



Raccolta di dati e monitoraggio delle performance dei carichi di lavoro

Active IQ Unified Manager 9.13

NetApp
December 18, 2023

Sommario

- Raccolta di dati e monitoraggio delle performance dei carichi di lavoro 1
 - Tipi di workload monitorati da Unified Manager 1
 - Valori di misurazione delle performance del carico di lavoro 2
 - Qual è la gamma di performance prevista 4
 - Come viene utilizzata la previsione di latenza nell'analisi delle performance 5
 - Come Unified Manager utilizza la latenza dei workload per identificare i problemi di performance 6
 - In che modo le operazioni del cluster possono influire sulla latenza del carico di lavoro 7
 - Monitoraggio delle performance delle configurazioni MetroCluster 8

Raccolta di dati e monitoraggio delle performance dei carichi di lavoro

Unified Manager raccoglie e analizza l'attività dei carichi di lavoro ogni 5 minuti per identificare gli eventi relativi alle performance e rileva le modifiche alla configurazione ogni 15 minuti. Conserva un massimo di 30 giorni di dati storici relativi alle performance e agli eventi di 5 minuti e utilizza questi dati per prevedere l'intervallo di latenza previsto per tutti i carichi di lavoro monitorati.

Unified Manager deve raccogliere un minimo di 3 giorni di attività del carico di lavoro prima che possa iniziare l'analisi e prima che la previsione di latenza per il tempo di risposta i/o possa essere visualizzata nella pagina workload Analysis e nella pagina Event Details. Durante la raccolta di questa attività, la previsione della latenza non visualizza tutte le modifiche che si verificano dall'attività del carico di lavoro. Dopo aver raccolto 3 giorni di attività, Unified Manager regola la previsione di latenza ogni 24 ore alle 12:00, per riflettere le modifiche dell'attività del carico di lavoro e stabilire una soglia di performance dinamica più precisa.

Durante i primi 4 giorni in cui Unified Manager sta monitorando un carico di lavoro, se sono trascorse più di 24 ore dall'ultima raccolta di dati, i grafici di latenza non visualizzano la previsione di latenza per quel carico di lavoro. Gli eventi rilevati prima dell'ultima raccolta sono ancora disponibili.



L'ora legale (DST) modifica l'ora del sistema, che modifica la previsione di latenza delle statistiche delle performance per i carichi di lavoro monitorati. Unified Manager inizia immediatamente a correggere la previsione di latenza, che richiede circa 15 giorni per essere completata. Durante questo periodo di tempo è possibile continuare a utilizzare Unified Manager, ma poiché Unified Manager utilizza la previsione della latenza per rilevare eventi dinamici, alcuni eventi potrebbero non essere precisi. Gli eventi rilevati prima del cambiamento di orario non vengono influenzati.

Tipi di workload monitorati da Unified Manager

È possibile utilizzare Unified Manager per monitorare le performance di due tipi di carichi di lavoro: Definiti dall'utente e definiti dal sistema.

• **workload definiti dall'utente**

Il throughput di i/o dalle applicazioni al cluster. Si tratta di processi coinvolti nelle richieste di lettura e scrittura. Un volume, LUN, condivisione NFS, condivisione SMB/CIFS e un carico di lavoro è un carico di lavoro definito dall'utente.



Unified Manager monitora solo l'attività del carico di lavoro sul cluster. Non esegue il monitoraggio delle applicazioni, dei client o dei percorsi tra le applicazioni e il cluster.

Se una o più delle seguenti affermazioni relative a un carico di lavoro sono vere, non possono essere monitorate da Unified Manager:

- Si tratta di una copia di protezione dei dati (DP) in modalità di sola lettura. (I volumi DP vengono monitorati per il traffico generato dall'utente).
- Si tratta di un clone dei dati offline.
- Si tratta di un volume mirrorato in una configurazione MetroCluster.

