Solutions de données de cloud hybride NetApp - Spark et Hadoop en fonction des cas d'utilisation clients

NetApp Solutions

NetApp
April 26, 2024

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

Sommaire

Solutions de données de cloud hybride NetApp - Spark et Hadoop en fonction des cas d'utilisation clients . . . 1
Tr-4657 : solutions de données de cloud hybride NetApp - Spark et Hadoop en fonction des cas d'utilisation clients . . . 1
Data Fabric optimisée par NetApp pour l'architecture Big Data . . . 2
Protection des données Hadoop et NetApp . . . 5
Présentation des cas d'utilisation de protection des données Hadoop . . . 6
Cas d'utilisation 1 : sauvegarde des données Hadoop . . . 7
Cas d'utilisation 2 : sauvegarde et reprise après incident du cloud vers les environnements sur site . . . 12
Cas d'utilisation 3 : activation de DevTest sur les données Hadoop existantes . . . 13
Cas d'utilisation 4 : protection des données et connectivité multicloud . . . 15
Cas d'utilisation 5 : accélération des workloads d'analytique . . . 16
Conclusion . . . 18

Solutions de données de cloud hybride NetApp - Spark et Hadoop en fonction des cas d'utilisation clients

Tr-4657 : solutions de données de cloud hybride NetApp - Spark et Hadoop en fonction des cas d'utilisation clients

Karthikeyan Nagalingam et Satish Thyagarajan, NetApp

Ce document présente les solutions de données de cloud hybride qui utilisent les systèmes de stockage NetApp AFF et FAS, NetApp Cloud Volumes ONTAP, les systèmes de stockage connectés NetApp et la technologie FlexClone pour Spark et Hadoop. Ces architectures de solution permettent aux clients de choisir la solution de protection des données adaptée à leur environnement. NetApp a conçu ces solutions en fonction des interactions avec les clients et de leurs utilisations. Ce document fournit les informations détaillées suivantes :

- Pourquoi nous devons nous doter d'une solution de protection des données pour les environnements Spark et Hadoop et les défis des clients.
- La Data Fabric optimisée par la vision NetApp, ainsi que ses éléments de base et ses services.
- Ces éléments de base peuvent être utilisés pour concevoir des flux de travail flexibles de protection des données.
- Les avantages et les inconvénients de plusieurs architectures basées sur des cas d'utilisation réels de clients. Chaque cas d'utilisation propose les composants suivants :
  - Scénarios clients
  - Besoins et défis
  - NetApp
  - Résumé des solutions

Pourquoi choisir la protection des données Hadoop ?

Dans un environnement Hadoop et Spark, les problèmes suivants doivent être résolus :

- **Pannes logicielles ou humaines.** une erreur humaine dans les mises à jour logicielles lors de l'exécution des opérations de données Hadoop peut entraîner un comportement défectueux qui peut entraîner des résultats inattendus de la tâche. Dans ce cas, nous devons protéger les données pour éviter les défaillances ou les résultats déraisonnables. Par exemple, suite à une mise à jour logicielle mal exécutée vers une application d'analyse de signal de trafic, une nouvelle fonction qui ne parvient pas à analyser correctement les données de signal de trafic sous forme de texte brut. Le logiciel analyse encore des formats JSON et d'autres formats de fichiers non texte, ce qui entraîne l'analyse en temps réel du contrôle du trafic, produisant des résultats de prédiction qui manquent de points de données. Cette situation peut entraîner des sorties défectueuses qui peuvent entraîner des accidents au niveau des signaux de circulation. La protection des données peut résoudre ce problème et permet de restaurer rapidement la version précédente de l'application de travail.
- **Taille et échelle.** la taille des données analytiques augmente jour après jour en raison du nombre toujours

croissant de sources de données et de volumes. Les médias sociaux, les applications mobiles, l'analytique et les plateformes de cloud computing sont les principales sources de données sur le marché actuel du Big Data, qui augmente très rapidement. Par conséquent, les données doivent être protégées pour assurer la précision des opérations.

- **La protection native des données de Hadoop.** Hadoop a une commande native pour protéger les données, mais cette commande n'assure pas la cohérence des données pendant la sauvegarde. Elle prend uniquement en charge la sauvegarde au niveau des répertoires. Les snapshots créés par Hadoop sont en lecture seule et ne peuvent pas être utilisés pour réutiliser directement les données de sauvegarde.

Défis posés par la protection des données pour les clients Hadoop et Spark

Le défi commun des clients d'Hadoop et d'Spark est de réduire le temps de sauvegarde et d'augmenter la fiabilité des sauvegardes sans nuire aux performances du cluster de production pendant la protection des données.

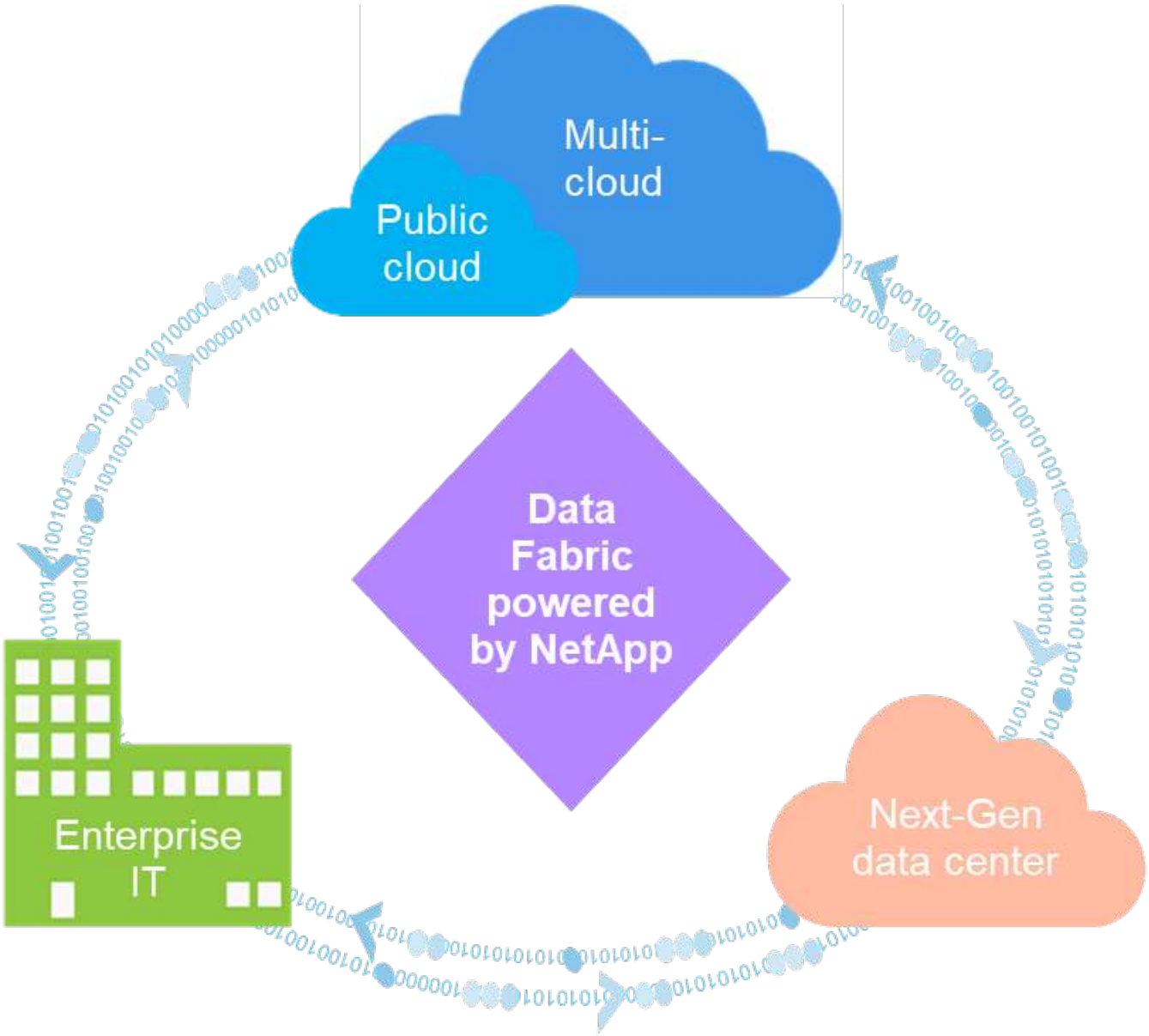
Les clients doivent également réduire les temps d'indisponibilité liés aux objectifs de point de récupération (RPO) et de durée de restauration (RTO), et contrôler leurs sites de reprise après incident sur site et dans le cloud pour assurer la continuité de l'activité. Ce contrôle vient généralement d'avoir des outils de gestion au niveau de l'entreprise.

Les environnements Hadoop et Spark se compliquent parce que non seulement le volume des données croît et exponentielle, mais que le débit de ces données augmente. Dans ce scénario, il est difficile de créer rapidement des environnements DevTest et QA efficaces et à jour à partir des données source. NetApp reconnaît ces défis et propose les solutions présentées dans ce livre blanc.

Data Fabric optimisée par NetApp pour l'architecture Big Data

La Data Fabric optimisée par NetApp simplifie et intègre la gestion des données entre les environnements cloud et sur site afin d'accélérer la transformation digitale.

