Solution hybride iceberg lakehouse de nouvelle génération de NetApp et Dremio

NetApp artificial intelligence solutions

NetApp
August 18, 2025

Sommaire

Solution hybride iceberg lakehouse de nouvelle génération de NetApp et Dremio

La solution hybride Iceberg Lakehouse de nouvelle génération de NetApp et Dremio

Dans ce document, nous discutons des détails de déploiement de Dremio avec différentes sources de données provenant de contrôleurs de stockage NetApp , notamment ONTAP S3, NAS et StorageGRID. Au cours du déploiement, nous avons utilisé l'outil d'analyse comparative TPC-DS pour exécuter 99 requêtes SQL sur diverses sources. Le document explore également les cas d'utilisation des clients au sein de NetApp, ainsi qu'un cas d'utilisation impliquant un client de vente de pièces automobiles.

Présentation de la solution

La solution Hybrid Iceberg Lakehouse offre des avantages uniques pour répondre aux défis auxquels sont confrontés les clients du lac de données. En tirant parti de la plate-forme Dremio Unified Lakehouse et des solutions NetApp ONTAP, StorageGRID et NetApp Cloud, les entreprises peuvent ajouter une valeur significative à leurs opérations commerciales. La solution fournit non seulement un accès à plusieurs sources de données, y compris les sources NetApp , mais améliore également les performances analytiques globales et aide les entreprises à générer des informations commerciales qui conduisent à la croissance de l'entreprise.

Présentation de NetApp

- Les offres de NetApp, telles que ONTAP et StorageGRID, permettent la séparation du stockage et du calcul, permettant une utilisation optimale des ressources en fonction d'exigences spécifiques. Cette flexibilité permet aux clients de faire évoluer leur stockage de manière indépendante à l'aide des solutions de stockage NetApp
- En exploitant les contrôleurs de stockage de NetApp, les clients peuvent efficacement fournir des données à leur base de données vectorielle à l'aide des protocoles NFS et S3. Ces protocoles facilitent le stockage des données client et gèrent l'index de la base de données vectorielle, éliminant ainsi le besoin de plusieurs copies de données accessibles via des méthodes de fichiers et d'objets.
- NetApp ONTAP fournit une prise en charge native du stockage NAS et d'objets auprès des principaux fournisseurs de services cloud tels qu'AWS, Azure et Google Cloud. Cette large compatibilité garantit une intégration transparente, permettant la mobilité des données client, l'accessibilité globale, la reprise après sinistre, l'évolutivité dynamique et des performances élevées.

StorageGRID

Notre stockage d'objets de pointe, storageGRID, offre un moteur de politique puissant pour le placement automatisé des données, des options de déploiement flexibles et une durabilité inégalée avec un codage d'effacement en couches. Il dispose d'une architecture évolutive prenant en charge des milliards d'objets et des pétaoctets de données dans un seul espace de noms. La solution permet l'intégration du cloud hybride, permettant la hiérarchisation des données vers les principales plates-formes cloud. Il a été reconnu comme

leader dans l'évaluation mondiale des fournisseurs basés sur les objets IDC Marketscape 2019.

De plus, storageGRID excelle dans la gestion de données non structurées à grande échelle avec un stockage d'objets défini par logiciel, une géo-redondance et des capacités multisites. Il intègre une gestion du cycle de vie des informations basée sur des politiques et offre des fonctionnalités d'intégration dans le cloud telles que la mise en miroir et la recherche. Elle dispose de diverses certifications, notamment Common Criteria, NF203 Digital Safe Component, ISO/IEC 25051, KPMG et Cohasset Compliance Assessment.

En résumé, NetApp storageGRID offre des fonctionnalités puissantes, une évolutivité, une intégration cloud hybride et des certifications de conformité pour une gestion efficace des données non structurées à grande échelle.

NetApp ONTAP

NetApp ONTAP est une solution de stockage robuste qui offre une large gamme de fonctionnalités d'entreprise. Il inclut Snapshot, qui fournit des sauvegardes instantanées cohérentes avec les applications et inviolables. SnapRestore permet une restauration quasi instantanée des sauvegardes à la demande, tandis que SnapMirror offre des fonctionnalités intégrées de sauvegarde à distance et de reprise après sinistre. La solution intègre également la protection autonome contre les ransomwares (ARP), garantissant la sécurité des données avec des fonctionnalités telles que la vérification multi-administrateur, le cryptage des données au repos avec certification FIPS, le cryptage des données en transit, l'authentification multifacteur (MFA) et le contrôle d'accès basé sur les rôles (RBAC). La journalisation complète, l'audit, la gestion des clés intégrées et externes, la purge sécurisée et la gestion sécurisée de plusieurs locataires améliorent encore la sécurité et la conformité des données.

NetApp ONTAP propose également SnapLock, qui offre une conservation des données conforme à la réglementation avec des niveaux élevés d'intégrité, de performances et de conservation à un faible coût total de possession. Il est entièrement intégré à NetApp ONTAP 9 et offre une protection contre les actes malveillants, les administrateurs malveillants et les ransomwares.

La solution comprend le cryptage NSE/NVE pour le cryptage en vol et des données au repos, l'accès administrateur multifacteur et la vérification multi-administrateur. Active IQ fournit des analyses prédictives et des mesures correctives basées sur l'IA, tandis que QoS garantit la qualité du contrôle de la charge de travail. L'intégration de la gestion et de l'automatisation est intuitive via SysMgr/GUI/CLI/API. FabricPool permet la hiérarchisation automatique des données et la solution offre une efficacité grâce à la compression, la déduplication et le compactage des données en ligne. NetApp garantit l'atteinte des objectifs d'efficacité de la charge de travail sans frais pour le client.

NetApp ONTAP prend en charge divers protocoles, notamment NVMe/FC, FC, NVMe/TCP, iSCSI, NFS, SMB et S3, ce qui en fait une solution de stockage unifiée. Dans l'ensemble, NetApp ONTAP offre des fonctionnalités d'entreprise étendues, une sécurité robuste, une conformité, une efficacité et une polyvalence pour répondre à divers besoins de stockage.

Présentation de Dremio

Dremio est la plateforme Lakehouse unifiée pour l'analyse en libre-service et l'IA. La plateforme d'analyse unifiée Dremio rapproche les utilisateurs des données avec la flexibilité, l'évolutivité et les performances du Lakehouse à une fraction du coût des solutions d'entrepôt de données existantes. Dremio permet des analyses « shift-left » pour éliminer l'intégration de données et l'ETL complexes et coûteux, offrant ainsi des analyses transparentes à l'échelle de l'entreprise sans aucun mouvement de données. Dremio propose également :

- Analyses en libre-service faciles à utiliser, activées grâce à une couche sémantique universelle et à un moteur de requête SQL hautement performant et étroitement intégré, facilitant la connexion, la gouvernance et l'analyse de toutes les données, à la fois dans le cloud et sur site.

- Les capacités de gestion de lakehouse natives d'Apache Iceberg de Dremio simplifient la découverte de données et automatisent l'optimisation des données, offrant des analyses hautes performances avec un versionnage de données inspiré de Git.
- Fondamentalement construit sur l'open source et les normes ouvertes, Dremio permet aux entreprises d'éviter le verrouillage et de rester positionnées pour l'innovation. Les entreprises font confiance à Dremio comme étant la plateforme lakehouse la plus simple à utiliser avec le meilleur rapport qualité-prix pour toutes les charges de travail.

Quelle valeur la solution Dremio et NetApp Hybrid Iceberg Lakehouse apporte-t-elle aux clients ?