- **workload definiti dal sistema**

I processi interni legati all'efficienza dello storage, alla replica dei dati e allo stato del sistema, tra cui:

- Efficienza dello storage, come la deduplica
- Integrità del disco, che include la ricostruzione RAID, lo scrubbing del disco e così via
- Replica dei dati, ad esempio le copie SnapMirror
- Attività di gestione
- Integrità del file system, che include varie attività di WAFL
- Scanner del file system, come ad esempio la scansione WAFL
- Offload delle copie, ad esempio operazioni di efficienza dello storage offload da host VMware
- Stato del sistema, ad esempio spostamenti di volumi, compressione dei dati e così via
- Volumi non monitorati

I dati sulle performance per i carichi di lavoro definiti dal sistema vengono visualizzati nella GUI solo quando il componente del cluster utilizzato da questi carichi di lavoro è in conflitto. Ad esempio, non è possibile cercare il nome di un carico di lavoro definito dal sistema per visualizzarne i dati sulle prestazioni nella GUI.



Valori di misurazione delle performance del carico di lavoro

Unified Manager misura le performance dei carichi di lavoro su un cluster in base a valori statistici storici e previsti, che formano la previsione di latenza dei valori per i carichi di lavoro. Confronta i valori statistici effettivi del carico di lavoro con la previsione di latenza per determinare quando le performance del carico di lavoro sono troppo alte o troppo basse. Un carico di lavoro che non funziona come previsto attiva un evento di performance dinamica per la notifica.

Nella seguente illustrazione, il valore effettivo, in rosso, rappresenta le statistiche effettive delle performance nel periodo di tempo. Il valore effettivo ha superato la soglia di performance, che è il limite superiore della previsione di latenza. Il picco è il valore effettivo più alto nell'intervallo di tempo. La deviazione misura la variazione tra i valori previsti (la previsione) e i valori effettivi, mentre la deviazione di picco indica la variazione maggiore tra i valori attesi e quelli effettivi.



La seguente tabella elenca i valori di misurazione delle performance del carico di lavoro.

Misurazione	Descrizione
Attività	<p>La percentuale del limite di QoS utilizzato dai carichi di lavoro nel gruppo di criteri.</p> <div style="border: 1px solid #ccc; padding: 10px; margin-top: 10px;">  <p>Se Unified Manager rileva una modifica a un gruppo di criteri, ad esempio l'aggiunta o la rimozione di un volume o la modifica del limite di QoS, i valori effettivi e previsti potrebbero superare il 100% del limite impostato. Se un valore supera il 100% del limite impostato, viene visualizzato come >100%. Se un valore è inferiore all'1% del limite impostato, viene visualizzato come <1%.</p> </div>
Effettivo	<p>Il valore misurato delle performance in un momento specifico per un determinato carico di lavoro.</p>
Deviazione	<p>Il cambiamento tra i valori previsti e quelli effettivi. Si tratta del rapporto tra il valore effettivo meno il valore previsto e il valore superiore dell'intervallo previsto meno il valore previsto.</p> <div style="border: 1px solid #ccc; padding: 10px; margin-top: 10px;">  <p>Un valore di deviazione negativo indica che le performance del carico di lavoro sono inferiori al previsto, mentre un valore di deviazione positivo indica che le performance del carico di lavoro sono superiori al previsto.</p> </div>
Previsto	<p>I valori previsti si basano sull'analisi dei dati storici delle performance per un determinato carico di lavoro. Unified Manager analizza questi valori statistici per determinare l'intervallo previsto (previsione di latenza) dei valori.</p>
Previsione di latenza (intervallo previsto)	<p>La previsione di latenza è una previsione dei valori di performance superiori e inferiori previsti in un momento specifico. Per la latenza del carico di lavoro, i valori superiori costituiscono la soglia di performance. Quando il valore effettivo supera la soglia di performance, Unified Manager attiva un evento di performance dinamico.</p>
Picco	<p>Il valore massimo misurato in un periodo di tempo.</p>

Misurazione	Descrizione
Deviazione di picco	Il valore di deviazione massimo misurato in un periodo di tempo.
Profondità della coda	Il numero di richieste i/o in sospeso che sono in attesa sul componente di interconnessione.
Utilizzo	Per l'elaborazione di rete, l'elaborazione dei dati e i componenti aggregati, la percentuale di tempo occupato per completare le operazioni dei carichi di lavoro in un determinato periodo di tempo. Ad esempio, la percentuale di tempo in cui i componenti di elaborazione dati o di elaborazione di rete elaborano una richiesta di i/o o un aggregato deve soddisfare una richiesta di lettura o scrittura.
Throughput in scrittura	La quantità di throughput in scrittura, espressa in megabyte al secondo (MB/s), dai carichi di lavoro su un cluster locale al cluster partner in una configurazione MetroCluster.