La Data Fabric optimisée par NetApp offre des services et des applications de gestion des données intégrés et cohérents (éléments de base) pour la visibilité, l'exploitation, l'accès, le contrôle ainsi que la protection et la sécurité des données, comme l'illustre la figure ci-dessous.



Cas d’utilisation prouvés pour Data Fabric

La Data Fabric optimisée par NetApp offre les neuf cas d’utilisation suivants :

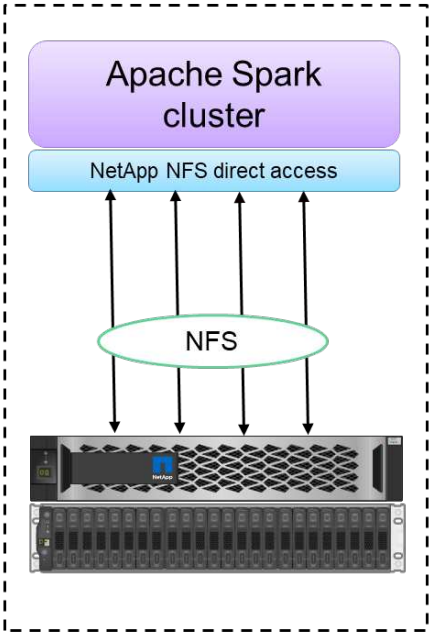
- Accélérez les workloads d’analytique
- Accélérez la transformation DevOps
- Créer une infrastructure d’hébergement cloud
- Intégrez des services de données cloud
- Protégez et sécurisez les données
- Optimiser les données non structurées
- Améliorez l’efficacité du data Center
- Exploitez et contrôlez vos données
- Simplifier et automatiser

Ce document couvre deux des neuf cas d'utilisation (ainsi que leurs solutions) :

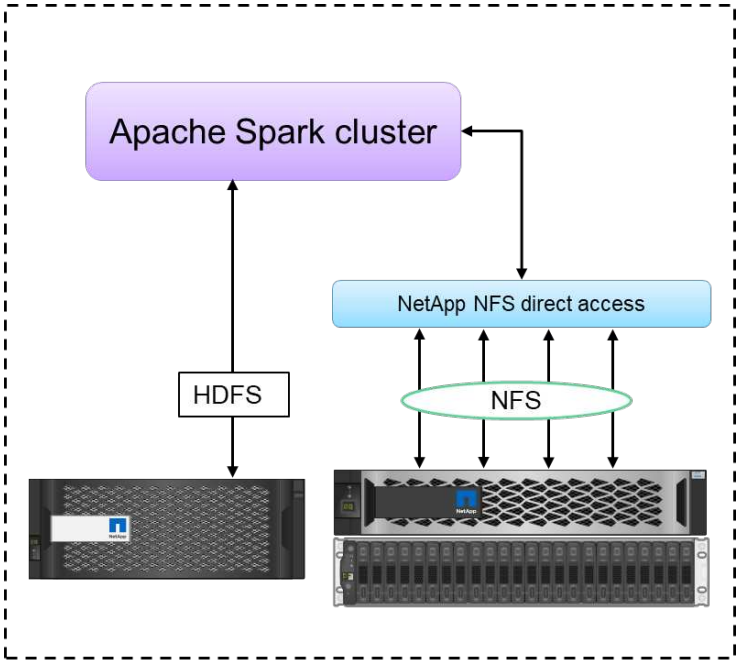
- Accélérez les workloads d'analytique
- Protégez et sécurisez les données

Accès direct NetApp NFS

Le système NetApp NFS permet aux clients d'exécuter des tâches d'analytique Big Data sur leurs données NFSv3 ou NFSv4 existantes ou nouvelles, sans déplacer ni copier les données. Elle empêche plusieurs copies de données et n'a plus besoin de synchroniser les données avec une source. Par exemple, dans le secteur financier, le transfert des données d'un endroit à un autre doit respecter les obligations légales, ce qui n'est pas une tâche facile. Dans ce scénario, l'accès direct NetApp NFS analyse les données financières à partir de leur emplacement d'origine. L'autre avantage clé est que l'accès direct NetApp NFS simplifie la protection des données Hadoop grâce aux commandes Hadoop natives et permet d'activer des workflows de protection des données à partir de la gamme riche de solutions NetApp de gestion des données.



Configuration 1: NFS as primary storage



Configuration 2: HDFS and NFS in single Spark cluster

L'accès direct NetApp NFS propose deux types d'options de déploiement pour les clusters Hadoop/Spark :

- Par défaut, les clusters Hadoop/Spark utilisent le système Hadoop Distributed File System (HDFS) pour le stockage des données et le système de fichiers par défaut. L'accès direct NetApp NFS peut remplacer le système HDFS par défaut par un stockage NFS comme système de fichiers par défaut, permettant ainsi des opérations d'analytique directe sur les données NFS.
- Dans une autre option de déploiement, l'accès direct NetApp NFS prend en charge la configuration de NFS en tant que stockage supplémentaire avec HDFS dans un seul cluster Hadoop/Spark. Dans ce cas, le client peut partager des données via les exports NFS et y accéder depuis le même cluster, ainsi que des données HDFS.

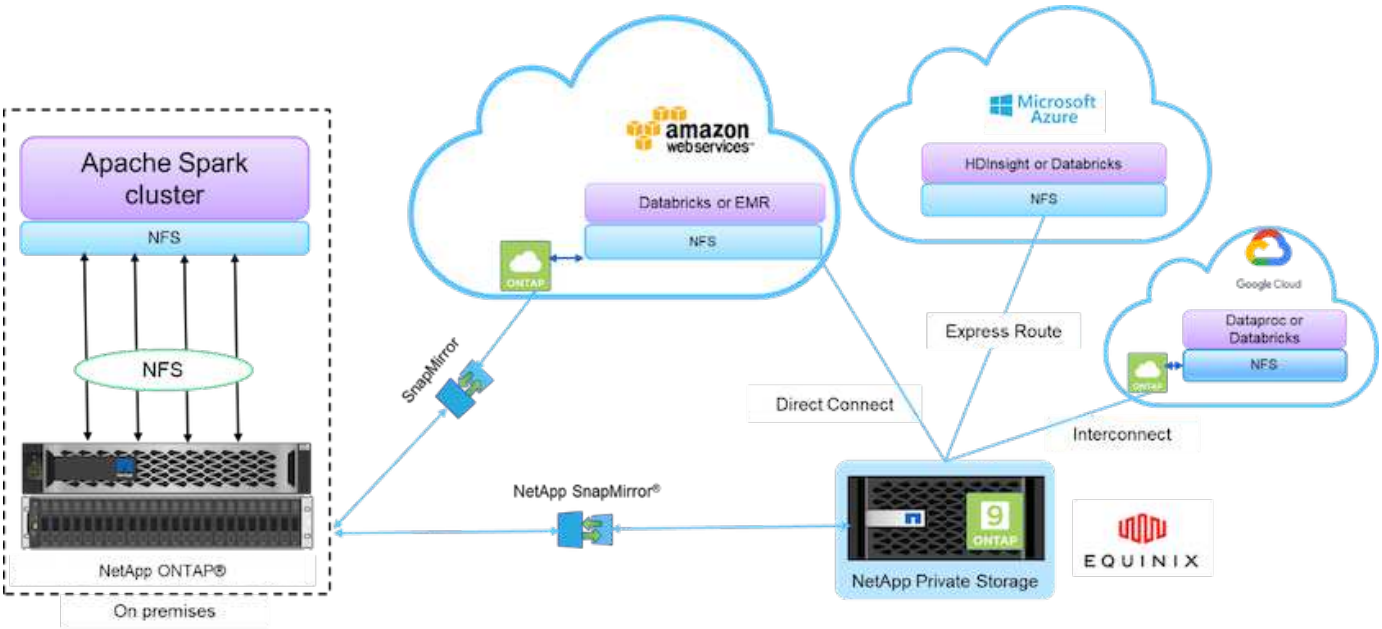
Voici les principaux avantages de l'accès direct NetApp NFS :

- Analyse les données de leur emplacement actuel, ce qui empêche la tâche fastidieuse de transférer des données analytiques vers une infrastructure Hadoop telle que HDFS.

- Réduit le nombre de répliques de trois à un.
- Permet aux utilisateurs de découpler les ressources de calcul et de stockage.
- Protège les données grâce aux fonctionnalités avancées de gestion d'ONTAP.
- Est certifiée avec la plateforme de données Hortonworks.
- Déploiements dans l'analytique hybride
- Réduit le temps de sauvegarde grâce à la fonctionnalité multithread dynamique.

Les éléments de base du Big Data

La Data Fabric optimisée par NetApp intègre des services et des applications de gestion des données (éléments de base) pour l'accès, le contrôle, la protection et la sécurité des données, comme l'illustre la figure ci-dessous.