- **Gestion et accessibilité des données améliorées** : Dremio est bien connu pour sa plateforme de data lakehouse qui permet aux organisations d'interroger les données directement à partir de leurs lacs de données à grande vitesse. NetApp, quant à lui, est un fournisseur leader de services de données cloud et de solutions de stockage de données. L'offre conjointe offre aux clients une solution complète pour stocker, gérer, accéder et analyser les données de leur entreprise de manière efficace et efficiente.
- **Optimisation des performances** : Grâce à l'expertise de NetApp en matière de stockage de données et aux capacités de Dremio en matière de traitement et d'optimisation des données, le partenariat offre une solution qui améliore les performances des opérations de données, réduit la latence et augmente la vitesse d'accès aux informations commerciales. Dremio a même apporté des avantages en termes de performances à l'infrastructure analytique informatique interne de NetApp.
- **Évolutivité** : Dremio et NetApp proposent tous deux une solution conçue pour évoluer. La solution commune offre aux clients des environnements de stockage de données, de gestion de données et d'analyse hautement évolutifs. Dans un environnement Hybrid Iceberg Lakehouse, le moteur de requête Dremio SQL associé à NetApp StorageGRID offre une évolutivité, une concurrence et des performances de requête inégalées, capables de gérer les besoins analytiques de toute entreprise.
- **Sécurité et gouvernance des données** : Les deux entreprises mettent fortement l'accent sur la sécurité et la gouvernance des données. Ensemble, ils offrent des fonctionnalités robustes de sécurité et de gouvernance des données, garantissant que les données sont protégées et que les exigences de gouvernance des données sont respectées. Des fonctionnalités telles que les contrôles d'accès basés sur les rôles et à granularité fine, l'audit complet, la lignée de données de bout en bout, la gestion unifiée des identités et l'authentification unique avec un cadre de conformité et de sécurité étendu garantissent que les environnements de données analytiques des entreprises sont sécurisés et gouvernés.
- **Efficacité des coûts** : En intégrant le moteur de lac de données de Dremio aux solutions de stockage de NetApp, les clients peuvent réduire les coûts associés à la gestion et au déplacement des données. Les organisations peuvent également passer d'environnements de lac de données hérités à une solution lakehouse plus moderne composée de NetApp et Dremio. Cette solution Hybrid Iceberg Lakehouse offre des performances de requête à grande vitesse et une concurrence de requêtes à la pointe du marché qui réduit le coût total de possession et le temps d'analyse de l'entreprise.

Exigences technologiques

Les configurations matérielles et logicielles décrites ci-dessous ont été utilisées pour les validations effectuées dans ce document. Ces configurations servent de guide pour vous aider à configurer votre environnement. Cependant, veuillez noter que les composants spécifiques peuvent varier en fonction des exigences individuelles des clients.

Configuration matérielle requise

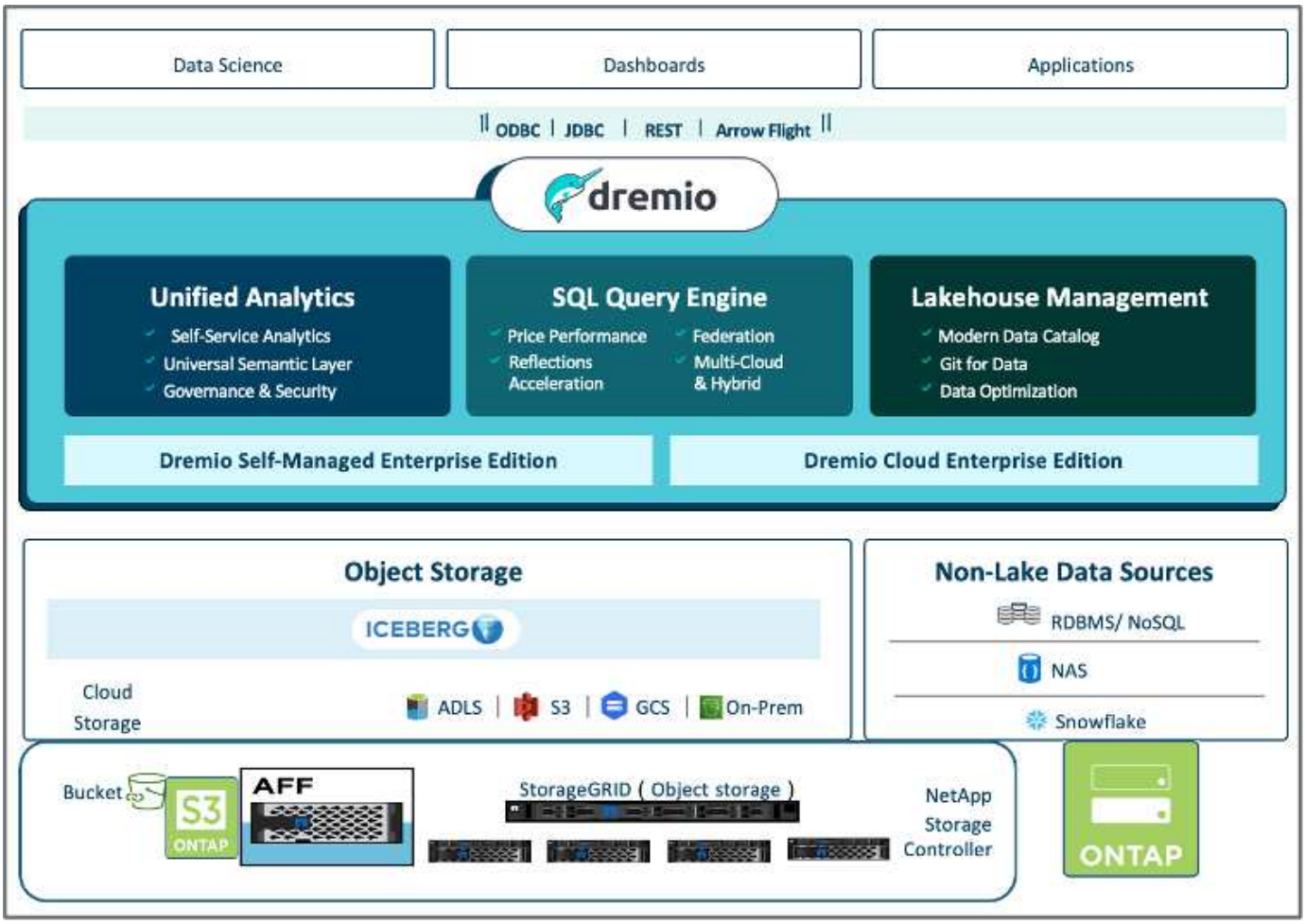
Matériel	Détails
Paire de baies de stockage NetApp AFF HA	• A800 • ONTAP 9.14.1 • 48 x 3,49 To SSD-NVM • Deux compartiments S3 : métadonnées Dremio et données client.
4 x FUJITSU PRIMERGY RX2540 M4	• 64 processeurs • Processeur Intel Xeon Gold 6142 à 2,60 GHz • 256 GM de mémoire physique • 1 port réseau 100 GbE
Réseautage	• 100 GbE
StorageGRID	* 1 x SG100, 3xSGF6024 * 3 x 24 x 7,68 To * Deux compartiments S3 : métadonnées Dremio et données client.

Configuration logicielle requise

Logiciels	Détails
Dremio	• version - 25.0.3-202405170357270647-d2042e1b • Édition Entreprise
Sur site	• Cluster Dremio à 5 nœuds • 1 maître coordinateur et 4 exécuteurs

Procédure de déploiement

Dans cette validation d’architecture de référence, nous avons utilisé une configuration Dremio composée d’un coordinateur et de quatre
exécuteurs



Configuration de NetApp

- Initialisation du système de stockage
- Création d'une machine virtuelle de stockage (SVM)
- Affectation des interfaces réseau logiques
- Configuration et licence NFS, S3

Veuillez suivre les étapes ci-dessous pour NFS (Network File System) : 1. Créez un volume Flex Group pour NFSv4 ou NFSv3. Dans notre configuration pour cette validation, nous avons utilisé 48 SSD, 1 SSD dédié au volume racine du contrôleur et 47 SSD répartis pour NFSv4]]. Vérifiez que la stratégie d'exportation NFS pour le volume Flex Group dispose d'autorisations de lecture/écriture pour le réseau des serveurs Dremio.

1. Sur tous les serveurs Dremio, créez un dossier et montez le volume Flex Group sur ce dossier via une interface logique (LIF) sur chaque serveur Dremio.

Veuillez suivre les étapes ci-dessous pour S3 (Simple Storage Service) :

1. Configurez un serveur de magasin d'objets avec HTTP activé et le statut d'administrateur défini sur « up » à l'aide de la commande « vserver object-store-server create ». Vous avez la possibilité d'activer HTTPS et de définir un port d'écoute personnalisé.
2. Créez un utilisateur object-store-server à l'aide de la commande « vserver object-store-server user create -user <username> ».
3. Pour obtenir la clé d'accès et la clé secrète, vous pouvez exécuter la commande suivante : « set diag;

vserver object-store-server user show -user <username> ». Cependant, à l'avenir, ces clés seront fournies lors du processus de création de l'utilisateur ou pourront être récupérées à l'aide d'appels d'API REST.

