



La soluzione ibrida iceberg lakehouse di nuova generazione di NetApp e Dremio

NetApp artificial intelligence solutions

NetApp
August 18, 2025

Sommario

- La soluzione ibrida iceberg lakehouse di nuova generazione di NetApp e Dremio 1
 - La soluzione ibrida Iceberg Lakehouse di nuova generazione di NetApp e Dremio 1
 - Panoramica della soluzione 1
 - Panoramica NetApp 1
 - Panoramica di Dremio 2
 - Quale valore offre ai clienti la soluzione Dremio e NetApp Hybrid Iceberg Lakehouse? 3
 - Requisiti tecnologici 3
 - Procedura di distribuzione 4
 - Panoramica della verifica della soluzione 15
 - Casi d'uso dei clienti 22
 - Caso d'uso di NetApp ActiveIQ 22
 - Caso d'uso del cliente nella vendita di ricambi auto 23
 - Conclusione 24
 - Dove trovare ulteriori informazioni 24

La soluzione ibrida iceberg lakehouse di nuova generazione di NetApp e Dremio

La soluzione ibrida Iceberg Lakehouse di nuova generazione di NetApp e Dremio

In questo documento, analizziamo i dettagli di distribuzione di Dremio con diverse fonti di dati provenienti dai controller di storage NetApp, tra cui ONTAP S3, NAS e StorageGRID. Durante l'implementazione, abbiamo utilizzato lo strumento di benchmarking TPC-DS per eseguire 99 query SQL su varie origini. Il documento esamina anche i casi d'uso dei clienti all'interno di NetApp, nonché un caso d'uso che coinvolge un cliente di vendita di ricambi auto.

Panoramica della soluzione

La soluzione Hybrid Iceberg Lakehouse offre vantaggi esclusivi per affrontare le sfide che i clienti dei data lake si trovano ad affrontare. Sfruttando la piattaforma Dremio Unified Lakehouse e le soluzioni NetApp ONTAP, StorageGRID e NetApp Cloud, le aziende possono aggiungere un valore significativo alle loro operazioni aziendali. La soluzione non solo fornisce accesso a più fonti di dati, tra cui quelle NetApp, ma migliora anche le prestazioni analitiche complessive e aiuta le aziende a ottenere informazioni aziendali che favoriscono la crescita aziendale.

Panoramica NetApp

- Le offerte di NetApp, come ONTAP e StorageGRID, consentono la separazione tra storage ed elaborazione, consentendo un utilizzo ottimale delle risorse in base a requisiti specifici. Questa flessibilità consente ai clienti di scalare in modo indipendente il proprio storage utilizzando le soluzioni di storage NetApp
- Sfruttando i controller di storage di NetApp, i clienti possono fornire in modo efficiente i dati al proprio database vettoriale utilizzando i protocolli NFS e S3. Questi protocolli facilitano l'archiviazione dei dati dei clienti e gestiscono l'indice del database vettoriale, eliminando la necessità di più copie dei dati a cui si accede tramite metodi di file e oggetti.
- NetApp ONTAP fornisce supporto nativo per NAS e storage di oggetti attraverso i principali provider di servizi cloud come AWS, Azure e Google Cloud. Questa ampia compatibilità garantisce un'integrazione perfetta, consentendo la mobilità dei dati dei clienti, l'accessibilità globale, il ripristino di emergenza, la scalabilità dinamica e le prestazioni elevate.

StorageGRID

Il nostro storageGRID, leader del settore per l'archiviazione di oggetti, offre un potente motore di policy per il posizionamento automatico dei dati, opzioni di distribuzione flessibili e una durata senza pari con codifica di cancellazione a livelli. Ha un'architettura scalabile che supporta miliardi di oggetti e petabyte di dati in un unico namespace. La soluzione consente l'integrazione del cloud ibrido, consentendo il tiering dei dati sulle principali piattaforme cloud. È stato riconosciuto come leader nell'IDC Marketscape Worldwide Object-Based Vendor Assessment del 2019.

Inoltre, storageGRID eccelle nella gestione di dati non strutturati su larga scala con archiviazione di oggetti definiti dal software, ridondanza geografica e funzionalità multi-sito. Incorpora la gestione del ciclo di vita delle informazioni basata su policy e offre funzionalità di integrazione cloud come mirroring e ricerca. Possiede diverse certificazioni, tra cui Common Criteria, NF203 Digital Safe Component, ISO/IEC 25051, KPMG e Cohasset Compliance Assessment.

In sintesi, NetApp storageGRID offre potenti funzionalità, scalabilità, integrazione con cloud ibrido e certificazioni di conformità per una gestione efficiente dei dati non strutturati su larga scala.

NetApp ONTAP

NetApp ONTAP è una soluzione di storage affidabile che offre un'ampia gamma di funzionalità aziendali. Include Snapshot, che fornisce backup istantanei coerenti con l'applicazione e a prova di manomissione. SnapRestore consente il ripristino quasi istantaneo dei backup su richiesta, mentre SnapMirror offre funzionalità integrate di backup remoto e disaster recovery. La soluzione integra anche Autonomous Ransomware Protection (ARP), garantendo la sicurezza dei dati con funzionalità quali la verifica multi-amministratore, la crittografia dei dati inattivi con certificazione FIPS, la crittografia dei dati in transito, l'autenticazione a più fattori (MFA) e il controllo degli accessi basato sui ruoli (RBAC). La registrazione completa, l'audit, la gestione delle chiavi interne ed esterne, l'eliminazione sicura e la gestione sicura di più tenant migliorano ulteriormente la sicurezza e la conformità dei dati.

NetApp ONTAP è inoltre dotato SnapLock, che garantisce la conservazione dei dati conforme alle normative con elevati livelli di integrità, prestazioni e conservazione a un basso costo totale di proprietà. È completamente integrato con NetApp ONTAP 9 e offre protezione contro azioni dannose, amministratori non autorizzati e ransomware.

La soluzione comprende la crittografia NSE/NVE per la crittografia in transito e dei dati a riposo, l'accesso amministrativo multifattoriale e la verifica multi-amministratore. Active IQ fornisce analisi predittive basate sull'intelligenza artificiale e azioni correttive, mentre QoS garantisce il controllo del carico di lavoro della qualità del servizio. L'integrazione di gestione e automazione è intuitiva tramite SysMgr/GUI/CLI/API. FabricPool consente la suddivisione automatica dei dati e la soluzione offre efficienza attraverso la compressione, la deduplicazione e la compattazione dei dati in linea. NetApp garantisce il raggiungimento degli obiettivi di efficienza del carico di lavoro senza alcun costo per il cliente.

NetApp ONTAP supporta vari protocolli, tra cui NVMe/FC, FC, NVMe/TCP, iSCSI, NFS, SMB e S3, il che lo rende una soluzione di storage unificata. Nel complesso, NetApp ONTAP offre ampie funzionalità aziendali, sicurezza avanzata, conformità, efficienza e versatilità per soddisfare diverse esigenze di storage.

Panoramica di Dremio

Dremio è la piattaforma unificata Lakehouse per l'analisi self-service e l'intelligenza artificiale. La piattaforma di analisi unificata Dremio avvicina gli utenti ai dati con flessibilità, scalabilità e prestazioni lakehouse a una frazione del costo delle soluzioni di data warehouse legacy. Dremio consente l'analisi "shift-left" per eliminare l'integrazione dei dati e l'ETL complessi e costosi, offrendo analisi fluide su scala aziendale senza spostamento dei dati. Dremio offre anche:

- Analisi self-service di facile utilizzo, abilitate tramite un livello semantico universale e un motore di query SQL altamente performante e strettamente integrato, che semplifica la connessione, la gestione e l'analisi di tutti i dati, sia nel cloud che in locale.
- Le funzionalità di gestione del lakehouse nativo di Apache Iceberg di Dremio semplificano la scoperta dei dati e automatizzano l'ottimizzazione degli stessi, offrendo analisi ad alte prestazioni con il controllo delle versioni dei dati ispirato a Git.
- Basato su standard e codice sorgente aperti, Dremio consente alle aziende di evitare il lock-in e di rimanere orientate all'innovazione. Le grandi aziende si affidano a Dremio perché è la piattaforma

lakehouse più semplice da usare e con il miglior rapporto qualità-prezzo per tutti i carichi di lavoro.

Quale valore offre ai clienti la soluzione Dremio e NetApp Hybrid Iceberg Lakehouse?

