Soluciones de datos para el cloud híbrido de NetApp: Spark y Hadoop en función de casos prácticos de clientes

NetApp Solutions

NetApp
April 26, 2024

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

Tabla de contenidos

Soluciones de datos para el cloud híbrido de NetApp: Spark y Hadoop en función de casos prácticos de clientes . . . 1
TR-4657: Soluciones de datos para el cloud híbrido de NetApp: Spark y Hadoop en función de casos prácticos de clientes . . . 1
Data Fabric con la tecnología de NetApp para la arquitectura de Big Data . . . 2
Protección de datos para Hadoop y NetApp . . . 6
Descripción general de los casos prácticos de protección de datos de Hadoop . . . 6
Caso de uso 1: Backup de datos de Hadoop . . . 7
Caso de uso 2: Backup y recuperación ante desastres del cloud a las instalaciones . . . 12
Caso de uso 3: Activación de DevTest en datos de Hadoop existentes . . . 13
Caso de uso 4: Protección de datos y conectividad multicloud . . . 15
Caso de uso 5: Acelerar las cargas de trabajo analíticas . . . 16
Conclusión . . . 18

Soluciones de datos para el cloud híbrido de NetApp: Spark y Hadoop en función de casos prácticos de clientes

TR-4657: Soluciones de datos para el cloud híbrido de NetApp: Spark y Hadoop en función de casos prácticos de clientes

Karthikeyan Nagalingam y Sathish Thyagarajan, NetApp

Este documento describe las soluciones de datos en el cloud híbrido mediante los sistemas de almacenamiento AFF y FAS de NetApp, Cloud Volumes ONTAP de NetApp, el almacenamiento conectado de NetApp y la tecnología FlexClone de NetApp para Spark y Hadoop. Estas arquitecturas de soluciones permiten a los clientes elegir una solución de protección de datos adecuada para su entorno. NetApp ha diseñado estas soluciones basándose en la interacción con los clientes y sus casos prácticos empresariales. Este documento proporciona información detallada siguiente:

- Por qué necesitamos protección de datos para entornos Spark y Hadoop y nuestros retos a los clientes.
- Data Fabric impulsado por la visión de NetApp y sus elementos básicos y servicios.
- Cómo se pueden utilizar estos elementos básicos para diseñar flujos de trabajo de protección de datos flexibles.
- Ventajas e inconvenientes de varias arquitecturas basadas en casos de uso de clientes reales. Cada caso de uso proporciona los siguientes componentes:
  - Situaciones de clientes
  - Requisitos y retos
  - NetApp
  - Resumen de las soluciones

¿Por qué la protección de datos de Hadoop?

En un entorno Hadoop y Spark, se deben solucionar las siguientes preocupaciones:

- **Fallos de software o humanos.** error humano en las actualizaciones de software mientras se realizan operaciones de datos de Hadoop puede provocar un comportamiento defectuoso que puede causar resultados inesperados del trabajo. En tal caso, necesitamos proteger los datos para evitar fallas o resultados irrazonables. Por ejemplo, como resultado de una actualización de software mal ejecutada en una aplicación de análisis de señales de tráfico, una nueva característica que no analiza correctamente los datos de señales de tráfico en forma de texto sin formato. El software sigue analizando JSON y otros formatos de archivos que no sean de texto, lo que provoca que el sistema de análisis del control de tráfico en tiempo real produzca resultados de predicción que faltan puntos de datos. Esta situación puede causar salidas defectuosas que podrían provocar accidentes en las señales de tráfico. La protección de datos puede resolver este problema proporcionando la capacidad de volver rápidamente a la versión anterior de la aplicación de trabajo.
- **Tamaño y escala.** el tamaño de los datos de análisis aumenta día a día debido al número cada vez mayor

de fuentes de datos y volumen. Las redes sociales, las aplicaciones móviles, los análisis de datos y las plataformas de cloud computing son las principales fuentes de datos del mercado actual de Big Data, que aumenta con gran rapidez y, por tanto, los datos deben protegerse para garantizar operaciones de datos precisas.

- **Protección de datos nativos de Hadoop.** Hadoop tiene un comando nativo para proteger los datos, pero este comando no proporciona coherencia de datos durante la copia de seguridad. Solo admite backups a nivel de directorio. Las copias Snapshot que crea Hadoop son de solo lectura y no se pueden utilizar para reutilizar los datos de backup directamente.

Retos de la protección de datos para los clientes de Hadoop y Spark

Un reto común para los clientes de Hadoop y Spark es reducir el tiempo de backup y aumentar la fiabilidad de éste sin que se vea afectado negativamente el rendimiento del clúster de producción durante la protección de datos.