Les éléments de base de la figure ci-dessus sont les suivants :

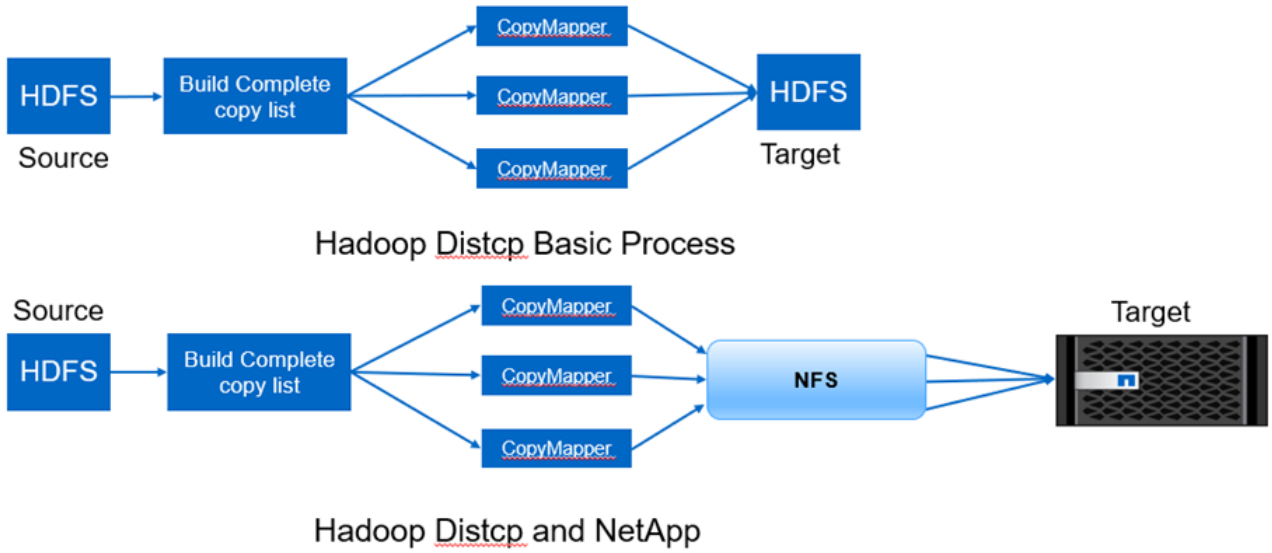
- **NetApp NFS Direct Access.** fournit les derniers clusters Hadoop et Spark avec un accès direct aux volumes NFS NetApp sans configuration logicielle ni pilote supplémentaire.
- **NetApp Cloud Volumes ONTAP et Cloud volumes Services.** stockage Software-defined basé sur ONTAP s'exécutant dans Amazon Web Services (AWS) ou Azure NetApp Files (ANF) dans les services cloud Microsoft Azure.
- **Technologie NetApp SnapMirror**. Fournit des fonctionnalités de protection des données entre les instances sur site et ONTAP Cloud ou NPS.
- **Fournisseurs de services cloud.** ces fournisseurs incluent AWS, Microsoft Azure, Google Cloud et IBM Cloud.
- **PaaS.** services d'analytique cloud tels qu'Amazon Elastic MapReduce (EMR) et Databricks dans AWS, ainsi que Microsoft Azure HDInsight et Azure Databricks.

Protection des données Hadoop et NetApp

Hadoop DistCp est un outil natif utilisé pour les copies intercluster et intracluster de

grande taille. Le processus de distribution Hadoop présenté dans la figure ci-dessous illustre un workflow de sauvegarde standard utilisant des outils natifs Hadoop tels que MapReduce pour copier les données Hadoop d'une source HDFS vers une cible correspondante.

L'accès direct NetApp NFS permet aux clients de définir NFS en tant que destination cible pour l'outil Hadoop DistCp afin de copier les données à partir d'une source HDFS vers un partage NFS via MapReduce. L'accès direct NFS NetApp joue le rôle de pilote NFS pour l'outil DistCp.



Présentation des cas d'utilisation de protection des données Hadoop

Cette section décrit de manière détaillée les cas d'utilisation de protection des données qui constituent le centre de ce document. Les sections restantes fournissent des informations détaillées sur chaque cas d'utilisation, telles que le problème d'un client (scénario), les exigences, les défis et les solutions.

Cas d'utilisation 1 : sauvegarde des données Hadoop

Pour ce cas d'utilisation, le volume NFS de NetApp a permis à une grande institution financière de réduire la durée de la fenêtre de sauvegarde, qui est passée de plus de 24 heures à seulement quelques heures.

Cas d'utilisation 2 : sauvegarde et reprise après incident du cloud vers les environnements sur site

En utilisant Data Fabric optimisé par NetApp comme éléments de base, une grande entreprise de diffusion a pu satisfaire aux exigences de sauvegarde instantanée des données cloud dans son data Center sur site, selon les différents modes de transfert de données, comme à la demande, Ou en fonction de la charge du cluster Hadoop/Spark.

Cas d’utilisation 3 : activation de DevTest sur les données Hadoop existantes

Les solutions NetApp ont permis à un distributeur de musique en ligne de créer rapidement plusieurs clusters Hadoop peu encombrants dans différentes succursales pour créer des rapports et exécuter des tâches quotidiennes DevTest à l’aide de règles planifiées.

Cas d’utilisation 4 : protection des données et connectivité multicloud

Un grand fournisseur de services a utilisé la Data Fabric optimisée par NetApp pour fournir des fonctionnalités d’analytique multicloud à ses clients à partir de différentes instances de cloud.

Cas d’utilisation 5 : accélération des workloads d’analytique

L’une des plus grandes banques d’investissement et de services financiers a eu recours à la solution de stockage NAS de NetApp pour réduire le temps d’attente des E/S et accélérer sa plateforme d’analytique financière quantitative.

Cas d’utilisation 1 : sauvegarde des données Hadoop

Scénario

Dans ce scénario, le client dispose d’un grand référentiel Hadoop sur site et souhaite le sauvegarder à des fins de reprise après incident. Toutefois, la solution de sauvegarde actuelle du client est coûteuse et nécessite une longue fenêtre de sauvegarde de plus de 24 heures.

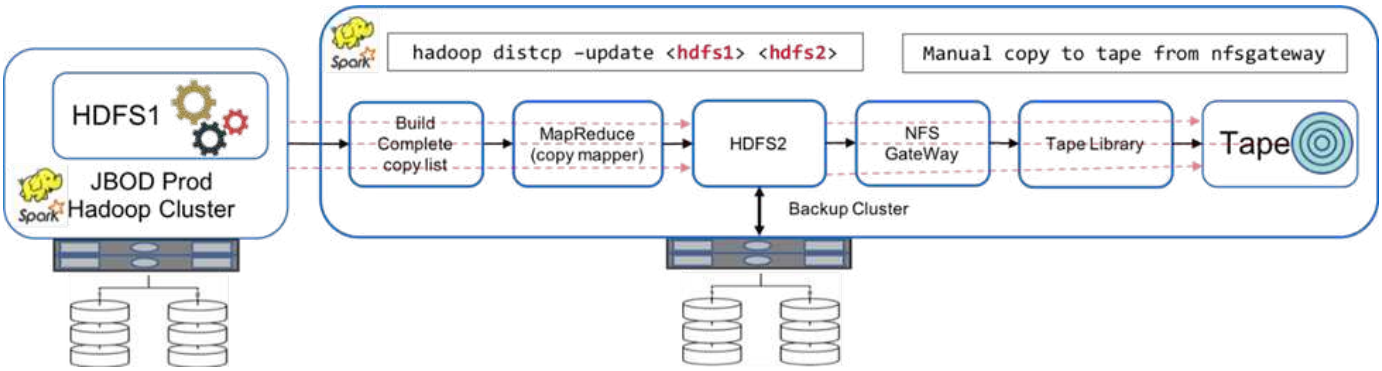
Besoins et défis

Voici les principaux défis et exigences de cette utilisation :

- Rétrocompatibilité logicielle :
  - La solution de sauvegarde proposée doit être compatible avec les versions logicielles actuellement utilisées dans le cluster Hadoop de production.
- Pour respecter les contrats de niveau de service (SLA) signés, la solution alternative proposée doit permettre d’atteindre des RPO et des RTO très faibles.
- La sauvegarde créée par la solution de sauvegarde NetApp peut être utilisée dans le cluster Hadoop intégré localement au data Center ainsi que dans le cluster Hadoop exécuté sur le site de reprise sur incident sur le site distant.
- La solution proposée doit être économique.
- La solution proposée doit réduire l’impact sur les performances des tâches d’analytique en production actuellement en cours d’exécution pendant les durées de sauvegarde.

Solution de sauvegarde existante du client x

La figure ci-dessous présente la solution de sauvegarde native Hadoop de départ.



Les données de production sont protégées sur bande via le cluster de sauvegarde intermédiaire :

- Les données HDFS1 sont copiées vers HDFS2 en exécutant le `hadoop distcp -update <hdfs1> <hdfs2>` commande.
- Le cluster de sauvegarde fait office de passerelle NFS et les données sont copiées manuellement sur bande via Linux `cp` commande via la bibliothèque de bandes.

Voici quelques avantages offerts par la solution de sauvegarde Hadoop native :

- La solution repose sur les commandes natives Hadoop qui évite à l'utilisateur d'apprendre à maîtriser de nouvelles procédures.
- La solution exploite l'architecture et le matériel standard.