- **Miglioramento della gestione e dell'accessibilità dei dati:** Dremio è noto per la sua piattaforma data lakehouse che consente alle organizzazioni di interrogare i dati direttamente dai loro data lake ad alta velocità. NetApp, d'altro canto, è un fornitore leader di servizi di dati cloud e soluzioni di archiviazione dati. L'offerta congiunta fornisce ai clienti una soluzione completa per archiviare, gestire, accedere e analizzare i dati della loro azienda in modo efficiente ed efficace.
- **Ottimizzazione delle prestazioni:** grazie all'esperienza di NetApp nell'archiviazione dei dati e alle capacità di Dremio nell'elaborazione e nell'ottimizzazione dei dati, la partnership offre una soluzione che migliora le prestazioni delle operazioni sui dati, riduce la latenza e aumenta la velocità di acquisizione di informazioni aziendali. Dremio ha addirittura migliorato le prestazioni dell'infrastruttura analitica IT interna di NetApp.
- **Scalabilità:** sia Dremio che NetApp offrono una soluzione progettata per essere scalabile. La soluzione congiunta fornisce ai clienti ambienti di archiviazione, gestione e analisi dei dati altamente scalabili. In un ambiente Hybrid Iceberg Lakehouse, il motore di query SQL di Dremio abbinato a NetApp StorageGRID offre scalabilità, concorrenza e prestazioni di query senza pari, in grado di gestire le esigenze analitiche di qualsiasi azienda.
- **Sicurezza e governance dei dati:** entrambe le aziende sono fortemente attente alla sicurezza e alla governance dei dati. Insieme, offrono solide funzionalità di sicurezza e governance dei dati, garantendo la protezione dei dati e il rispetto dei requisiti di governance dei dati. Funzionalità quali controlli di accesso basati sui ruoli e dettagliati, auditing completo, derivazione dei dati end-to-end, gestione unificata delle identità e SSO con un ampio framework di conformità e sicurezza garantiscono che gli ambienti di dati analitici delle aziende siano sicuri e governati.
- **Efficienza dei costi:** integrando il motore data lake di Dremio con le soluzioni di storage di NetApp, i clienti possono ridurre i costi associati alla gestione e allo spostamento dei dati. Le organizzazioni possono anche passare da ambienti data lake legacy a una soluzione lakehouse più moderna composta da NetApp e Dremio. Questa soluzione Hybrid Iceberg Lakehouse offre prestazioni di query ad alta velocità e una concorrenza di query leader di mercato, che riduce il TCO e riduce il tempo necessario per ottenere informazioni aziendali.

Requisiti tecnologici

Per le convalide eseguite in questo documento sono state utilizzate le configurazioni hardware e software descritte di seguito. Queste configurazioni servono come linee guida per aiutarti a configurare il tuo ambiente. Tuttavia, tieni presente che i componenti specifici possono variare a seconda delle esigenze individuali del cliente.

Requisiti hardware

Hardware	Dettagli
Coppia HA di array di storage NetApp AFF	• A800 • ONTAP 9.14.1 • 48 SSD-NVM da 3,49 TB • Due bucket S3: metadati Dremio e dati dei clienti.

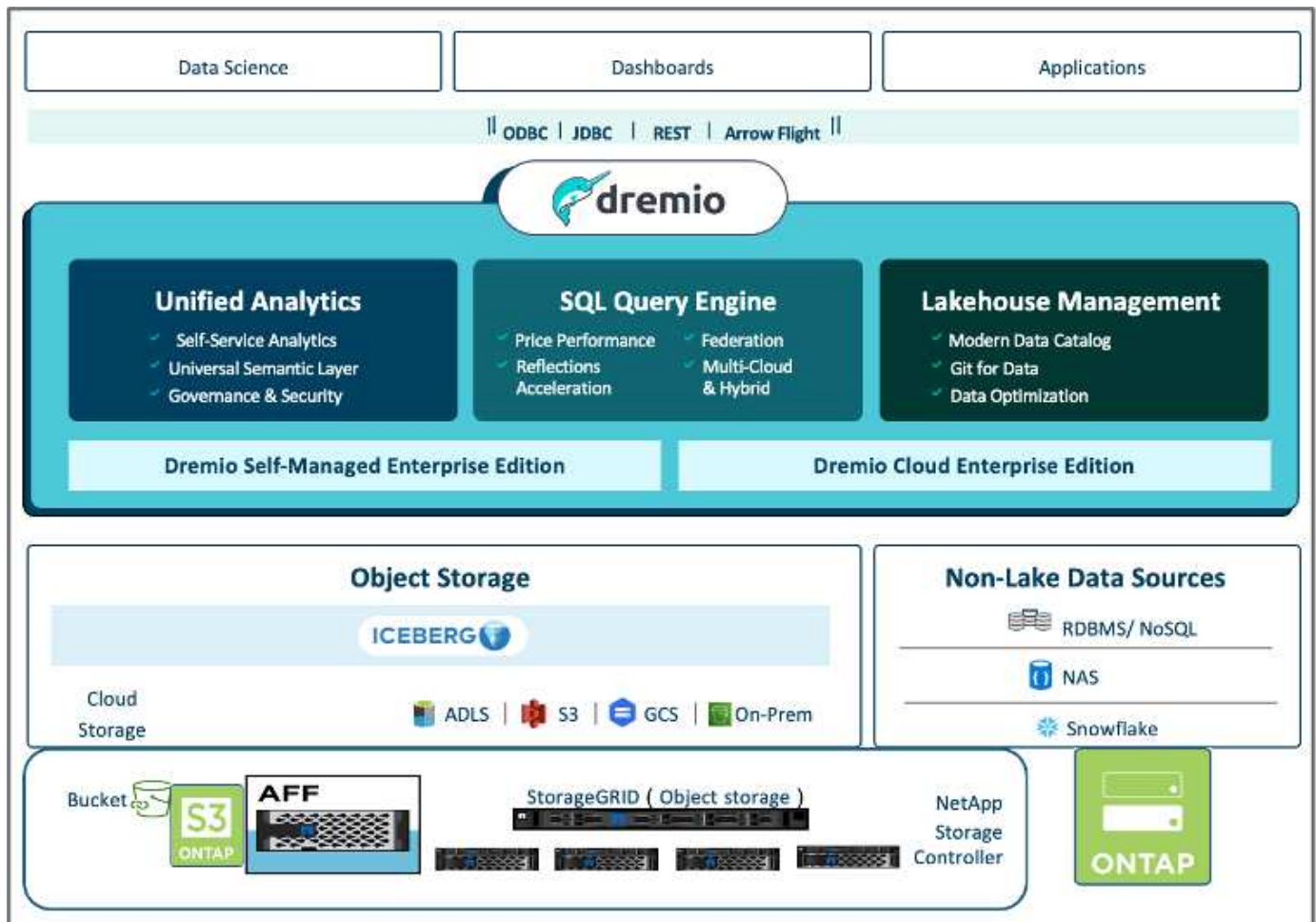
Hardware	Dettagli
4 x FUJITSU PRIMERGY RX2540 M4	• 64 CPU • CPU Intel Xeon Gold 6142 a 2,60 GHz • 256 GM di memoria fisica • 1 porta di rete da 100 GbE
Networking	• 100 GbE
StorageGRID	* 1 x SG100, 3xSGF6024 * 3 x 24 x 7,68 TB * Due bucket S3: metadati Dremio e dati dei clienti.

Requisiti software

Software	Dettagli
Dremio	• versione - 25.0.3-202405170357270647-d2042e1b • Edizione Enterprise
In sede	• Cluster Dremio a 5 nodi • 1 coordinatore principale e 4 esecutori

Procedura di distribuzione

In questa convalida dell'architettura di riferimento, abbiamo utilizzato una configurazione Dremio composta da un coordinatore e quattro esecutori



Configurazione NetApp

- Inizializzazione del sistema di archiviazione
- Creazione di una macchina virtuale di archiviazione (SVM)
- Assegnazione delle interfacce di rete logiche
- Configurazione e licenza NFS, S3

Per NFS (Network File System) procedere come segue: 1. Creare un volume Flex Group per NFSv4 o NFSv3. Nella nostra configurazione per questa convalida, abbiamo utilizzato 48 SSD, 1 SSD dedicato al volume root del controller e 47 SSD distribuiti per NFSv4]]. Verificare che il criterio di esportazione NFS per il volume Flex Group disponga di autorizzazioni di lettura/scrittura per la rete dei server Dremio.

1. Su tutti i server Dremio, creare una cartella e montare il volume Flex Group su questa cartella tramite un'interfaccia logica (LIF) su ciascun server Dremio.