Qual è la gamma di performance prevista

La previsione di latenza è una previsione dei valori di performance superiori e inferiori previsti in un momento specifico. Per la latenza del carico di lavoro, i valori superiori costituiscono la soglia di performance. Quando il valore effettivo supera la soglia di performance, Unified Manager attiva un evento di performance dinamico.

Ad esempio, durante le normali ore di lavoro tra le 9:00 e alle 17:00, la maggior parte dei dipendenti potrebbe controllare la posta elettronica tra le 9:00 e alle 10:30. L'aumento della domanda sui server di posta elettronica comporta un aumento dell'attività dei carichi di lavoro sullo storage back-end durante questo periodo. I dipendenti potrebbero notare tempi di risposta lenti dai propri clienti di posta elettronica.

Durante l'ora di pranzo tra le 12:00 e alle 13:00 e alla fine della giornata lavorativa dopo le 17:00, la maggior parte dei dipendenti è probabilmente lontana dai computer. La domanda sui server di posta elettronica in genere diminuisce, diminuendo anche la domanda sullo storage back-end. In alternativa, potrebbero essere pianificate operazioni di carico di lavoro, come backup dello storage o scansione virus, che iniziano dopo le 17:00 e aumentare l'attività sullo storage back-end.

Nel corso di diversi giorni, l'aumento e la diminuzione dell'attività del carico di lavoro determina l'intervallo previsto (previsione di latenza) dell'attività, con limiti superiori e inferiori per un carico di lavoro. Quando l'attività effettiva del carico di lavoro di un oggetto si trova al di fuori dei limiti superiori o inferiori e rimane al di fuori dei limiti per un certo periodo di tempo, ciò potrebbe indicare che l'oggetto è stato utilizzato in eccesso o sottoutilizzato.

Come si forma la previsione di latenza

Unified Manager deve raccogliere un minimo di 3 giorni di attività del carico di lavoro prima che possa iniziare l'analisi e prima che la previsione di latenza per il tempo di risposta i/o possa essere visualizzata nella GUI. La raccolta dati minima richiesta non tiene conto di tutte le modifiche che si verificano dall'attività del carico di

lavoro. Dopo aver raccolto i primi 3 giorni di attività, Unified Manager regola la previsione di latenza ogni 24 ore alle 12:00 riflettere le modifiche dell'attività del carico di lavoro e stabilire una soglia di performance dinamica più precisa.



L'ora legale (DST) modifica l'ora del sistema, che modifica la previsione di latenza delle statistiche delle performance per i carichi di lavoro monitorati. Unified Manager inizia immediatamente a correggere la previsione di latenza, che richiede circa 15 giorni per essere completata. Durante questo periodo di tempo è possibile continuare a utilizzare Unified Manager, ma poiché Unified Manager utilizza la previsione della latenza per rilevare eventi dinamici, alcuni eventi potrebbero non essere precisi. Gli eventi rilevati prima del cambiamento di orario non vengono influenzati.

Come viene utilizzata la previsione di latenza nell'analisi delle performance

Unified Manager utilizza la previsione della latenza per rappresentare la tipica attività di latenza i/o (tempo di risposta) per i carichi di lavoro monitorati. Ti avvisa quando la latenza effettiva per un carico di lavoro supera i limiti superiori della previsione di latenza, che attiva un evento di performance dinamica, in modo da poter analizzare il problema delle performance e intraprendere azioni correttive per risolverlo.

La previsione della latenza definisce la linea di base delle performance per il carico di lavoro. Nel corso del tempo, Unified Manager apprende dalle precedenti misurazioni delle performance per prevedere i livelli di performance e attività previsti per il carico di lavoro. Il limite superiore dell'intervallo previsto stabilisce la soglia di performance dinamica. Unified Manager utilizza la linea di base per determinare quando la latenza effettiva è superiore o inferiore a una soglia o al di fuori dei limiti previsti. Il confronto tra i valori effettivi e quelli previsti crea un profilo di performance per il carico di lavoro.