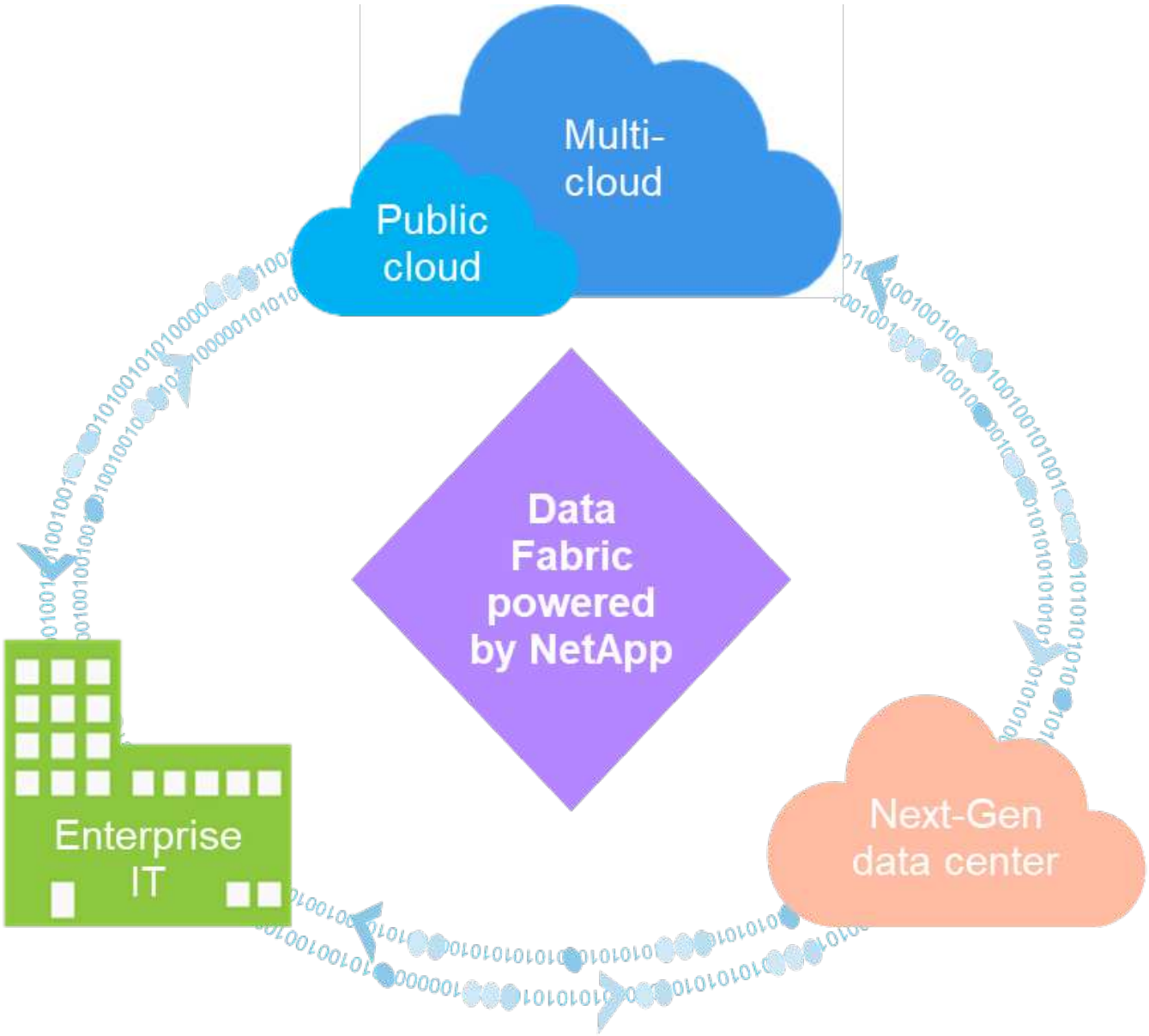
Los clientes también deben minimizar el objetivo de punto de recuperación (RPO) y el tiempo de inactividad del objetivo de tiempo de recuperación (RTO) y controlar sus sitios de recuperación ante desastres internos y basados en el cloud para obtener una continuidad empresarial óptima. Normalmente, este control se debe a que dispone de herramientas de gestión de nivel empresarial.

Los entornos Hadoop y Spark son complicados porque no solo el volumen de datos es enorme y está creciendo, sino que la velocidad a la que llegan estos datos está aumentando. Esta situación dificulta la creación rápida de entornos eficientes y actualizados de DevTest y QA a partir de los datos de origen. NetApp reconoce estos retos y ofrece las soluciones presentadas en este documento.

Data Fabric con la tecnología de NetApp para la arquitectura de Big Data

El Data Fabric con tecnología de NetApp simplifica e integra la gestión de datos en entornos cloud y en las instalaciones para acelerar la transformación digital.

El Data Fabric con tecnología de NetApp proporciona servicios y aplicaciones de gestión de datos consistentes e integrados (elementos básicos) para ofrecer visibilidad, información, acceso, control, protección y seguridad, como se muestra en la figura siguiente.