Per S3 (Simple Storage Service) seguire i passaggi indicati di seguito:

1. Impostare un object-store-server con HTTP abilitato e lo stato di amministrazione impostato su "attivo" utilizzando il comando "vserver object-store-server create". Hai la possibilità di abilitare HTTPS e impostare una porta di ascolto personalizzata.
2. Creare un utente object-store-server utilizzando il comando "vserver object-store-server user create -user <username>".
3. Per ottenere la chiave di accesso e la chiave segreta, è possibile eseguire il seguente comando: "set diag;

vserver object-store-server user show -user <nomeutente>". Tuttavia, in futuro, queste chiavi verranno fornite durante il processo di creazione dell'utente oppure potranno essere recuperate tramite chiamate API REST.

4. Creare un gruppo object-store-server utilizzando l'utente creato nel passaggio 2 e concedere l'accesso. In questo esempio abbiamo fornito "FullAccess".
5. Crea due bucket S3 impostandone il tipo su "S3". Uno per la configurazione di Dremio e uno per i dati del cliente.

Configurazione del guardiano dello zoo

Puoi utilizzare la configurazione zookeeper fornita da Dremio. In questa convalida, abbiamo utilizzato uno zookeeper separato. Abbiamo seguito i passaggi menzionati in questo collegamento web <https://medium.com/@ahmetfurkandemir/distributed-hadoop-cluster-1-spark-with-all-dependencies-03c8ec616166>

Configurazione di Dremio

Abbiamo seguito questo collegamento web per installare Dremio tramite tarball.

1. Crea un gruppo Dremio.

```
sudo groupadd -r dremio
```

2. Crea un utente dremio.

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

3. Crea directory 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. Scarica il file tar da <https://download.dremio.com/community-server/>
5. Scompattare Dremio nella directory /opt/dremio.

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

6. Creare un collegamento simbolico per la cartella di configurazione.

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


7. Imposta la configurazione del servizio (configurazione SystemD).

- a. Copiare il file dell'unità per il demone dremio da /opt/dremio/share/dremio.service a /etc/systemd/system/dremio.service.
- b. Riavviare il sistema

```
sudo systemctl daemon-reload
```

- c. Abilita l'avvio di dremio all'avvio.

```
sudo systemctl enable dremio
```

8. Configurare Dremio sul coordinatore. Per maggiori informazioni, vedere Configurazione di Dremio

- 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. -->

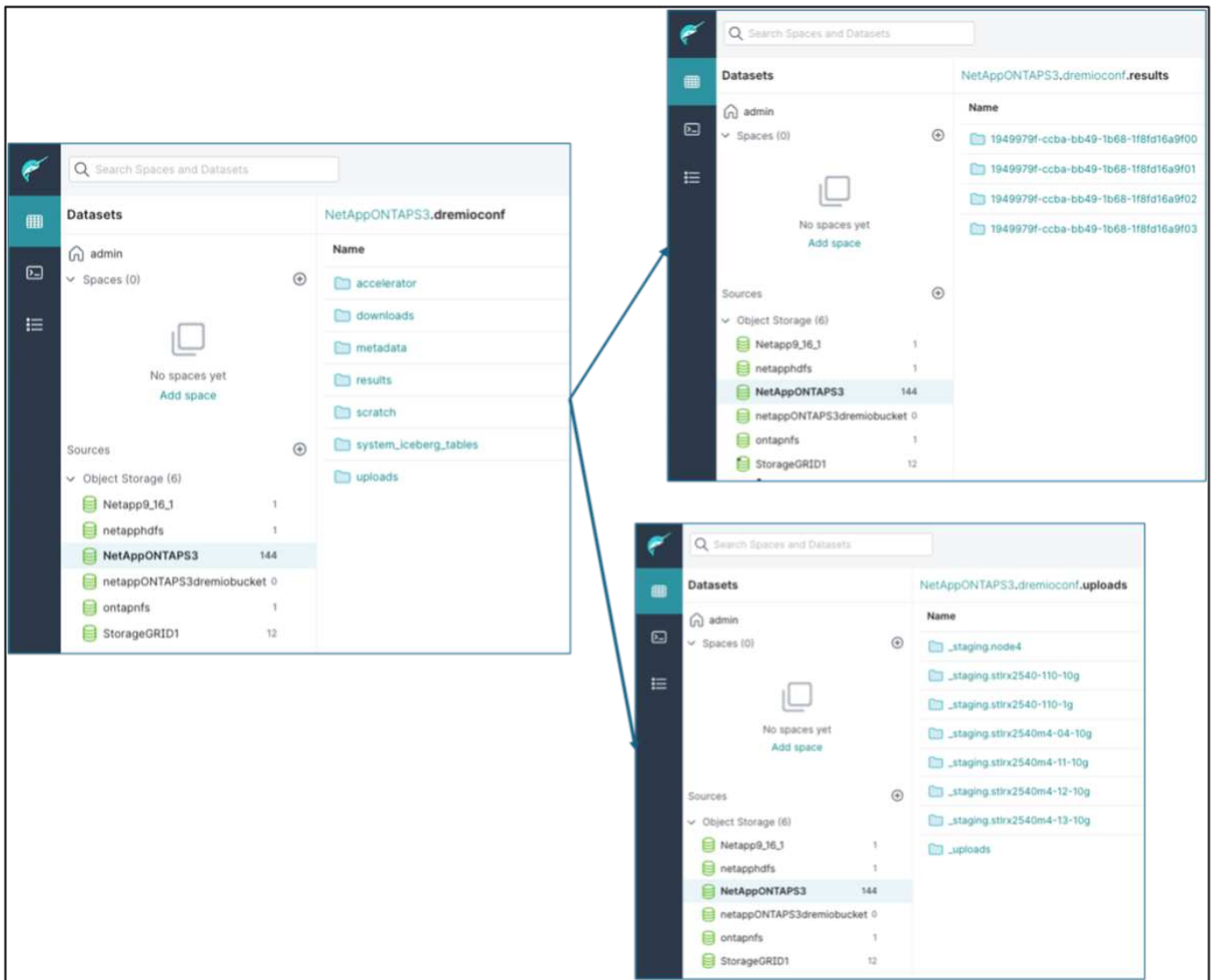
```
<configuration>
  <property>
    <name>fs.dremioS3.impl</name>
    <value>com.dremio.plugins.s3.store.S3FileSystem</value>
  </property>
  <property>
    <name>fs.s3a.access.key</name>
    <value>24G4C1316APP2BIPDE5S</value>
  </property>
  <property>
    <name>fs.s3a.endpoint</name>
    <value>10.63.150.69:80</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 configurazione di Dremio è archiviata nell'archivio oggetti NetApp . Nella nostra convalida, il bucket "dremioconf" risiede in un bucket S3 ontap. L'immagine sottostante mostra alcuni dettagli delle cartelle "scratch" e "uploads" del bucket S3 "dremioconf".



1. Configurare Dremio sugli esecutori. Nella nostra configurazione abbiamo 3 esecutori.
 - 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: uguale alla configurazione del coordinatore.



NetApp consiglia StorageGRID come soluzione di archiviazione di oggetti primaria per gli ambienti Datalake e Lakehouse. Inoltre, NetApp ONTAP viene utilizzato per la dualità file/oggetto. Nel contesto di questo documento, abbiamo condotto dei test su ONTAP S3 in risposta alla richiesta di un cliente, e il sistema funziona correttamente come fonte di dati.

Configurazione di più sorgenti

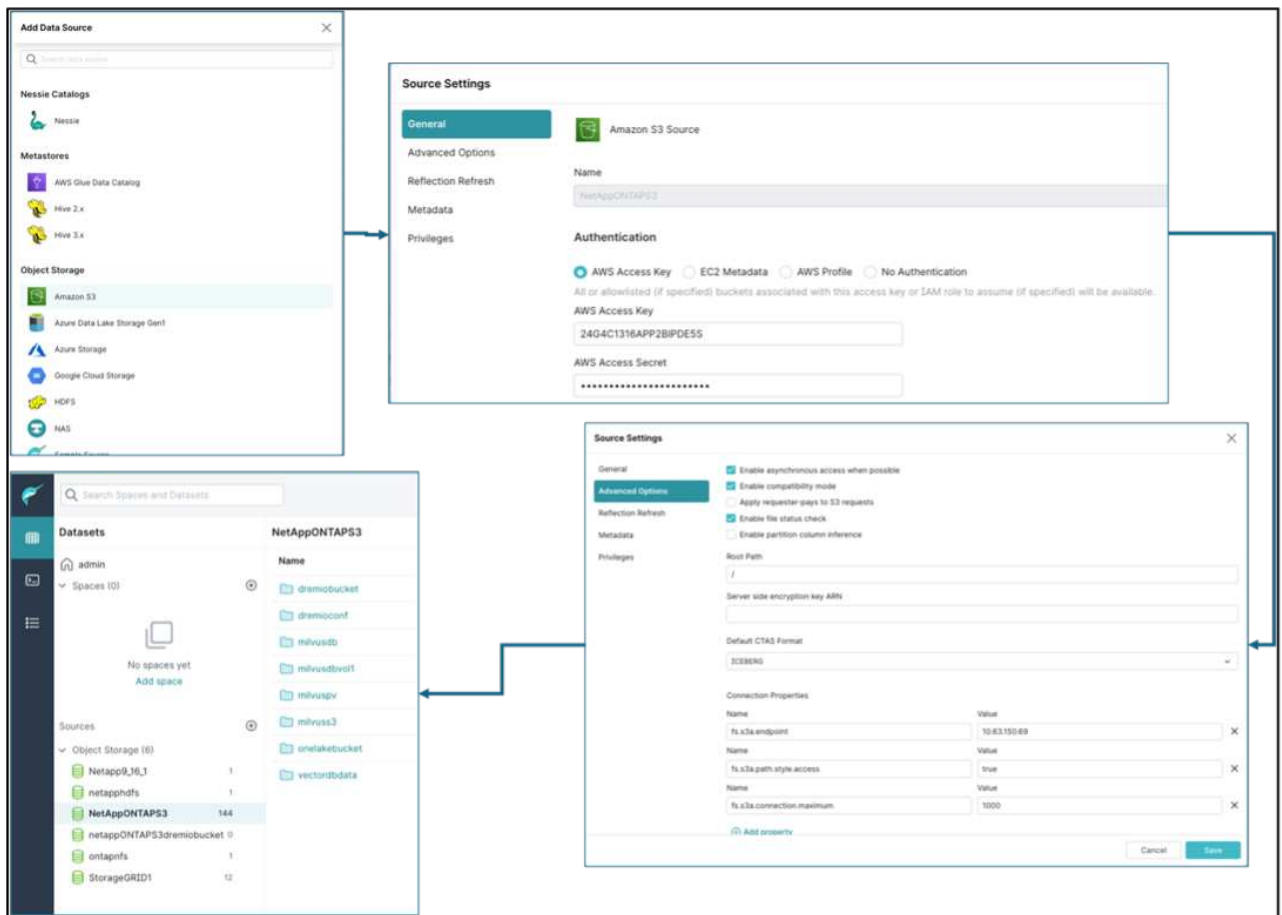
1. Configurare ONTAP S3 e storageGRID come sorgente S3 in Dremio.
 - a. Dashboard di Dremio → set di dati → fonti → aggiungi fonte.
 - b. Nella sezione generale, aggiorna l'accesso AWS e la chiave segreta
 - c. Nell'opzione avanzata, abilita la modalità compatibilità e aggiorna le proprietà di connessione con i dettagli seguenti. L'IP/nome dell'endpoint dal controller di archiviazione NetApp da ontap S3 o storageGRID.

