permiten la separación del almacenamiento y la computación, lo que posibilita una utilización óptima de los recursos según los requisitos específicos. Esta flexibilidad permite a los clientes escalar de forma independiente su almacenamiento utilizando soluciones de almacenamiento de NetApp.
- Al aprovechar los controladores de almacenamiento de NetApp, los clientes pueden servir datos de manera eficiente a su base de datos vectorial utilizando los protocolos NFS y S3. Estos protocolos facilitan el almacenamiento de datos de clientes y administran el índice de la base de datos vectorial, eliminando la necesidad de múltiples copias de datos a los que se accede a través de métodos de archivos y objetos.
- NetApp ONTAP proporciona soporte nativo para NAS y almacenamiento de objetos en los principales proveedores de servicios de nube como AWS, Azure y Google Cloud. Esta amplia compatibilidad garantiza una integración perfecta, lo que permite la movilidad de los datos del cliente, la accesibilidad global, la recuperación ante desastres, la escalabilidad dinámica y el alto rendimiento.

StorageGRID

Nuestro almacenamiento de objetos líder en la industria, storageGRID, ofrece un potente motor de políticas para la ubicación automatizada de datos, opciones de implementación flexibles y una durabilidad inigualable con codificación de borrado en capas. Tiene una arquitectura escalable que admite miles de millones de objetos y petabytes de datos en un solo espacio de nombres. La solución permite la integración de la nube híbrida, lo que permite la organización de datos en niveles en las principales plataformas de nube. Ha sido reconocido como líder en la Evaluación mundial de proveedores basados en objetos de IDC Marketscape 2019.

Además, storageGRID se destaca en la gestión de datos no estructurados a escala con almacenamiento de objetos definido por software, redundancia geográfica y capacidades de múltiples sitios. Incorpora gestión del ciclo de vida de la información basada en políticas y ofrece funciones de integración en la nube como duplicación y búsqueda. Cuenta con diversas certificaciones, entre ellas Common Criteria, NF203 Digital Safe Component, ISO/IEC 25051, KPMG y Cohasset Compliance Assessment.

En resumen, NetApp storageGRID ofrece potentes funciones, escalabilidad, integración de nube híbrida y certificaciones de cumplimiento para una gestión eficiente de datos no estructurados a escala.

ONTAP de NetApp

NetApp ONTAP es una solución de almacenamiento sólida que ofrece una amplia gama de funciones empresariales. Incluye Snapshot, que proporciona copias de seguridad instantáneas, consistentes con la aplicación y a prueba de manipulaciones. SnapRestore permite la restauración casi instantánea de copias de seguridad a pedido, mientras que SnapMirror ofrece capacidades integradas de copia de seguridad remota y recuperación ante desastres. La solución también incorpora Autonomous Ransomware Protection (ARP), que garantiza la seguridad de los datos con funciones como verificación de múltiples administradores, cifrado de datos en reposo con certificación FIPS, cifrado de datos en tránsito, autenticación multifactor (MFA) y control de acceso basado en roles (RBAC). El registro integral, la auditoría, la gestión de claves integradas y externas, la purga segura y la gestión segura de múltiples inquilinos mejoran aún más la seguridad y el cumplimiento de los datos.

NetApp ONTAP también cuenta con SnapLock, que proporciona retención de datos que cumple con las normativas con altos niveles de integridad, rendimiento y retención a un bajo costo total de propiedad. Está completamente integrado con NetApp ONTAP 9 y ofrece protección contra actos maliciosos, administradores deshonestos y ransomware.

La solución incluye cifrado NSE/NVE para cifrado en tránsito y de datos en reposo, acceso de administrador multifactor y verificación de múltiples administradores. Active IQ proporciona análisis predictivos y acciones correctivas basados en IA, mientras que QoS garantiza el control de la carga de trabajo de la calidad del servicio. La integración de la gestión y la automatización es intuitiva a través de SysMgr/GUI/CLI/API. FabricPool permite la clasificación automática de datos y la solución ofrece eficiencia a través de la compresión, deduplicación y compactación de datos en línea. NetApp garantiza el cumplimiento de los objetivos de eficiencia de la carga de trabajo sin coste para el cliente.

NetApp ONTAP admite varios protocolos, incluidos NVMe/FC, FC, NVMe/TCP, iSCSI, NFS, SMB y S3, lo que lo convierte en una solución de almacenamiento unificada. En general, NetApp ONTAP ofrece amplias funciones empresariales, seguridad sólida, cumplimiento, eficiencia y versatilidad para satisfacer diversas necesidades de almacenamiento.

Descripción general de Dremio

Dremio es la plataforma unificada de Lakehouse para análisis de autoservicio e IA. La plataforma de análisis unificado Dremio acerca a los usuarios a los datos con la flexibilidad, escalabilidad y rendimiento de un lago a una fracción del costo de las soluciones de almacenamiento de datos tradicionales. Dremio permite el análisis "shift-left" para eliminar la integración de datos y ETL complejos y costosos, brindando un análisis a escala empresarial sin inconvenientes y sin movimiento de datos. Dremio también incluye:

- Análisis de autoservicio fácil de usar habilitado a través de una capa semántica universal y un motor de consulta SQL altamente integrado y de alto rendimiento, lo que facilita la conexión, la gestión y el análisis de todos los datos, tanto en la nube como en las instalaciones.
- Las capacidades de administración de lakehouse nativas de Apache Iceberg de Dremio simplifican el descubrimiento de datos y automatizan la optimización de datos, brindando análisis de alto rendimiento con versiones de datos inspiradas en Git.

- Dremio, basado en código abierto y estándares abiertos, permite a las empresas evitar el estancamiento y permanecer posicionadas para la innovación. Las empresas confían en Dremio como la plataforma lakehouse más fácil de usar y con la mejor relación precio-rendimiento en todas las cargas de trabajo.

¿Qué valor ofrece la solución híbrida Iceberg Lakehouse de Dremio y NetApp a los clientes?

- **Gestión y accesibilidad de datos mejoradas**: Dremio es conocido por su plataforma de lago de datos que permite a las organizaciones consultar datos directamente desde sus lagos de datos a alta velocidad. NetApp, por otro lado, es un proveedor líder de servicios de datos en la nube y soluciones de almacenamiento de datos. La oferta conjunta proporciona a los clientes una solución integral para almacenar, gestionar, acceder y analizar los datos de su empresa de forma eficiente y eficaz.
- **Optimización del rendimiento**: Con la experiencia de NetApp en almacenamiento de datos y las capacidades de Dremio en procesamiento y optimización de datos, la asociación ofrece una solución que mejora el rendimiento de las operaciones de datos, reduce la latencia y aumenta la velocidad para obtener información empresarial. Dremio incluso ha proporcionado beneficios de rendimiento a la propia infraestructura analítica de TI interna de NetApp.
- **Escalabilidad**: Tanto Dremio como NetApp ofrecen una solución diseñada para escalar. La solución conjunta proporciona a los clientes entornos de almacenamiento, gestión y análisis de datos altamente escalables. En un entorno híbrido Iceberg Lakehouse, el motor de consulta SQL de Dremio combinado con NetApp StorageGRID ofrece escalabilidad, simultaneidad y rendimiento de consultas incomparables, capaz de gestionar las necesidades analíticas de cualquier empresa.
- **Seguridad y gobernanza de datos**: Ambas empresas tienen un fuerte enfoque en la seguridad y gobernanza de datos. Juntos, ofrecen funciones robustas de seguridad y gobernanza de datos, lo que garantiza que los datos estén protegidos y que se cumplan los requisitos de gobernanza de datos. Características como controles de acceso basados en roles y de grano fino, auditoría integral, linaje de datos de extremo a extremo, gestión de identidad unificada y SSO con un amplio marco de cumplimiento y seguridad garantizan que los entornos de datos analíticos de las empresas sean seguros y estén gobernados.
- **Relación costo-eficiencia**: Al integrar el motor de lago de datos de Dremio con las soluciones de almacenamiento de NetApp, los clientes pueden reducir los costos asociados con la gestión y el movimiento de datos. Las organizaciones también pueden migrar de entornos de data lake tradicionales a una solución de lakehouse más moderna compuesta por NetApp y Dremio. Esta solución híbrida Iceberg Lakehouse ofrece un rendimiento de consultas de alta velocidad y una concurrencia de consultas líder en el mercado que reduce el TCO y el tiempo para obtener información comercial.