Quando la latenza effettiva per un carico di lavoro supera la soglia di performance dinamica, a causa di un conflitto su un componente del cluster, la latenza è elevata e il carico di lavoro funziona più lentamente del previsto. Anche le performance di altri carichi di lavoro che condividono gli stessi componenti del cluster potrebbero essere più lente del previsto.

Unified Manager analizza l'evento di superamento della soglia e determina se l'attività è un evento di performance. Se l'elevata attività del carico di lavoro rimane costante per un lungo periodo di tempo, ad esempio diverse ore, Unified Manager considera l'attività normale e regola dinamicamente la previsione di latenza per formare la nuova soglia di performance dinamica.

Alcuni carichi di lavoro potrebbero avere un'attività costantemente bassa, dove la previsione di latenza per la latenza non ha un elevato tasso di cambiamento nel tempo. Per ridurre al minimo il numero di eventi durante l'analisi degli eventi delle performance, Unified Manager attiva un evento solo per volumi a bassa attività le cui operazioni e latenze sono molto più elevate del previsto.



In questo esempio, la latenza per un volume ha una previsione di latenza, in grigio, di 3.5 millisecondi per operazione (ms/op) al minimo e di 5.5 ms/op al massimo. Se la latenza effettiva, in blu, aumenta improvvisamente a 10 ms/op, a causa di un picco intermittente nel traffico di rete o di un conflitto su un componente del cluster, supera la previsione di latenza e supera la soglia di performance dinamica.

Quando il traffico di rete è diminuito o il componente del cluster non è più in conflitto, la latenza ritorna entro la previsione di latenza. Se la latenza rimane pari o superiore a 10 ms/op per un lungo periodo di tempo, potrebbe essere necessario intraprendere un'azione correttiva per risolvere l'evento.

Come Unified Manager utilizza la latenza dei workload per identificare i problemi di performance

La latenza del carico di lavoro (tempo di risposta) è il tempo necessario a un volume di un cluster per rispondere alle richieste di i/o provenienti dalle applicazioni client. Unified Manager utilizza la latenza per rilevare e avvisare gli utenti in caso di eventi relativi alle performance.

Un'elevata latenza significa che le richieste dalle applicazioni a un volume su un cluster richiedono più tempo del solito. La causa dell'elevata latenza potrebbe essere il cluster stesso, a causa di conflitti su uno o più componenti del cluster. L'elevata latenza potrebbe essere causata anche da problemi esterni al cluster, come colli di bottiglia della rete, problemi con il client che ospita le applicazioni o problemi con le applicazioni stesse.

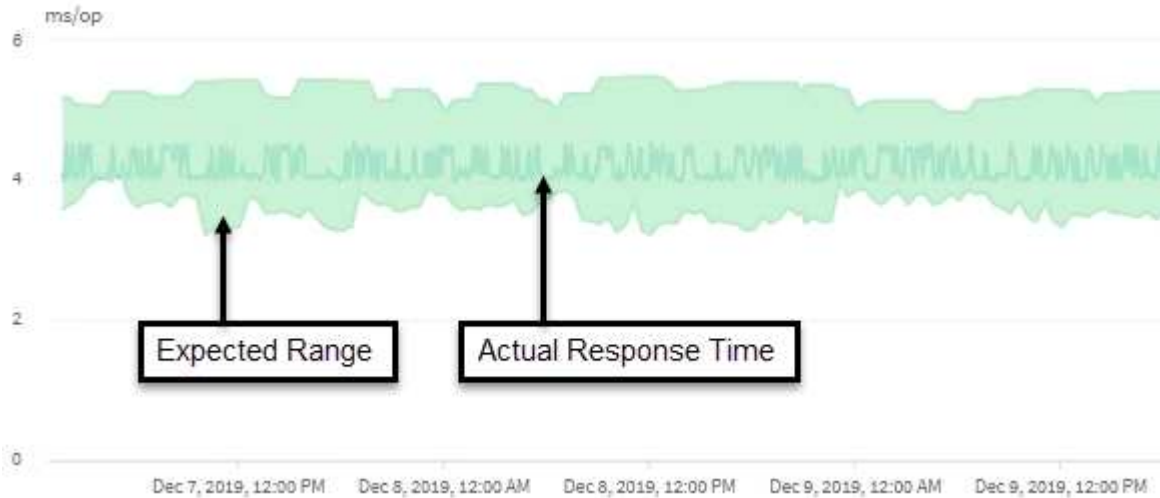


Unified Manager monitora solo l'attività del carico di lavoro sul cluster. Non esegue il monitoraggio delle applicazioni, dei client o dei percorsi tra le applicazioni e il cluster.