```

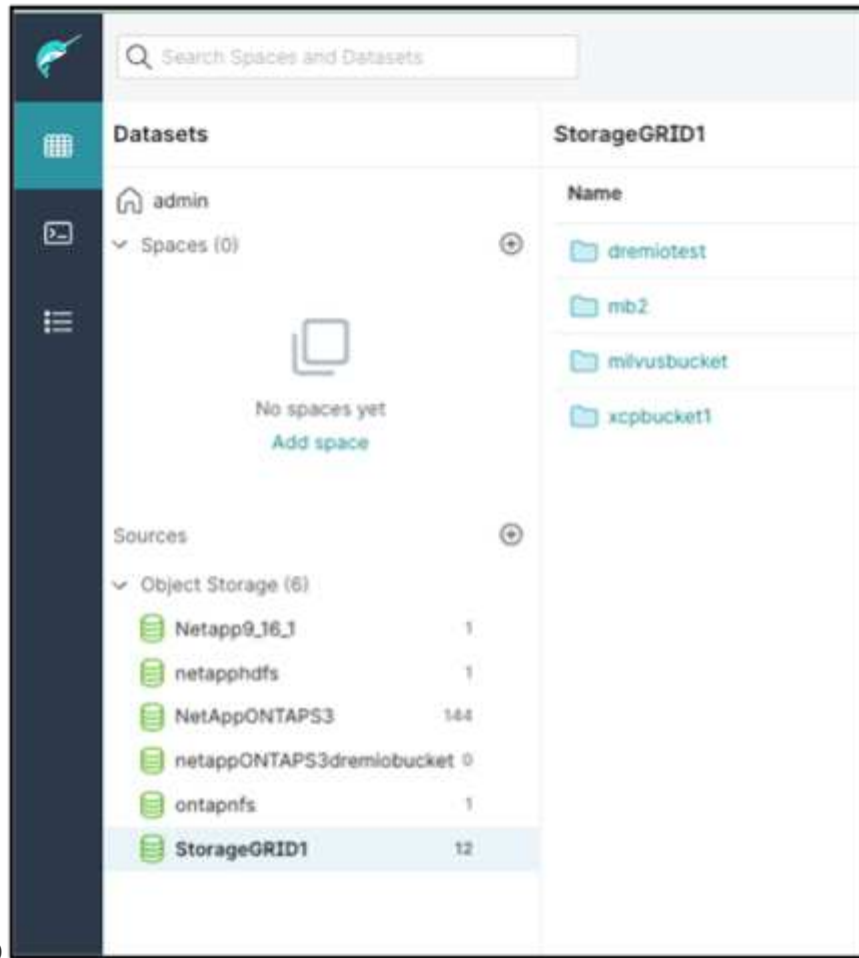
fs.s3a.endpoint = 10.63.150.69
fs.s3a.path.style.access = true
fs.s3a.connection.maximum=1000

```

- d. Abilita la memorizzazione nella cache locale quando possibile, percentuale massima della cache totale disponibile da utilizzare quando possibile = 100
- e. Quindi visualizza l'elenco dei bucket dall'archiviazione di oggetti NetApp

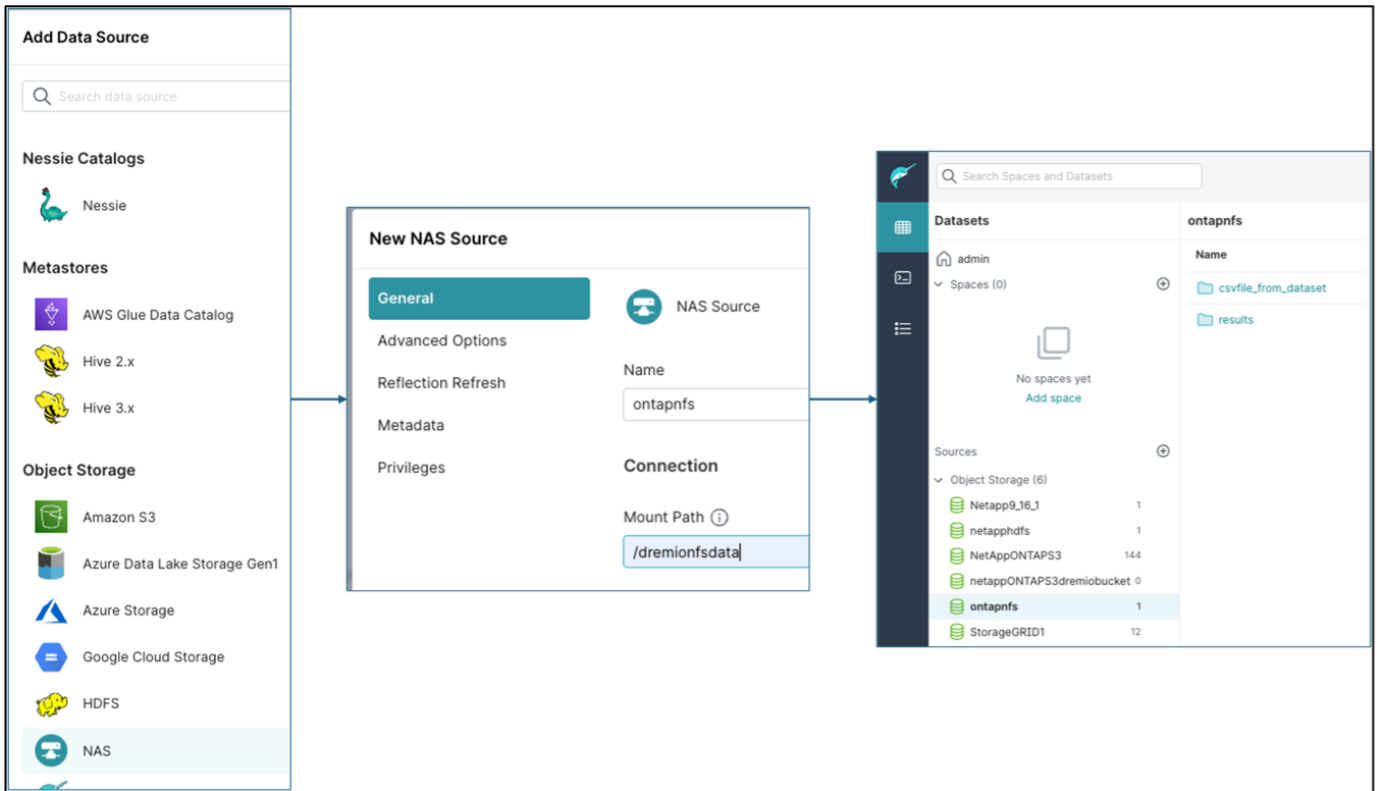


f. Visualizzazione di esempio dei dettagli del bucket



storageGRID

2. Configurare NAS (in particolare NFS) come sorgente in Dremio.
 - a. Dashboard di Dremio → set di dati → fonti → aggiungi fonte.
 - b. Nella sezione generale, immettere il nome e il percorso di montaggio NFS. Assicurarsi che il percorso di montaggio NFS sia montato sulla stessa cartella su tutti i nodi del cluster Dremio.