Requisitos tecnológicos

Las configuraciones de hardware y software que se describen a continuación se utilizaron para las validaciones realizadas en este documento. Estas configuraciones sirven como guía para ayudarle a configurar su entorno. Sin embargo, tenga en cuenta que los componentes específicos pueden variar según los requisitos individuales del cliente.

Requisitos de hardware

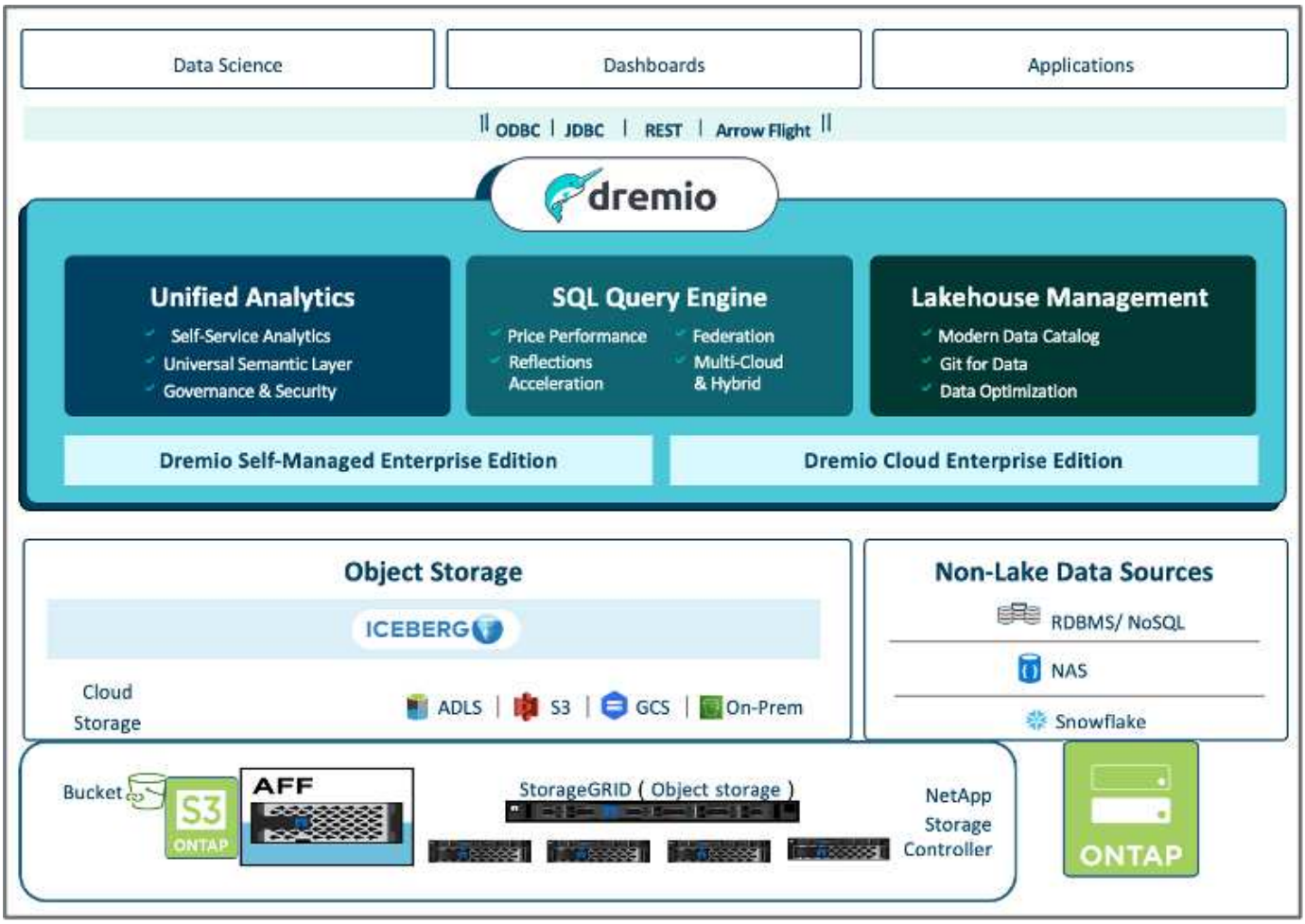
Hardware	Detalles
Par HA de matriz de almacenamiento AFF de NetApp	• A800 • ONTAP 9.14.1 • 48 unidades SSD-NVM de 3,49 TB • Dos depósitos S3: metadatos de Dremio y datos de clientes.
4 x FUJITSU PRIMERGY RX2540 M4	• 64 CPU • Procesador Intel Xeon Gold 6142 a 2,60 GHz • Memoria física de 256 GM • 1 puerto de red de 100 GbE
Redes	• 100 GbE
StorageGRID	* 1 x SG100, 3 x SGF6024 * 3 x 24 x 7,68 TB * Dos buckets S3: metadatos de Dremio y datos de clientes.

Requisitos de software

Software	Detalles
Dremio	• versión - 25.0.3-202405170357270647-d2042e1b • Edición empresarial
En las instalaciones	• Clúster Dremio de 5 nodos • 1 coordinador maestro y 4 ejecutores

Procedimiento de implementación

En esta validación de arquitectura de referencia, utilizamos una configuración de Dremio compuesta por un coordinador y cuatro
ejecutores



Configuración de NetApp

- Inicialización del sistema de almacenamiento
- Creación de una máquina virtual de almacenamiento (SVM)
- Asignación de interfaces de red lógicas
- Configuración y licencias de NFS, S3

Siga los pasos a continuación para NFS (sistema de archivos de red): 1. Cree un volumen de grupo flexible para NFSv4 o NFSv3. En nuestra configuración para esta validación, hemos utilizado 48 SSD, 1 SSD dedicado al volumen raíz del controlador y 47 SSD distribuidos para NFSv4. Verifique que la política de exportación NFS para el volumen Flex Group tenga permisos de lectura y escritura para la red de servidores Dremio.

1. En todos los servidores Dremio, cree una carpeta y monte el volumen Flex Group en esta carpeta a través de una interfaz lógica (LIF) en cada servidor Dremio.

Siga los pasos a continuación para S3 (Servicio de almacenamiento simple):

1. Configure un servidor de almacén de objetos con HTTP habilitado y el estado de administrador establecido en "activo" mediante el comando "vserver object-store-server create". Tiene la opción de habilitar HTTPS y configurar un puerto de escucha personalizado.
2. Cree un usuario de servidor de almacén de objetos mediante el comando "vserver object-store-server user create -user <nombre de usuario>".

3. Para obtener la clave de acceso y la clave secreta, puede ejecutar el siguiente comando: "set diag; vserver object-store-server user show -user <username>". Sin embargo, en el futuro, estas claves se proporcionarán durante el proceso de creación del usuario o se podrán recuperar mediante llamadas a la API REST.
4. Establezca un grupo de servidores de almacén de objetos utilizando el usuario creado en el paso 2 y otorgue acceso. En este ejemplo, proporcionamos "Acceso completo".
5. Cree dos depósitos S3 configurando su tipo en "S3". Uno para la configuración de Dremio y otro para los datos del cliente.

Configuración de Zookeeper

Puede utilizar la configuración del zoológico proporcionada por Dremio. En esta validación, utilizamos un zookeeper independiente. Seguimos los pasos mencionados en este enlace web. https://medium.com/@ahmetfurkandemir/distributed-hadoop-cluster-1-spark-with-all-dependincies-03c8ec616166

Configuración de Dremio

Seguimos este enlace web para instalar Dremio mediante paquete tar.

1. Crea un grupo Dremio.

```
sudo groupadd -r dremio
```

2. Crear un usuario dremio.

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

3. Crear directorios de 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. Descargue el archivo tar desde https://download.dremio.com/community-server/
5. Desempaquete Dremio en el directorio /opt/dremio.

