Soluzioni dati di cloud ibrido NetApp: Spark e Hadoop in base ai casi di utilizzo dei clienti

NetApp Solutions

NetApp
April 26, 2024

This PDF was generated from https://docs.netapp.com/it-it/netapp-solutions/data-analytics/hdcs-sh-solution-overview.html on April 26, 2024. Always check docs.netapp.com for the latest.

Sommario

Soluzioni dati di cloud ibrido NetApp: Spark e Hadoop in base ai casi di utilizzo dei clienti . . . 1
TR-4657: Soluzioni dati di cloud ibrido NetApp - Spark e Hadoop in base ai casi di utilizzo dei clienti . . . 1
Data fabric basato su NetApp per l'architettura dei big data . . . 2
Protezione dei dati Hadoop e NetApp . . . 5
Panoramica dei casi di utilizzo della protezione dei dati Hadoop . . . 6
Caso d'utilizzo 1: Backup dei dati Hadoop . . . 7
Caso d'utilizzo 2: Backup e disaster recovery dal cloud all'on-premise . . . 11
Caso d'utilizzo 3: Abilitare DevTest sui dati Hadoop esistenti . . . 13
Caso d'utilizzo 4: Protezione dei dati e connettività multicloud . . . 15
Caso d'utilizzo 5: Accelerare i carichi di lavoro di analisi . . . 16
Conclusione . . . 17

Soluzioni dati di cloud ibrido NetApp: Spark e Hadoop in base ai casi di utilizzo dei clienti

TR-4657: Soluzioni dati di cloud ibrido NetApp - Spark e Hadoop in base ai casi di utilizzo dei clienti

Karthikeyan Nagalingam e Sathish Thyagarajan, NetApp

Questo documento descrive le soluzioni dati del cloud ibrido che utilizzano i sistemi storage NetApp AFF e FAS, NetApp Cloud Volumes ONTAP, NetApp Connected Storage e la tecnologia FlexClone per Spark e Hadoop. Queste architetture di soluzioni consentono ai clienti di scegliere una soluzione di protezione dei dati appropriata per il proprio ambiente. NetApp ha progettato queste soluzioni in base all'interazione con i clienti e ai casi di utilizzo aziendali. Questo documento fornisce le seguenti informazioni dettagliate:

- Perché abbiamo bisogno della protezione dei dati per gli ambienti Spark e Hadoop e per le sfide dei clienti.
- Il data fabric basato sulla visione di NetApp e i suoi elementi di base e servizi.
- Come questi building block possono essere utilizzati per progettare flussi di lavoro flessibili per la protezione dei dati.
- I pro e i contro di diverse architetture basate su casi di utilizzo reali dei clienti. Ogni caso d'utilizzo fornisce i seguenti componenti:
  - Scenari dei clienti
  - Requisiti e sfide
  - Soluzioni
  - Riepilogo delle soluzioni

Perché la protezione dei dati Hadoop?

In un ambiente Hadoop e Spark, è necessario risolvere i seguenti problemi:

- **Errori software o umani.** un errore umano negli aggiornamenti software durante l'esecuzione delle operazioni dei dati Hadoop può portare a comportamenti errati che possono causare risultati imprevisti dal lavoro. In tal caso, dobbiamo proteggere i dati per evitare errori o risultati irragionevoli. Ad esempio, a causa di un aggiornamento software eseguito in modo non corretto per un'applicazione di analisi del segnale di traffico, una nuova funzionalità che non riesce ad analizzare correttamente i dati del segnale di traffico sotto forma di testo normale. Il software continua ad analizzare JSON e altri formati di file non di testo, con il risultato che il sistema di analisi del controllo del traffico in tempo reale produce risultati di previsione che sono punti dati mancanti. Questa situazione può causare uscite guaste che potrebbero causare incidenti ai segnali stradali. La protezione dei dati può risolvere questo problema fornendo la possibilità di eseguire rapidamente il rollback alla versione dell'applicazione precedente.
- **Dimensione e scalabilità.** la dimensione dei dati di analisi cresce giorno per giorno a causa del numero sempre crescente di origini dati e volumi. I social media, le app mobili, l'analisi dei dati e le piattaforme di cloud computing sono le principali fonti di dati nell'attuale mercato dei big data, che sta crescendo molto rapidamente, e quindi i dati devono essere protetti per garantire operazioni accurate dei dati.

- **Protezione dei dati nativa di Hadoop.** Hadoop ha un comando nativo per proteggere i dati, ma questo comando non fornisce coerenza dei dati durante il backup. Supporta solo il backup a livello di directory. Le snapshot create da Hadoop sono di sola lettura e non possono essere utilizzate per riutilizzare direttamente i dati di backup.

Problemi di protezione dei dati per i clienti Hadoop e Spark

Una sfida comune per i clienti di Hadoop e Spark consiste nel ridurre i tempi di backup e aumentare l'affidabilità del backup senza influire negativamente sulle performance del cluster di produzione durante la protezione dei dati.

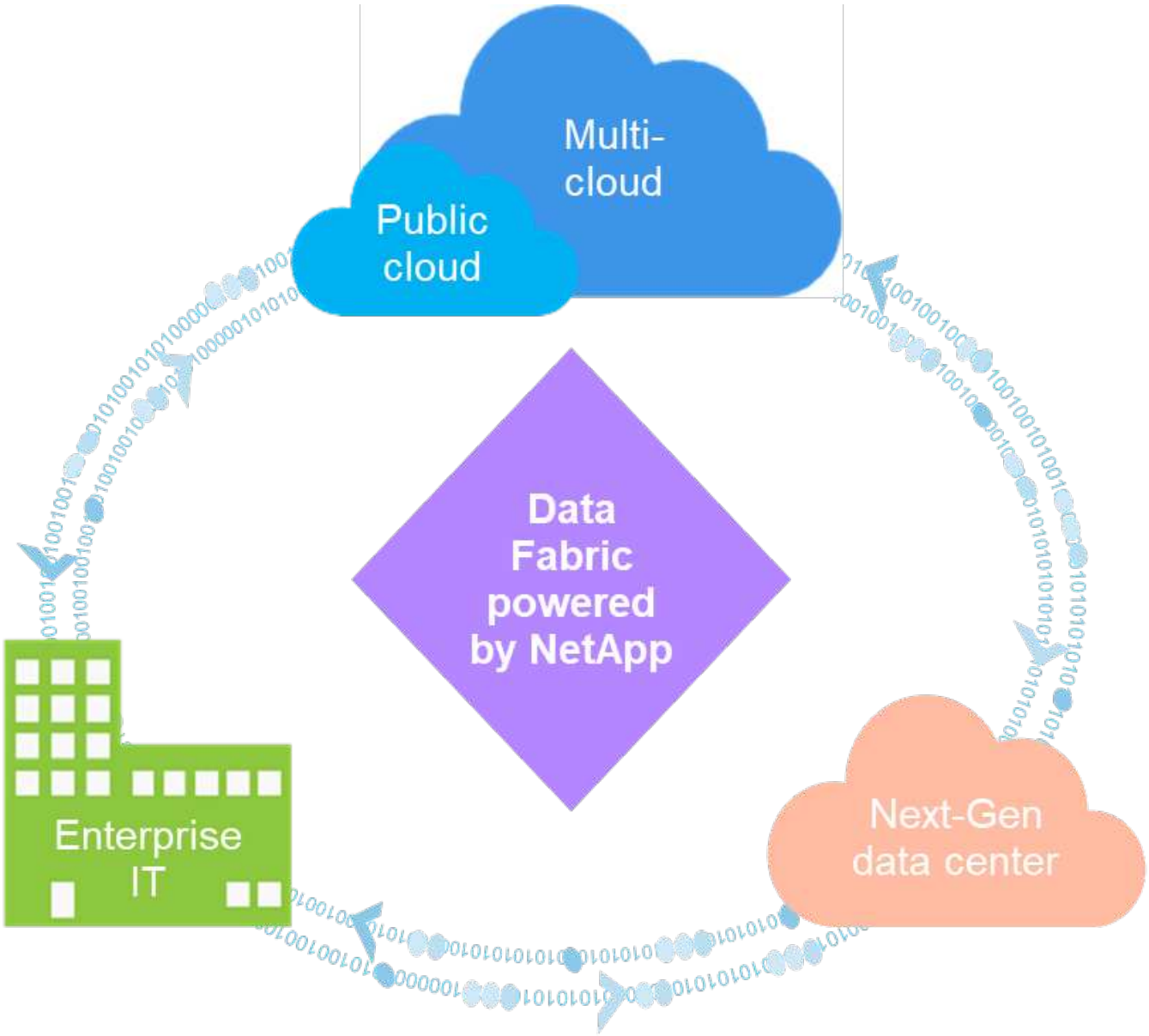
I clienti devono anche ridurre al minimo i tempi di inattività dell'obiettivo del punto di ripristino (RPO) e dell'obiettivo del tempo di ripristino (RTO) e controllare i siti di disaster recovery on-premise e basati sul cloud per una business continuity ottimale. Questo controllo deriva in genere dalla disponibilità di strumenti di gestione di livello Enterprise.