+


```

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:~#

```

Panoramica della verifica della soluzione

In questa sezione abbiamo eseguito query di test SQL da più fonti per verificare la funzionalità, testare e verificare lo spillover sullo storage NetApp .

Query SQL su Object Storage

1. Imposta la memoria a 250 GB per server in 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. Controllare la posizione di spillover (`${DREMIO_HOME}/dremiocache`) nel file `dremio.conf` e i dettagli di archiviazione.

```

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

```

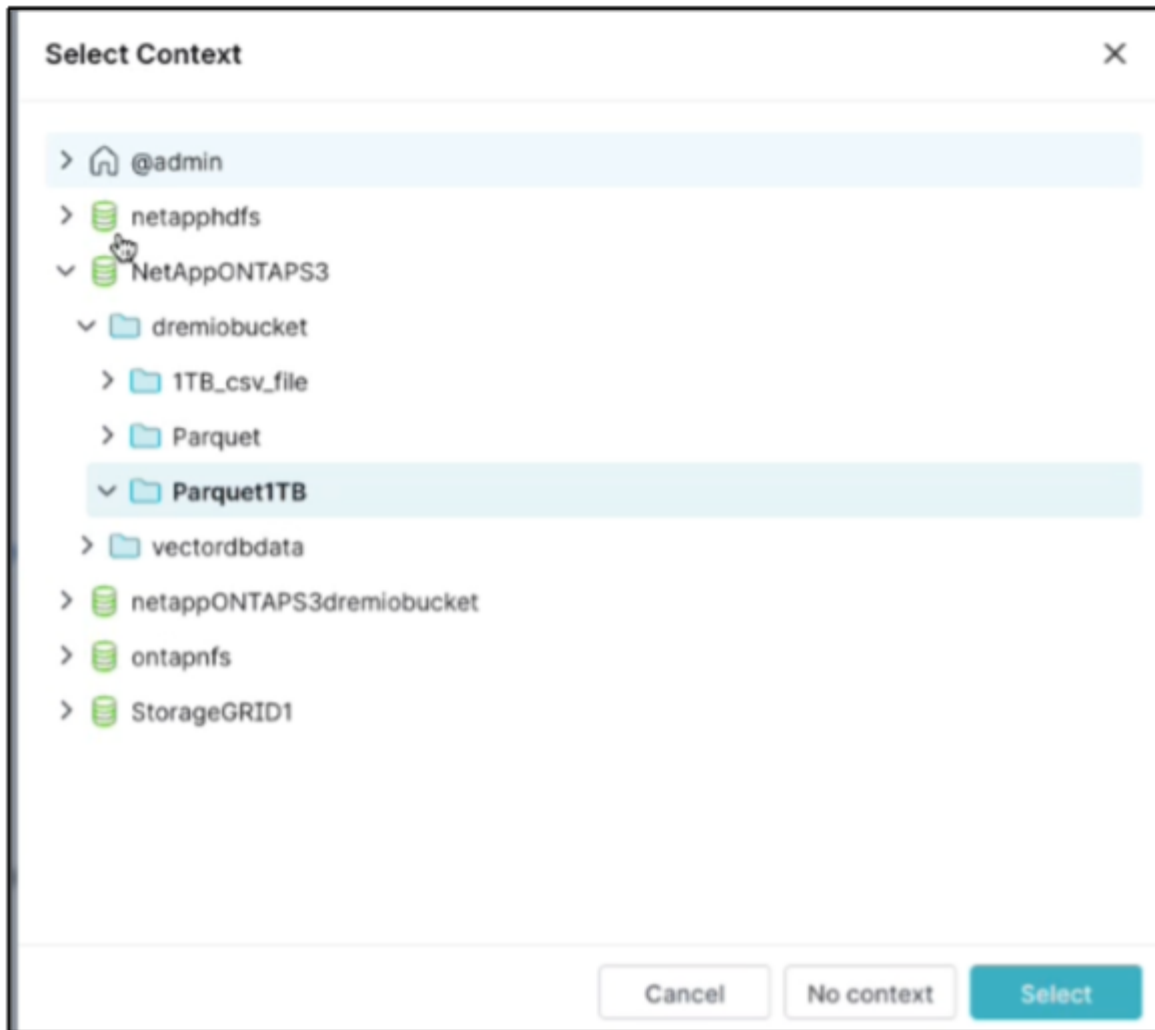
3. Indicare la posizione di spillover di Dremio allo storage NFS NetApp

```

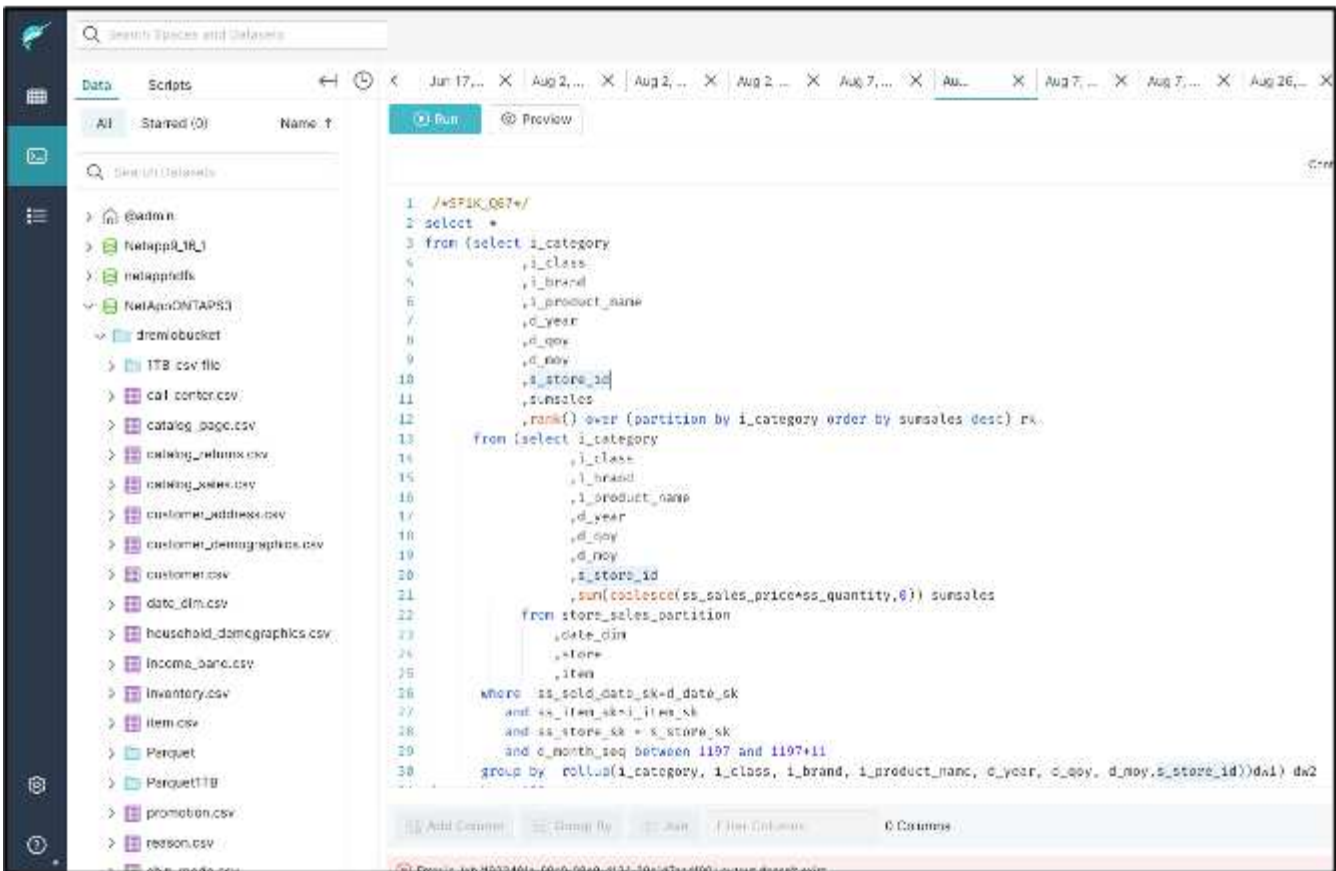
root@hadoopnode1:~# ls -ltrh /dremiocache
total 4.0K
drwx----- 3 nobody nogroup 4.0K Sep 13 16:00 spilling_stlrx2540m4-12-
10g_45678
root@hadoopnode1:~# ls -ltrh /opt/dremio/dremiocache/
total 8.0K
drwxr-xr-x 3 dremio dremio 4.0K Aug 22 18:19 spill_old
drwxr-xr-x 4 dremio dremio 4.0K Aug 22 18:19 cm
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. Seleziona il contesto. Nel nostro test, abbiamo eseguito il test sui file parquet generati da TPCDS residenti in ONTAP S3. Dashboard Dremio → SQL runner → contesto → NetAppONTAPS3→Parquet1TB