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

6. Crea un enlace simbólico para la carpeta de configuración.

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

7. Configure la configuración de su servicio (configuración de SystemD).
   a. Copie el archivo de unidad del demonio dremio desde /opt/dremio/share/dremio.service a /etc/systemd/system/dremio.service.
   b. Reiniciar el sistema

```
sudo systemctl daemon-reload
```

   c. Habilitar dremio para que se inicie durante el arranque.

```
sudo systemctl enable dremio
```

8. Configurar Dremio en el coordinador. Consulte Configuración de Dremio para obtener más información
   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#
```

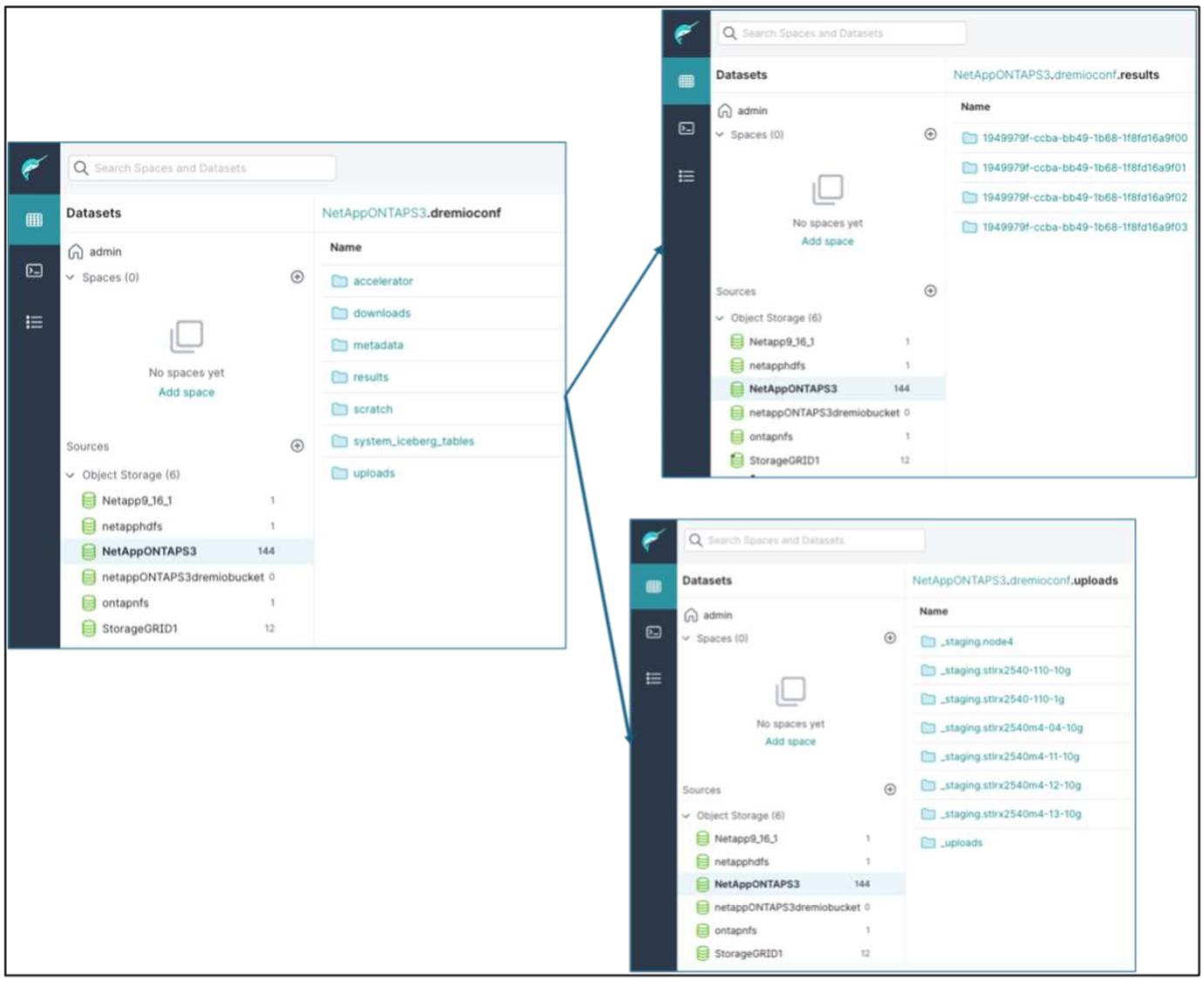
   b. Sitio principal.xml

```
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. La configuración de Dremio se almacena en el almacenamiento de objetos de NetApp . En nuestra validación, el bucket "dremioconf" reside en un bucket S3 de ontap. La siguiente imagen muestra algunos detalles de las carpetas "scratch" y "uploads" del bucket S3 "dremioconf".



1. Configurar Dremio en ejecutores. En nuestra configuración, tenemos 3 ejecutores.
    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: igual que la configuración del coordinador.

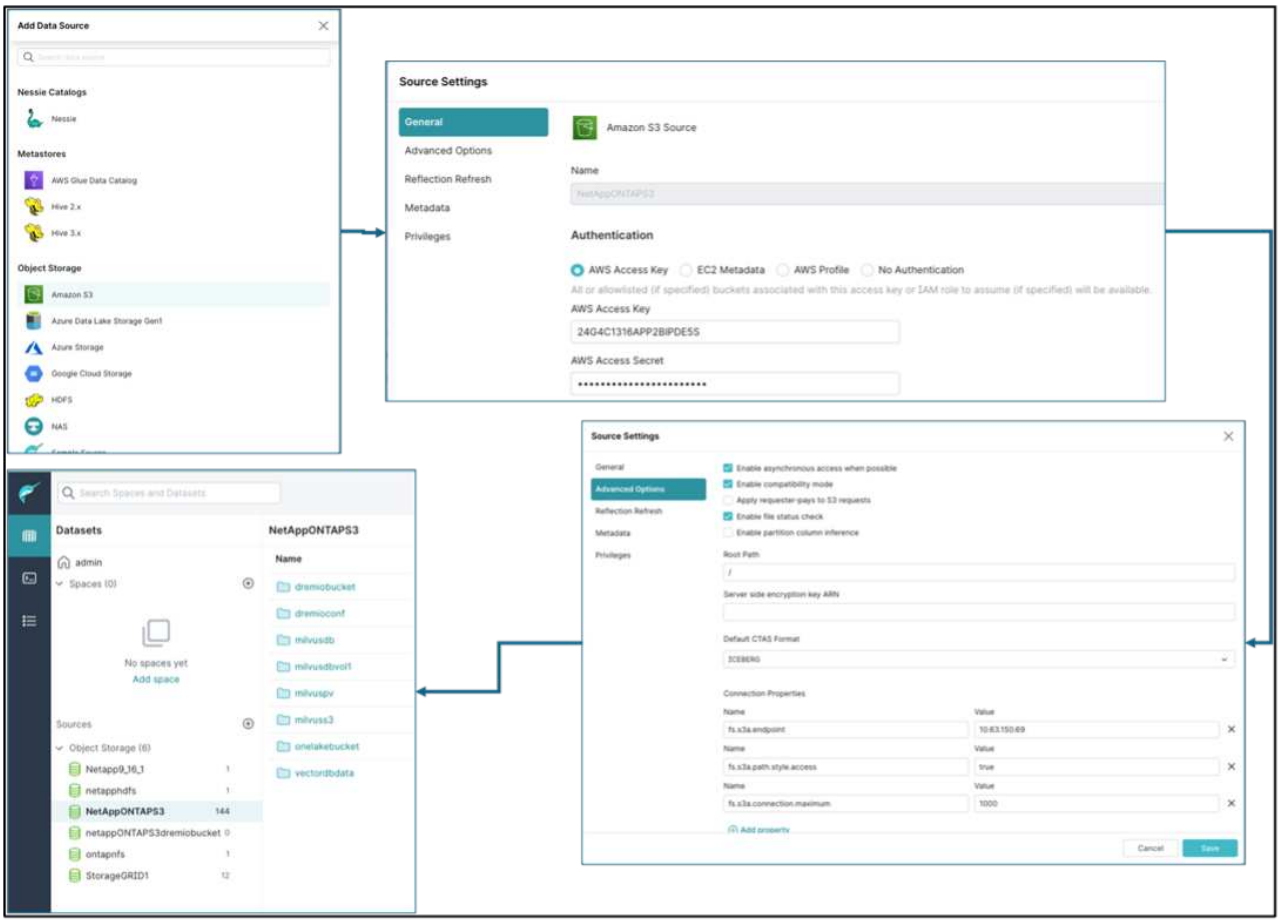
> NetApp recomienda StorageGRID como su principal solución de almacenamiento de objetos para entornos Datalake y Lakehouse. Además, NetApp ONTAP se utiliza para la dualidad archivo/objeto. En el contexto de este documento, hemos realizado pruebas en ONTAP S3 en respuesta a una solicitud del cliente y funciona exitosamente como fuente de datos.