4. Créez un groupe object-store-server à l'aide de l'utilisateur créé à l'étape 2 et accordez l'accès. Dans cet exemple, nous avons fourni « FullAccess ».
5. Créez deux compartiments S3 en définissant leur type sur « S3 ». Un pour la configuration Dremio et un pour les données client.

Configuration du gardien de zoo

Vous pouvez utiliser la configuration zookeeper fournie par Dremio. Dans cette validation, nous avons utilisé un zookeeper distinct. Nous avons suivi les étapes mentionnées dans ce lien Web. https://medium.com/@ahmetfurkandemir/distributed-hadoop-cluster-1-spark-with-all-dependincies-03c8ec616166

Configuration de Dremio

Nous avons suivi ce lien Web pour installer Dremio via tar ball.

1. Créez un groupe Dremio.

```
sudo groupadd -r dremio
```

2. Créez un utilisateur dremio.

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

3. Créez des répertoires 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. Téléchargez le fichier tar depuis https://download.dremio.com/community-server/
5. Décompressez Dremio dans le répertoire /opt/dremio.

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

6. Créez un lien symbolique pour le dossier de configuration.

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

7. Configurez votre configuration de service (configuration SystemD).

a. Copiez le fichier d'unité pour le démon dremio de /opt/dremio/share/dremio.service vers /etc/systemd/system/dremio.service.

b. Redémarrer le système

```
sudo systemctl daemon-reload
```

c. Activer dremio pour démarrer au démarrage.

```
sudo systemctl enable dremio
```

8. Configurer Dremio sur le coordinateur. Voir Configuration Dremio pour plus d'informations

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#
```

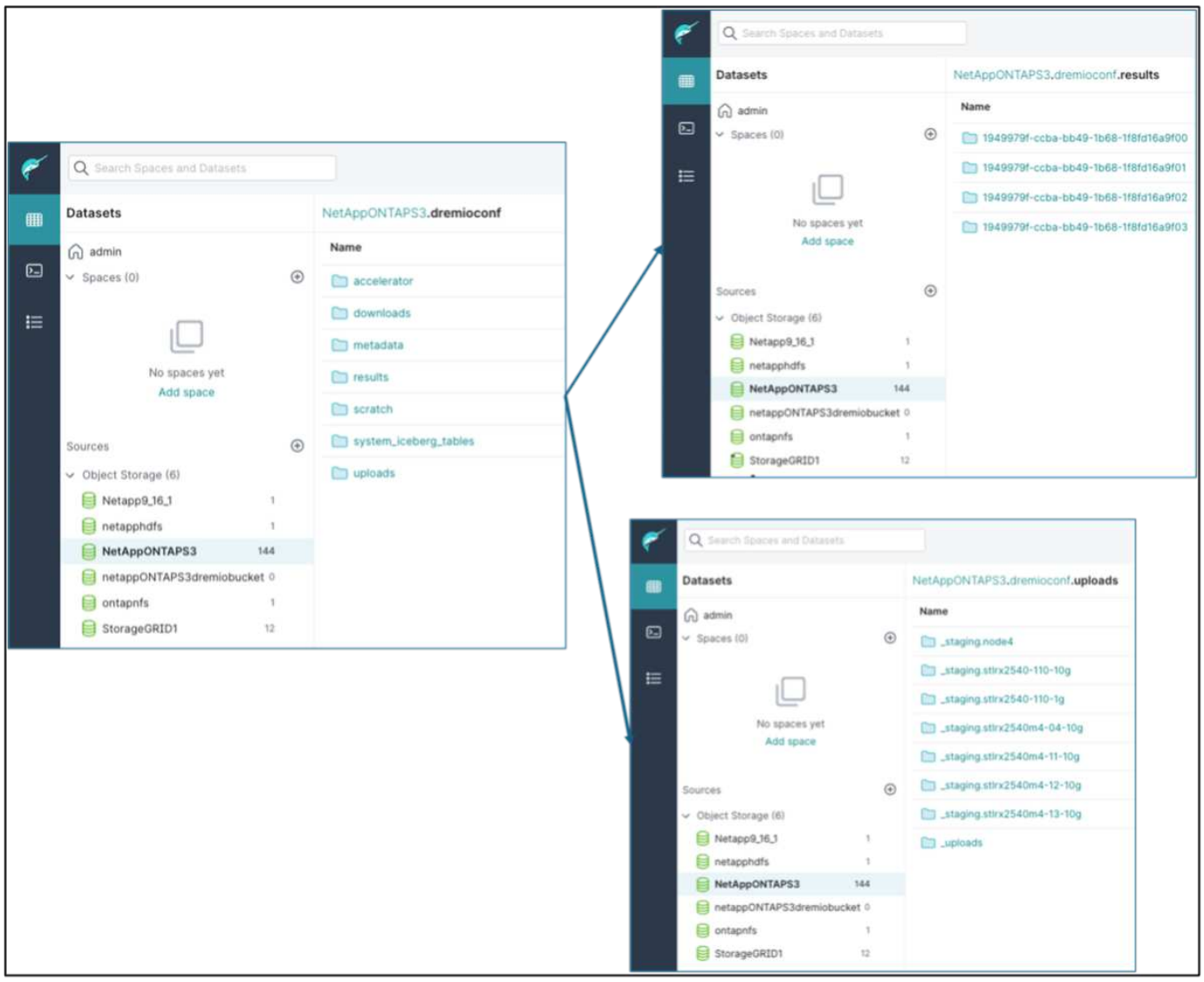
b. Core-site.xml

```
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. La configuration Dremio est stockée dans le stockage d'objets NetApp . Dans notre validation, le bucket « dremioconf » réside dans un bucket S3 ontap. L'image ci-dessous montre quelques détails des dossiers « scratch » et « uploads » du bucket S3 « dremioconf ».



1. Configurer Dremio sur les exécuteurs. Dans notre configuration, nous avons 3 exécuteurs.
    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 – identique à la configuration du coordinateur.

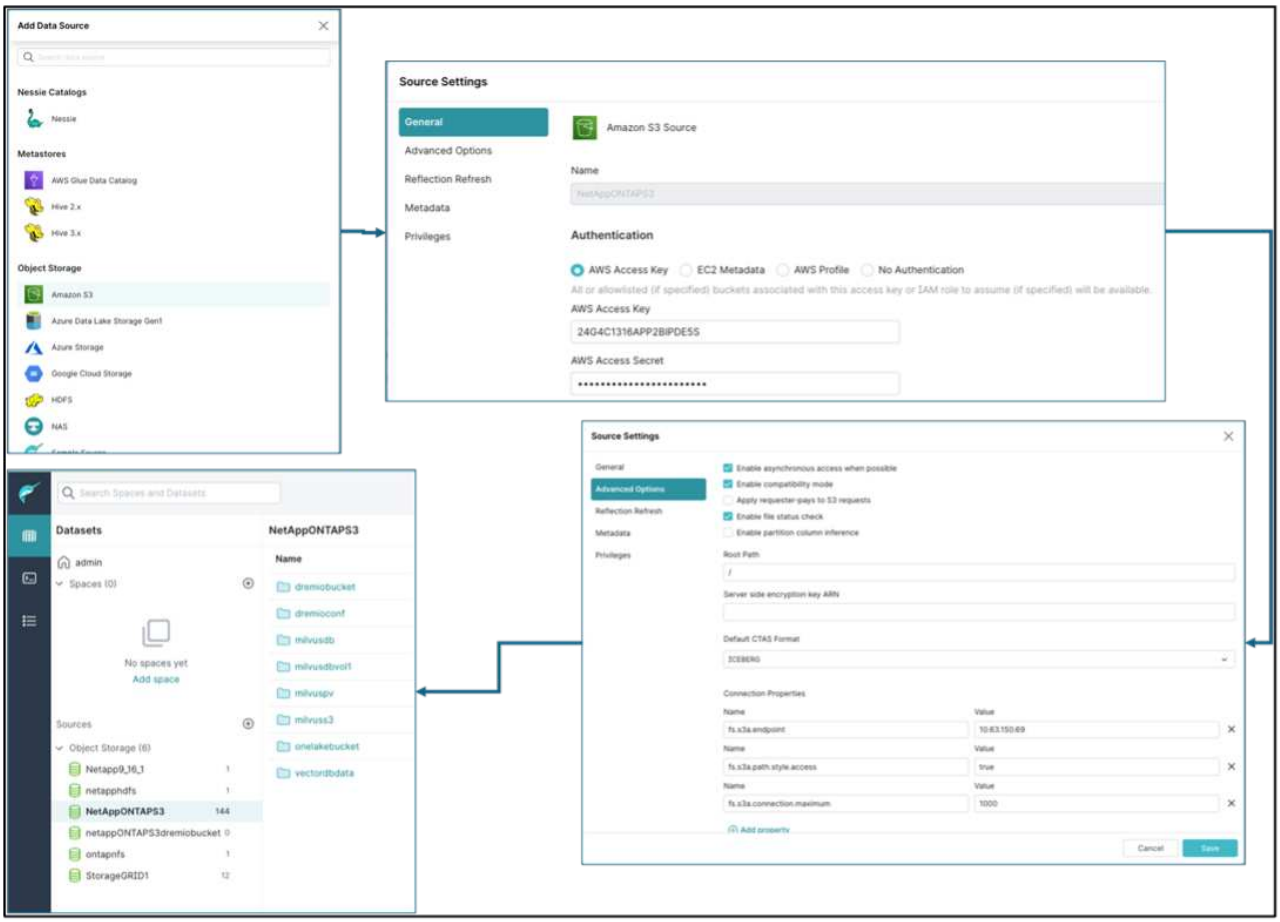
NetApp recommande StorageGRID comme solution de stockage d'objets principale pour les environnements Datalake et Lakehouse. De plus, NetApp ONTAP est utilisé pour la dualité fichier/objet. Dans le cadre de ce document, nous avons effectué des tests sur ONTAP S3 en réponse à une demande client, et il fonctionne avec succès comme source de données.