Casos prácticos de clientes de Data Fabric demostrados

El Data Fabric con tecnología de NetApp ofrece los siguientes nueve casos prácticos contrastados a los clientes:

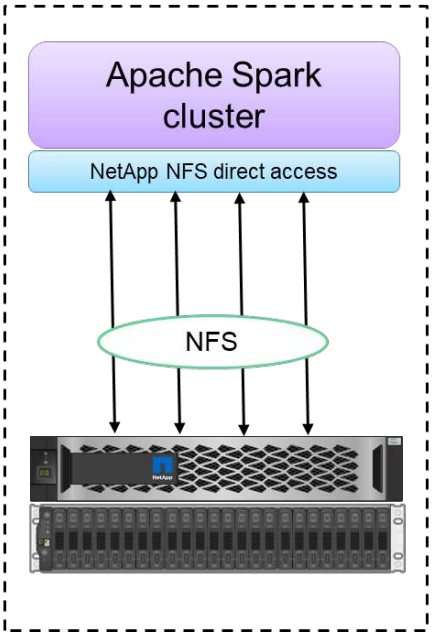
- Acelere las cargas de trabajo de análisis
- Acelere la transformación DevOps
- Cree una infraestructura de host en el cloud
- Integre los servicios de datos en el cloud
- Proteja y asegure sus datos
- Optimice los datos no estructurados
- Aumente la eficiencia del centro de datos
- Proporcione información y control de los datos
- Simplifique y automatice

En este documento se tratan dos de los nueve casos prácticos (junto con sus soluciones):

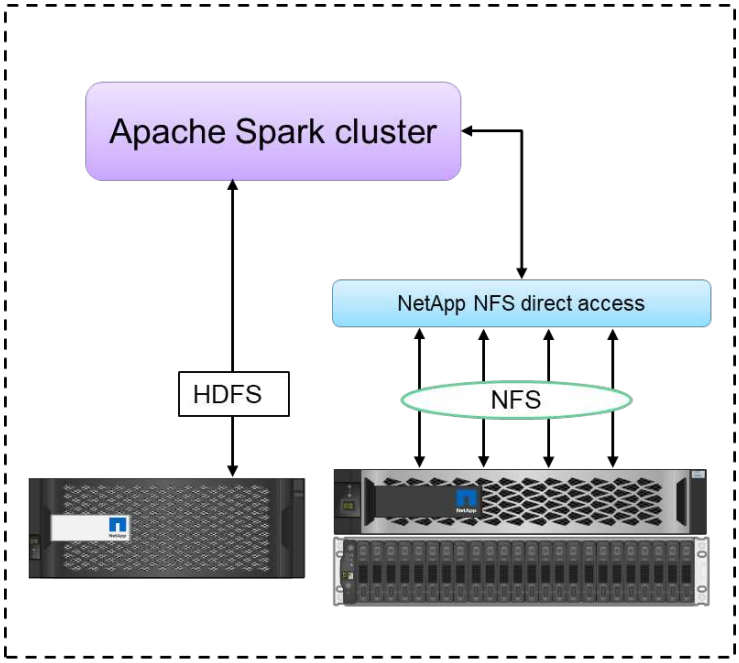
- Acelere las cargas de trabajo de análisis
- Proteja y asegure sus datos

Acceso directo de NFS de NetApp

NFS de NetApp permite a los clientes ejecutar trabajos de análisis de Big Data en sus datos de NFSv3 o NFSv4 nuevos o existentes sin necesidad de mover ni copiar los datos. Evita múltiples copias de datos y elimina la necesidad de sincronizar los datos con un origen. Por ejemplo, en el sector financiero, el movimiento de datos de un lugar a otro debe cumplir con las obligaciones legales, lo que no es una tarea fácil. En esta situación, el acceso directo NFS de NetApp analiza los datos financieros desde su ubicación original. Otra ventaja clave es que el uso del acceso directo NFS de NetApp simplifica la protección de los datos de Hadoop mediante comandos nativos de Hadoop y permite que los flujos de trabajo de protección de datos aprovechen la amplia cartera de gestión de datos de NetApp.



Configuration 1: NFS as primary storage



Configuration 2: HDFS and NFS in single Spark cluster

El acceso directo de NFS de NetApp proporciona dos tipos de opciones de puesta en marcha para clústeres Hadoop/Spark:

- De forma predeterminada, los clústeres de Hadoop/Spark utilizan el sistema de archivos distribuidos Hadoop (HDFS) para el almacenamiento de datos y el sistema de archivos predeterminado. El acceso directo NFS de NetApp puede sustituir el HDFS predeterminado por el almacenamiento NFS como sistema de archivos predeterminado, lo que permite operaciones de análisis directos en datos NFS.
- En otra opción de puesta en marcha, el acceso directo de NFS de NetApp admite la configuración de NFS como almacenamiento adicional junto con HDFS en un único clúster de Hadoop/Spark. En este caso, el cliente puede compartir datos a través de las exportaciones NFS y acceder a ellos desde el mismo clúster junto con los datos de HDFS.