1. Eseguire la query TPC-DS67 dalla dashboard di Dremio



1. Verificare che il lavoro sia in esecuzione su tutti gli esecutori. Dashboard di Dremio → lavori → <jobid> → profilo raw → seleziona EXTERNAL_SORT → Nome host

Raw Profile

04-xx-04 - FILTER

04-xx-05 - WINDOW

04-xx-06 - EXTERNAL_SORT

Thread	Setup Time	Process Time	Wait Time	Max Batches	Max Records	Peak Memory	Hostname	Record Processing Rate	Operator State	Last Schedule Time
04-00-06	0.000s	0.000s	0.000s	0	0	128KB	stnx2540-110-10g	0	CAN_CONSUME	16:35:54
04-01-06	0.000s	0.000s	0.000s	0	0	128KB	stnx2540m4-04-10g	0	CAN_CONSUME	16:35:54
04-02-06	0.000s	0.000s	0.000s	0	0	128KB	stnx2540m4-12-10g	0	CAN_CONSUME	16:35:54
04-03-06	0.017s	0.000s	0.000s	0	0	128KB	stnx2540m4-13-10g	0	CAN_CONSUME	16:35:54
04-04-06	0.000s	0.000s	0.000s	0	0	128KB	stnx2540-110-10g	0	CAN_CONSUME	16:35:54
04-05-06	0.000s	0.000s	0.000s	0	0	128KB	stnx2540m4-04-10g	0	CAN_CONSUME	16:35:54
04-06-06	0.027s	0.000s	0.000s	0	0	128KB	stnx2540m4-12-10g	0	CAN_CONSUME	16:35:54
04-07-06	0.000s	0.000s	0.000s	0	0	128KB	stnx2540m4-13-10g	0	CAN_CONSUME	16:35:54

1. Quando si esegue la query SQL, è possibile controllare la cartella divisa per la memorizzazione dei dati nella cache nel controller di archiviazione 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 query SQL è stata completata con spill over

Job ID	User	Dataset	Query Type	Queue	Start Time	Duration	SQL
19383301-5cd9-0a48-1e38-e2f5b414d000	admin	store_sales_partition	JDBC Client	High Cost User Q...	08/22/2024, 19:42:54	00:08:25 R	/SF1K_Q67/ select * from (select i_category, i_class, i_brand, i_product_name, d_year, d_qoy, d_m...
19384af3-285d-e07c-5277-48d88168d200	admin	store_sales_partition	JDBC Client	High Cost User Q...	08/22/2024, 18:00:44	00:08:23 R	/SF1K_Q67/ select * from (select i_category, i_class, i_brand, i_product_name, d_year, d_qoy, d_m...
1938650f-0f9d-e205-6ea3-673aaa3c7d00	admin	store_sales_partition	JDBC Client	High Cost User Q...	08/22/2024, 16:09:20	00:08:26 R	/SF1K_Q67/ select * from (select i_category, i_class, i_brand, i_product_name, d_year, d_qoy, d_m...
19387983-2031-1d4f-cd9e-57c6c287bd00	admin	store_sales_partition	UI (run)	High Cost User Q...	08/22/2024, 14:42:04	00:07:26 R	/SF1K_Q67/ select * from (select i_category, i_class, i_brand, i_product_name, d_year, d_qoy, d_m...
19387d04-3dc3-34bd-13e5-d7538fa4a000	admin	store_sales_partition	UI (run)	High Cost User Q...	08/22/2024, 14:22:51	00:07:48 R	/SF1K_Q67/ select * from (select i_category, i_class, i_brand, i_product_name, d_year, d_qoy, d_m...

3. Riepilogo del completamento del

Jobs

19335115-a0a5-9dab-2b16-e2ec24459900

Overview

SQL

Summary

Status:

COMPLETED

Total Memory:

267.16 GB

CPU Used:

02h:18m:52s

Query Type:

UI (run)

Start Time:

08/26/2024 12:35:53

Duration:

08m:25s

Wait on Client:

<1s

User:

admin

Queue:

High Cost User Queries

Input:

21.32 GB / 563.2M Rows

Output:

6.92 KB / 100 Rows

Total Execution Time

08m:25s (100%)

Pending

2ms (0.00%)

Metadata Retrieval

22ms (0.00%)

Planning

140ms (0.03%)

Queued

30ms (0.01%)

Execution Planning

116ms (0.02%)

Starting

569ms (0.11%)

Running

8m:24s (99.83%)

Submitted SQL

1

/*SF1K_Q67*/

2

select *

3

from (select i_category

4

,i_class

5

,i_brand

6

,i_product_name

7

,d_year

8

,d_qoy

9

,d_moy

Queried Datasets

store_sales_partition

NetAppONTAPS3.dremiobucket.Parquet1TB

date_dim

NetAppONTAPS3.dremiobucket.Parquet1TB

store

NetAppONTAPS3.dremiobucket.Parquet1TB

Show more

Scans

store_sales_partition

date_dim

store

item

lavoro.