Configuración de múltiples fuentes

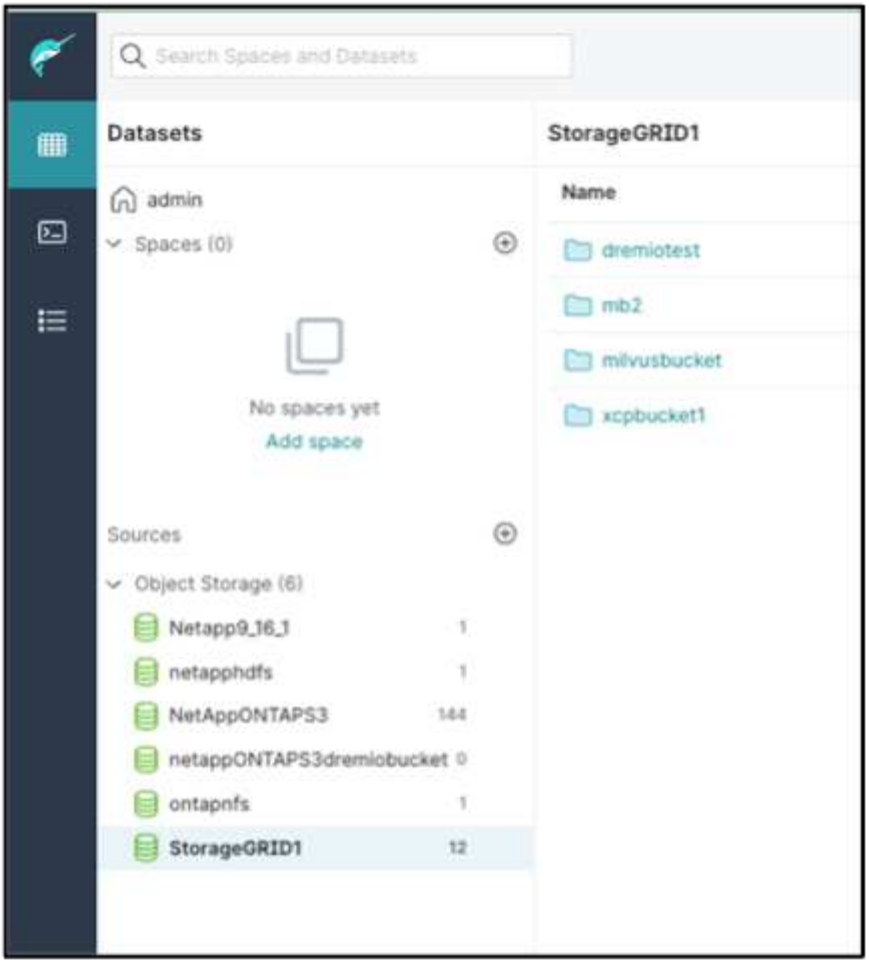
1. Configurar ONTAP S3 y storageGRID como una fuente s3 en Dremio.
   a. Panel de Dremio → conjuntos de datos → fuentes → agregar fuente.
   b. En la sección general, actualice el acceso y la clave secreta de AWS.
   c. En la opción avanzada, habilite el modo de compatibilidad, actualice las propiedades de conexión con los siguientes detalles. La IP/nombre del punto final del controlador de almacenamiento de NetApp , ya sea de ontap S3 o storageGRID.

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

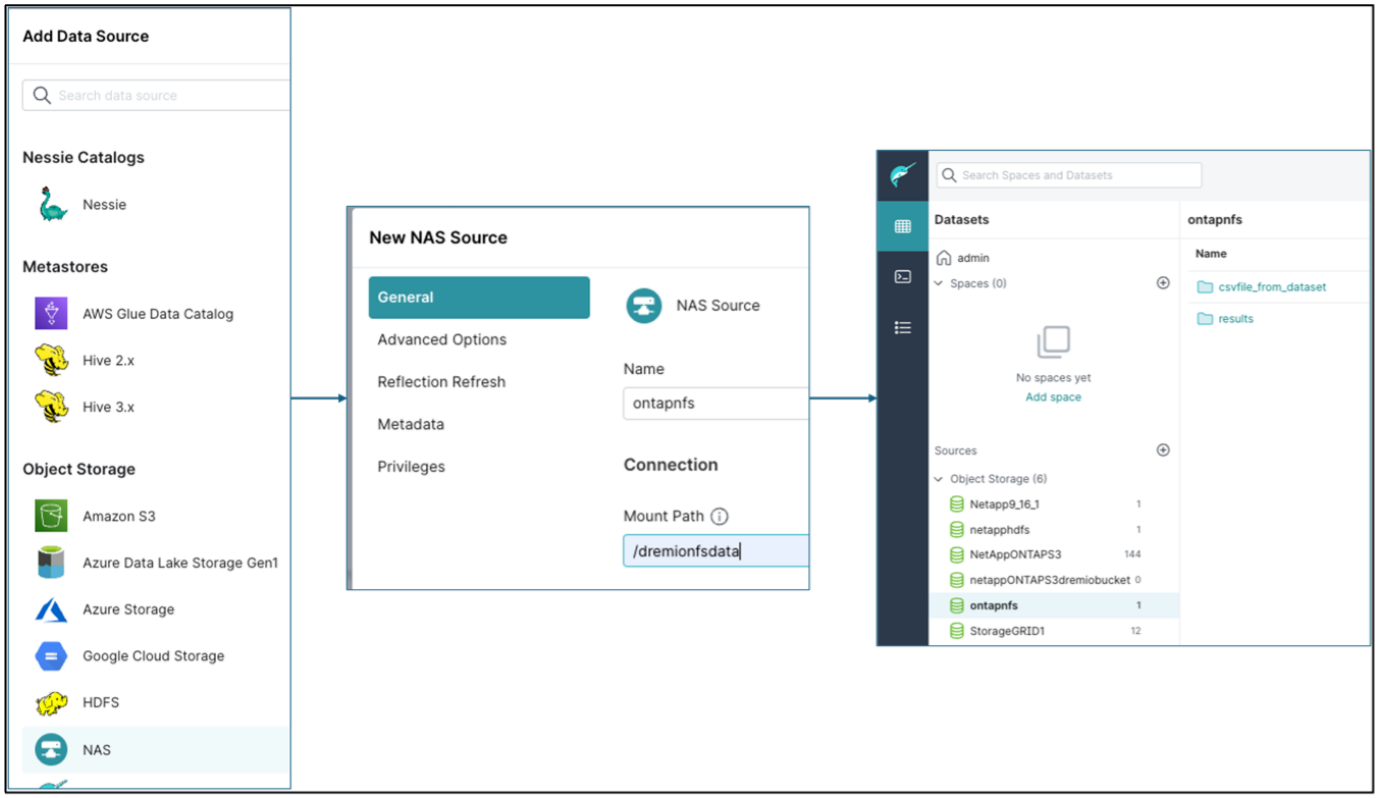
   d. Habilitar el almacenamiento en caché local cuando sea posible. Porcentaje máximo del caché total disponible para usar cuando sea posible = 100
   e. Luego vea la lista de depósitos del almacenamiento de objetos de NetApp
   .



f. Vista de muestra de los detalles del depósito de

storageGRID 

2. Configurar NAS (específicamente NFS) como fuente en Dremio.
   a. Panel de Dremio → conjuntos de datos → fuentes → agregar fuente.
   b. En la sección general, ingrese el nombre y la ruta de montaje NFS. Asegúrese de que la ruta de montaje de NFS esté montada en la misma carpeta en todos los nodos del clúster 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:~#
```

Descripción general de la verificación de la solución

En esta sección, hemos ejecutado consultas de prueba SQL desde múltiples fuentes para verificar la funcionalidad, probar y verificar el derrame al almacenamiento de NetApp .

Consulta SQL sobre almacenamiento de objetos

1. Establezca la memoria en 250 GB por servidor en 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:~#
```