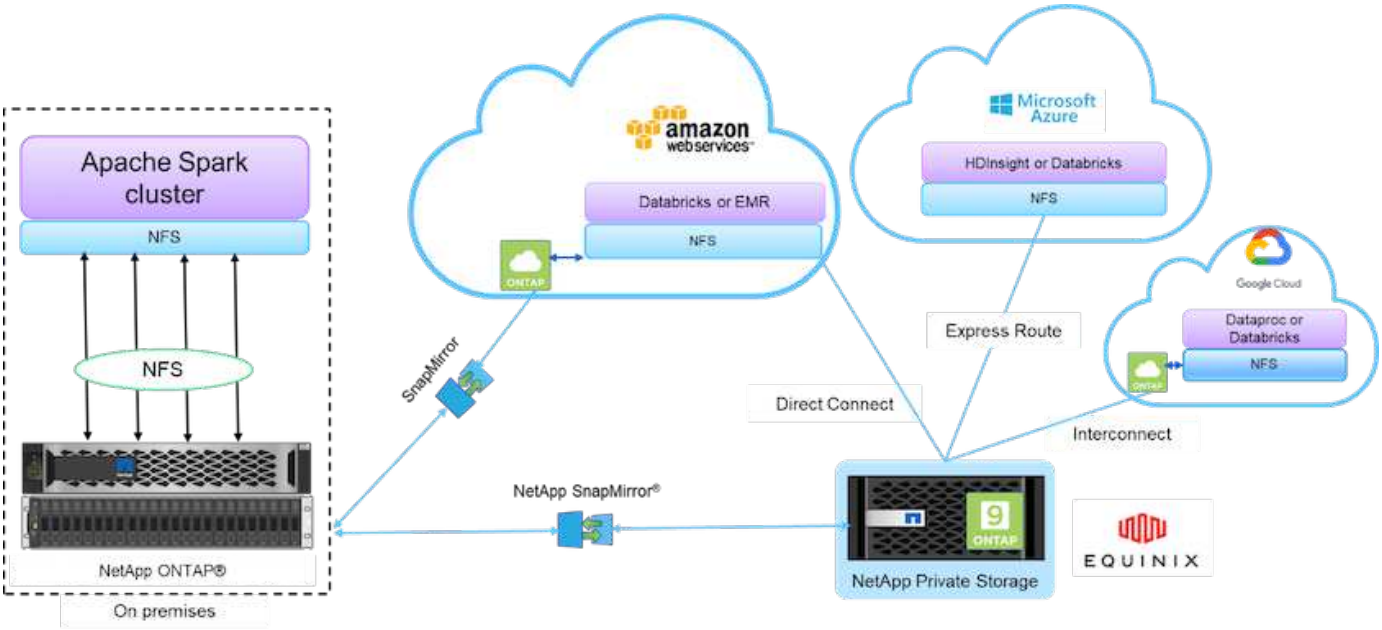
Las ventajas clave del uso del acceso directo de NFS de NetApp incluyen:

- Analiza los datos desde su ubicación actual, lo que evita la tarea que requiere tiempo y rendimiento, consistente en mover datos de análisis a una infraestructura de Hadoop como HDFS.

- Reduce el número de réplicas de tres a una.
- Permite a los usuarios desacoplar los recursos informáticos y del almacenamiento para escalar los recursos de forma independiente.
- Proporciona protección de datos empresariales aprovechando las funcionalidades de gestión de datos enriquecidos de ONTAP.
- Está certificado con la plataforma de datos Hortonworks.
- Habilita la puesta en marcha de análisis de datos híbridos.
- Reduce el tiempo de copia de seguridad aprovechando la capacidad de multisubproceso dinámica.

Elementos básicos para Big Data

El Data Fabric con tecnología de NetApp integra servicios y aplicaciones de gestión de datos (elementos básicos) para el acceso, control, protección y seguridad de los datos, como se muestra en la siguiente figura.



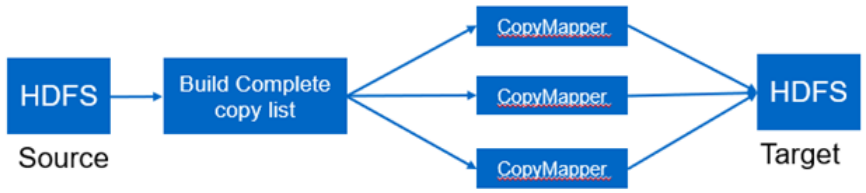
Los elementos básicos de la figura anterior incluyen:

- **Acceso directo NFS de NetApp.** proporciona los últimos clústeres Hadoop y Spark con acceso directo a volúmenes NFS de NetApp sin requisitos adicionales de software o controlador.
- **Servicios de volúmenes de cloud y Cloud Volumes ONTAP de NetApp.** almacenamiento conectado definido por software basado en ONTAP que se ejecuta en Amazon Web Services (AWS) o Azure NetApp Files (ANF) en servicios de cloud de Microsoft Azure.
- **La tecnología SnapMirror de NetApp**. Proporciona funcionalidades de protección de datos entre instancias locales y de ONTAP Cloud o NPS.
- **Proveedores de servicios en la nube.** estos proveedores incluyen AWS, Microsoft Azure, Google Cloud e IBM Cloud.
- **PaaS.** Servicios de análisis basados en cloud como Amazon Elastic MapReduce (EMR) y Databricks en AWS, así como Microsoft Azure HDInsight y Azure Databricks.