Gli ambienti Hadoop e Spark sono complicati perché non solo il volume di dati è enorme e in crescita, ma la velocità di arrivo dei dati è in aumento. Questo scenario rende difficile creare rapidamente ambienti DevTest e QA efficienti e aggiornati dai dati di origine. NetApp riconosce queste sfide e offre le soluzioni presentate in questo documento.

Data fabric basato su NetApp per l'architettura dei big data

Il data fabric basato su NetApp semplifica e integra la gestione dei dati in ambienti cloud e on-premise per accelerare la trasformazione digitale.

Il data fabric basato su NetApp offre servizi e applicazioni per la gestione dei dati coerenti e integrati (building block) per visibilità e approfondimenti dei dati, accesso e controllo dei dati, protezione e sicurezza dei dati, come mostrato nella figura seguente.



Casi di utilizzo comprovati per i clienti del data fabric

Il data fabric basato su NetApp offre ai clienti i seguenti nove casi di utilizzo comprovati:

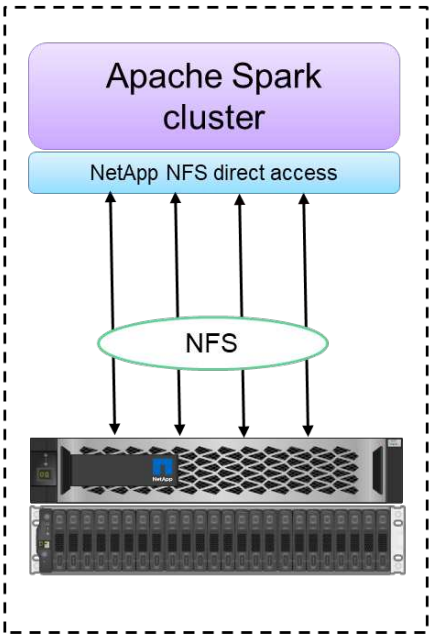
- Accelera i carichi di lavoro di analisi
- Accelera la trasformazione DevOps
- Costruire un'infrastruttura di cloud hosting
- Integra i servizi dati cloud
- Proteggere e proteggere i dati
- Ottimizza i dati non strutturati
- Ottieni efficienze nel data center
- Offri informazioni e controllo sui dati
- Semplifica e automatizza

Questo documento tratta due dei nove casi di utilizzo (insieme alle relative soluzioni):

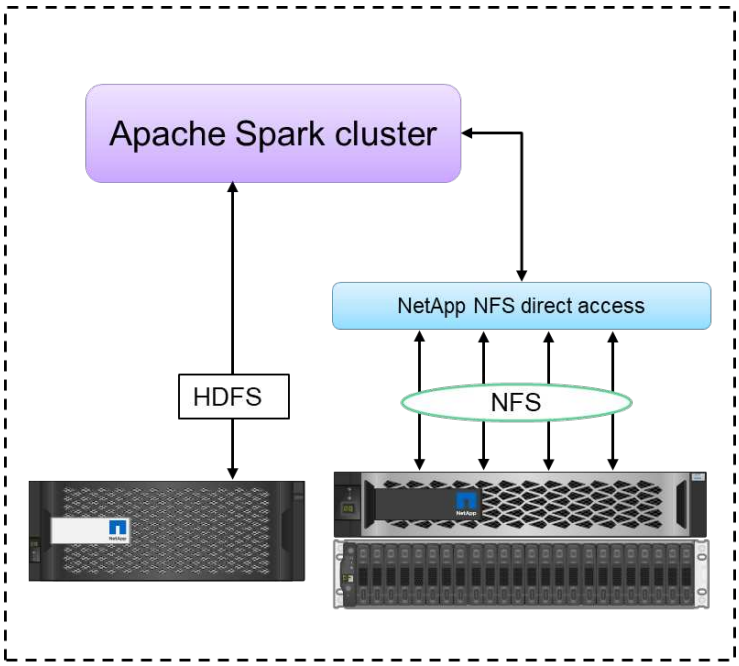
- Accelera i carichi di lavoro di analisi
- Proteggere e proteggere i dati

Accesso diretto NetApp NFS

NFS NetApp consente ai clienti di eseguire job di analytics dei big data sui propri NFSv4 dati NFSv3 o nuovi, senza spostare o copiare i dati. Impedisce più copie di dati ed elimina la necessità di sincronizzare i dati con un'origine. Ad esempio, nel settore finanziario, lo spostamento dei dati da un luogo all'altro deve soddisfare obblighi legali, il che non è un compito facile. In questo scenario, l'accesso diretto NetApp NFS analizza i dati finanziari dalla posizione originale. Un altro vantaggio chiave è che l'utilizzo dell'accesso diretto NetApp NFS semplifica la protezione dei dati Hadoop utilizzando comandi Hadoop nativi e abilita i flussi di lavoro per la protezione dei dati sfruttando il ricco portfolio di gestione dei dati di NetApp.



Configuration 1: NFS as primary storage



Configuration 2: HDFS and NFS in single Spark cluster

L'accesso diretto NetApp NFS offre due tipi di opzioni di implementazione per i cluster Hadoop/Spark:

- Per impostazione predefinita, i cluster Hadoop/Spark utilizzano Hadoop Distributed file System (HDFS) per lo storage dei dati e il file system predefinito. L'accesso diretto NetApp NFS può sostituire l'HDFS predefinito con lo storage NFS come file system predefinito, consentendo operazioni di analisi dirette sui dati NFS.
- In un'altra opzione di implementazione, l'accesso diretto NetApp NFS supporta la configurazione di NFS come storage aggiuntivo insieme a HDFS in un singolo cluster Hadoop/Spark. In questo caso, il cliente può condividere i dati attraverso le esportazioni NFS e accedervi dallo stesso cluster insieme ai dati HDFS.

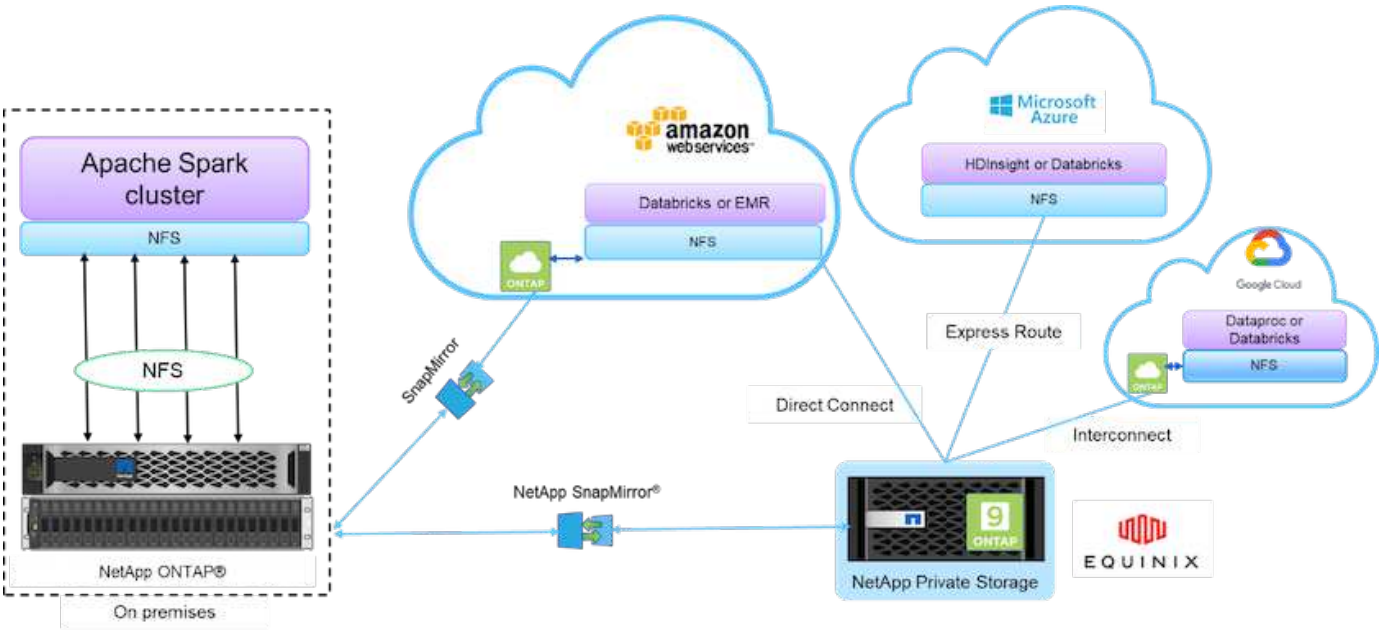
I vantaggi principali dell'utilizzo dell'accesso diretto NetApp NFS includono:

- Analizza i dati dalla posizione corrente, impedendo il dispendio di tempo e performance dello spostamento dei dati di analisi in un'infrastruttura Hadoop come HDFS.
- Riduce il numero di repliche da tre a uno.