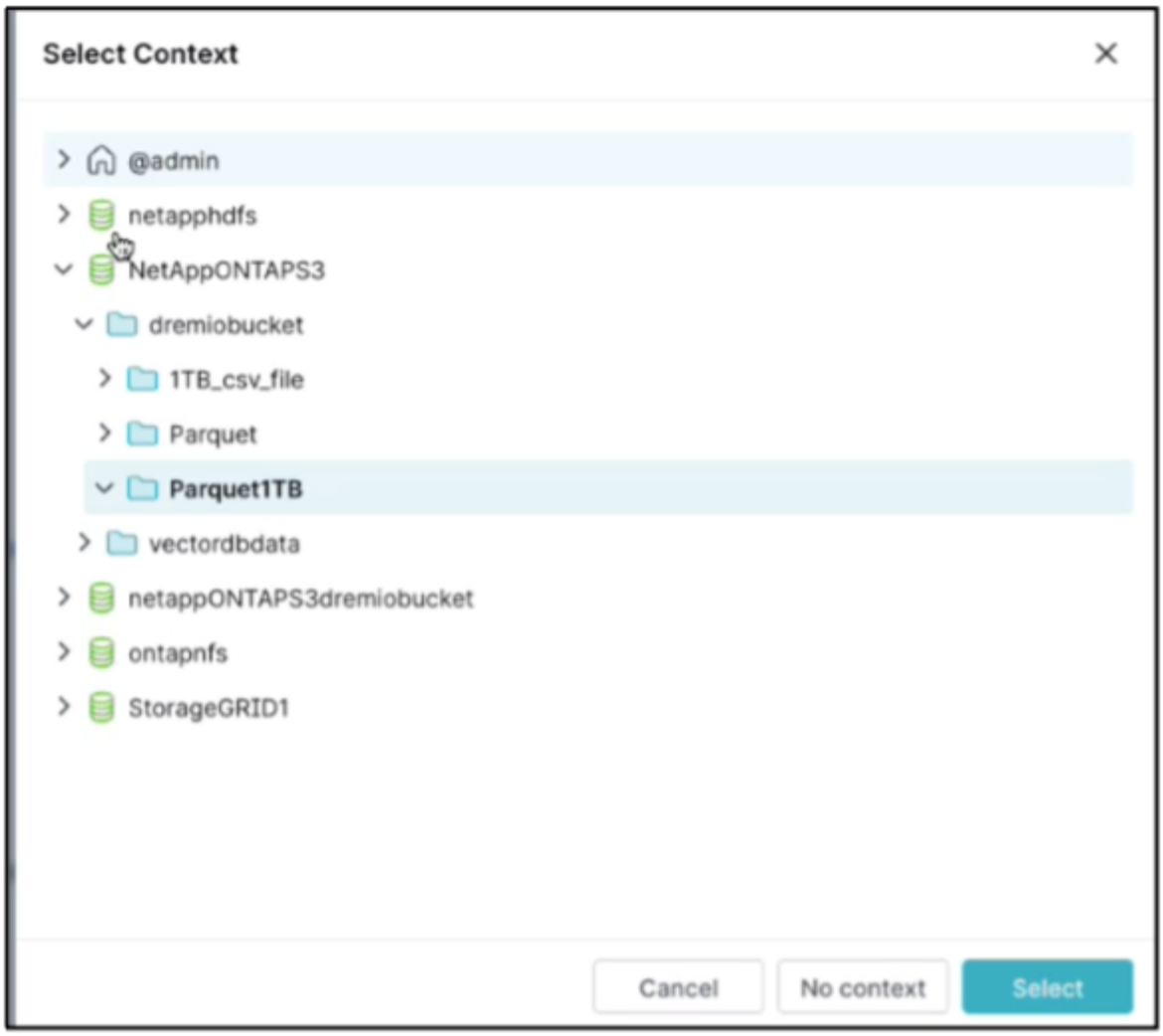
2. Verifique la ubicación del derrame (${DREMIO_HOME}"/dremiocache) en el archivo dremio.conf y los detalles de almacenamiento.

```
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
```

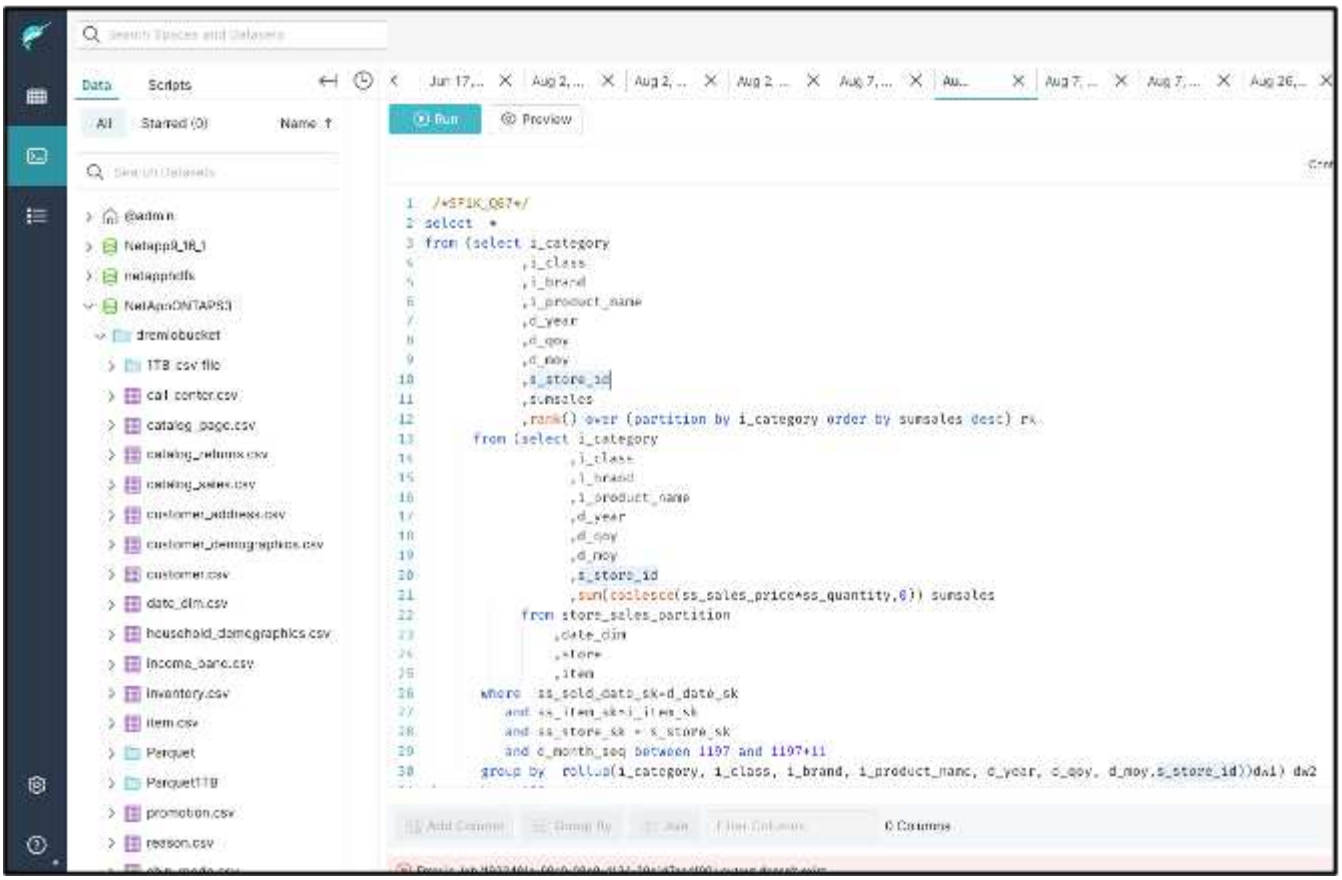
3. Apunte la ubicación de derrame de Dremio al almacenamiento NFS de NetApp

```
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 /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:~#
```