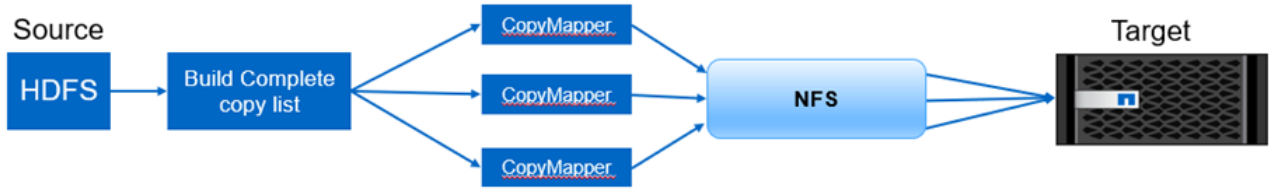
Protección de datos para Hadoop y NetApp

Hadoop DistCp es una herramienta nativa que se usa para copiar dentro de los clústeres y crear clústeres de gran tamaño. El proceso básico de Hadoop DistCp que se muestra en la siguiente figura es un flujo de trabajo de backup típico que utiliza herramientas nativas de Hadoop, como MapReduce, para copiar datos de Hadoop desde un origen de HDFS a un destino correspondiente.

El acceso directo NFS de NetApp permite a los clientes establecer NFS como destino para la herramienta Hadoop DistCp para copiar los datos desde el origen HDFS en un recurso compartido de NFS a través de MapReduce. El acceso directo para NFS de NetApp actúa como controlador NFS de la herramienta DistCp.



Hadoop Distcp Basic Process



Hadoop Distcp and NetApp

Descripción general de los casos prácticos de protección de datos de Hadoop

Esta sección proporciona una descripción detallada de los casos de uso de protección de datos que constituyen el centro de este documento. En el resto de secciones, se proporcionan más detalles sobre cada caso práctico, como el problema del cliente (escenario), los requisitos y retos, y las soluciones.

Caso de uso 1: Backup de datos de Hadoop

Para este caso de uso, el volumen NFS de NetApp ayudó a una gran institución financiera a reducir el tiempo de tiempo de backup prolongado de más de 24 horas a tan solo unas pocas horas.

Caso de uso 2: Backup y recuperación ante desastres del cloud a las instalaciones

Al utilizar Data Fabric con la tecnología de NetApp como elementos básicos, una gran empresa de retransmisión pudo cumplir los requisitos de realizar backups de los datos del cloud en su centro de datos en las instalaciones en función de los diferentes modos de transferencia de datos, como, por ejemplo, bajo demanda, de forma instantánea. O basado en la carga del clúster de Hadoop/Spark.

Caso de uso 3: Activación de DevTest en datos de Hadoop existentes

Las soluciones de NetApp ayudaron a un distribuidor de música en línea a crear rápidamente varios clústeres de Hadoop con una gestión eficiente del espacio en diferentes sucursales, con el fin de crear informes y ejecutar tareas diarias de DevTest mediante políticas programadas.

Caso de uso 4: Protección de datos y conectividad multicloud

Un gran proveedor de servicios utilizó el Data Fabric con tecnología de NetApp para proporcionar análisis multicloud a sus clientes de diferentes instancias de cloud.

Caso de uso 5: Acelerar las cargas de trabajo analíticas

Uno de los mayores bancos de inversión y servicios financieros utilizó la solución de almacenamiento conectado a la red de NetApp para reducir el tiempo de espera de I/o y acelerar su plataforma de análisis financiero cuantitativo.

Caso de uso 1: Backup de datos de Hadoop

Situación

En este escenario, el cliente tiene un gran repositorio local de Hadoop y desea realizar un backup para fines de recuperación ante desastres. Sin embargo, la solución de backup actual del cliente es costosa y padece una larga duración de backup de más de 24 horas.

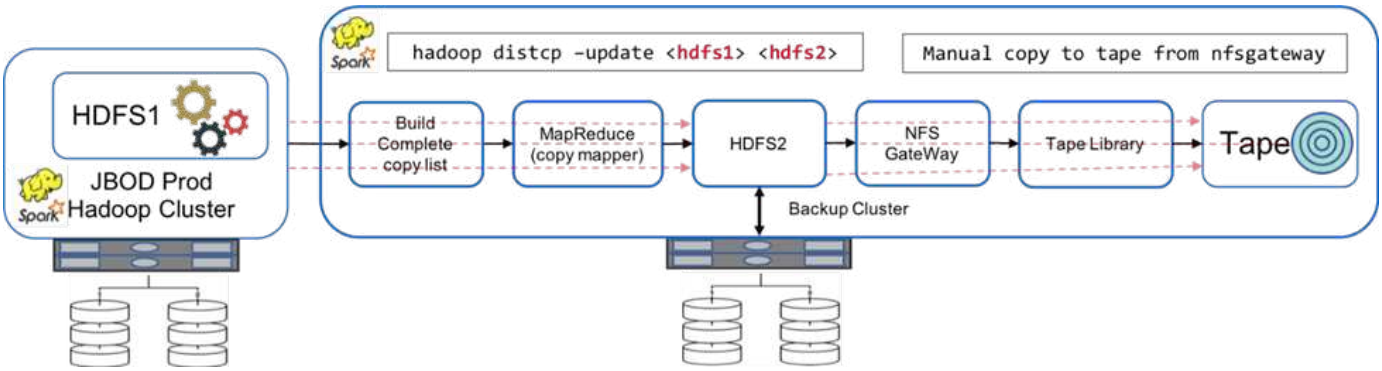
Requisitos y retos

Los principales requisitos y retos de este caso de uso son:

- Compatibilidad con versiones anteriores del software:
  - La solución de backup alternativa propuesta debe ser compatible con las versiones de software que se estén ejecutando actualmente y que se utilicen en el clúster de Hadoop de producción.
- Con el fin de cumplir los acuerdos de nivel de servicio comprometidos, la solución alternativa propuesta debería alcanzar unos objetivos de punto de recuperación y de tiempo de recuperación muy bajos.
- El backup creado por la solución de backup de NetApp se puede usar en el clúster de Hadoop, creado localmente en el centro de datos, así como en el clúster de Hadoop que se ejecuta en la ubicación de recuperación ante desastres del sitio remoto.
- La solución propuesta debe ser rentable.
- La solución propuesta debe reducir el efecto sobre el rendimiento de las tareas de análisis en producción que se están ejecutando actualmente durante los tiempos de backup.

Solución de backup existente del cliente

La siguiente ilustración muestra la solución original de backup nativo de Hadoop.



Los datos de producción se protegen a cinta a través del clúster de backup intermedio:

- Los datos HDFS1 se copian en HDFS2 ejecutando el `hadoop distcp -update <hdfs1> <hdfs2>` comando.
- El clúster de backup actúa como una puerta de enlace NFS y los datos se copian manualmente en cinta a través de Linux `cp` a través de la biblioteca de cintas.

Las ventajas de la solución original de backup nativo de Hadoop incluyen:

- La solución se basa en comandos nativos de Hadoop, que evitan que el usuario tenga que aprender nuevos procedimientos.
- La solución aprovecha el hardware y la arquitectura estándar del sector.

Las desventajas de la solución original de backup nativo de Hadoop son las siguientes:

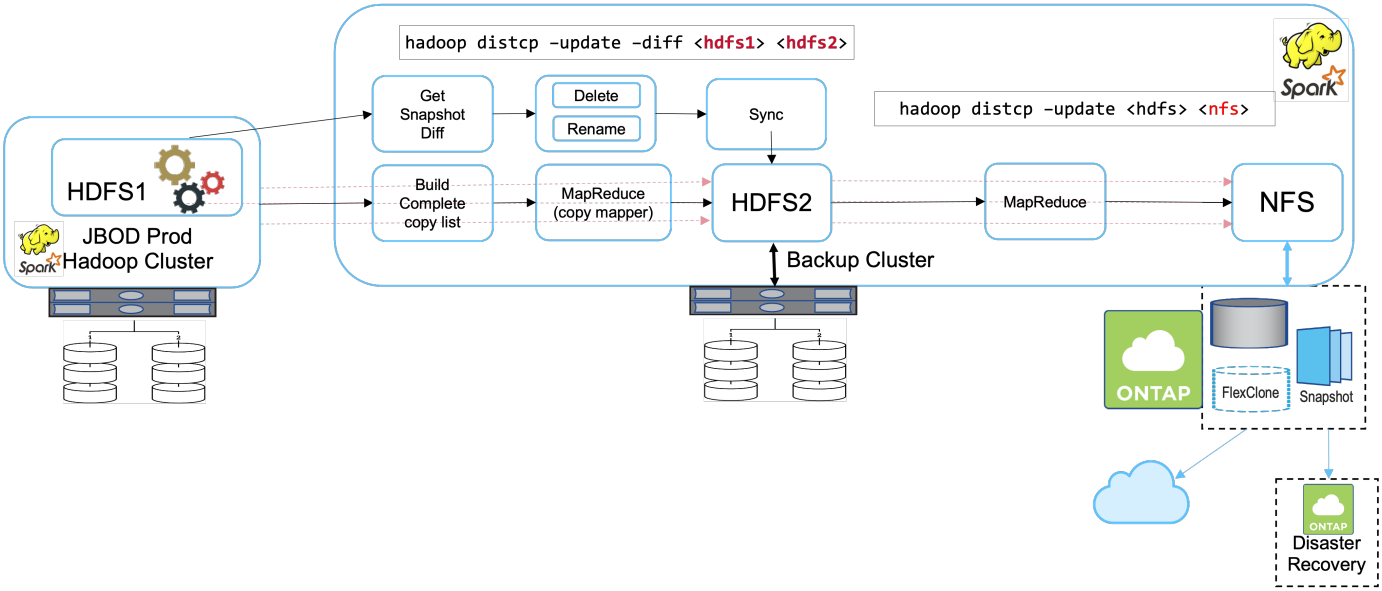
- El tiempo de backup prolongado supera las 24 horas, lo que hace que los datos de producción sean vulnerables.
- Degradación significativa del rendimiento del clúster durante los tiempos de backup.
- Copiar en cinta es un proceso manual.
- La solución de backup es costosa en términos de hardware necesario y de las horas humanas necesarias para los procesos manuales.