4. Controllare la dimensione dei dati

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

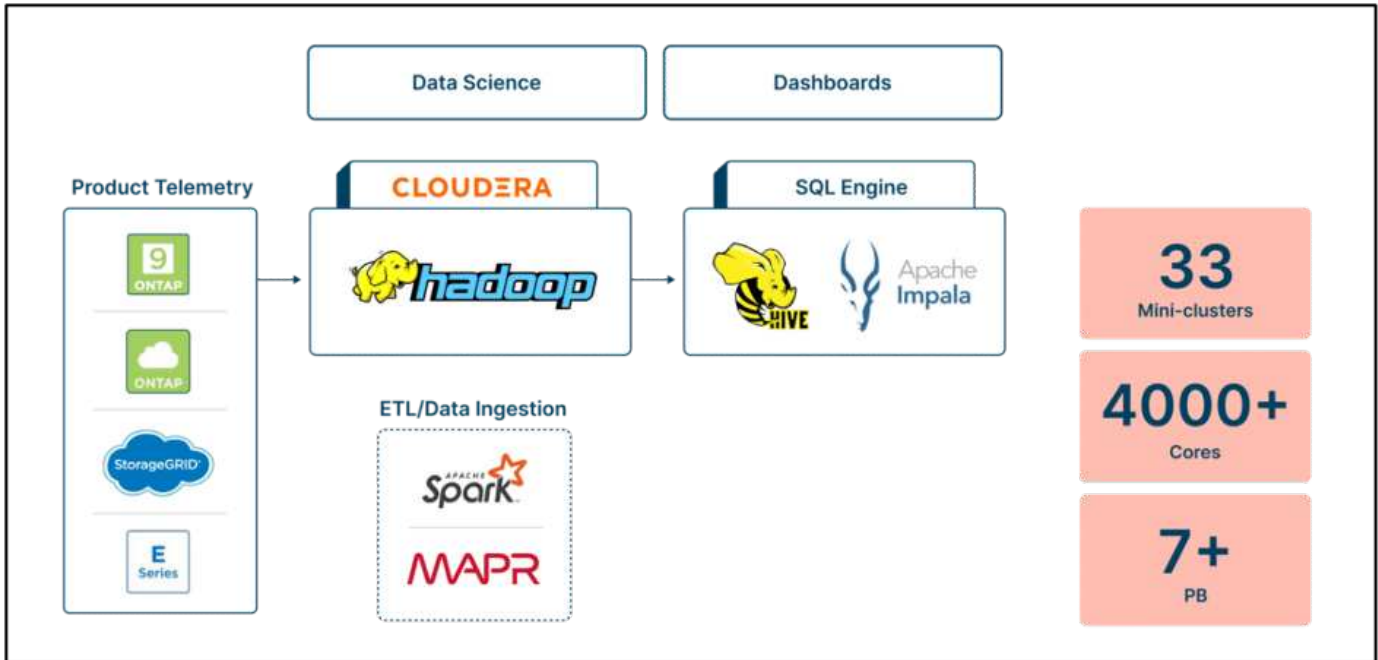
Operator Statistics

Merge Time Nanos:	0ns
Spill Count:	360
Spill Time Nanos:	37.68m
Total Spilled Data Size:	20,339,702,765
Batches Spilled:	97,854

La stessa procedura è applicabile per l'archiviazione di oggetti NAS e StorageGRID .

Casi d'uso dei clienti

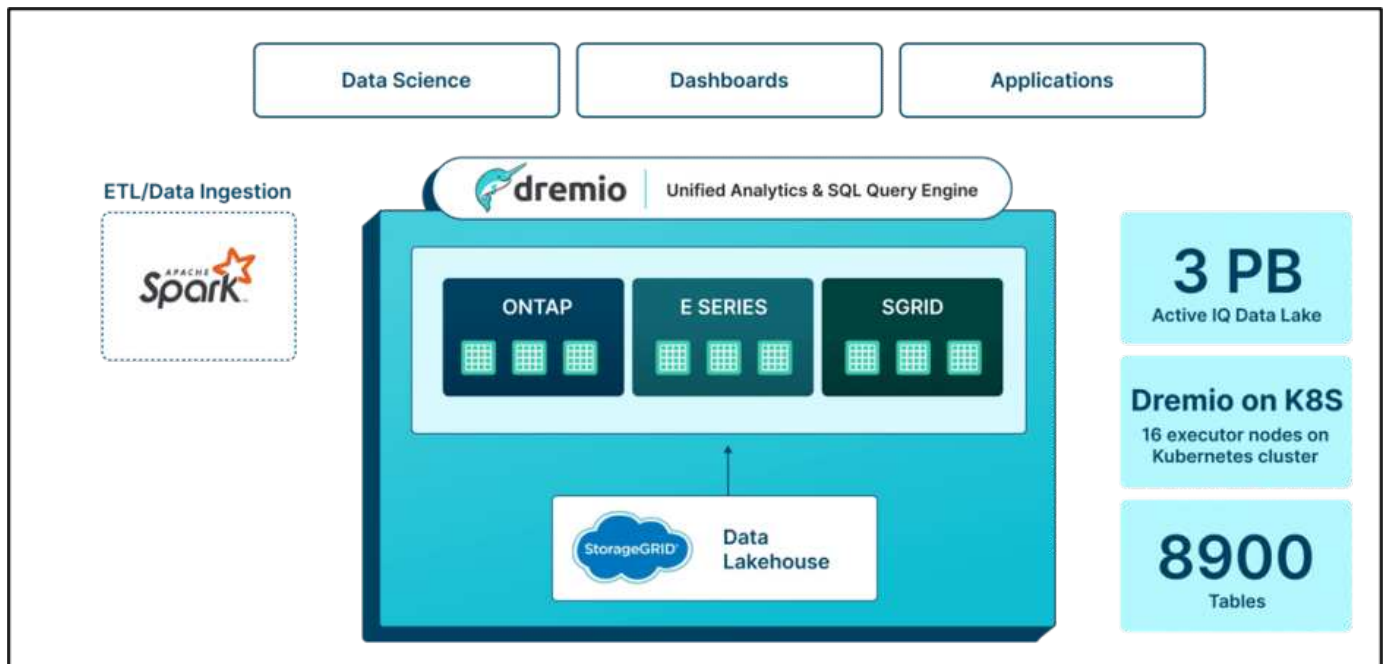
Caso d'uso di NetApp ActiveIQ



Sfida: la soluzione Active IQ interna di NetApp, inizialmente progettata per supportare numerosi casi d'uso, si è evoluta in un'offerta completa sia per gli utenti interni che per i clienti. Tuttavia, l'infrastruttura backend sottostante basata su Hadoop/MapR ha posto sfide in termini di costi e prestazioni, a causa della rapida crescita dei dati e della necessità di un accesso efficiente ai dati. L'aumento dello storage ha comportato l'aggiunta di risorse di elaborazione non necessarie, con conseguente aumento dei costi.

Inoltre, la gestione del cluster Hadoop richiedeva molto tempo e competenze specialistiche. Problemi di gestione e prestazioni dei dati hanno complicato ulteriormente la situazione, con query che richiedevano in media 45 minuti e carenza di risorse a causa di configurazioni errate. Per affrontare queste sfide, NetApp ha cercato un'alternativa all'ambiente Hadoop legacy esistente e ha deciso che una nuova soluzione moderna basata su Dremio avrebbe ridotto i costi, separato storage ed elaborazione, migliorato le prestazioni, semplificato la gestione dei dati, offerto controlli dettagliati e fornito funzionalità di disaster recovery.

Soluzione:



Dremio ha consentito a NetApp di modernizzare la propria infrastruttura dati basata su Hadoop con un approccio graduale, fornendo una roadmap per l'analisi unificata. A differenza di altri fornitori che hanno richiesto modifiche significative all'elaborazione dei dati, Dremio si è integrato perfettamente con le pipeline esistenti, risparmiando tempo e denaro durante la migrazione. Passando a un ambiente completamente containerizzato, NetApp ha ridotto i costi di gestione, migliorato la sicurezza e aumentato la resilienza. L'adozione da parte di Dremio di ecosistemi aperti come Apache Iceberg e Arrow ha garantito sicurezza, trasparenza ed estensibilità per il futuro.

In sostituzione dell'infrastruttura Hadoop/Hive, Dremio offriva funzionalità per casi d'uso secondari attraverso il livello semantico. Mentre i meccanismi ETL e di acquisizione dati basati su Spark esistenti sono rimasti invariati, Dremio ha fornito un livello di accesso unificato per una più semplice esplorazione e scoperta dei dati senza duplicazioni. Questo approccio ha ridotto significativamente i fattori di replicazione dei dati e ha disaccoppiato l'archiviazione e l'elaborazione.

Vantaggi: Con Dremio, NetApp ha ottenuto notevoli riduzioni dei costi riducendo al minimo il consumo di elaborazione e i requisiti di spazio su disco nei propri ambienti dati. Il nuovo Active IQ Data Lake è composto da 8.900 tabelle contenenti 3 petabyte di dati, rispetto agli oltre 7 petabyte della precedente infrastruttura. La migrazione a Dremio ha comportato anche il passaggio da 33 mini-cluster e 4.000 core a 16 nodi esecutori su cluster Kubernetes. Nonostante le significative riduzioni delle risorse di elaborazione, NetApp ha registrato notevoli miglioramenti nelle prestazioni. Accedendo direttamente ai dati tramite Dremio, il tempo di esecuzione delle query è diminuito da 45 minuti a 2 minuti, con un conseguente risparmio del 95% sui tempi di acquisizione di informazioni per la manutenzione predittiva e l'ottimizzazione. La migrazione ha inoltre prodotto una riduzione di oltre il 60% nei costi di elaborazione, query più veloci di oltre 20 volte e un risparmio di oltre il 30% nel costo totale di proprietà (TCO).