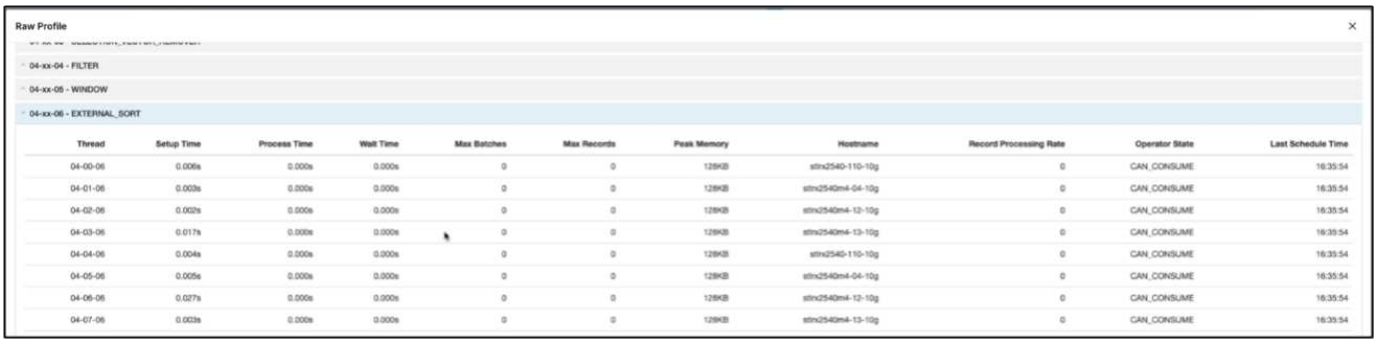
4. Seleccione el contexto. En nuestra prueba, ejecutamos la prueba contra archivos parquet generados por TPCDS que residen en ONTAP S3. Panel de control de Dremio → Ejecutor SQL → contexto → NetAppONTAPS3→Parquet1TB



1. Ejecute la consulta TPC-DS67 desde el panel de Dremio



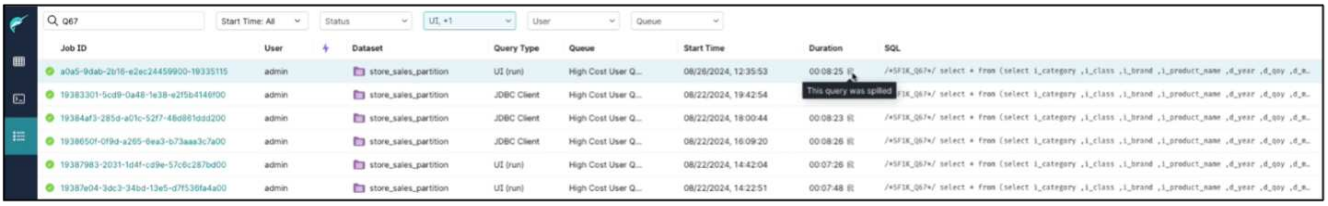
1. Verifique que el trabajo se esté ejecutando en todos los ejecutores. Panel de Dremio → trabajos → <jobid> → perfil sin procesar → seleccionar EXTERNAL_SORT → Nombre de host



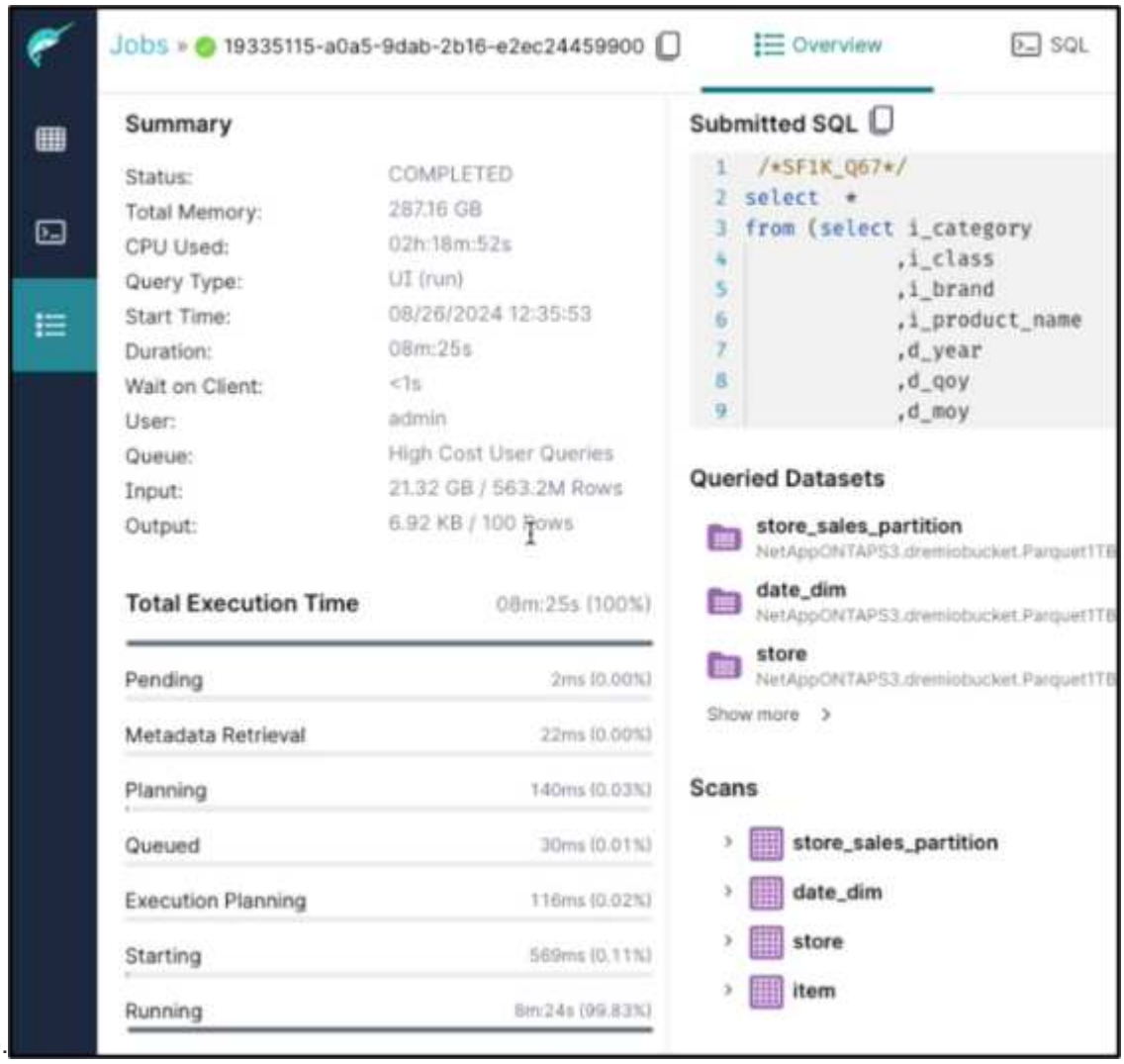
1. Cuando se ejecuta la consulta SQL, puede verificar la carpeta dividida para el almacenamiento en caché de datos en el controlador de almacenamiento de NetApp .

```
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
```