Anche le operazioni sul cluster, come la creazione di backup o l'esecuzione della deduplica, che aumentano la domanda di componenti del cluster condivisi da altri carichi di lavoro possono contribuire a un'elevata latenza. Se la latenza effettiva supera la soglia di performance dinamica dell'intervallo previsto (previsione di latenza), Unified Manager analizza l'evento per determinare se si tratta di un evento di performance che potrebbe essere necessario risolvere. La latenza viene misurata in millisecondi per operazione (ms/op).

Nel grafico latenza totale della pagina analisi del carico di lavoro, è possibile visualizzare un'analisi delle statistiche di latenza per vedere come l'attività dei singoli processi, come le richieste di lettura e scrittura, si confronta con le statistiche di latenza complessive. Il confronto consente di determinare quali operazioni hanno l'attività più elevata o se operazioni specifiche hanno attività anomale che influiscono sulla latenza di un volume. Quando si analizzano gli eventi delle performance, è possibile utilizzare le statistiche di latenza per

determinare se un evento è stato causato da un problema nel cluster. È inoltre possibile identificare le attività specifiche del carico di lavoro o i componenti del cluster coinvolti nell'evento.



Questo esempio mostra il grafico della latenza. L'attività del tempo di risposta effettivo (latenza) è una linea blu e la previsione di latenza (intervallo previsto) è verde.

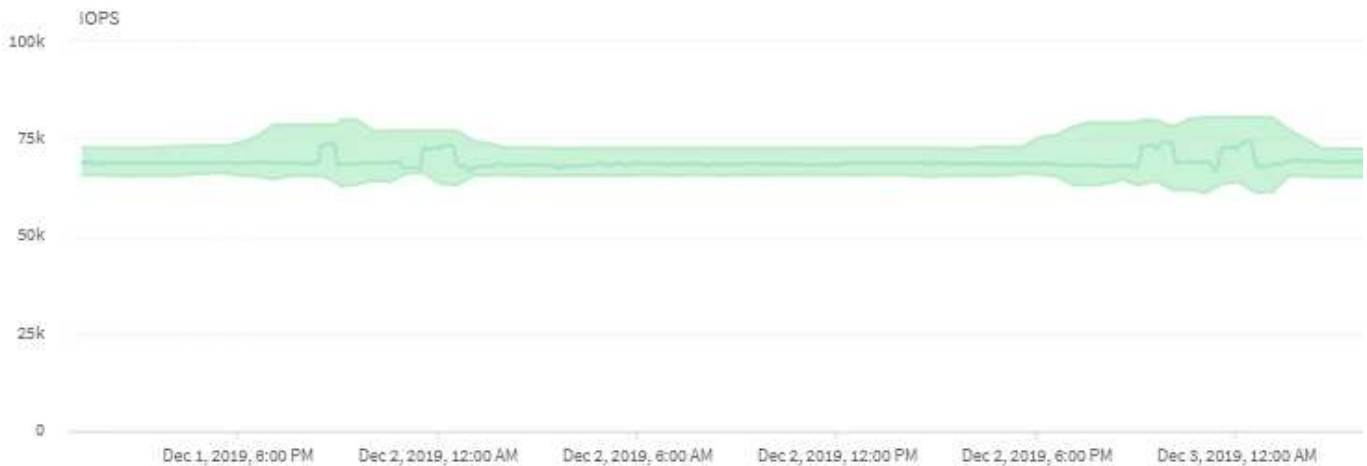


Se Unified Manager non è in grado di raccogliere i dati, la linea blu può presentare delle lacune. Ciò può verificarsi perché il cluster o il volume non era raggiungibile, Unified Manager è stato disattivato durante tale periodo o la raccolta richiede più tempo del periodo di raccolta di 5 minuti.