Backup y recuperación de datos

Teniendo en cuenta estos retos y requisitos, y teniendo en cuenta el sistema de backup existente, se sugirieron tres posibles soluciones de backup. En las siguientes subsecciones se describe cada una de estas tres soluciones de backup distintas, se describe una solución A través de la solución C.

Solución A

En la solución A, el clúster de backup de Hadoop envía los backups secundarios a los sistemas de almacenamiento NFS de NetApp, eliminando los requisitos de cinta, como se muestra en la siguiente figura.



Las tareas detalladas de la solución A incluyen:

- El clúster de producción de Hadoop tiene los datos de análisis del cliente en HDFS que requiere protección.
- El clúster de Hadoop de backup con HDFS actúa como una ubicación intermedia de los datos. Un solo grupo de discos (JBOD) proporciona almacenamiento para HDFS tanto en los clústeres de producción como en los de Hadoop de backup.
- Proteja los datos de producción de Hadoop están protegidos del clúster de producción HDFS al clúster de backup HDFS con la ejecución de `Hadoop distcp -update -diff <hdfs1> <hdfs2>` comando.

> La snapshot de Hadoop se usa para proteger los datos de producción para el clúster de Hadoop de backup.

- La controladora de almacenamiento ONTAP de NetApp proporciona un volumen exportado de NFS, que se aprovisiona para el clúster de Hadoop de backup.
- Ejecutando el `Hadoop distcp` Con MapReduce y varios asignadores, los datos de análisis están protegidos del clúster de Hadoop de backup a NFS.

  Una vez que los datos se almacenan en NFS en el sistema de almacenamiento de NetApp, las tecnologías Snapshot, SnapRestore y FlexClone de NetApp se utilizan para realizar backups, restaurar y duplicar los datos de Hadoop según sea necesario.

> Los datos de Hadoop pueden protegerse en el cloud y en ubicaciones de recuperación ante desastres con la tecnología SnapMirror.

Las ventajas de la solución A incluyen:

- Los datos de producción de Hadoop se protegen del clúster de backup.
- Los datos HDFS están protegidos por NFS, lo que permite la protección en el cloud y en ubicaciones de recuperación ante desastres.
- Mejora el rendimiento mediante la descarga de las operaciones de backup al cluster de backup.
- Elimina las operaciones manuales en cintas

- Permite funciones de gestión empresarial mediante herramientas de NetApp.
- Requiere unos cambios mínimos en el entorno existente.
- Es una solución rentable.

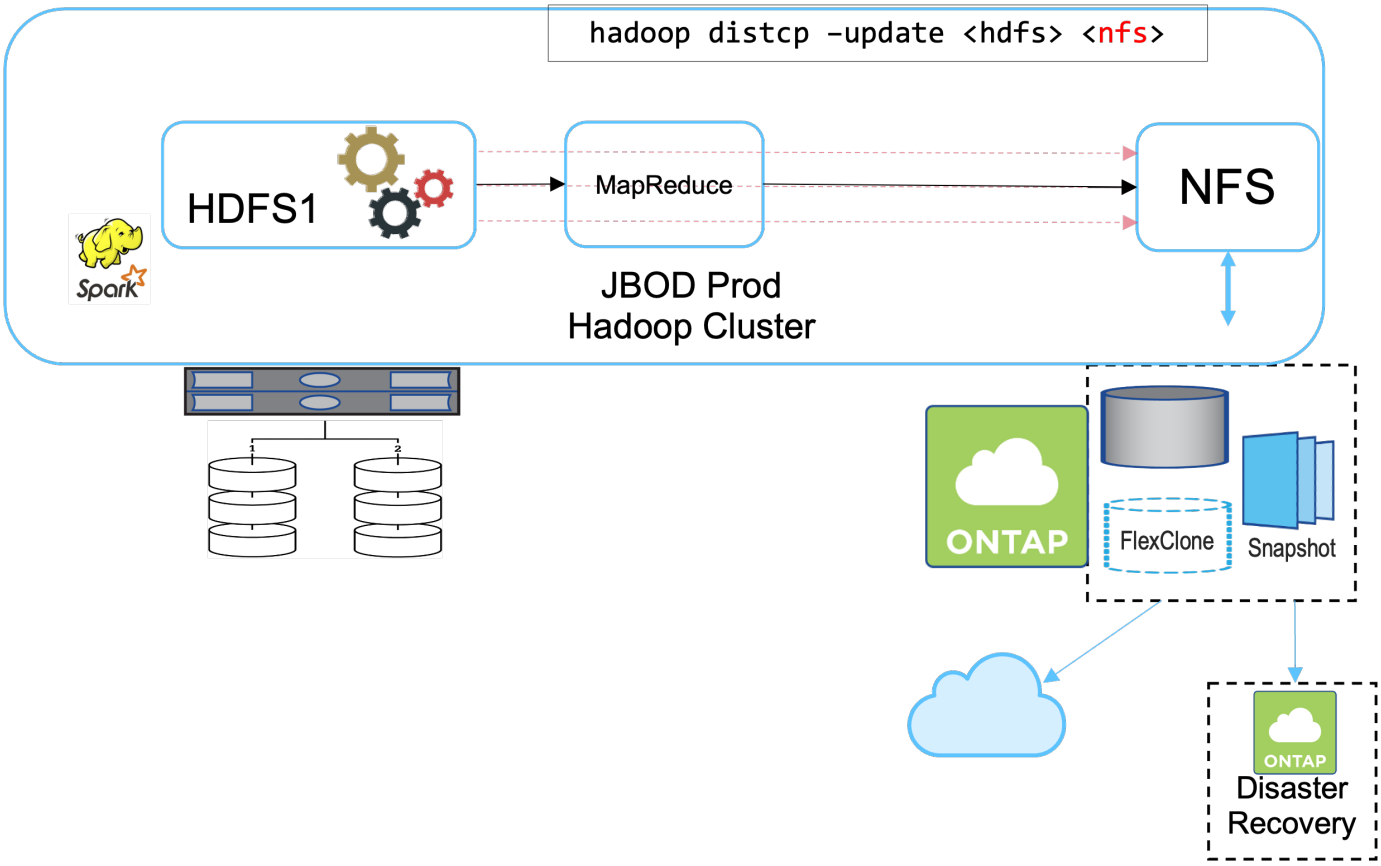
La desventaja de esta solución es que requiere un clúster de backup y asignadores adicionales para mejorar el rendimiento.