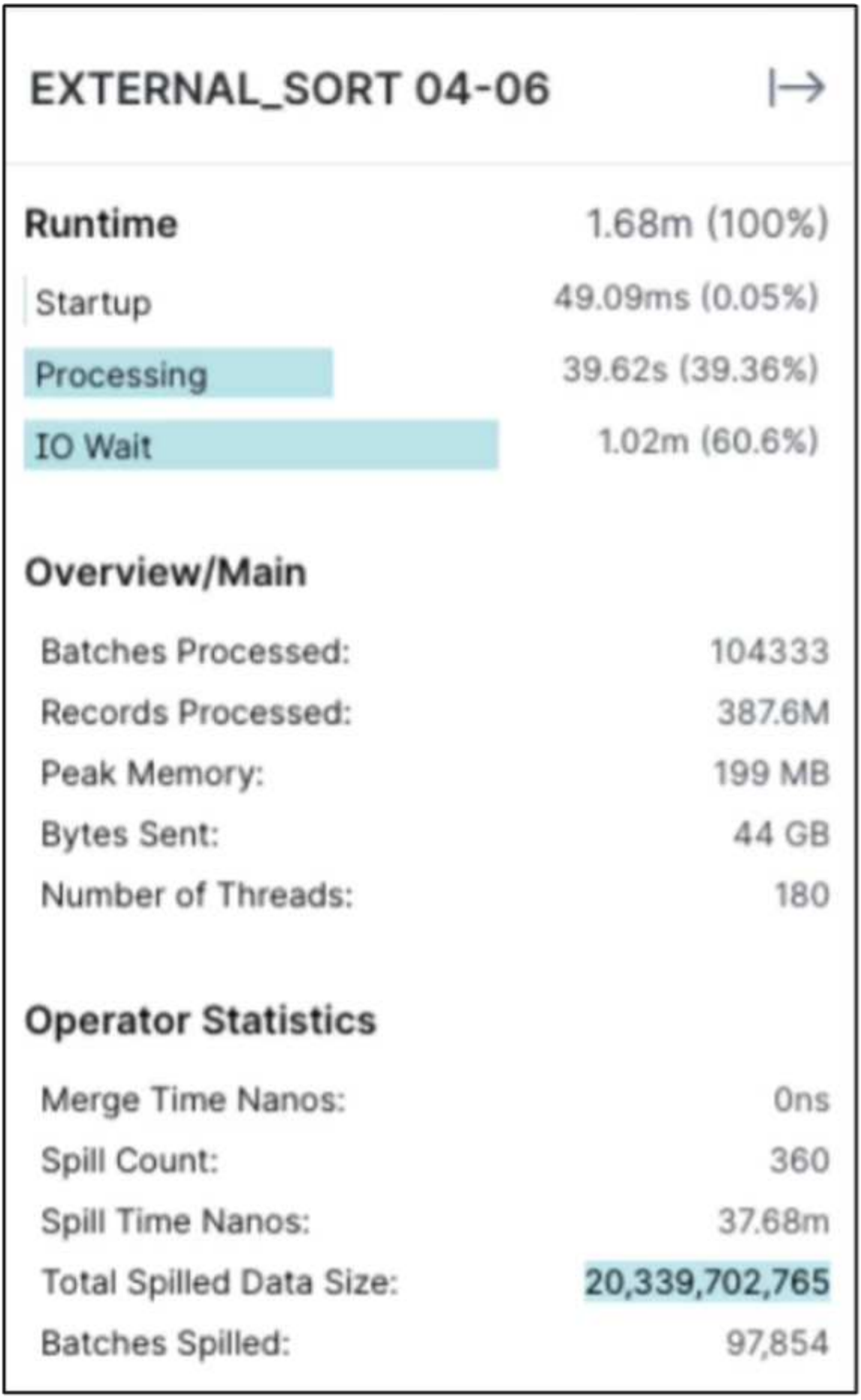
2. La consulta SQL se completó con derrame



3. Resumen de finalización del trabajo.



4. Verifique el tamaño de los datos

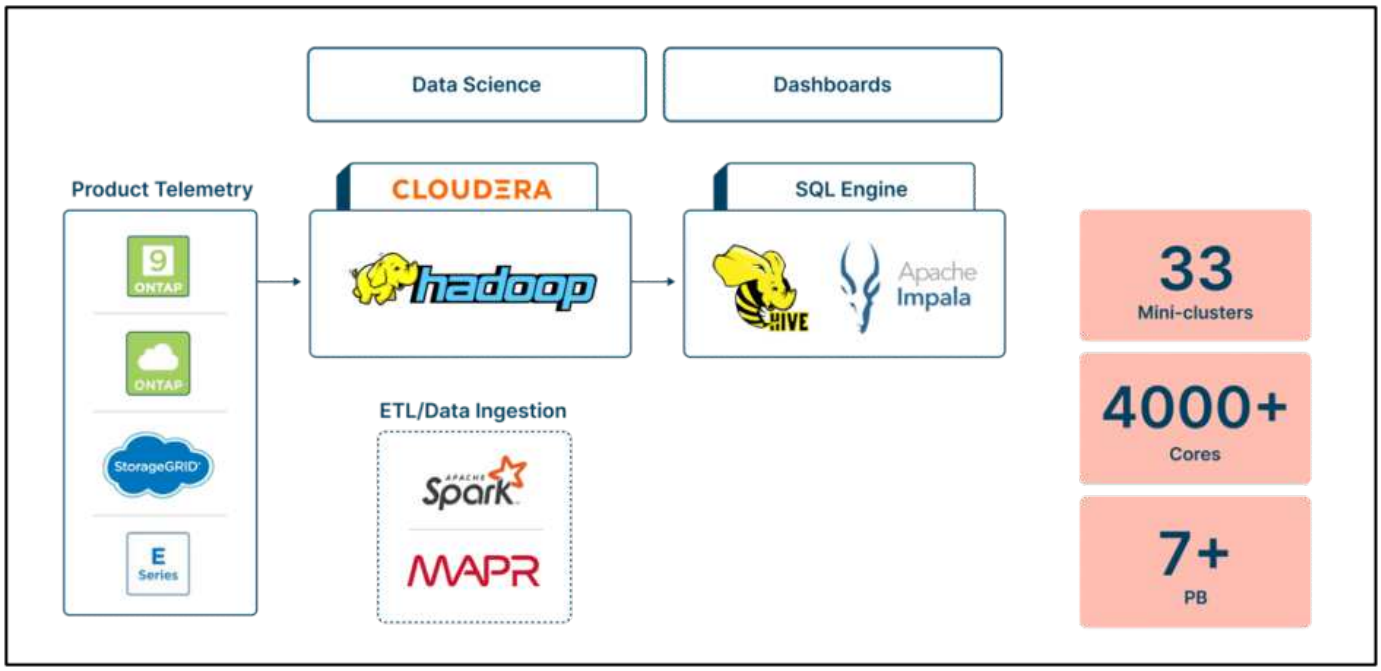

EXTERNAL_SORT 04-06
Runtime
1.68m (100%)
Startup
49.09ms (0.05%)
Processing
39.62s (39.36%)
IO Wait
1.02m (60.6%)
Overview/Main
Batches Processed:
104333
Records Processed:
387.6M
Peak Memory:
199 MB
Bytes Sent:
44 GB
Number of Threads:
180
Operator Statistics
Merge Time Nanos:
0ns
Spill Count:
360
Spill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

derramados

El mismo procedimiento se aplica para el almacenamiento de objetos NAS y StorageGRID .