Configuration de sources multiples

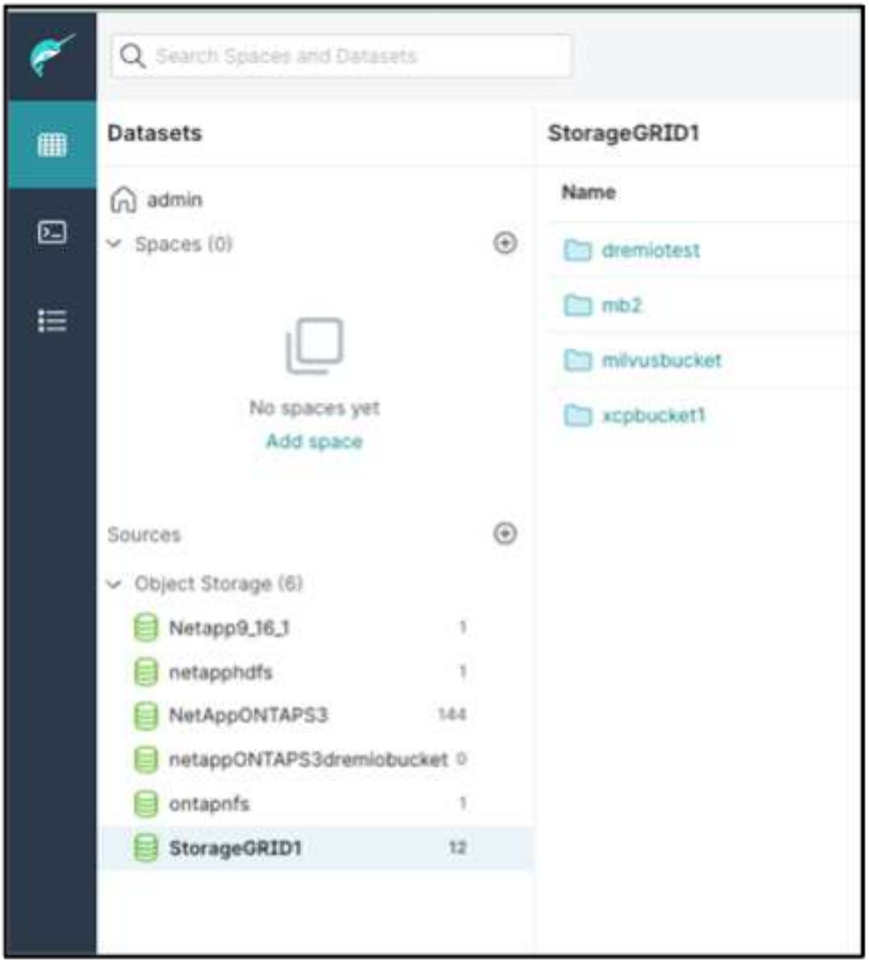
1. Configurez ONTAP S3 et storageGRID comme source s3 dans Dremio.
   a. Tableau de bord Dremio → ensembles de données → sources → ajouter une source.
   b. Dans la section générale, veuillez mettre à jour l'accès AWS et la clé secrète
   c. Dans l'option avancée, activez le mode de compatibilité, mettez à jour les propriétés de connexion avec les détails ci-dessous. L'adresse IP/le nom du point de terminaison du contrôleur de stockage NetApp provenant d'ontap S3 ou de storageGRID.

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

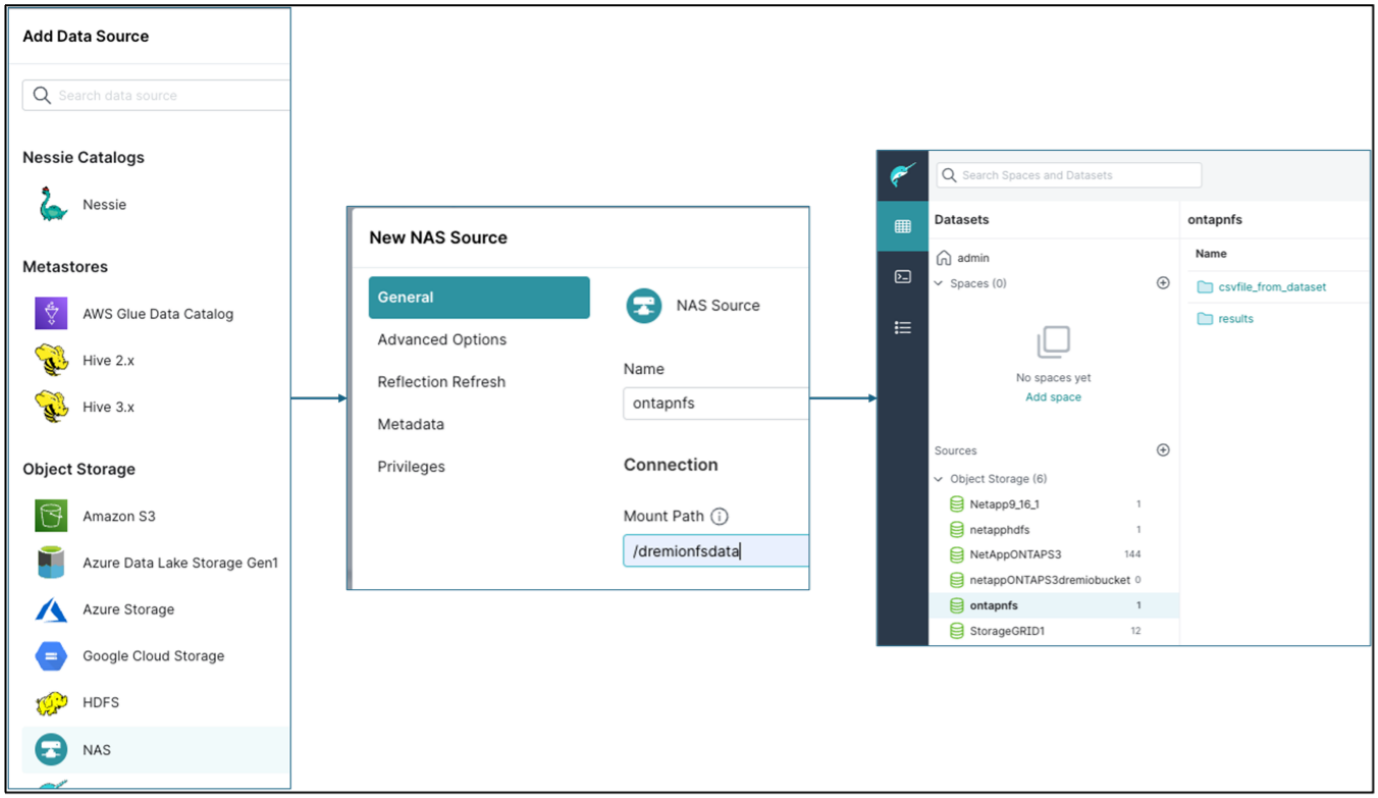
   d. Activer la mise en cache locale lorsque cela est possible, pourcentage maximal du cache total disponible à utiliser lorsque cela est possible = 100
   e. Affichez ensuite la liste des buckets du stockage d'objets NetApp
   .