- Consente agli utenti di separare il calcolo e lo storage per scalare in modo indipendente.
- Offre protezione dei dati aziendali sfruttando le funzionalità di gestione dei dati avanzate di ONTAP.
- È certificato con la piattaforma dati Hortonworks.
- Consente implementazioni di analisi dei dati ibride.
- Riduce i tempi di backup sfruttando la funzionalità multithread dinamica.

Elementi di base per i big data

Il data fabric basato su NetApp integra i servizi e le applicazioni di gestione dei dati (building block) per l'accesso, il controllo, la protezione e la sicurezza dei dati, come mostrato nella figura seguente.



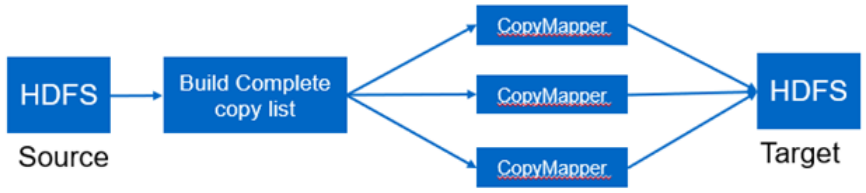
Gli elementi di base della figura precedente includono:

- **Accesso diretto NetApp NFS.** offre i più recenti cluster Hadoop e Spark con accesso diretto ai volumi NetApp NFS senza requisiti aggiuntivi di software o driver.
- **NetApp Cloud Volumes ONTAP e servizi di volume cloud.** storage connesso definito tramite software basato su ONTAP eseguito in AWS (Amazon Web Services) o Azure NetApp Files (ANF) nei servizi cloud Microsoft Azure.
- **Tecnologia NetApp SnapMirror**. Offre funzionalità di protezione dei dati tra istanze cloud on-premise e ONTAP o NPS.
- **Cloud service provider.** questi provider includono AWS, Microsoft Azure, Google Cloud e IBM Cloud.
- **PaaS.** servizi di analisi basati sul cloud come Amazon Elastic MapReduce (EMR) e Databricks in AWS, Microsoft Azure HDInsight e Azure Databricks.

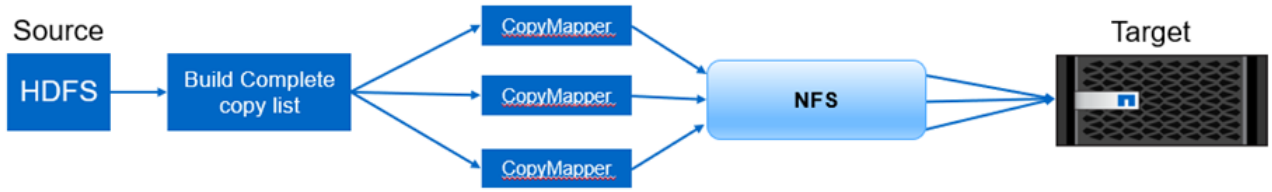
Protezione dei dati Hadoop e NetApp

Hadoop DistCp è uno strumento nativo utilizzato per la copia di intercluster e intracluster di grandi dimensioni. Il processo di base di Hadoop DistCp illustrato nella figura seguente è un tipico workflow di backup che utilizza strumenti nativi di Hadoop come MapReduce per copiare i dati di Hadoop da un'origine HDFS a una destinazione corrispondente.

L'accesso diretto NetApp NFS consente ai clienti di impostare NFS come destinazione per lo strumento Hadoop DistCp per copiare i dati dall'origine HDFS in una condivisione NFS tramite MapReduce. L'accesso diretto NetApp NFS funge da driver NFS per lo strumento DistCp.



Hadoop Distcp Basic Process



Hadoop Distcp and NetApp

Panoramica dei casi di utilizzo della protezione dei dati Hadoop

In questa sezione viene fornita una descrizione dettagliata dei casi di utilizzo della protezione dei dati, che costituiscono l'oggetto del presente documento. Le restanti sezioni forniscono ulteriori dettagli per ciascun caso di utilizzo, ad esempio il problema del cliente (scenario), i requisiti, le sfide e le soluzioni.

Caso d'utilizzo 1: Backup dei dati Hadoop

Per questo caso d'utilizzo, il volume NFS di NetApp ha aiutato un grande istituto finanziario a ridurre l'estesa finestra di backup da più di 24 ore a poco meno di poche ore.

Caso d'utilizzo 2: Backup e disaster recovery dal cloud all'on-premise

Utilizzando il data fabric basato su NetApp come elementi di base, una grande azienda di radiodiffusione è stata in grado di soddisfare il proprio requisito di backup dei dati cloud nel proprio data center on-premise, a seconda delle diverse modalità di trasferimento dei dati, come on-demand, istantaneo, O in base al carico del cluster Hadoop/Spark.

Caso d'utilizzo 3: Abilitare DevTest sui dati Hadoop esistenti

Le soluzioni NetApp hanno aiutato un distributore di musica online a creare rapidamente più cluster Hadoop efficienti in termini di spazio in diverse filiali per creare report ed eseguire attività DevTest quotidiane utilizzando policy pianificate.

Caso d’utilizzo 4: Protezione dei dati e connettività multicloud

Un grande provider di servizi ha utilizzato il data fabric basato su NetApp per fornire ai propri clienti analytics multicloud da diverse istanze di cloud.

Caso d’utilizzo 5: Accelerare i carichi di lavoro di analisi

Una delle più grandi banche di investimento e servizi finanziari ha utilizzato la soluzione di storage NAS di NetApp per ridurre i tempi di attesa i/o e accelerare la piattaforma di analisi finanziaria quantitativa.

Caso d’utilizzo 1: Backup dei dati Hadoop

Scenario

In questo scenario, il cliente dispone di un grande repository Hadoop on-premise e desidera eseguirne il backup a scopo di disaster recovery. Tuttavia, l’attuale soluzione di backup del cliente è costosa e presenta una lunga finestra di backup di oltre 24 ore.

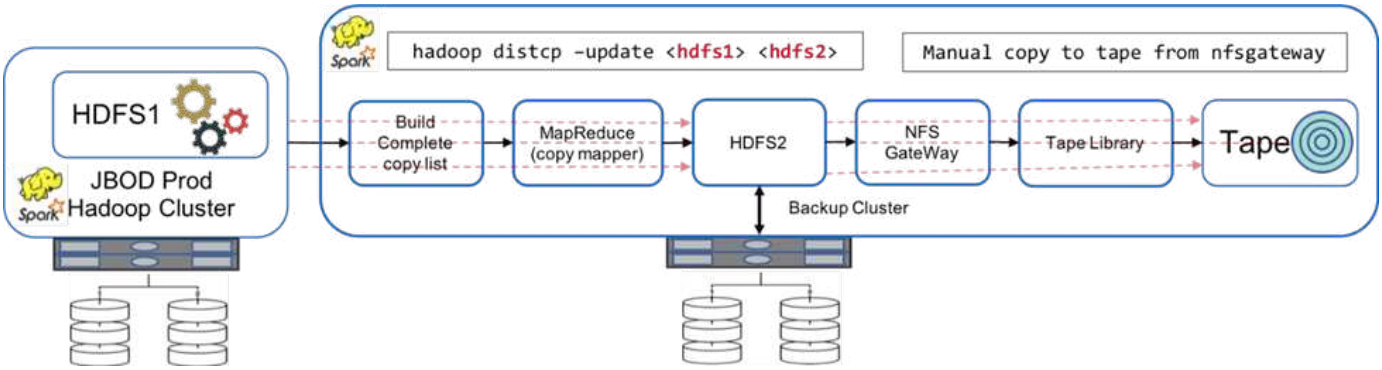
Requisiti e sfide

I requisiti e le sfide principali per questo caso di utilizzo includono:

- Compatibilità con le versioni precedenti del software:
  - La soluzione di backup alternativa proposta deve essere compatibile con le versioni software attualmente in esecuzione utilizzate nel cluster Hadoop di produzione.
- Per soddisfare gli SLA impegnati, la soluzione alternativa proposta dovrebbe raggiungere RPO e RTO molto bassi.
- Il backup creato dalla soluzione di backup NetApp può essere utilizzato nel cluster Hadoop costruito localmente nel data center e nel cluster Hadoop in esecuzione nella posizione di disaster recovery presso il sito remoto.
- La soluzione proposta deve essere conveniente.
- La soluzione proposta deve ridurre l’effetto delle performance sui processi di analisi in produzione attualmente in esecuzione durante i tempi di backup.