Les inconvénients de la solution de sauvegarde native Hadoop initiale sont les suivants :

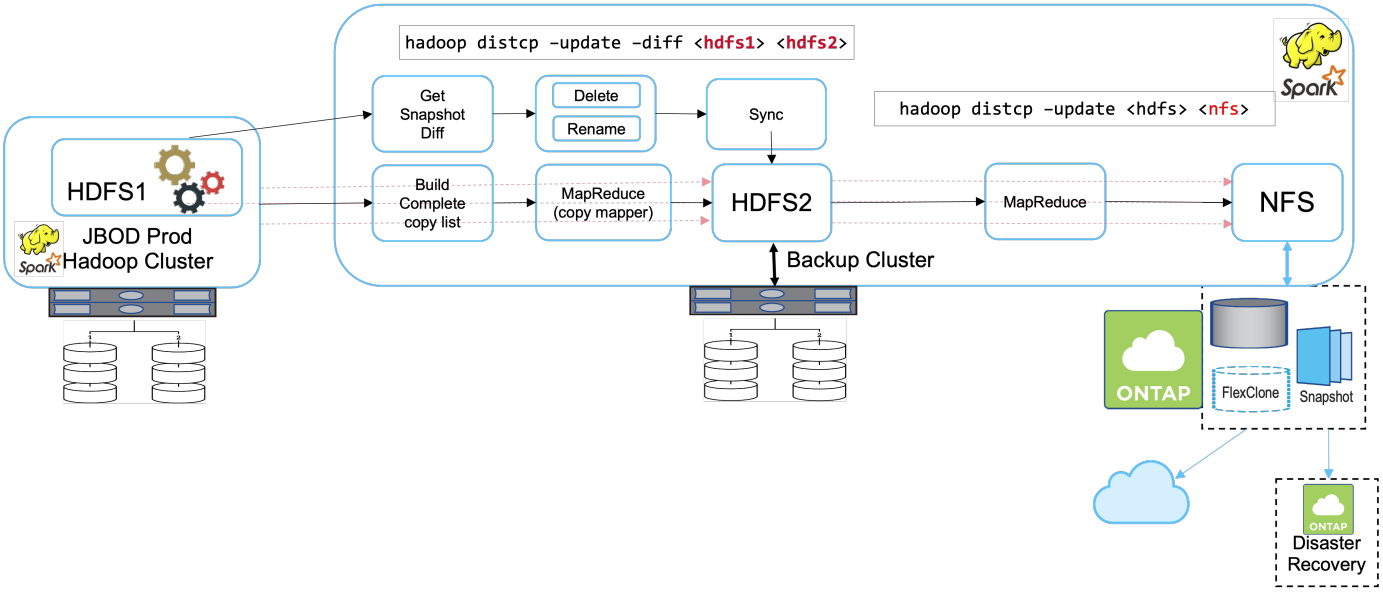
- La longue fenêtre de sauvegarde dépasse 24 heures, ce qui rend les données de production vulnérables.
- Une dégradation significative des performances du cluster durant les heures de sauvegarde
- La copie sur bande est un processus manuel.
- La solution de sauvegarde est coûteuse en termes de matériel et d'heures de travail nécessaires aux processus manuels.

Solutions de sauvegarde

En fonction de ces défis et exigences, et en tenant compte du système de sauvegarde existant, trois solutions de sauvegarde possibles ont été suggérées. Les sous-sections suivantes décrivent chacune de ces trois solutions de sauvegarde, intitulées solution A à la solution C.

Solution A

Dans la solution A, le cluster Hadoop de sauvegarde envoie les sauvegardes secondaires aux systèmes de stockage NetApp NFS, ce qui élimine les besoins en bande, comme illustré dans la figure ci-dessous.



Les tâches détaillées de la solution A incluent :

- Le cluster Hadoop de production dispose des données d'analytique de l'entreprise dans le système HDFS qui requiert une protection.
- Le cluster Hadoop de sauvegarde avec HDFS sert d'emplacement intermédiaire pour les données. Une simple concaténation de disques durs (JBOD) permet de stocker le système HDFS à la fois dans les clusters Hadoop de production et de sauvegarde.
- Protéger les données de production Hadoop est protégé du cluster HDFS de production vers le cluster de sauvegarde HDFS en exécutant le `Hadoop distcp -update -diff <hdfs1> <hdfs2>` commande.

> Le snapshot Hadoop protège les données de la production vers le cluster Hadoop de sauvegarde.

- Le contrôleur de stockage NetApp ONTAP fournit un volume exporté NFS, provisionné vers le cluster Hadoop de sauvegarde.
- En exécutant le `Hadoop distcp` Commande tirant parti de MapReduce et de plusieurs mappeurs, les données d'analytique sont protégées du cluster Hadoop de sauvegarde vers NFS.

  Une fois les données stockées dans NFS sur le système de stockage NetApp, les technologies NetApp Snapshot, SnapRestore et FlexClone sont utilisées pour sauvegarder, restaurer et dupliquer les données Hadoop selon les besoins.

> Les données Hadoop peuvent être protégées sur le cloud et les emplacements de reprise après incident grâce à la technologie SnapMirror.

La solution A présente plusieurs avantages :

- Les données de production Hadoop sont protégées du cluster de sauvegarde.
- Les données HDFS sont protégées par le biais de NFS qui assure la protection des données dans les sites cloud et de reprise après incident.
- Améliore les performances en redirigeant les opérations de sauvegarde vers le cluster de sauvegarde.
- Élimine les opérations manuelles de bande

- Fonctions de gestion d'entreprise via les outils NetApp
- Elle ne nécessite que des changements minimes de l'environnement existant.
- Est une solution économique.

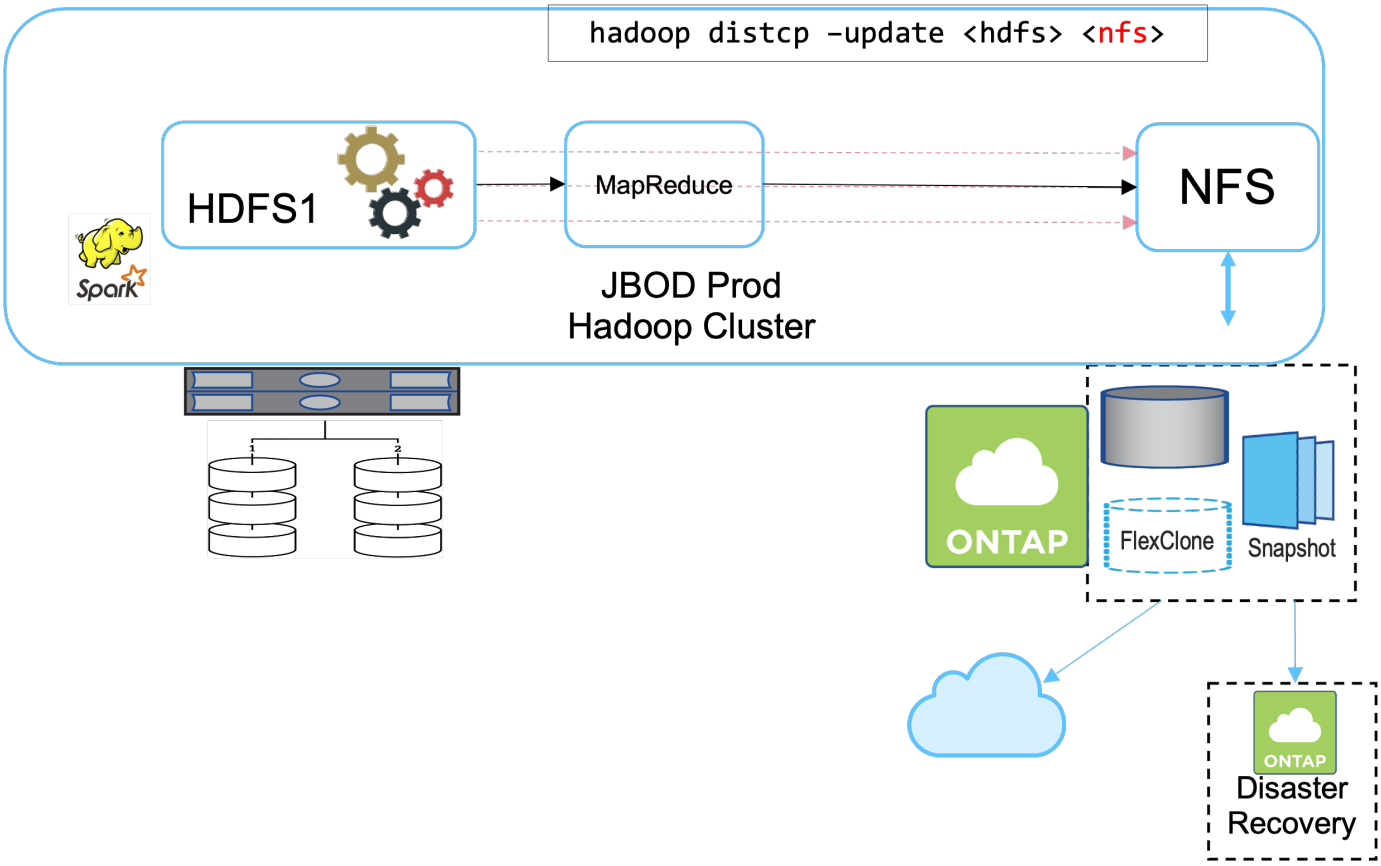
L'inconvénient de cette solution est qu'elle nécessite un cluster de sauvegarde et des mappeurs supplémentaires pour améliorer les performances.