f. Exemple de vue des détails du bucket

storageGRID 

2. Configurer NAS (en particulier NFS) comme source dans Dremio.
   a. Tableau de bord Dremio → ensembles de données → sources → ajouter une source.
   b. Dans la section générale, entrez le nom et le chemin de montage NFS. Assurez-vous que le chemin de montage NFS est monté sur le même dossier sur tous les nœuds du 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:~#
```

Présentation de la vérification des solutions

Dans cette section, nous avons exécuté des requêtes de test SQL à partir de plusieurs sources pour vérifier la fonctionnalité, tester et vérifier le débordement vers le stockage NetApp .

Requête SQL sur le stockage d'objets

1. Définissez la mémoire à 250 Go par serveur dans 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:~#
```

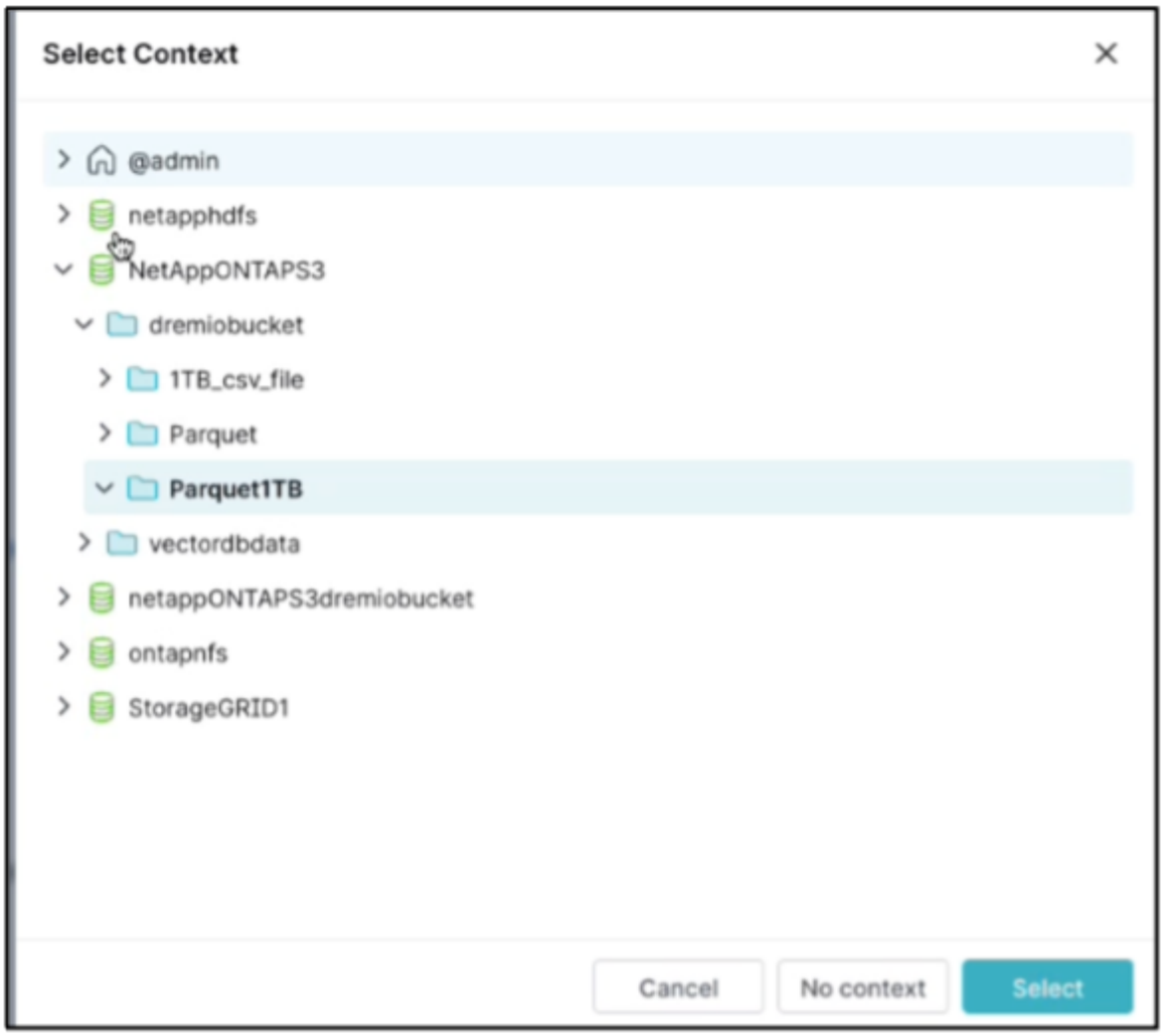
2. Vérifiez l'emplacement de débordement (${DREMIO_HOME}"/dremiocache) dans le fichier dremio.conf et les détails de stockage.

```
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
```

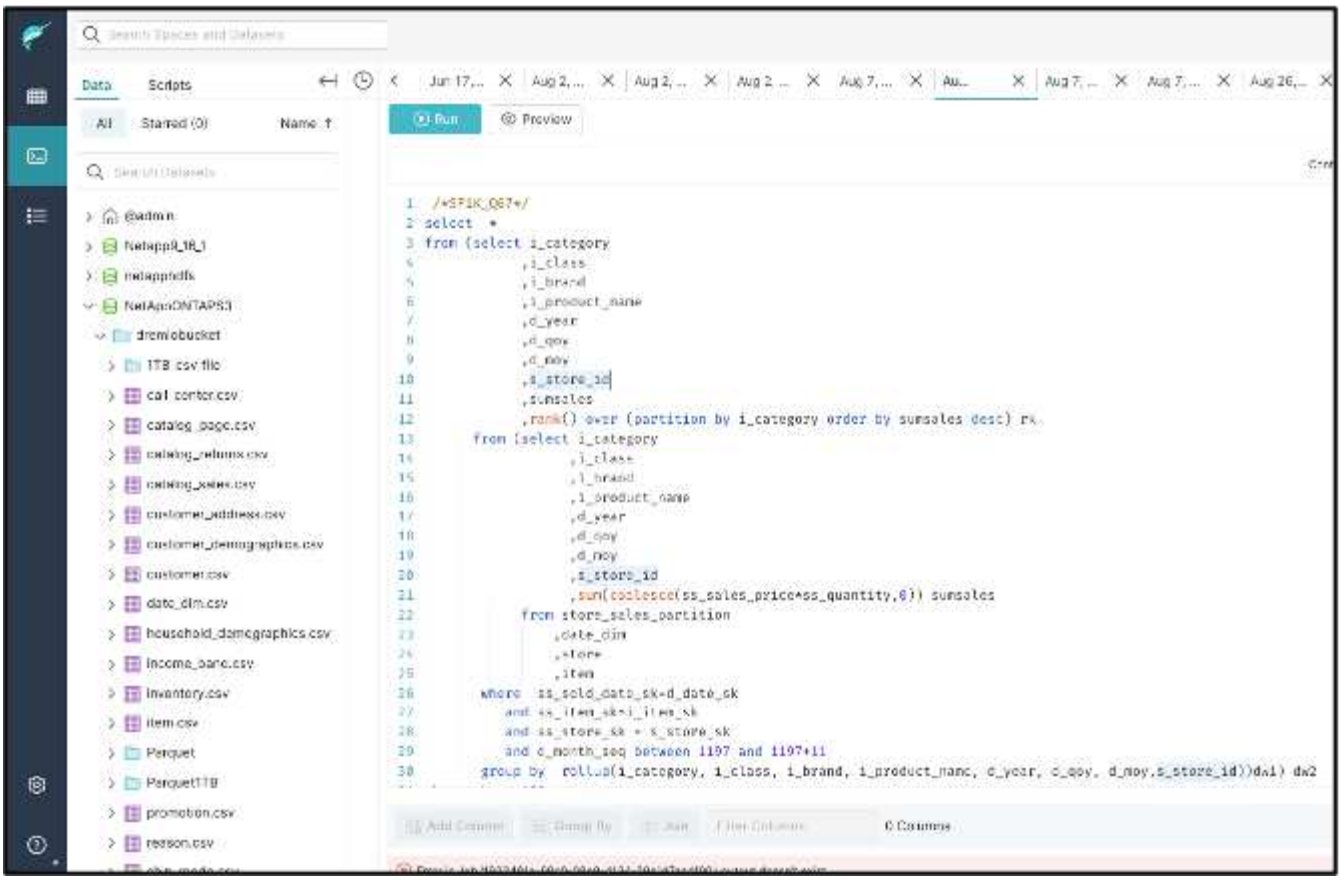
3. Dirigez l'emplacement de débordement de Dremio vers le stockage NetApp NFS

```
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:~#
```