La soluzione di backup esistente del cliente x

La figura seguente mostra la soluzione di backup nativa Hadoop originale.



I dati di produzione sono protetti su nastro attraverso il cluster di backup intermedio:

- I dati HDFS1 vengono copiati in HDFS2 eseguendo il `hadoop distcp -update <hdfs1> <hdfs2>` comando.
- Il cluster di backup funge da gateway NFS e i dati vengono copiati manualmente su nastro attraverso Linux `cp` tramite la libreria di nastri.

I vantaggi della soluzione di backup nativa Hadoop originale includono:

- La soluzione si basa sui comandi nativi di Hadoop, che consentono all'utente di non dover apprendere nuove procedure.
- La soluzione sfrutta l'architettura e l'hardware standard di settore.

Gli svantaggi della soluzione di backup nativa Hadoop originale includono:

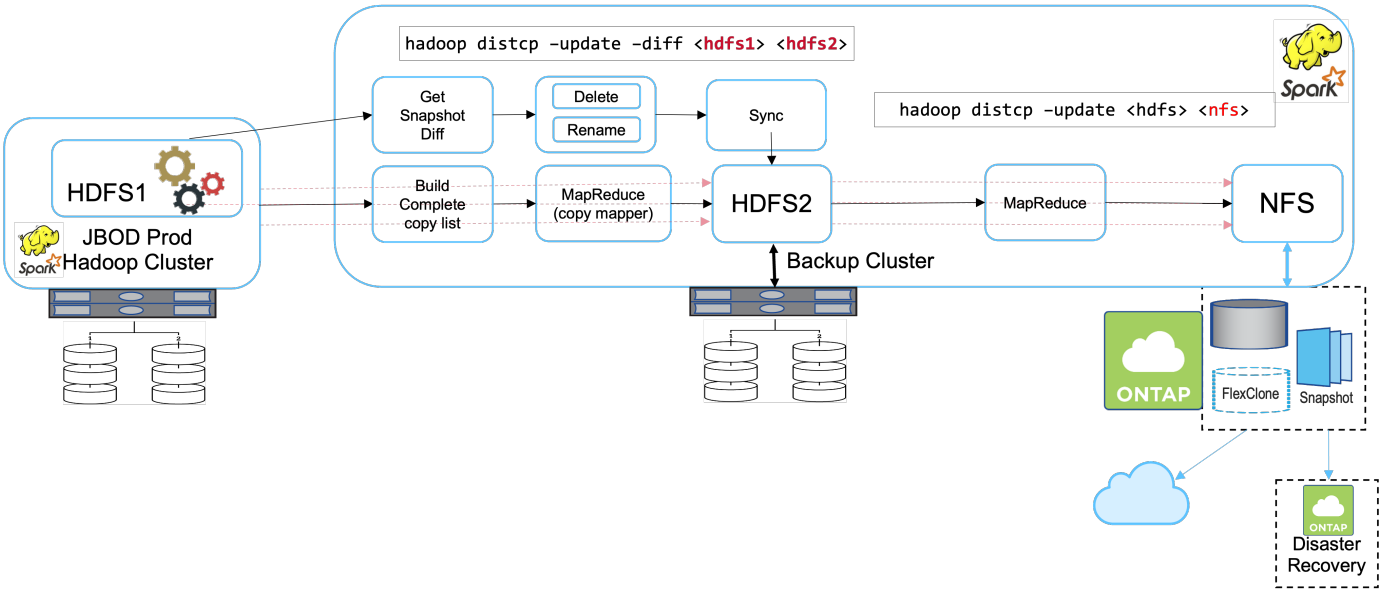
- La lunga finestra di backup supera le 24 ore, rendendo vulnerabili i dati di produzione.
- Peggioramento significativo delle performance del cluster durante i tempi di backup.
- La copia su nastro è un processo manuale.
- La soluzione di backup è costosa in termini di hardware richiesto e di ore umane richieste per i processi manuali.

Soluzioni di backup

In base a queste sfide e requisiti e tenendo conto del sistema di backup esistente, sono state suggerite tre possibili soluzioni di backup. Le seguenti sottosezioni descrivono ciascuna di queste tre diverse soluzioni di backup, etichettate dalla soluzione A alla soluzione C.

Soluzione A

Nella soluzione A, il cluster Hadoop di backup invia i backup secondari ai sistemi storage NFS NetApp, eliminando i requisiti su nastro, come mostrato nella figura sotto.



Le attività dettagliate per la soluzione A includono:

- Il cluster di produzione Hadoop contiene i dati di analisi del cliente nell'HDFS che richiede protezione.

- Il cluster di backup Hadoop con HDFS funge da posizione intermedia per i dati. Solo un gruppo di dischi (JBOD) fornisce lo storage per HDFS sia nei cluster Hadoop di produzione che di backup.
- Proteggere i dati di produzione di Hadoop dal cluster di produzione HDFS al cluster di backup HDFS mediante l'esecuzione di `Hadoop distcp -update -diff <hdfs1> <hdfs2>` comando.

> Lo snapshot Hadoop viene utilizzato per proteggere i dati dalla produzione al cluster di backup Hadoop.

- Il controller di storage NetApp ONTAP fornisce un volume esportato NFS, che viene fornito al cluster di backup Hadoop.
- Eseguendo il `Hadoop distcp` Comando che sfrutta MapReduce e mappatori multipli, i dati degli analytics sono protetti dal cluster Hadoop di backup su NFS.

  Una volta archiviati i dati in NFS sul sistema storage NetApp, le tecnologie NetApp Snapshot, SnapRestore e FlexClone vengono utilizzate per eseguire il backup, il ripristino e la duplicazione dei dati Hadoop in base alle necessità.

> I dati Hadoop possono essere protetti nel cloud e nelle posizioni di disaster recovery utilizzando la tecnologia SnapMirror.

I vantaggi della soluzione A includono:

- I dati di produzione di Hadoop sono protetti dal cluster di backup.
- I dati HDFS sono protetti tramite NFS, consentendo la protezione in ambienti cloud e di disaster recovery.
- Migliora le performance trasferendo le operazioni di backup nel cluster di backup.
- Elimina le operazioni manuali su nastro
- Consente funzioni di gestione aziendale tramite gli strumenti NetApp.
- Richiede modifiche minime all'ambiente esistente.
- È una soluzione conveniente.

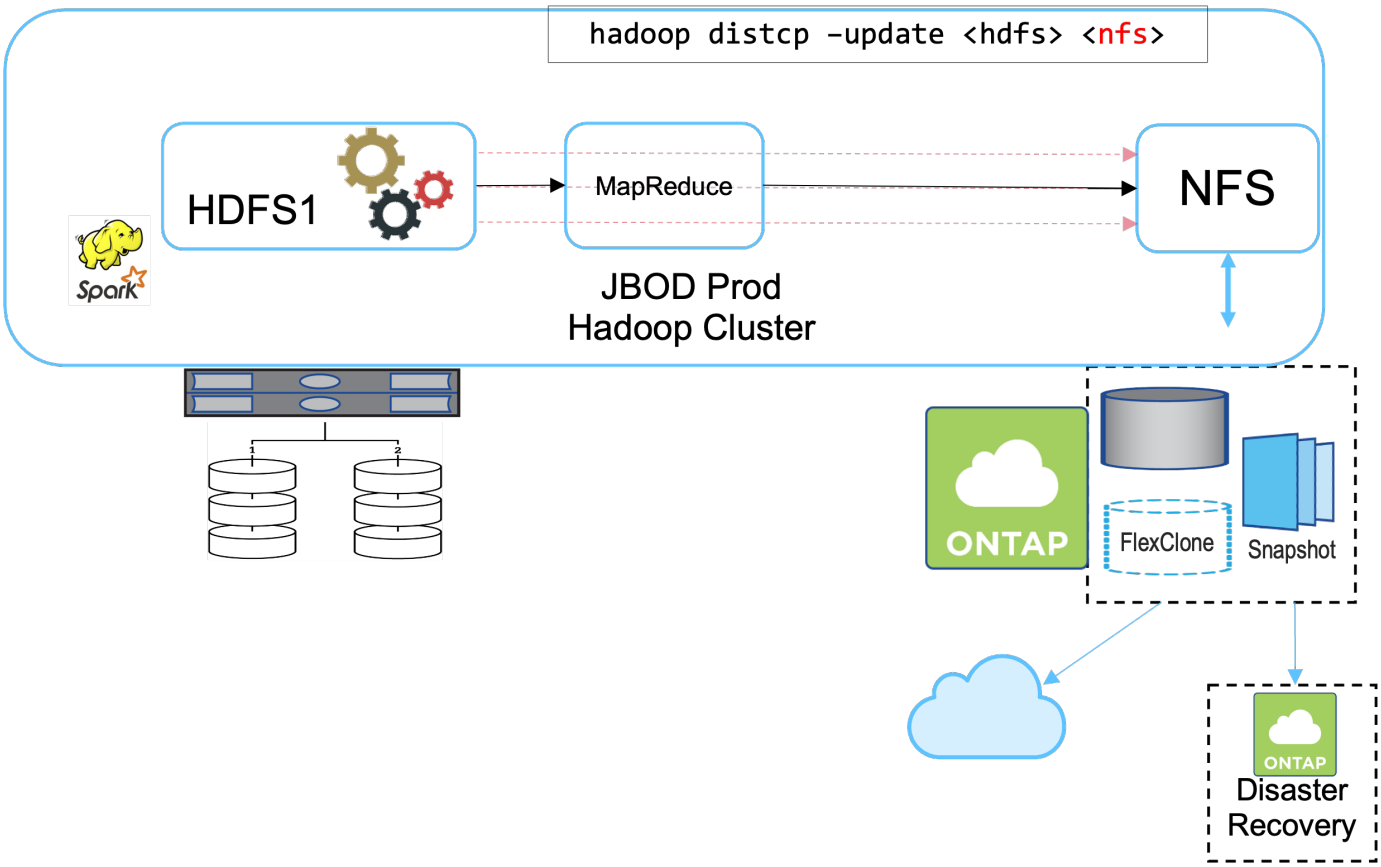
Lo svantaggio di questa soluzione è che richiede un cluster di backup e mappatori aggiuntivi per migliorare le performance.