Le client a récemment déployé la solution A en raison de sa simplicité, de son coût et de ses performances globales.

Avec cette solution, les disques SAN de ONTAP peuvent être utilisés à la place d'un JBOD. Cette option décharge la charge du stockage du cluster de sauvegarde vers ONTAP, mais l'inconvénient est que des commutateurs de structure SAN sont nécessaires.

Solution B

La solution B ajoute un volume NFS au cluster Hadoop de production, ce qui évite d'avoir à utiliser le cluster Hadoop de sauvegarde, comme illustré dans la figure ci-dessous.



Les tâches détaillées de la solution B incluent :

- Le contrôleur de stockage NetApp ONTAP provisionne l'exportation NFS vers le cluster Hadoop de production.

  Hadoop natif `hadoop distcp` Protège les données Hadoop du cluster de production HDFS vers NFS.

- Une fois les données stockées dans NFS sur le système de stockage NetApp, les technologies Snapshot, SnapRestore et FlexClone sont utilisées pour sauvegarder, restaurer et dupliquer les données Hadoop

selon les besoins.

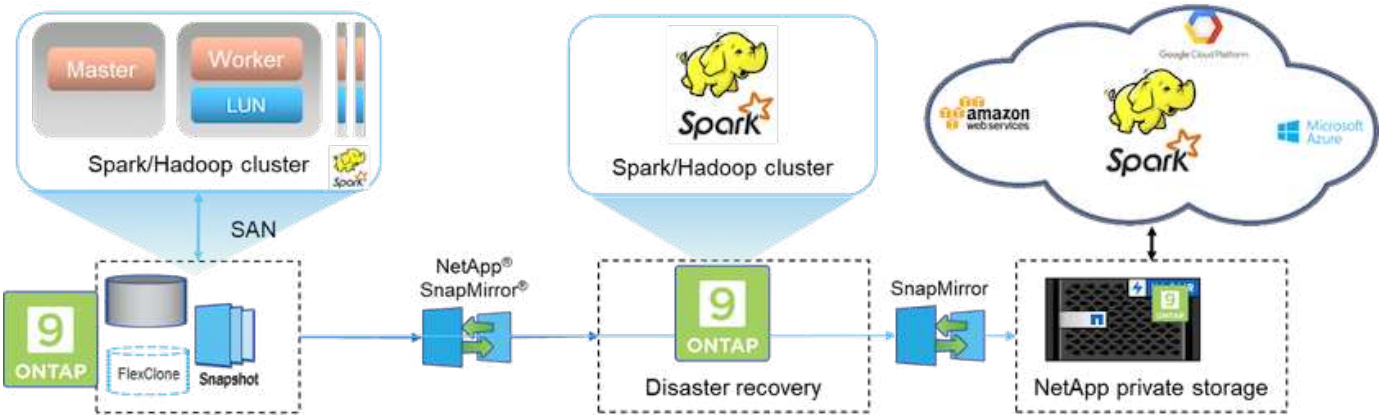
La solution B présente plusieurs avantages :

- Le cluster de production est légèrement modifié pour la solution de sauvegarde, ce qui simplifie l'implémentation et réduit les coûts d'infrastructure supplémentaires.
- Aucun cluster de sauvegarde n'est requis pour l'opération de sauvegarde.
- Dans la conversion des données NFS, les données de production HDFS sont protégées.
- La solution permet de gérer l'entreprise à l'aide des outils NetApp.

L'inconvénient de cette solution est qu'elle est implémentée dans le cluster de production, ce qui peut ajouter des tâches d'administrateur supplémentaires dans le cluster de production.

Solution C

Dans la solution C, les volumes SAN NetApp sont directement provisionnés vers le cluster de production Hadoop pour le stockage HDFS, comme illustré dans la figure ci-dessous.



Les étapes détaillées de la solution C incluent :

- Le stockage SAN NetApp ONTAP est provisionné au niveau du cluster Hadoop de production pour le stockage des données HDFS.
- Les technologies NetApp Snapshot et SnapMirror sont utilisées pour sauvegarder les données HDFS à partir du cluster Hadoop de production.
- La sauvegarde n'a aucun impact sur les performances de production du cluster Hadoop/Spark au cours du processus de sauvegarde de copie Snapshot, car elle se trouve au niveau de la couche de stockage.

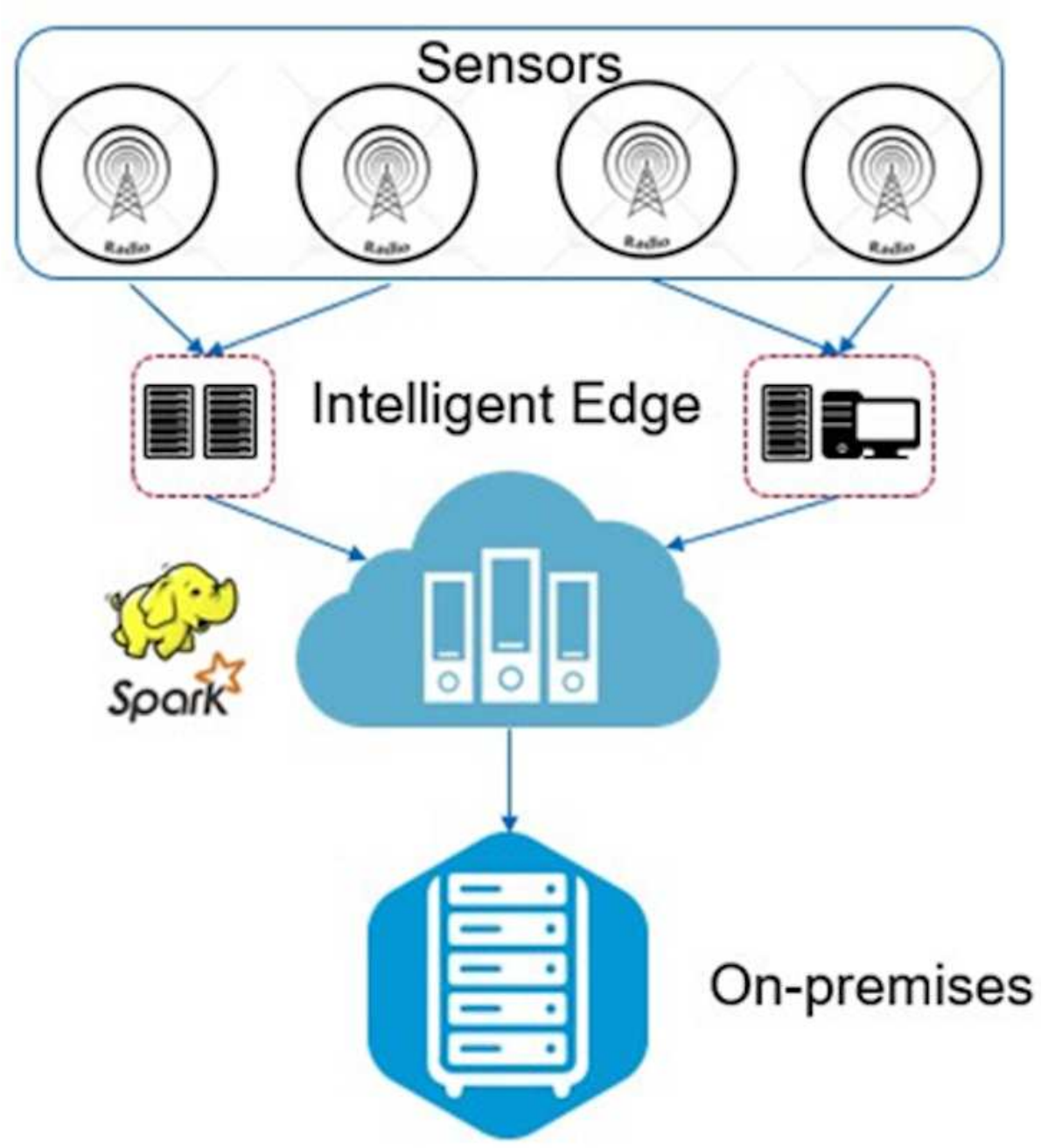
> La technologie Snapshot effectue des sauvegardes en quelques secondes, quelle que soit la taille des données.

La solution C présente plusieurs avantages :

- La technologie Snapshot permet de créer des sauvegardes compactes.
- Fonctions de gestion d'entreprise via les outils NetApp

Cas d’utilisation 2 : sauvegarde et reprise après incident du cloud vers les environnements sur site

Ce cas d’utilisation est basé sur un client de radiodiffusion qui doit sauvegarder les données d’analytique cloud vers son data Center sur site, comme l’illustre la figure ci-dessous.



Scénario

Dans ce scénario, les données de capteurs de l’IoT sont ingérées sur le cloud et analysées à l’aide d’un cluster Apache Spark open source dans AWS. Une nécessité consiste à sauvegarder les données traitées depuis le cloud vers une infrastructure sur site.