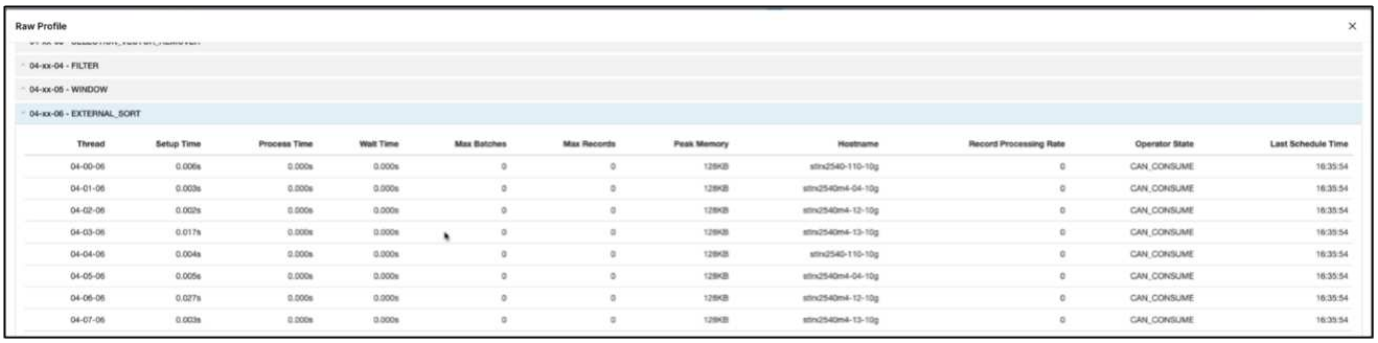
4. Sélectionnez le contexte. Dans notre test, nous avons exécuté le test sur les fichiers parquet générés par TPCDS résidant dans ONTAP S3. Tableau de bord Dremio → Exécuteur SQL → Contexte → NetAppONTAPS3 → Parquet1TB



1. Exécutez la requête TPC-DS67 à partir du tableau de bord Dremio



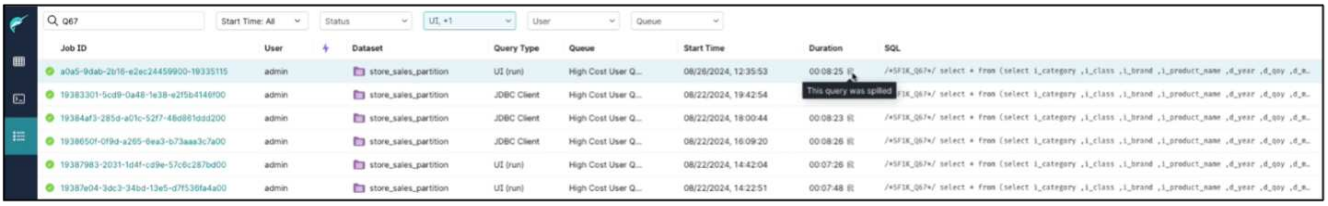
1. Vérifiez que le travail est en cours d'exécution sur tous les exécuteurs. Tableau de bord Dremio → tâches → <jobid> → profil brut → sélectionnez EXTERNAL_SORT → Nom d'hôte



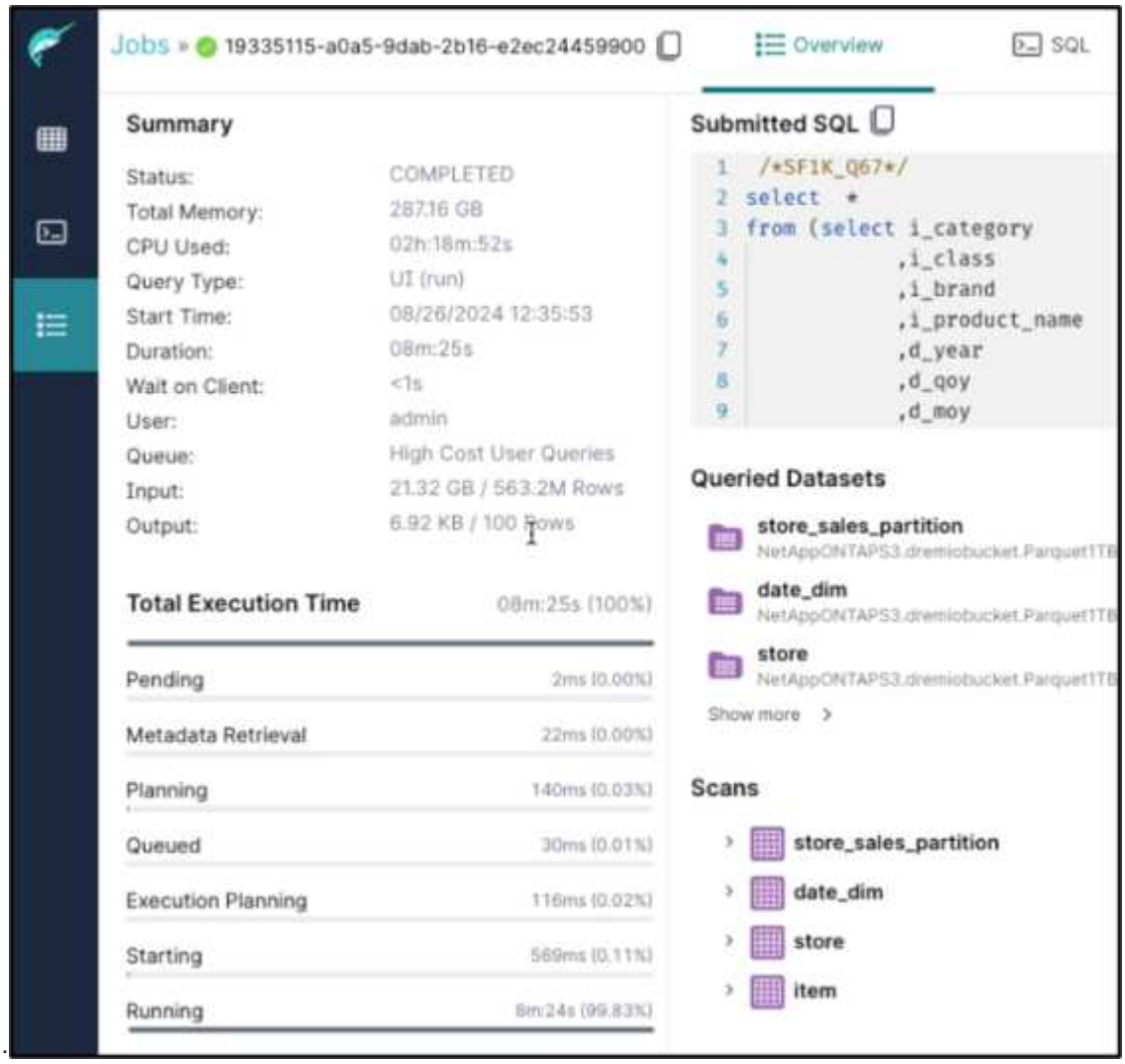
1. Lorsque la requête SQL est en cours d'exécution, vous pouvez vérifier le dossier divisé pour la mise en cache des données dans le contrôleur de stockage NetApp .

```
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
```