Caso d'uso del cliente nella vendita di ricambi auto.

Sfide: all'interno di questa azienda globale di vendita di ricambi auto, i gruppi di pianificazione e analisi finanziaria aziendale e dirigenziale non sono riusciti a ottenere una visione consolidata dei report sulle vendite e sono stati costretti a leggere i report sulle metriche di vendita delle singole linee di business e a tentare di consolidarli. Ciò ha portato i clienti a prendere decisioni sulla base di dati risalenti ad almeno un giorno prima. I tempi di elaborazione per ottenere nuove informazioni analitiche normalmente superano le quattro settimane. La risoluzione dei problemi relativi alle pipeline di dati richiederebbe ancora più tempo, aggiungendo altri tre giorni o più alla già lunga tempistica. Il lento processo di sviluppo dei report e le relative prestazioni hanno costretto la comunità degli analisti ad attendere continuamente che i dati venissero elaborati o caricati, anziché

consentire loro di trovare nuove informazioni aziendali e di promuovere nuovi comportamenti aziendali. Questi ambienti problematici erano composti da numerosi database diversi per diverse linee di business, con il risultato di creare numerosi silos di dati. L'ambiente lento e frammentato complicava la governance dei dati, poiché gli analisti avevano troppi modi per elaborare la propria versione della verità anziché ricorrere a un'unica fonte di verità. L'approccio è costato oltre 1,9 milioni di dollari in costi di piattaforma dati e personale. Per mantenere la piattaforma legacy e soddisfare le richieste di dati erano necessari sette Field Technical Engineer (FTE) all'anno. Con l'aumento delle richieste di dati, il team di data intelligence non è riuscito a ridimensionare l'ambiente legacy per soddisfare le esigenze future

Soluzione: archiviare e gestire in modo conveniente grandi tabelle Iceberg in NetApp Object Store. Crea domini di dati utilizzando il livello semantico di Dremio, consentendo agli utenti aziendali di creare, cercare e condividere facilmente prodotti di dati.

Vantaggi per il cliente: • Architettura dei dati esistente migliorata e ottimizzata e riduzione del tempo per ottenere informazioni da quattro settimane a poche ore • Tempo di risoluzione dei problemi ridotto da tre giorni a poche ore • Costi di gestione e piattaforma dati ridotti di oltre \$ 380.000 • (2) FTE di sforzi di Data Intelligence risparmiati all'anno

Conclusione

In conclusione, questo rapporto tecnico ha fornito dettagli completi sulla distribuzione di q Hybrid Iceberg Lakehouse con Dremio in combinazione con varie fonti di dati dai controller di storage NetApp , tra cui ONTAP S3, NAS e StorageGRID. Il processo di distribuzione è stato eseguito con successo e lo strumento di benchmarking TPC-DS è stato utilizzato per eseguire 99 query SQL su diverse fonti di dati. Il rapporto ha inoltre esaminato i casi d'uso dei clienti all'interno di NetApp, dimostrando la versatilità e l'efficacia di Dremio nel soddisfare diverse esigenze aziendali. Inoltre, è stato esaminato un caso d'uso specifico che coinvolge un cliente di vendita di ricambi auto, evidenziando l'applicazione pratica e i vantaggi derivanti dall'utilizzo di Dremio per l'analisi dei dati e le informazioni dettagliate.

Nel complesso, questo documento rappresenta una risorsa preziosa per comprendere l'implementazione e l'utilizzo di Dremio con i controller di storage NetApp , evidenziandone le capacità e il potenziale per favorire l'ottimizzazione e il processo decisionale basato sui dati in vari settori.

Dove trovare ulteriori informazioni

Per saperne di più sulle informazioni descritte nel presente documento, consultare i seguenti documenti e/o siti web:

- Installazione di Zookeeper

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

- Dremio

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

- Configurazione di Dremio con storageGRID

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

- Caso d'uso NetApp

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

Informazioni sul copyright

Copyright © 2025 NetApp, Inc. Tutti i diritti riservati. Stampato negli Stati Uniti d'America. Nessuna porzione di questo documento soggetta a copyright può essere riprodotta in qualsiasi formato o mezzo (grafico, elettronico o meccanico, inclusi fotocopie, registrazione, nastri o storage in un sistema elettronico) senza previo consenso scritto da parte del detentore del copyright.

Il software derivato dal materiale sottoposto a copyright di NetApp è soggetto alla seguente licenza e dichiarazione di non responsabilità:

IL PRESENTE SOFTWARE VIENE FORNITO DA NETAPP "COSÌ COM'È" E SENZA QUALSIVOGLIA TIPO DI GARANZIA IMPLICITA O ESPRESSA FRA CUI, A TITOLO ESEMPLIFICATIVO E NON ESAUSTIVO, GARANZIE IMPLICITE DI COMMERCIALIZZABILITÀ E IDONEITÀ PER UNO SCOPO SPECIFICO, CHE VENGONO DECLINATE DAL PRESENTE DOCUMENTO. NETAPP NON VERRÀ CONSIDERATA RESPONSABILE IN ALCUN CASO PER QUALSIVOGLIA DANNO DIRETTO, INDIRETTO, ACCIDENTALE, SPECIALE, ESEMPLARE E CONSEGUENZIALE (COMPRESI, A TITOLO ESEMPLIFICATIVO E NON ESAUSTIVO, PROCUREMENT O SOSTITUZIONE DI MERCI O SERVIZI, IMPOSSIBILITÀ DI UTILIZZO O PERDITA DI DATI O PROFITTI OPPURE INTERRUZIONE DELL'ATTIVITÀ AZIENDALE) CAUSATO IN QUALSIVOGLIA MODO O IN RELAZIONE A QUALUNQUE TEORIA DI RESPONSABILITÀ, SIA ESSA CONTRATTUALE, RIGOROSA O DOVUTA A INSOLVENZA (COMPRESA LA NEGLIGENZA O ALTRO) INSORTA IN QUALSIASI MODO ATTRAVERSO L'UTILIZZO DEL PRESENTE SOFTWARE ANCHE IN PRESENZA DI UN PREAVVISO CIRCA L'EVENTUALITÀ DI QUESTO TIPO DI DANNI.

NetApp si riserva il diritto di modificare in qualsiasi momento qualunque prodotto descritto nel presente documento senza fornire alcun preavviso. NetApp non si assume alcuna responsabilità circa l'utilizzo dei prodotti o materiali descritti nel presente documento, con l'eccezione di quanto concordato espressamente e per iscritto da NetApp. L'utilizzo o l'acquisto del presente prodotto non comporta il rilascio di una licenza nell'ambito di un qualche diritto di brevetto, marchio commerciale o altro diritto di proprietà intellettuale di NetApp.

Il prodotto descritto in questa guida può essere protetto da uno o più brevetti degli Stati Uniti, esteri o in attesa di approvazione.

LEGENDA PER I DIRITTI SOTTOPOSTI A LIMITAZIONE: l'utilizzo, la duplicazione o la divulgazione da parte degli enti governativi sono soggetti alle limitazioni indicate nel sottoparagrafo (b)(3) della clausola Rights in Technical Data and Computer Software del DFARS 252.227-7013 (FEB 2014) e FAR 52.227-19 (DIC 2007).

I dati contenuti nel presente documento riguardano un articolo commerciale (secondo la definizione data in FAR 2.101) e sono di proprietà di NetApp, Inc. Tutti i dati tecnici e il software NetApp forniti secondo i termini del presente Contratto sono articoli aventi natura commerciale, sviluppati con finanziamenti esclusivamente privati. Il governo statunitense ha una licenza irrevocabile limitata, non esclusiva, non trasferibile, non cedibile, mondiale, per l'utilizzo dei Dati esclusivamente in connessione con e a supporto di un contratto governativo statunitense in base al quale i Dati sono distribuiti. Con la sola esclusione di quanto indicato nel presente documento, i Dati non possono essere utilizzati, divulgati, riprodotti, modificati, visualizzati o mostrati senza la previa approvazione scritta di NetApp, Inc. I diritti di licenza del governo degli Stati Uniti per il Dipartimento della Difesa sono limitati ai diritti identificati nella clausola DFARS 252.227-7015(b) (FEB 2014).

Informazioni sul marchio commerciale

NETAPP, il logo NETAPP e i marchi elencati alla pagina <http://www.netapp.com/TM> sono marchi di NetApp, Inc. Gli altri nomi di aziende e prodotti potrebbero essere marchi dei rispettivi proprietari.