Il cliente ha recentemente implementato la soluzione A per la sua semplicità, i costi e le performance complessive.

In questa soluzione, è possibile utilizzare i dischi SAN di ONTAP invece di JBOD. Questa opzione trasferisce il carico dello storage del cluster di backup su ONTAP; tuttavia, il downside è che sono richiesti switch fabric SAN.

Soluzione B

La soluzione B aggiunge un volume NFS al cluster Hadoop di produzione, che elimina la necessità di un cluster Hadoop di backup, come mostrato nella figura sotto.



Le attività dettagliate per la soluzione B includono:

- Lo storage controller NetApp ONTAP fornisce l'esportazione NFS al cluster Hadoop di produzione.

  Nativo di Hadoop `hadoop distcp` Command protegge i dati Hadoop dal cluster di produzione HDFS a NFS.

- Una volta archiviati i dati in NFS sul sistema storage NetApp, le tecnologie Snapshot, SnapRestore e FlexClone vengono utilizzate per eseguire il backup, il ripristino e la duplicazione dei dati Hadoop in base alle necessità.

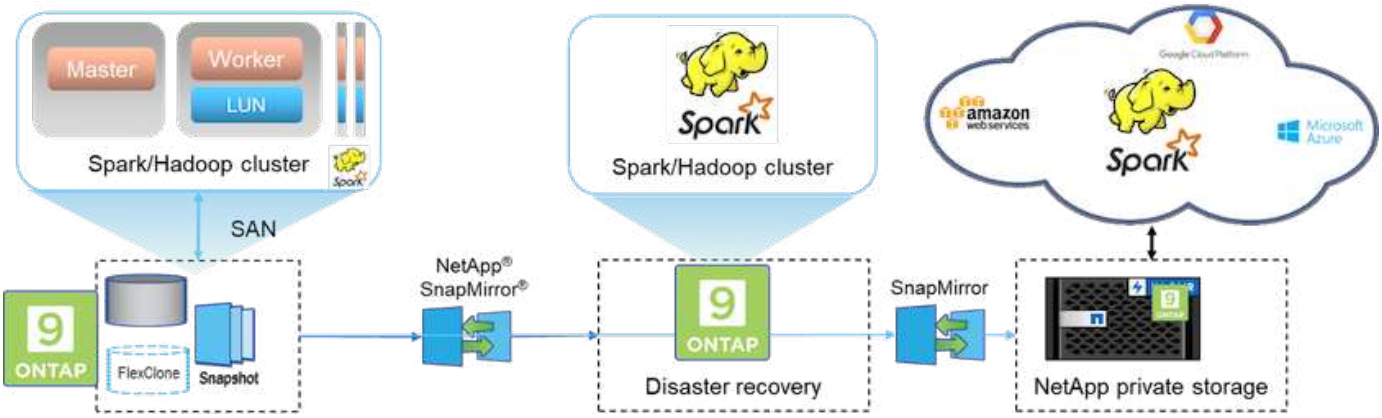
I vantaggi della soluzione B includono:

- Il cluster di produzione viene leggermente modificato per la soluzione di backup, semplificando l'implementazione e riducendo i costi aggiuntivi dell'infrastruttura.
- Non è necessario un cluster di backup per l'operazione di backup.
- I dati di produzione HDFS sono protetti nella conversione in dati NFS.
- La soluzione consente funzioni di gestione aziendale tramite gli strumenti NetApp.

Lo svantaggio di questa soluzione è che è implementata nel cluster di produzione, che può aggiungere ulteriori attività di amministratore nel cluster di produzione.

Soluzione C

Nella soluzione C, il provisioning dei volumi SAN NetApp viene eseguito direttamente nel cluster di produzione Hadoop per lo storage HDFS, come illustrato nella figura seguente.



I passaggi dettagliati per la soluzione C includono:

- Lo storage SAN NetApp ONTAP viene fornito nel cluster di produzione Hadoop per lo storage dei dati HDFS.
- Le tecnologie NetApp Snapshot e SnapMirror vengono utilizzate per eseguire il backup dei dati HDFS dal cluster Hadoop di produzione.
- Durante il processo di backup della copia Snapshot non si verificano effetti sulle performance per il cluster Hadoop/Spark, poiché il backup si trova a livello di storage.

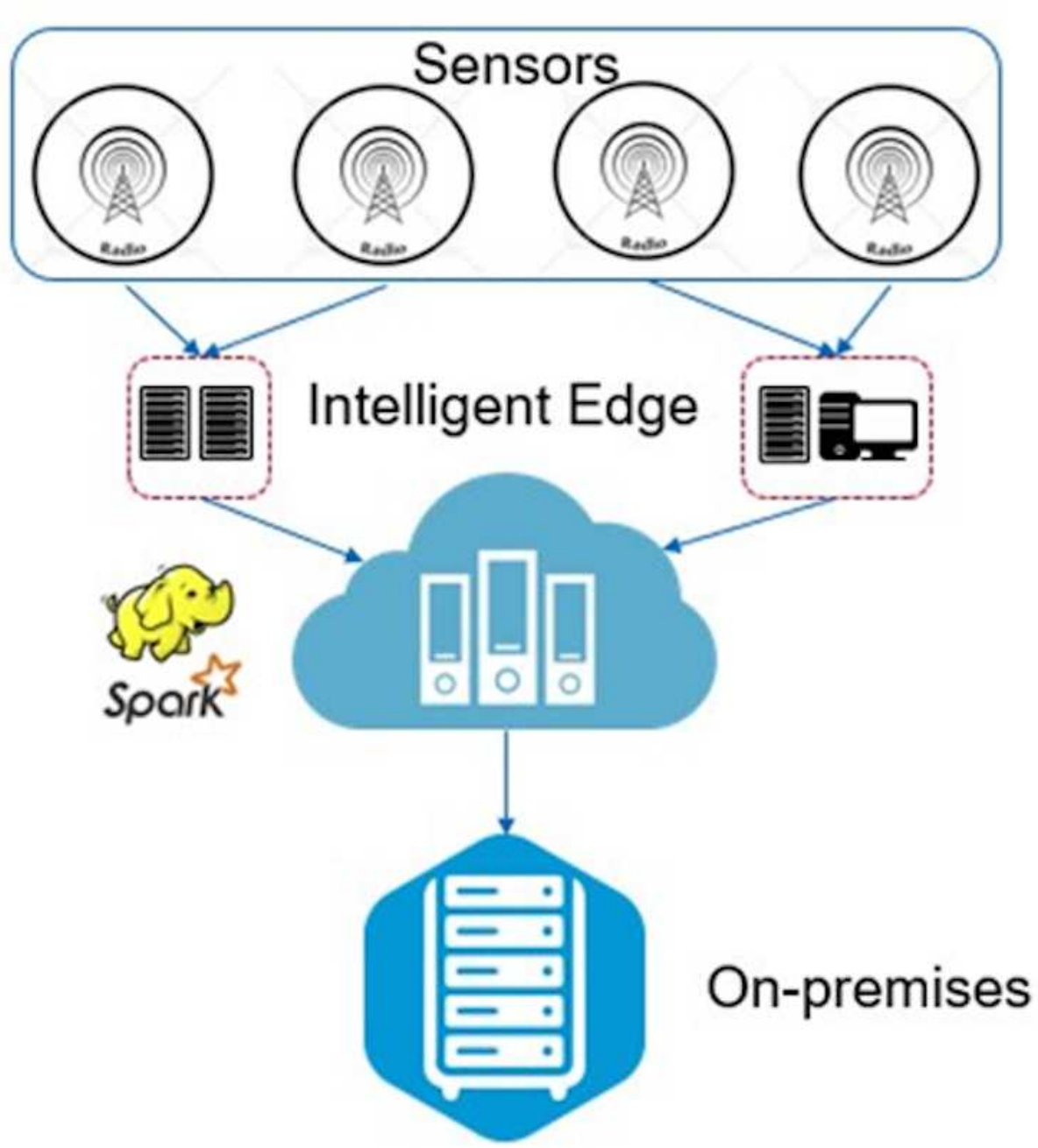
> La tecnologia Snapshot offre backup completi in pochi secondi, indipendentemente dalle dimensioni dei dati.