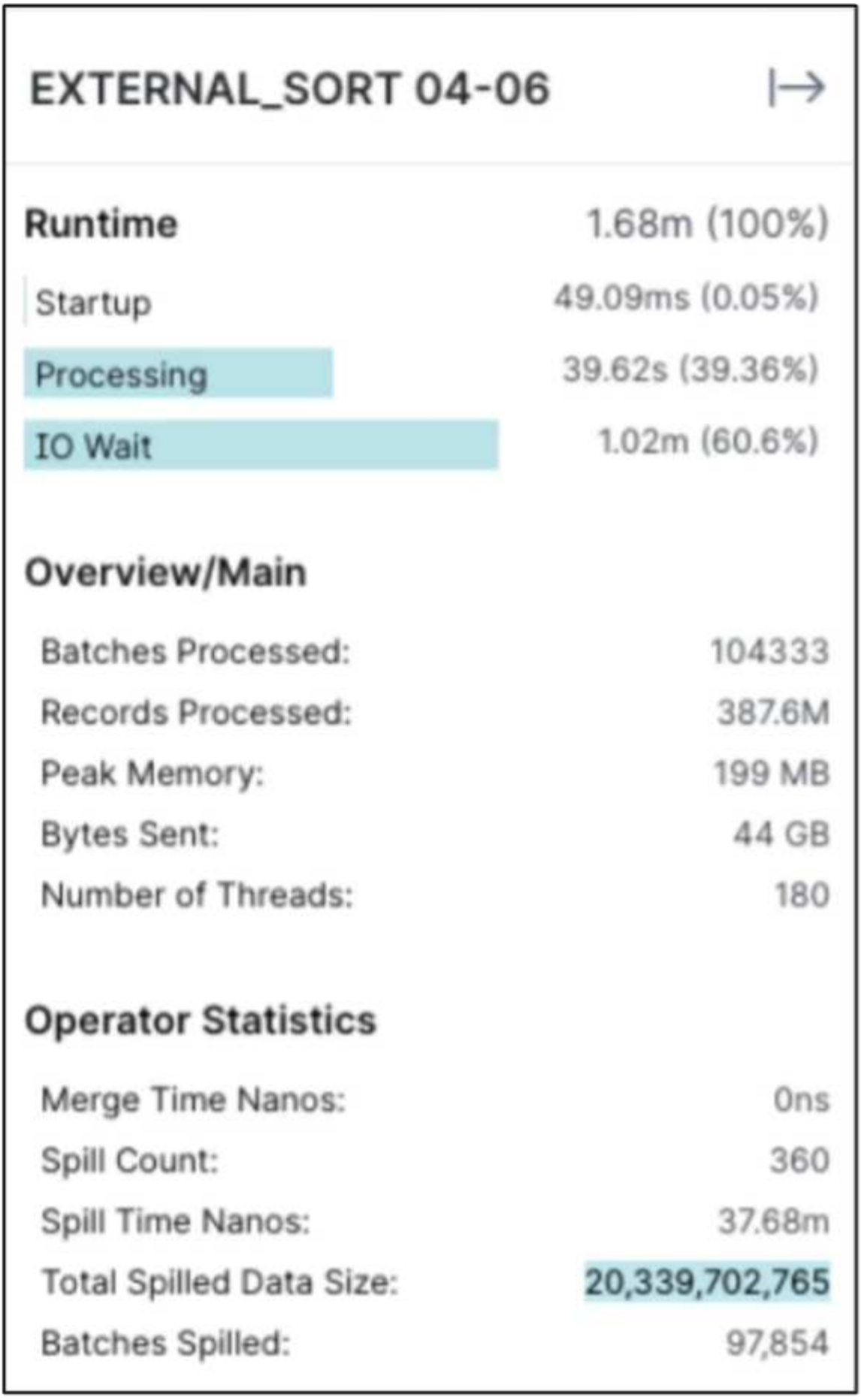
2. La requête SQL complétée avec débordement



3. Résumé de l'achèvement du travail.



4. Vérifiez la taille des données

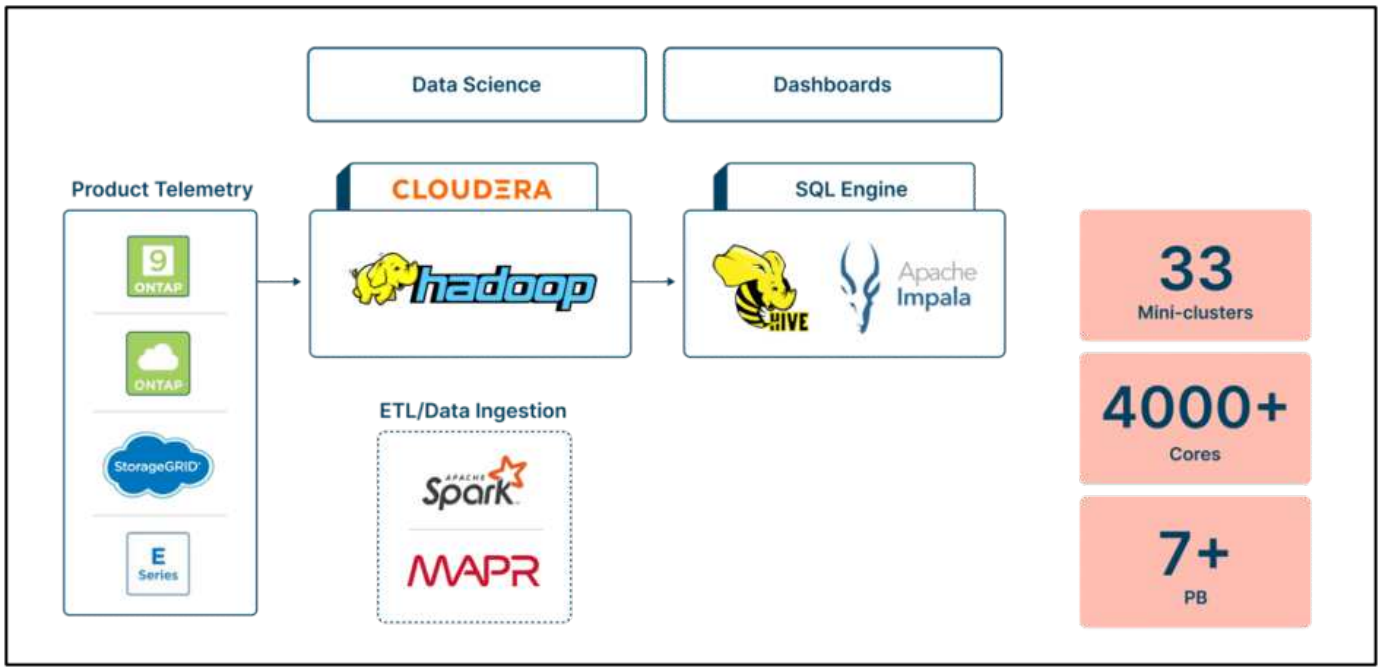

EXTERNAL_SORT 04-06
Runtime
1.68m (100%)
Startup
49.09ms (0.05%)
Processing
39.62s (39.36%)
IO Wait
1.02m (60.6%)
Overview/Main
Batches Processed:
104333
Records Processed:
387.6M
Peak Memory:
199 MB
Bytes Sent:
44 GB
Number of Threads:
180
Operator Statistics
Merge Time Nanos:
0ns
Spill Count:
360
Spill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

renversées

La même procédure s'applique au stockage d'objets NAS et StorageGRID .