Besoins et défis

Voici les principaux défis et exigences de cette utilisation :

- L'activation de la protection des données ne doit pas affecter les performances du cluster Spark/Hadoop de production dans le cloud.
- Les données de Cloud Sensor doivent être déplacées et protégées sur site de manière efficace et sécurisée.
- Flexibilité du transfert des données depuis le cloud vers un environnement sur site dans des conditions différentes, telles que sur demande, instantanément et pendant les faibles temps de chargement du cluster.

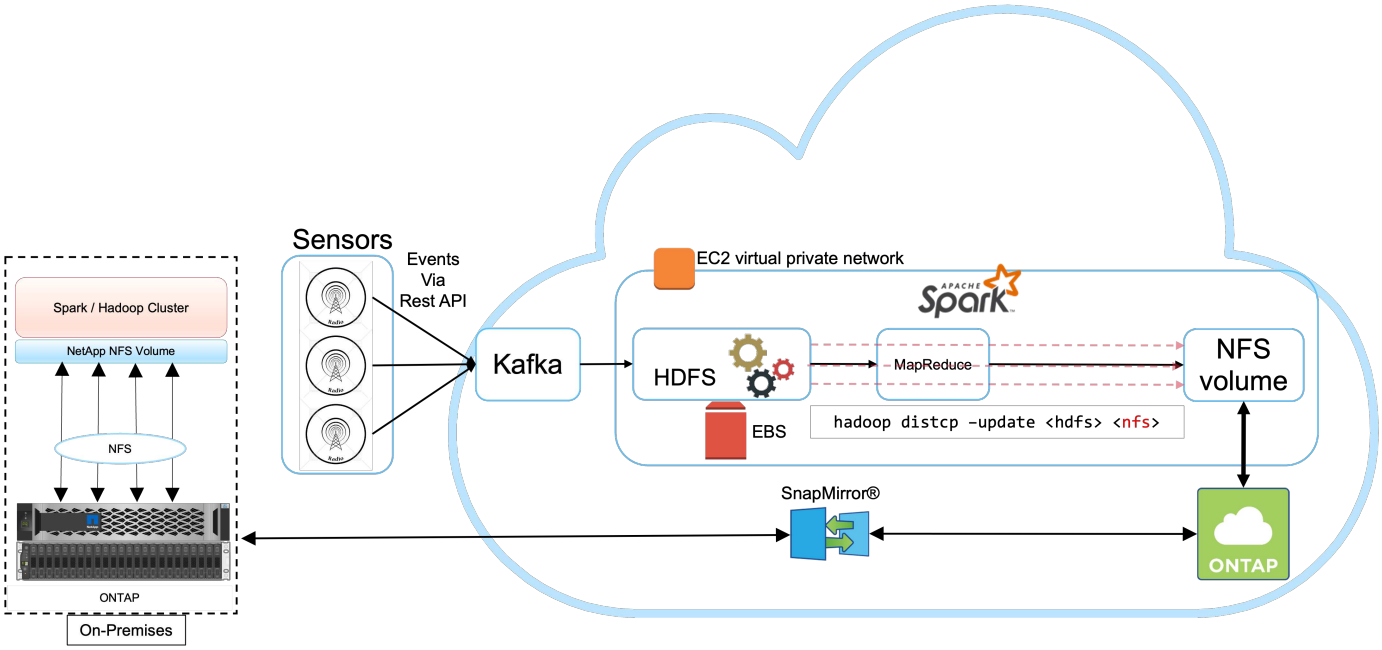
Solution

Le client utilise AWS Elastic Block Store (EBS) pour le stockage HDFS du cluster Spark afin de recevoir et d'ingérer les données provenant de capteurs distants via Kafka. Par conséquent, le stockage HDFS agit comme source des données de sauvegarde.

Pour y répondre, NetApp ONTAP Cloud est déployé dans AWS, et un partage NFS est créé pour servir de cible de sauvegarde pour le cluster Spark/Hadoop.

Une fois le partage NFS créé, copiez les données du stockage HDFS EBS dans le partage NFS ONTAP. Une fois que les données sont stockées dans le système NFS dans ONTAP Cloud, la technologie SnapMirror peut être utilisée pour mettre en miroir les données du cloud vers du stockage sur site selon les besoins, de manière sécurisée et efficace.

Cette image présente la sauvegarde et la reprise d'activité entre le cloud et la solution sur site.



Cas d'utilisation 3 : activation de DevTest sur les données Hadoop existantes

Dans ce cas d'utilisation, l'entreprise a besoin de créer rapidement et efficacement de

nouveaux clusters Hadoop/Spark basés sur un cluster Hadoop existant contenant un grand nombre de données d'analytique pour DevTest et la création de rapports dans le même data Center et sur des sites distants.

Scénario

Dans ce scénario, plusieurs clusters Spark/Hadoop sont conçus à partir d'une implémentation d'un data Lake Hadoop volumineux sur site et dans des sites de reprise après incident.

Besoins et défis

Voici les principaux défis et exigences de cette utilisation :

- Créez plusieurs clusters Hadoop pour le DevTest, l'assurance qualité ou tout autre objectif nécessitant l'accès aux mêmes données de production. Le défi ici est de cloner un cluster Hadoop de très grande taille plusieurs fois instantanément et de façon très compacte.
- Synchronisation des données Hadoop avec les équipes de DevTest et de création de rapports pour une efficacité opérationnelle optimale
- Distribution des données Hadoop à l'aide des mêmes identifiants sur les nouveaux clusters et environnements de production.
- Utilisez des règles planifiées pour créer efficacement des clusters d'assurance qualité sans affecter le cluster de production.

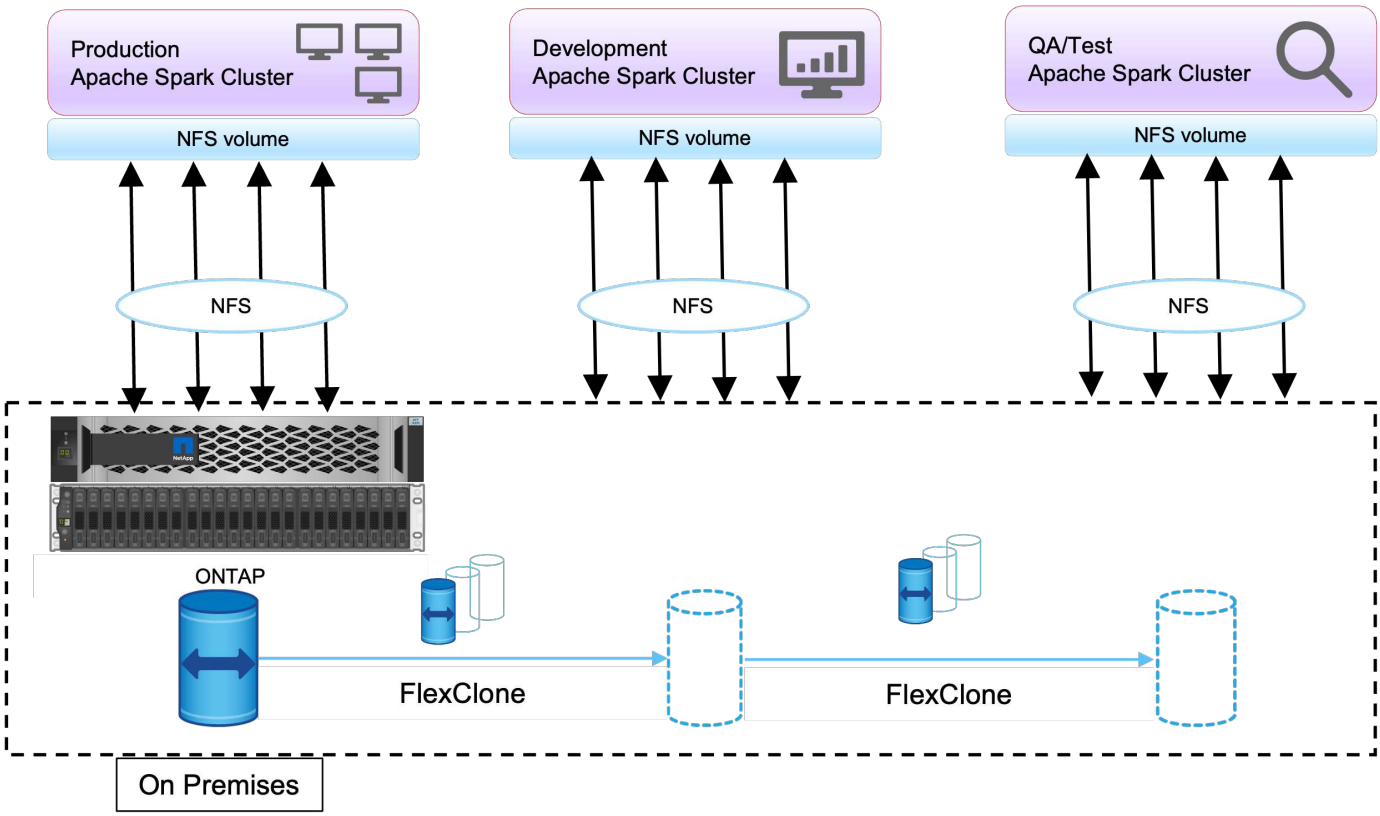
Solution

La technologie FlexClone est utilisée pour répondre aux exigences décrites précédemment. La technologie FlexClone constitue la copie de lecture/écriture d'une copie Snapshot. Il lit les données de la copie Snapshot parent et consomme uniquement de l'espace supplémentaire pour les blocs nouveaux/modifiés. Elle est rapide et compacte.