I vantaggi della soluzione C includono:

- È possibile creare backup efficienti in termini di spazio utilizzando la tecnologia Snapshot.
- Consente funzioni di gestione aziendale tramite gli strumenti NetApp.

Caso d'utilizzo 2: Backup e disaster recovery dal cloud all'on-premise

Questo caso di utilizzo si basa su un cliente che trasmette dati che deve eseguire il backup dei dati di analisi basati sul cloud nel proprio data center on-premise, come illustrato nella figura seguente.



Scenario

In questo scenario, i dati dei sensori IoT vengono acquisiti nel cloud e analizzati utilizzando un cluster open source Apache Spark all'interno di AWS. Il requisito è di eseguire il backup dei dati elaborati dal cloud a on-premise.

Requisiti e sfide

I requisiti e le sfide principali per questo caso di utilizzo includono:

- L'attivazione della protezione dei dati non dovrebbe causare alcun effetto sulle performance del cluster Spark/Hadoop di produzione nel cloud.
- I dati dei sensori cloud devono essere spostati e protetti on-premise in modo efficiente e sicuro.

- Flessibilità per il trasferimento dei dati dal cloud all'on-premise in diverse condizioni, ad esempio on-demand, istantaneo e durante i bassi tempi di carico del cluster.

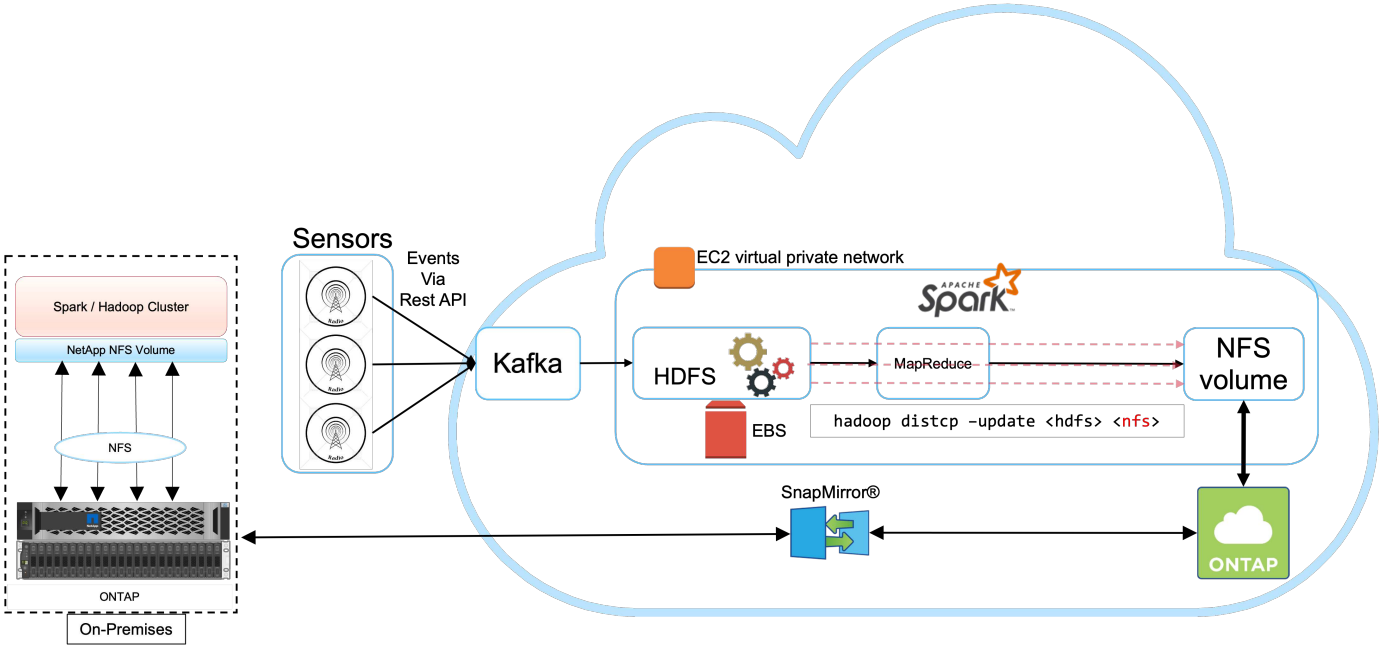
Soluzione

Il cliente utilizza AWS Elastic Block Store (EBS) per lo storage HDFS del cluster Spark per ricevere e acquisire dati dai sensori remoti tramite Kafka. Di conseguenza, lo storage HDFS funge da origine per i dati di backup.

Per soddisfare questi requisiti, il cloud NetApp ONTAP viene implementato in AWS e viene creata una condivisione NFS per fungere da destinazione di backup per il cluster Spark/Hadoop.

Una volta creata la share NFS, copia i dati dallo storage HDFS EBS nella share NFS ONTAP. Una volta che i dati risiedono in NFS nel cloud ONTAP, la tecnologia SnapMirror può essere utilizzata per eseguire il mirroring dei dati dal cloud allo storage on-premise in modo sicuro ed efficiente.

Questa immagine mostra il backup e il disaster recovery dal cloud alla soluzione on-premise.



Caso d'utilizzo 3: Abilitare DevTest sui dati Hadoop esistenti

In questo caso di utilizzo, il requisito del cliente è quello di creare rapidamente ed efficientemente nuovi cluster Hadoop/Spark basati su un cluster Hadoop esistente contenente una grande quantità di dati di analisi per DevTest e scopi di reporting nello stesso data center e in sedi remote.

Scenario

In questo scenario, i cluster Spark/Hadoop multipli sono costruiti da un'implementazione di data Lake Hadoop di grandi dimensioni on-premise e in ubicazioni di disaster recovery.

Requisiti e sfide

I requisiti e le sfide principali per questo caso di utilizzo includono:

- Creare più cluster Hadoop per DevTest, QA o qualsiasi altro scopo che richieda l'accesso agli stessi dati di produzione. La sfida consiste nel clonare un cluster Hadoop molto grande più volte istantaneamente e in modo molto efficiente in termini di spazio.
- Sincronizza i dati di Hadoop con DevTest e i team di reporting per l'efficienza operativa.
- Distribuire i dati Hadoop utilizzando le stesse credenziali in produzione e nei nuovi cluster.
- Utilizza policy pianificate per creare in modo efficiente cluster di QA senza influire sul cluster di produzione.