In che modo le operazioni del cluster possono influire sulla latenza del carico di lavoro

Le operazioni (IOPS) rappresentano l'attività di tutti i carichi di lavoro definiti dall'utente e dal sistema su un cluster. Le statistiche IOPS consentono di determinare se i processi del cluster, come l'esecuzione di backup o la deduplica, influiscono sulla latenza del carico di lavoro (tempo di risposta) o potrebbero aver causato o contribuito a un evento di performance.

Quando si analizzano gli eventi relativi alle performance, è possibile utilizzare le statistiche IOPS per determinare se un evento relativo alle performance è stato causato da un problema nel cluster. È possibile identificare le attività specifiche dei carichi di lavoro che potrebbero aver contribuito in maniera determinante all'evento delle performance. Gli IOPS vengono misurati in operazioni al secondo (Ops/sec).



Questo esempio mostra il grafico IOPS. Le statistiche effettive delle operazioni sono una linea blu e le previsioni IOPS delle statistiche delle operazioni sono verdi.



In alcuni casi in cui un cluster è sovraccarico, Unified Manager potrebbe visualizzare il messaggio `Data collection is taking too long on Cluster cluster_name`. Ciò significa che non sono state raccolte statistiche sufficienti per l'analisi di Unified Manager. È necessario ridurre le risorse utilizzate dal cluster in modo da poter raccogliere le statistiche.

Monitoraggio delle performance delle configurazioni MetroCluster

Unified Manager consente di monitorare il throughput di scrittura tra i cluster in una configurazione MetroCluster per identificare i carichi di lavoro con un'elevata quantità di throughput in scrittura.

Se questi carichi di lavoro dalle performance elevate causano elevati tempi di risposta i/o per altri volumi nel cluster locale, Unified Manager attiva gli eventi relativi alle performance per ricevere una notifica.



Unified Manager tratta i cluster in una configurazione MetroCluster come singoli cluster. Non distingue i cluster che sono partner o correlano il throughput di scrittura da ciascun cluster.

Quando un cluster locale in una configurazione MetroCluster esegue il mirroring dei dati nel cluster partner, i dati vengono scritti nella NVRAM e quindi trasferiti attraverso i collegamenti interswitch (ISL) agli aggregati remoti. Unified Manager analizza la NVRAM per identificare i carichi di lavoro il cui throughput di scrittura elevato sta utilizzando la NVRAM in eccesso, mettendo la NVRAM in conflitto.

I carichi di lavoro la cui deviazione nel tempo di risposta ha superato la soglia di performance sono denominati *vittime* e i carichi di lavoro la cui deviazione nel throughput di scrittura nella NVRAM è superiore al solito, causando il conflitto, sono denominati *bullies*. Poiché solo le richieste di scrittura vengono mirrorate al cluster partner, Unified Manager non analizza il throughput in lettura.

È possibile visualizzare il throughput di qualsiasi cluster in una configurazione MetroCluster analizzando i carichi di lavoro dei LUN e dei volumi corrispondenti dalle seguenti schermate. È possibile filtrare i risultati in base al cluster. Dal riquadro di navigazione di sinistra:

- Vista **Storage > Clusters > Performance: Tutti i Clusters**. Vedere

- Vista **Storage > Volumes > Performance: All Volumes.**
- **Storage > LUN > Performance: Vista di tutte le LUN.**
- **Analisi del carico di lavoro > tutti i carichi di lavoro**

Informazioni correlate

["Analisi e notifica degli eventi relativi alle performance"](#)

["Analisi degli eventi di performance per una configurazione MetroCluster"](#)

["Ruoli dei carichi di lavoro coinvolti in un evento di performance"](#)

["Identificazione dei carichi di lavoro delle vittime coinvolti in un evento di performance"](#)

["Identificazione dei carichi di lavoro ingombranti coinvolti in un evento di performance"](#)

["Identificazione dei carichi di lavoro di Shark coinvolti in un evento di performance"](#)

Informazioni sul copyright

Copyright © 2023 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 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.