Tout d'abord, une copie Snapshot du cluster existant a été créée à l'aide d'un groupe de cohérence NetApp.

Copies snapshot dans NetApp System Manager ou l'invite d'administrateur du stockage. Les copies Snapshot de groupe de cohérence sont des copies Snapshot de groupe cohérentes au niveau des applications et le volume FlexClone est créé à partir des copies Snapshot de groupe de cohérence. Il est utile de mentionner qu'un volume FlexClone hérite des règles d'exportation NFS du volume parent. Une fois la copie Snapshot créée, un nouveau cluster Hadoop doit être installé à des fins de DevTest et de création de rapports, comme illustré dans la figure ci-dessous. Le volume NFS cloné depuis le nouveau cluster Hadoop accède aux données NFS.

Cette image représente le cluster Hadoop pour DevTest.



Cas d'utilisation 4 : protection des données et connectivité multicloud

Ce cas d'utilisation s'pertinentes pour les partenaires de services clouds chargés d'assurer une connectivité multicloud pour les données d'analytique Big Data des clients.

Scénario

Dans ce scénario, des données IoT reçues dans AWS de différentes sources sont stockées dans un emplacement central dans NPS. Le stockage NPS est connecté aux clusters Spark/Hadoop situés dans AWS et Azure, ce qui permet aux applications d'analytique Big Data exécutées dans plusieurs clouds d'accéder aux mêmes données.

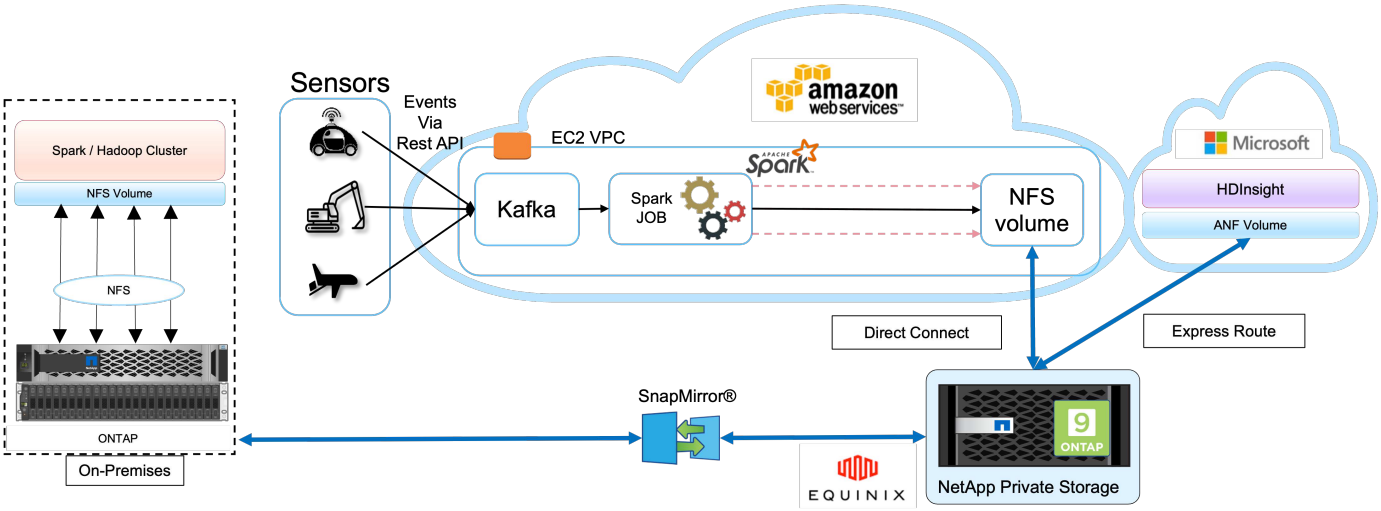
Besoins et défis

Voici les principaux défis et exigences de cette utilisation :

- Les clients veulent exécuter les tâches d'analytique sur les mêmes données à l'aide de plusieurs clouds.
- Les données doivent être reçues de différentes sources, telles que les données sur site et dans le cloud, par le biais de différents capteurs et concentrateurs.
- La solution doit être efficace et économique.
- Le principal défi consiste à concevoir une solution efficace et économique qui propose des services d'analytique hybride entre les environnements sur site et cloud.

Solution

Cette image illustre la solution de protection des données et de connectivité multicloud.



Comme illustré dans la figure ci-dessus, les données provenant des capteurs sont transmises et ingérées sur le cluster AWS Spark via Kafka. Les données sont stockées dans un partage NFS hébergé sur NPS, situé en dehors du fournisseur cloud au sein d'un data Center Equinix. Étant donné que le NPS NetApp est connecté à Amazon AWS et Microsoft Azure via des connexions Direct Connect et Express route, les clients peuvent accéder aux données NFS à partir des clusters analytiques Amazon et AWS. Cette approche résout l'analytique cloud de plusieurs hyperscalers.

Par conséquent, comme le stockage sur site et NPS exécute le logiciel ONTAP, SnapMirror peut mettre en miroir les données NPS dans le cluster sur site, pour une analytique de cloud hybride entre les environnements sur site et clouds.

Pour optimiser les performances, NetApp recommande généralement d'utiliser plusieurs interfaces réseau et des routes directes/express pour accéder aux données à partir des instances cloud.

Cas d'utilisation 5 : accélération des workloads d'analytique

Dans ce scénario, une plateforme d'analyse des services financiers et des banques d'investissement a été modernisée avec la solution de stockage NFS NetApp afin d'améliorer considérablement l'analyse des risques d'investissement et des produits dérivés de sa gestion des actifs et de sa business unit quantitative.

Scénario

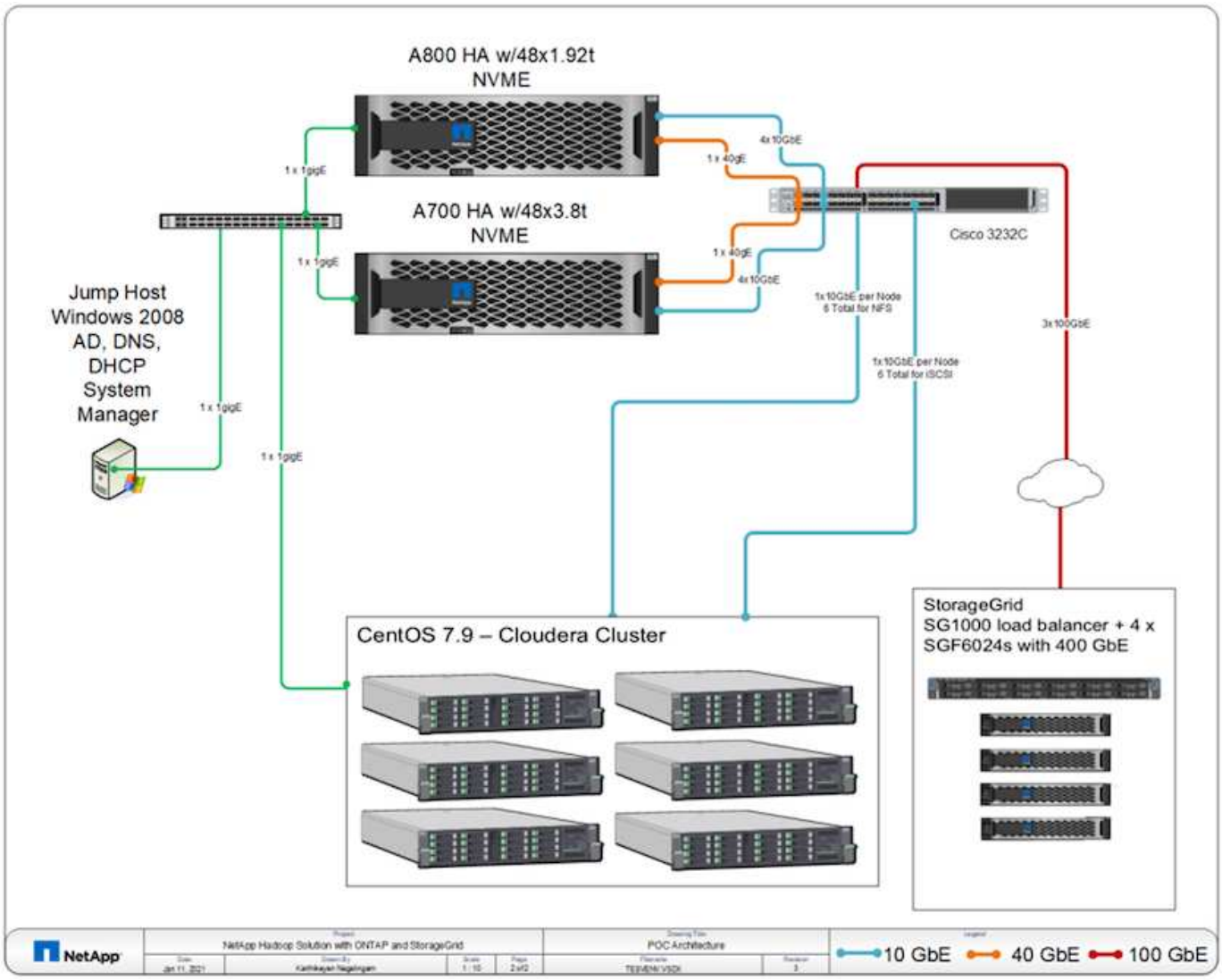
Dans l'environnement existant du client, l'infrastructure Hadoop utilisée pour la plateforme d'analytique exploite le stockage interne des serveurs Hadoop. En raison de la nature exclusive de l'environnement JBOD, de nombreux clients internes de l'entreprise n'ont pas pu tirer parti de leur modèle quantitatif Monte Carlo, une simulation qui s'appuie sur les échantillons récurrents de données en temps réel. La capacité sous-optimale de comprendre les effets de l'incertitude dans les mouvements du marché était de ne pas être favorable à l'unité commerciale de gestion quantitative des actifs.