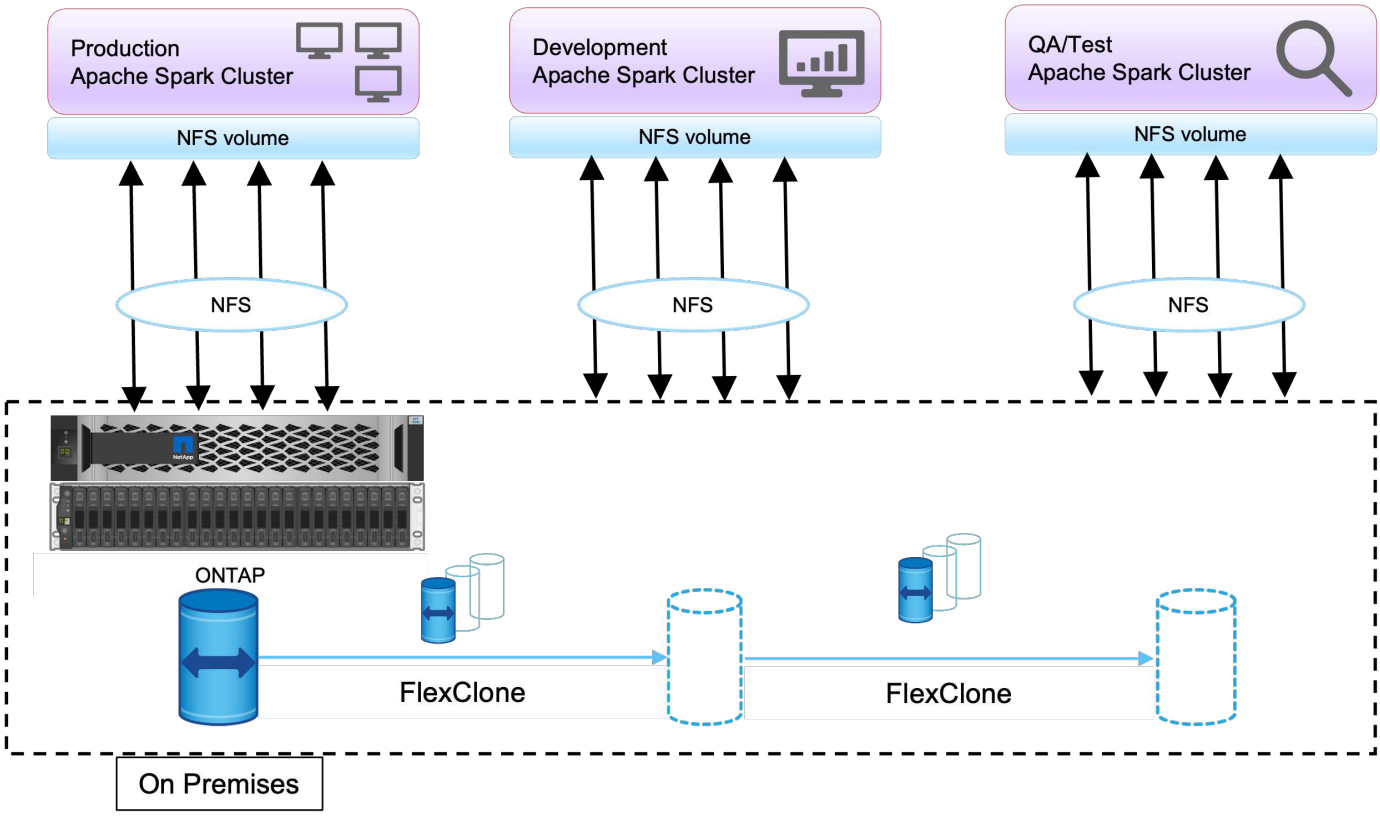
Soluzione

La tecnologia FlexClone viene utilizzata per rispondere ai requisiti appena descritti. La tecnologia FlexClone è la copia in lettura/scrittura di una copia Snapshot. Legge i dati dai dati di copia Snapshot padre e consuma solo spazio aggiuntivo per i blocchi nuovi/modificati. È veloce ed efficiente in termini di spazio.

Innanzitutto, è stata creata una copia Snapshot del cluster esistente utilizzando un gruppo di coerenza NetApp.

Copie Snapshot all'interno di NetApp System Manager o al prompt di amministrazione dello storage. Il gruppo di coerenza copie Snapshot sono copie Snapshot di gruppo coerenti con l'applicazione e il volume FlexClone viene creato in base alle copie Snapshot del gruppo di coerenza. Vale la pena ricordare che un volume FlexClone eredita la policy di esportazione NFS del volume padre. Una volta creata la copia Snapshot, è necessario installare un nuovo cluster Hadoop per DevTest e per la creazione di report, come illustrato nella figura seguente. Il volume NFS clonato dal nuovo cluster Hadoop accede ai dati NFS.

Questa immagine mostra il cluster Hadoop per DevTest.



Caso d'utilizzo 4: Protezione dei dati e connettività multicloud

Questo caso di utilizzo è importante per un partner di servizi cloud che ha il compito di fornire connettività multi-cloud per i dati di analisi dei big data dei clienti.

Scenario

In questo scenario, i dati IoT ricevuti in AWS da diverse origini vengono memorizzati in una posizione centrale in NPS. Lo storage NPS è connesso ai cluster Spark/Hadoop situati in AWS e Azure, consentendo l'esecuzione di applicazioni di analisi dei big data in più cloud che accedono agli stessi dati.

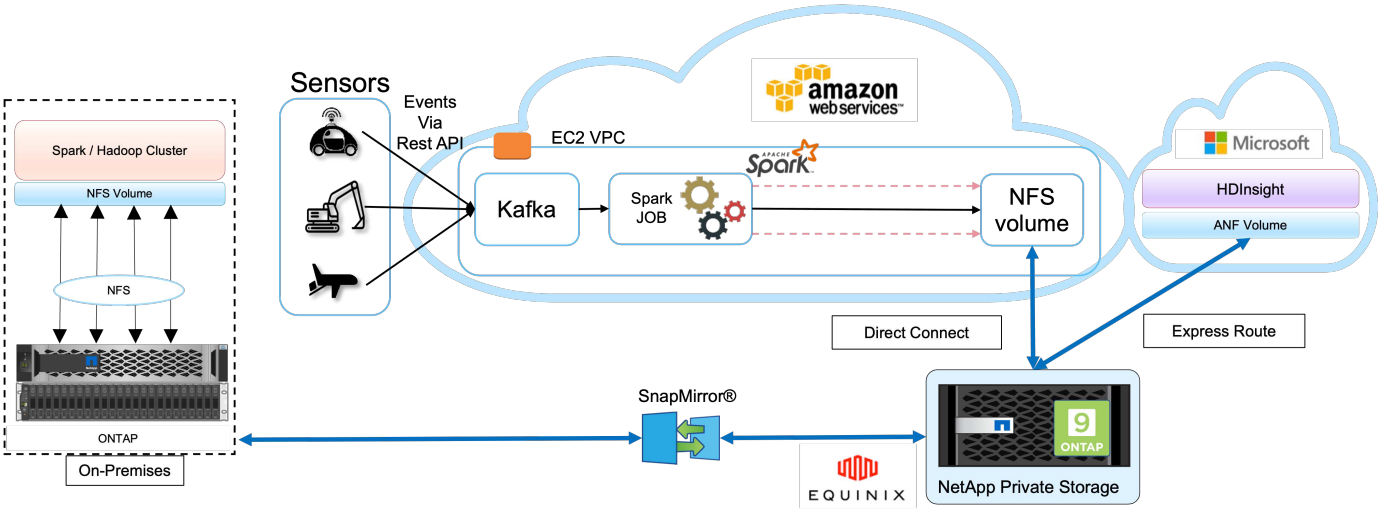
Requisiti e sfide

I requisiti e le sfide principali per questo caso di utilizzo includono:

- I clienti vogliono eseguire lavori di analisi sugli stessi dati utilizzando più cloud.
- I dati devono essere ricevuti da fonti diverse, ad esempio on-premise e cloud, attraverso diversi sensori e hub.
- La soluzione deve essere efficiente e conveniente.
- La sfida principale è quella di creare una soluzione conveniente ed efficiente in grado di offrire servizi di analisi ibridi tra cloud diversi e on-premise.

Soluzione

Questa immagine illustra la soluzione per la protezione dei dati e la connettività multicloud.



Come mostrato nella figura precedente, i dati provenienti dai sensori vengono trasmessi e acquisiti nel cluster AWS Spark tramite Kafka. I dati vengono memorizzati in una condivisione NFS residente in NPS, che si trova all'esterno del cloud provider all'interno di un data center Equinix. Poiché NetApp NetApp Private Storage è connesso ad Amazon AWS e Microsoft Azure tramite connessioni Direct Connect ed Express Route, i clienti possono accedere ai dati NFS dai cluster di analytics di Amazon e AWS. Questo approccio risolve l'utilizzo di analytics nel cloud in più hyperscaler.

Di conseguenza, poiché sia lo storage on-premise che NPS esegue il software ONTAP, SnapMirror può

eseguire il mirroring dei dati NPS nel cluster on-premise, fornendo analisi del cloud ibrido tra cloud on-premise e multipli.

Per ottenere le migliori performance, NetApp consiglia di utilizzare interfacce di rete multiple e percorsi di connessione diretta/express per accedere ai dati dalle istanze cloud.

Caso d’utilizzo 5: Accelerare i carichi di lavoro di analisi

In questo scenario, è stata modernizzata una grande piattaforma di analytics per i servizi finanziari e le banche di investimento utilizzando la soluzione di storage NetApp NFS per ottenere un miglioramento significativo nell’analisi dei rischi di investimento e dei derivati per la sua business unit quantitativa e di gestione delle risorse.