Casos de uso de clientes

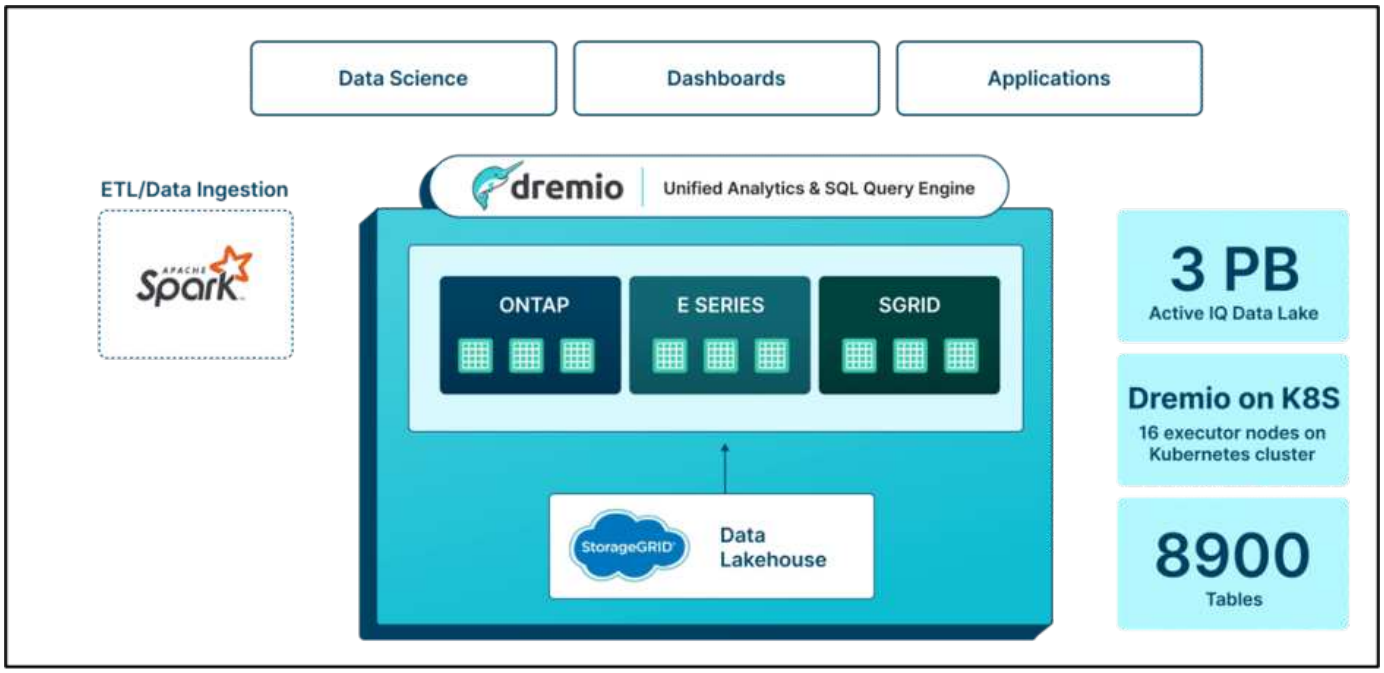
Caso de uso de NetApp ActiveIQ



Desafío: La solución interna Active IQ de NetApp, diseñada inicialmente para soportar numerosos casos de uso, había evolucionado hasta convertirse en una oferta integral tanto para usuarios internos como para clientes. Sin embargo, la infraestructura subyacente de backend basada en Hadoop/MapR planteó desafíos en términos de costos y rendimiento, debido al rápido crecimiento de los datos y la necesidad de un acceso eficiente a los mismos. Ampliar el almacenamiento implicaba añadir recursos informáticos innecesarios, lo que generaba mayores costos.

Además, gestionar el clúster Hadoop consumía mucho tiempo y requería conocimientos especializados. Los problemas de rendimiento y gestión de datos complicaron aún más la situación: las consultas tardaban un promedio de 45 minutos y los recursos se agotaban debido a configuraciones incorrectas. Para abordar estos desafíos, NetApp buscó una alternativa al entorno Hadoop existente y determinó que una nueva solución moderna basada en Dremio reduciría costos, desacoplaría el almacenamiento y el cómputo, mejoraría el rendimiento, simplificaría la gestión de datos, ofrecería controles detallados y proporcionaría capacidades de recuperación ante desastres.

Solución:



Dremio permitió a NetApp modernizar su infraestructura de datos basada en Hadoop en un enfoque gradual, proporcionando una hoja de ruta para análisis unificados. A diferencia de otros proveedores que requirieron cambios significativos en el procesamiento de datos, Dremio se integró perfectamente con las tuberías existentes, ahorrando tiempo y gastos durante la migración. Al realizar la transición a un entorno completamente en contenedores, NetApp redujo la sobrecarga de administración, mejoró la seguridad y mejoró la resiliencia. La adopción por parte de Dremio de ecosistemas abiertos como Apache Iceberg y Arrow garantizó la protección futura, la transparencia y la extensibilidad.

Como reemplazo de la infraestructura Hadoop/Hive, Dremio ofreció funcionalidad para casos de uso secundarios a través de la capa semántica. Si bien los mecanismos de ingesta de datos y ETL basados en Spark existentes se mantuvieron, Dremio proporcionó una capa de acceso unificada para facilitar el descubrimiento y la exploración de datos sin duplicación. Este enfoque redujo significativamente los factores de replicación de datos y desacopló el almacenamiento de la computación.

Beneficios: Con Dremio, NetApp logró importantes reducciones de costos al minimizar el consumo de cómputo y los requisitos de espacio en disco en sus entornos de datos. El nuevo Active IQ Data Lake está compuesto por 8.900 tablas que contienen 3 petabytes de datos, en comparación con la infraestructura anterior con más de 7 petabytes. La migración a Dremio también implicó la transición de 33 miniclústeres y 4000 núcleos a 16 nodos ejecutores en clústeres de Kubernetes. Incluso con disminuciones significativas en los recursos informáticos, NetApp experimentó mejoras notables en el rendimiento. Al acceder directamente a los datos a través de Dremio, el tiempo de ejecución de las consultas se redujo de 45 minutos a 2 minutos, lo que resultó en un tiempo 95 % más rápido para obtener información para el mantenimiento predictivo y la optimización. La migración también produjo una reducción de más del 60% en los costos de procesamiento, consultas más de 20 veces más rápidas y un ahorro de más del 30% en el costo total de propiedad (TCO).

Caso de uso de un cliente de venta de autopartes.

Desafíos: Dentro de esta empresa global de ventas de autopartes, los grupos de planificación y análisis financiero ejecutivo y corporativo no pudieron obtener una visión consolidada de los informes de ventas y se vieron obligados a leer los informes de métricas de ventas de cada línea de negocios e intentar consolidarlos. Esto dio lugar a que los clientes tomaran decisiones con datos que tenían al menos un día de antigüedad. Los plazos para obtener nuevos conocimientos analíticos normalmente demoran más de cuatro semanas. La solución de problemas en las canalizaciones de datos requeriría incluso más tiempo, sumando tres días o más al ya extenso cronograma. El lento proceso de desarrollo de informes, así como su rendimiento, obligaron a la comunidad de analistas a esperar continuamente a que los datos se procesaran o cargaran, en lugar de

permitirles descubrir nuevos conocimientos comerciales e impulsar nuevos comportamientos comerciales. Estos entornos problemáticos estaban compuestos por numerosas bases de datos diferentes para distintas líneas de negocio, lo que daba lugar a numerosos silos de datos. El entorno lento y fragmentado complicó la gobernanza de los datos, ya que había demasiadas formas para que los analistas elaboraran su propia versión de la verdad frente a una única fuente de verdad. El enfoque costó más de 1,9 millones de dólares en costos de plataforma de datos y personal. El mantenimiento de la plataforma heredada y el cumplimiento de las solicitudes de datos requerían siete ingenieros técnicos de campo (ETC) por año. Con el aumento de las solicitudes de datos, el equipo de inteligencia de datos no pudo escalar el entorno heredado para satisfacer las necesidades futuras.

Solución: Almacenar y administrar de forma rentable tablas Iceberg grandes en NetApp Object Store. Cree dominios de datos utilizando la capa semántica de Dremio, lo que permite a los usuarios comerciales crear, buscar y compartir productos de datos fácilmente.

Beneficios para el cliente: • Arquitectura de datos existente mejorada y optimizada y tiempo reducido para obtener información de cuatro semanas a solo horas • Tiempo de resolución de problemas reducido de tres días a solo horas • Costos de plataforma y administración de datos reducidos en más de $380,000 • (2) FTE de esfuerzo de inteligencia de datos ahorrados por año

Conclusión

En conclusión, este informe técnico ha proporcionado detalles de implementación completos de q Hybrid Iceberg Lakehouse con Dremio junto con varias fuentes de datos de controladores de almacenamiento de NetApp , incluidos ONTAP S3, NAS y StorageGRID. El proceso de implementación se ejecutó con éxito y se utilizó la herramienta de evaluación comparativa TPC-DS para realizar 99 consultas SQL en las diferentes fuentes de datos. El informe también ha explorado casos de uso de clientes dentro de NetApp, demostrando la versatilidad y eficacia de Dremio para satisfacer diversos requisitos comerciales. Además, se examinó un caso de uso específico que involucraba a un cliente de ventas de autopartes, destacando la aplicación práctica y los beneficios de aprovechar Dremio para el análisis y la obtención de información de datos.

En general, este documento sirve como un recurso valioso para comprender la implementación y el uso de Dremio con controladores de almacenamiento NetApp , mostrando sus capacidades y potencial para impulsar la toma de decisiones basada en datos y la optimización en diversas industrias.

Dónde encontrar información adicional

Para obtener más información sobre la información que se describe en este documento, revise los siguientes documentos y/o sitios web:

- Instalación de Zookeeper

  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/

- Configuración de Dremio con storageGRID

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

- Caso de uso de NetApp

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

Información de copyright

Información de la marca comercial