Besoins et défis

L'unité commerciale quantitative de la banque voulait une méthode de prévision efficace pour obtenir des prévisions précises et opportunes. Elle a ainsi reconnu la nécessité de moderniser l'infrastructure, de réduire le temps d'attente des E/S existantes et d'améliorer les performances des applications d'analytique telles que Hadoop et Spark pour simuler efficacement des modèles d'investissement, mesurer les gains potentiels et analyser les risques.

Solution

Le client disposait d'une solution JBOD pour sa solution Spark existante. NetApp ONTAP, NetApp StorageGRID et MiniO Gateway to NFS ont ensuite été utilisés pour réduire le temps d'attente des E/S du groupe financier quantitatif de la banque qui exécute des simulations et des analyses sur des modèles d'investissement permettant d'évaluer les gains et risques potentiels. Cette image présente la solution Spark avec du stockage NetApp.



Comme illustré ci-dessus, les systèmes AFF A800, A700 et StorageGRID ont été déployés pour accéder aux fichiers parquet via les protocoles NFS et S3 dans un cluster Hadoop à six nœuds avec Spark, et LES services de métadonnées YARN et Hive pour les opérations d'analytique.

Dans l'ancien environnement de client, une solution DAS (Direct-Attached Storage) avait des inconvénients :

faire évoluer indépendamment les ressources de calcul et de stockage. Avec la solution NetApp ONTAP pour Spark, l'unité commerciale dédiée aux analyses financières de la banque a pu dissocier le stockage des ressources de calcul et ajouter de manière transparente les ressources d'infrastructure en fonction des besoins.

Grâce à ONTAP avec NFS, les processeurs du serveur de calcul étaient presque pleinement utilisés pour les tâches Spark SQL et le temps d'attente en E/S a été réduit de près de 70 %, ce qui optimise la puissance de calcul et les performances des charges de travail Spark. Par la suite, l'augmentation de l'utilisation du CPU a également permis au client d'exploiter des GPU, tels que GPUDirect, pour poursuivre la modernisation de la plateforme. En outre, StorageGRID offre une option de stockage économique pour les charges de travail Spark et la passerelle MiniO permet un accès sécurisé aux données NFS via le protocole S3. Pour les données dans le cloud, NetApp recommande Cloud Volumes ONTAP, Azure NetApp Files et NetApp Cloud Volumes Service.

Conclusion

Cette section présente un récapitulatif des utilisations et des solutions proposées par NetApp pour répondre à différents besoins de protection de données Hadoop. Grâce à la Data Fabric optimisée par NetApp, les clients peuvent :

- Choisissez les solutions de protection des données adaptées en tirant parti des fonctionnalités avancées de gestion des données de NetApp et en intégrant les flux de travail natifs Hadoop.
- Réduire la durée des fenêtres de sauvegarde des clusters Hadoop de près de 70 %.
- Éliminez les effets des sauvegardes de clusters Hadoop sur les performances.
- Protégez les données multicloud et assurez l'accès aux données provenant de différents fournisseurs cloud simultanément à une seule source d'analytique.
- Création de copies de clusters Hadoop rapides et compactes à l'aide de la technologie FlexClone

Où trouver des informations complémentaires

Pour en savoir plus sur les informations fournies dans ce document, consultez ces documents et/ou sites web :

- Solutions NetApp d'analytique Big Data

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

- Charge de travail Apache Spark avec le stockage NetApp

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

- Solutions de stockage NetApp pour Apache Spark

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

- Apache Hadoop sur l'environnement Data Fabric optimisé par NetApp

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

Remerciements

- Paul Burland, Ingénieur commercial, ventes pour le district ANZ Victoria District, NetApp

- Hoseb Dermanilian, responsable du développement commercial, NetApp
- Lee Dorrier, Directeur MPSG, NetApp
- David Thiessen, Ingénieur système, ANZ Victoria District se, NetApp

Historique des versions

Version	Date	Historique des versions du document
Version 1.0	Janvier 2018	Version initiale
Version 2.0	Octobre 2021	Mise à jour avec le cas d'utilisation n° 5 : accélération de la charge de travail analytique
Version 3.0	Novembre 2023	Suppression des détails NIPAM

Informations sur le copyright

Copyright © 2024 NetApp, Inc. Tous droits réservés. Imprimé aux États-Unis. Aucune partie de ce document protégé par copyright ne peut être reproduite sous quelque forme que ce soit ou selon quelque méthode que ce soit (graphique, électronique ou mécanique, notamment par photocopie, enregistrement ou stockage dans un système de récupération électronique) sans l'autorisation écrite préalable du détenteur du droit de copyright.

Les logiciels dérivés des éléments NetApp protégés par copyright sont soumis à la licence et à l'avis de non-responsabilité suivants :

CE LOGICIEL EST FOURNI PAR NETAPP « EN L'ÉTAT » ET SANS GARANTIES EXPRESSES OU TACITES, Y COMPRIS LES GARANTIES TACITES DE QUALITÉ MARCHANDE ET D'ADÉQUATION À UN USAGE PARTICULIER, QUI SONT EXCLUES PAR LES PRÉSENTES. EN AUCUN CAS NETAPP NE SERA TENU POUR RESPONSABLE DE DOMMAGES DIRECTS, INDIRECTS, ACCESSOIRES, PARTICULIERS OU EXEMPLAIRES (Y COMPRIS L'ACHAT DE BIENS ET DE SERVICES DE SUBSTITUTION, LA PERTE DE JOUISSANCE, DE DONNÉES OU DE PROFITS, OU L'INTERRUPTION D'ACTIVITÉ), QUELLES QU'EN SOIENT LA CAUSE ET LA DOCTRINE DE RESPONSABILITÉ, QU'IL S'AGISSE DE RESPONSABILITÉ CONTRACTUELLE, STRICTE OU DÉLICTUELLE (Y COMPRIS LA NÉGLIGENCE OU AUTRE) DÉCOULANT DE L'UTILISATION DE CE LOGICIEL, MÊME SI LA SOCIÉTÉ A ÉTÉ INFORMÉE DE LA POSSIBILITÉ DE TELS DOMMAGES.

NetApp se réserve le droit de modifier les produits décrits dans le présent document à tout moment et sans préavis. NetApp décline toute responsabilité découlant de l'utilisation des produits décrits dans le présent document, sauf accord explicite écrit de NetApp. L'utilisation ou l'achat de ce produit ne concède pas de licence dans le cadre de droits de brevet, de droits de marque commerciale ou de tout autre droit de propriété intellectuelle de NetApp.

Le produit décrit dans ce manuel peut être protégé par un ou plusieurs brevets américains, étrangers ou par une demande en attente.

LÉGENDE DE RESTRICTION DES DROITS : L'utilisation, la duplication ou la divulgation par le gouvernement sont sujettes aux restrictions énoncées dans le sous-paragraphe (b)(3) de la clause Rights in Technical Data-Noncommercial Items du DFARS 252.227-7013 (février 2014) et du FAR 52.227-19 (décembre 2007).

Les données contenues dans les présentes se rapportent à un produit et/ou service commercial (tel que défini par la clause FAR 2.101). Il s'agit de données propriétaires de NetApp, Inc. Toutes les données techniques et tous les logiciels fournis par NetApp en vertu du présent Accord sont à caractère commercial et ont été exclusivement développés à l'aide de fonds privés. Le gouvernement des États-Unis dispose d'une licence limitée irrévocable, non exclusive, non cessible, non transférable et mondiale. Cette licence lui permet d'utiliser uniquement les données relatives au contrat du gouvernement des États-Unis d'après lequel les données lui ont été fournies ou celles qui sont nécessaires à son exécution. Sauf dispositions contraires énoncées dans les présentes, l'utilisation, la divulgation, la reproduction, la modification, l'exécution, l'affichage des données sont interdits sans avoir obtenu le consentement écrit préalable de NetApp, Inc. Les droits de licences du Département de la Défense du gouvernement des États-Unis se limitent aux droits identifiés par la clause 252.227-7015(b) du DFARS (février 2014).

Informations sur les marques commerciales

NETAPP, le logo NETAPP et les marques citées sur le site http://www.netapp.com/TM sont des marques déposées ou des marques commerciales de NetApp, Inc. Les autres noms de marques et de produits sont des marques commerciales de leurs propriétaires respectifs.