Cas d'utilisation client

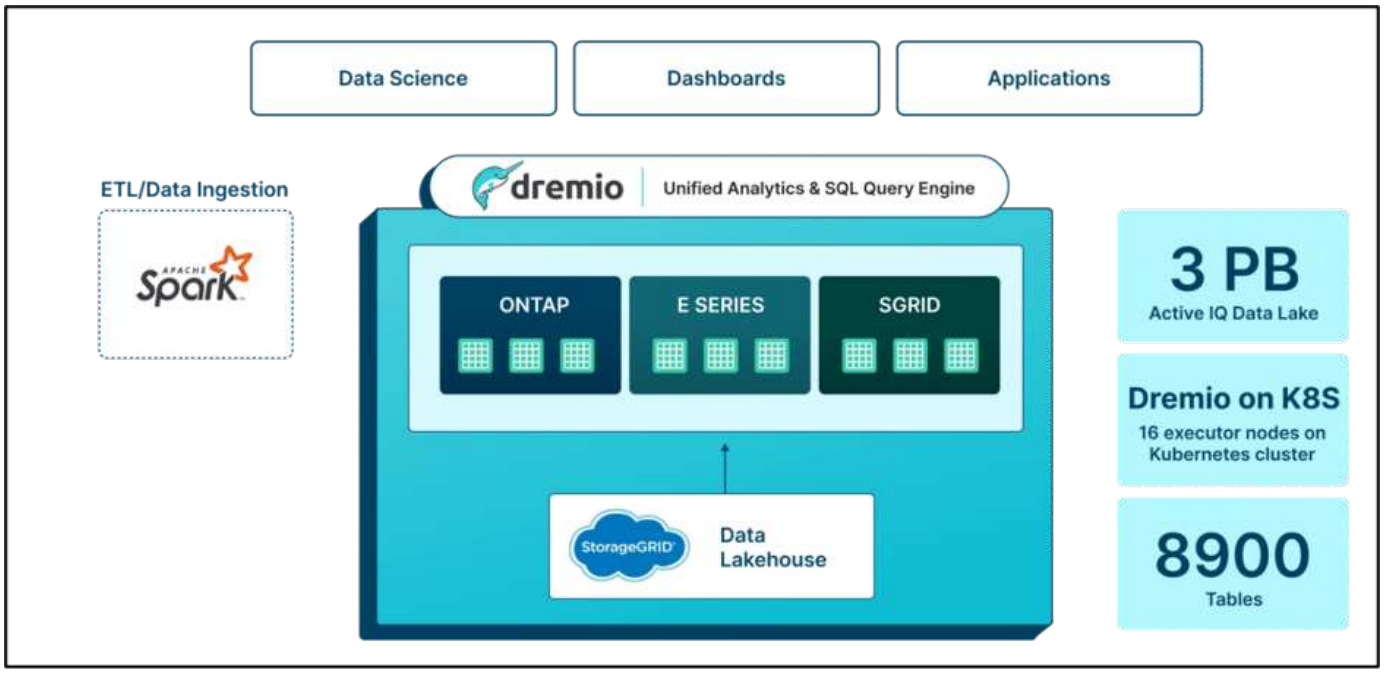
Cas d'utilisation de NetApp ActiveIQ



Défi : La solution interne Active IQ de NetApp, initialement conçue pour prendre en charge de nombreux cas d'utilisation, a évolué vers une offre complète destinée aux utilisateurs internes et aux clients. Cependant, l'infrastructure back-end sous-jacente basée sur Hadoop/MapR posait des problèmes de coût et de performance, en raison de la croissance rapide des données et du besoin d'un accès efficace aux données. La mise à l'échelle du stockage impliquait l'ajout de ressources informatiques inutiles, ce qui entraînait une augmentation des coûts.

De plus, la gestion du cluster Hadoop prenait du temps et nécessitait une expertise spécialisée. Les problèmes de performance et de gestion des données ont encore compliqué la situation, les requêtes prenant en moyenne 45 minutes et les ressources étant limitées en raison de mauvaises configurations. Pour relever ces défis, NetApp a cherché une alternative à l'environnement Hadoop existant et a déterminé qu'une nouvelle solution moderne basée sur Dremio réduirait les coûts, découplerait le stockage et le calcul, améliorerait les performances, simplifierait la gestion des données, offrirait des contrôles précis et fournirait des capacités de reprise après sinistre.

Solution:



Dremio a permis à NetApp de moderniser son infrastructure de données basée sur Hadoop selon une approche progressive, fournissant une feuille de route pour une analyse unifiée. Contrairement à d'autres fournisseurs qui nécessitaient des modifications importantes du traitement des données, Dremio s'est parfaitement intégré aux pipelines existants, économisant ainsi du temps et des dépenses lors de la migration. En passant à un environnement entièrement conteneurisé, NetApp a réduit les frais de gestion, amélioré la sécurité et renforcé la résilience. L'adoption par Dremio d'écosystèmes ouverts tels qu'Apache Iceberg et Arrow a assuré la pérennité, la transparence et l'extensibilité.

En remplacement de l'infrastructure Hadoop/Hive, Dremio offrait des fonctionnalités pour les cas d'utilisation secondaires via la couche sémantique. Bien que les mécanismes d'ETL et d'ingestion de données basés sur Spark existants soient restés, Dremio a fourni une couche d'accès unifiée pour faciliter la découverte et l'exploration des données sans duplication. Cette approche a considérablement réduit les facteurs de réplication des données et découplé le stockage et le calcul.

Avantages : Avec Dremio, NetApp a réalisé des réductions de coûts significatives en minimisant la consommation de calcul et les besoins en espace disque dans ses environnements de données. Le nouveau lac de données Active IQ est composé de 8 900 tables contenant 3 pétaoctets de données, contre plus de 7 pétaoctets pour l'infrastructure précédente. La migration vers Dremio impliquait également la transition de 33 mini-clusters et 4 000 cœurs vers 16 nœuds exécuteurs sur des clusters Kubernetes. Même avec des diminutions significatives des ressources informatiques, NetApp a connu des améliorations de performances remarquables. En accédant directement aux données via Dremio, le temps d'exécution des requêtes est passé de 45 minutes à 2 minutes, ce qui a permis d'obtenir des informations 95 % plus rapidement pour la maintenance prédictive et l'optimisation. La migration a également permis une réduction de plus de 60 % des coûts de calcul, des requêtes plus de 20 fois plus rapides et des économies de plus de 30 % sur le coût total de possession (TCO).

Cas d'utilisation client de vente de pièces automobiles.

Défis : Au sein de cette société mondiale de vente de pièces automobiles, les groupes de planification et d'analyse financière de la direction et de l'entreprise n'étaient pas en mesure d'obtenir une vue consolidée des rapports de vente et étaient obligés de lire les rapports de mesures de vente de chaque secteur d'activité et de tenter de les consolider. Cela a conduit les clients à prendre des décisions avec des données datant d'au moins un jour. Les délais d'obtention de nouvelles informations analytiques prennent généralement plus de quatre semaines. Le dépannage des pipelines de données nécessiterait encore plus de temps, ajoutant trois jours ou plus au délai déjà long. Le processus lent de développement des rapports ainsi que les performances

des rapports ont obligé la communauté des analystes à attendre continuellement que les données soient traitées ou chargées, plutôt que de leur permettre de trouver de nouvelles perspectives commerciales et de stimuler de nouveaux comportements commerciaux. Ces environnements perturbés étaient composés de nombreuses bases de données différentes pour différents secteurs d'activité, ce qui a donné lieu à de nombreux silos de données. L'environnement lent et fragmenté a compliqué la gouvernance des données, car les analystes disposaient de trop de moyens pour élaborer leur propre version de la vérité par rapport à une source unique de vérité. L'approche a coûté plus de 1,9 million de dollars en frais de plateforme de données et de personnel. La maintenance de la plateforme existante et le traitement des demandes de données nécessitaient sept ingénieurs techniques de terrain (ETP) par an. Avec l'augmentation des demandes de données, l'équipe de renseignement sur les données n'a pas pu adapter l'environnement existant pour répondre aux besoins futurs.

Solution : Stockez et gérez de manière rentable de grandes tables Iceberg dans NetApp Object Store. Créez des domaines de données à l'aide de la couche sémantique de Dremio, permettant aux utilisateurs professionnels de créer, rechercher et partager facilement des produits de données.

Avantages pour le client : • Architecture de données existante améliorée et optimisée et temps d'obtention d'informations réduit de quatre semaines à quelques heures seulement • Temps de dépannage réduit de trois jours à quelques heures seulement • Coûts de gestion et de plateforme de données réduits de plus de 380 000 $ • (2) ETP d'efforts de renseignement sur les données économisés par an

Conclusion

En conclusion, ce rapport technique a fourni des détails de déploiement complets de q Hybrid Iceberg Lakehouse avec Dremio en conjonction avec diverses sources de données provenant de contrôleurs de stockage NetApp , notamment ONTAP S3, NAS et StorageGRID. Le processus de déploiement a été exécuté avec succès et l'outil d'analyse comparative TPC-DS a été utilisé pour effectuer 99 requêtes SQL sur les différentes sources de données. Le rapport a également exploré les cas d'utilisation des clients au sein de NetApp, démontrant la polyvalence et l'efficacité de Dremio pour répondre à diverses exigences commerciales. De plus, un cas d'utilisation spécifique impliquant un client de vente de pièces automobiles a été examiné, mettant en évidence l'application pratique et les avantages de l'exploitation de Dremio pour l'analyse et la compréhension des données.

Dans l'ensemble, ce document constitue une ressource précieuse pour comprendre le déploiement et l'utilisation de Dremio avec les contrôleurs de stockage NetApp , en présentant ses capacités et son potentiel pour favoriser la prise de décision et l'optimisation basées sur les données dans divers secteurs.

Où trouver des informations supplémentaires

Pour en savoir plus sur les informations décrites dans ce document, consultez les documents et/ou sites Web suivants :

- Installation du gardien de zoo

  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/

- Configuration de Dremio avec storageGRID

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

- Cas d'utilisation de NetApp

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

Informations sur le copyright

Informations sur les marques commerciales