Scenario

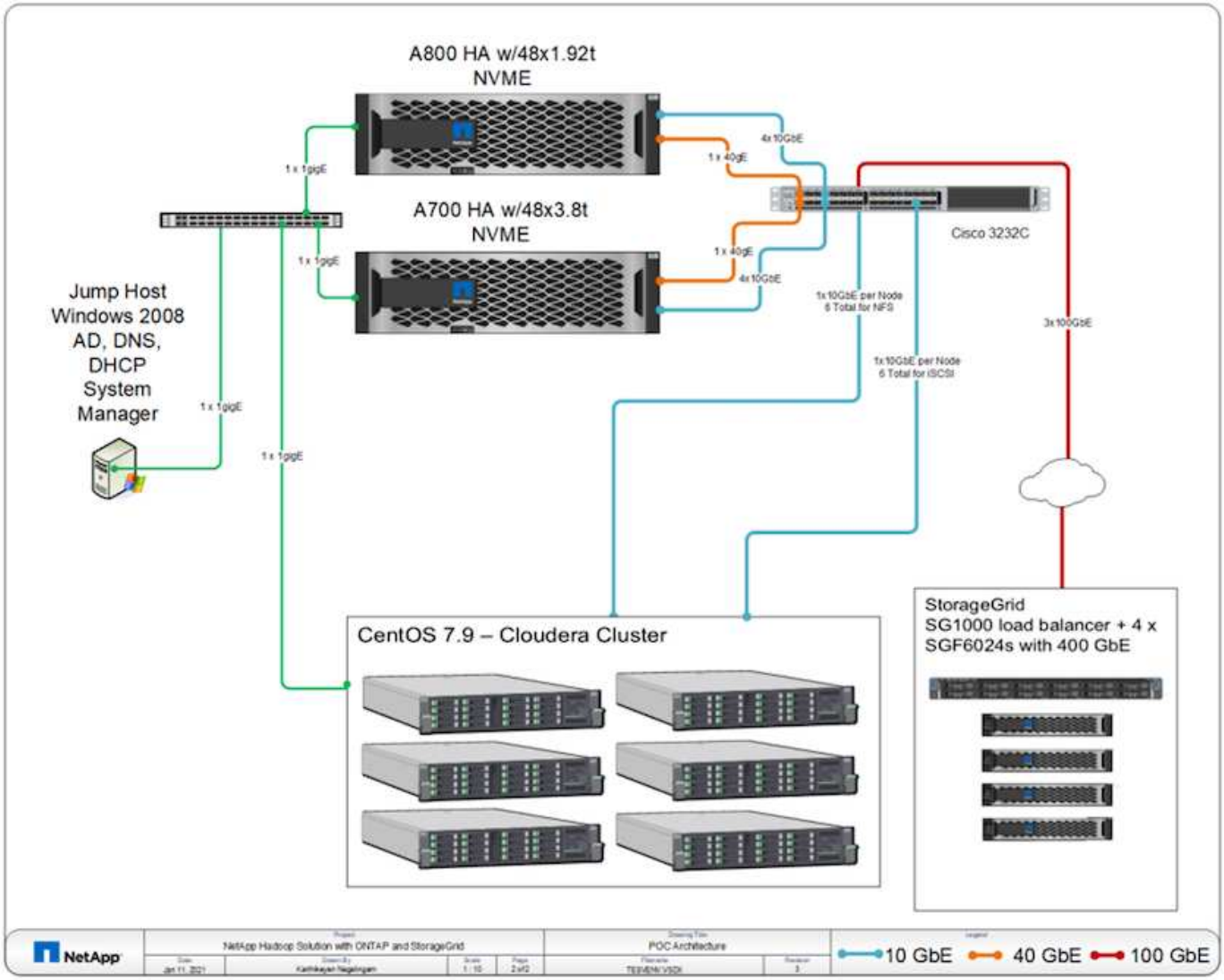
Nell’ambiente esistente del cliente, l’infrastruttura Hadoop utilizzata per la piattaforma di analisi ha sfruttato lo storage interno dei server Hadoop. A causa della natura proprietaria dell’ambiente JBOD, molti clienti interni all’interno dell’organizzazione non sono stati in grado di sfruttare il proprio modello quantitativo Monte Carlo, una simulazione che si basa sui campioni ricorrenti di dati in tempo reale. La capacità non ottimale di comprendere gli effetti dell’incertezza nei movimenti di mercato serviva in modo sfavorevole per la business unit di gestione quantitativa delle risorse.

Requisiti e sfide

La business unit quantitativa della banca desiderava un metodo di previsione efficiente per ottenere previsioni precise e tempestive. A tale scopo, il team ha riconosciuto la necessità di modernizzare l’infrastruttura, ridurre i tempi di attesa i/o esistenti e migliorare le performance delle applicazioni di analisi come Hadoop e Spark per simulare in modo efficiente i modelli di investimento, misurare i potenziali guadagni e analizzare i rischi.

Soluzione

Il cliente disponeva di JBOD per la soluzione Spark esistente. NetApp ONTAP, NetApp StorageGRID e MinIO Gateway to NFS sono stati quindi sfruttati per ridurre i tempi di attesa i/o per il gruppo finanziario quantitativo della banca che esegue simulazioni e analisi su modelli di investimento che valutano potenziali guadagni e rischi. Questa immagine mostra la soluzione Spark con lo storage NetApp.



Come mostrato nella figura precedente, i sistemi AFF A800, A700 e StorageGRID sono stati implementati per accedere ai file di parquet attraverso i protocolli NFS e S3 in un cluster Hadoop a sei nodi con Spark e i servizi di metadati YARN e Hive per le operazioni di analisi dei dati.

Una soluzione DAS (Direct-Attached Storage) nel vecchio ambiente del cliente ha avuto lo svantaggio di scalare calcolo e storage in modo indipendente. Con la soluzione NetApp ONTAP per Spark, la business unit di analisi finanziaria della banca è stata in grado di separare lo storage dal calcolo e di portare le risorse dell'infrastruttura in modo più efficace secondo necessità.

Utilizzando ONTAP con NFS, le CPU dei server di calcolo sono state quasi completamente utilizzate per i job SQL di Spark e il tempo di attesa i/o è stato ridotto di quasi il 70%, fornendo quindi una potenza di calcolo e un incremento delle performance migliori per i carichi di lavoro di Spark. In seguito, l'aumento dell'utilizzo della CPU ha anche consentito al cliente di sfruttare GPU, come GPUDirect, per un'ulteriore modernizzazione della piattaforma. Inoltre, StorageGRID offre un'opzione di storage a basso costo per i carichi di lavoro Spark e il gateway MinIO fornisce un accesso sicuro ai dati NFS attraverso il protocollo S3. Per i dati nel cloud, NetApp consiglia Cloud Volumes ONTAP, Azure NetApp Files e NetApp Cloud Volumes Service.

Conclusione

Questa sezione fornisce un riepilogo dei casi di utilizzo e delle soluzioni fornite da NetApp per soddisfare i vari requisiti di protezione dei dati Hadoop. Utilizzando il data

fabric basato su NetApp, i clienti possono:

- Avere la flessibilità di scegliere le giuste soluzioni per la protezione dei dati sfruttando le ricche funzionalità di gestione dei dati di NetApp e l'integrazione con i flussi di lavoro nativi di Hadoop.
- Ridurre i tempi di backup del cluster Hadoop di quasi il 70%.
- Elimina qualsiasi effetto sulle performance derivante dai backup del cluster Hadoop.
- Protezione dei dati multicloud e accesso ai dati da diversi cloud provider contemporaneamente a una singola fonte di dati di analisi.
- Crea copie cluster Hadoop veloci ed efficienti in termini di spazio utilizzando la tecnologia FlexClone.

Dove trovare ulteriori informazioni

Per ulteriori informazioni sulle informazioni descritte in questo documento, consultare i seguenti documenti e/o siti Web:

- Soluzioni NetApp per l'analisi dei big data

  "https://www.netapp.com/us/solutions/applications/big-data-analytics/index.aspx"

- Apache Spark workload con lo storage NetApp

  https://www.netapp.com/pdf.html?item=/media/26877-nva-1157-deploy.pdf

- Soluzioni storage NetApp per Apache Spark

  "https://www.netapp.com/media/16864-tr-4570.pdf"

- Apache Hadoop su data fabric abilitato da NetApp

  "https://www.netapp.com/media/16877-tr-4529.pdf"

Ringraziamenti

- Paul Burland, Sales Rep, ANZ Victoria District Sales, NetApp
- Hoseb Dermanilian, Business Development Manager, NetApp
- Lee Dorrier, Director MPSG, NetApp
- David Thiessen, Systems Engineer, ANZ Victoria District se, NetApp

Cronologia delle versioni

Versione	Data	Cronologia delle versioni del documento
Versione 1.0	Gennaio 2018	Release iniziale
Versione 2.0	Ottobre 2021	Aggiornato con il caso di utilizzo n. 5: Accelerare il carico di lavoro di analisi
Versione 3.0	Novembre 2023	Dettagli NIPAM rimossi

Informazioni sul copyright

Copyright © 2024 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 COMMERCIABILITÀ 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 CONSEQUENZIALE (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.