El cliente ha puesto en marcha recientemente una solución debido a su simplicidad, coste y rendimiento general.

En esta solución, los discos SAN de ONTAP pueden utilizarse en vez de JBOD. Esta opción transfiere la carga de almacenamiento del clúster de backup a ONTAP; sin embargo, la desventaja es que se necesitan los switches de la estructura SAN.

Solución B

La solución B añade el volumen NFS al clúster Hadoop de producción, lo que elimina la necesidad de tener que utilizar el clúster Hadoop de backup, como se muestra en la siguiente figura.



Entre las tareas detalladas de la solución B se incluyen las siguientes:

- La controladora de almacenamiento ONTAP de NetApp aprovisiona la exportación NFS al clúster Hadoop de producción.

  Hadoop de forma nativa `hadoop distcp` El comando protege los datos de Hadoop desde HDFS del clúster de producción a NFS.

- Una vez que los datos se almacenan en NFS en el sistema de almacenamiento de NetApp, las

tecnologías Snapshot, SnapRestore y FlexClone se utilizan para realizar backups, restaurar y duplicar los datos de Hadoop según sea necesario.

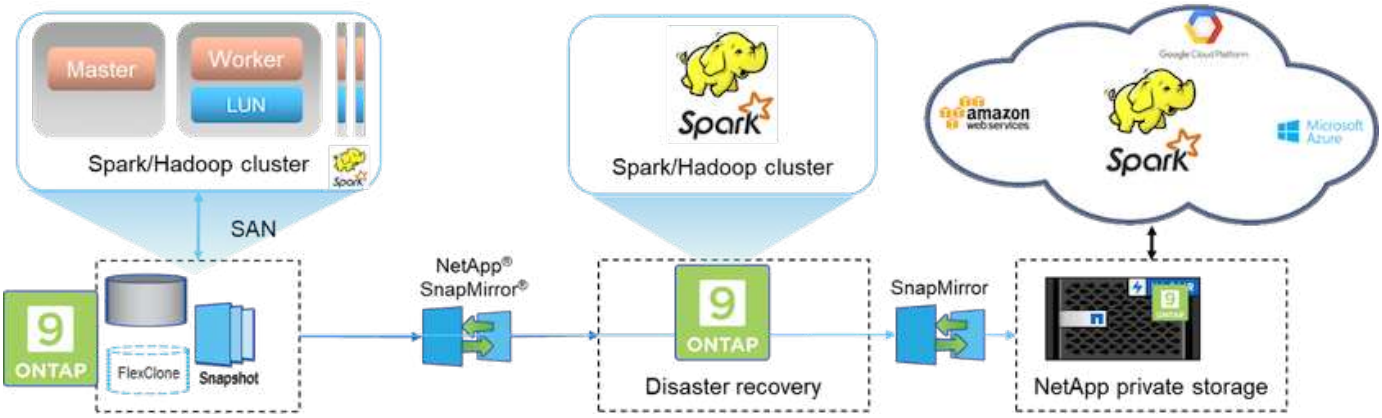
Las ventajas de la solución B incluyen:

- El clúster de producción está ligeramente modificado para la solución de backup, lo que simplifica la implementación y reduce los costes adicionales de infraestructura.
- No se necesita un clúster de backup para la operación de backup.
- Los datos de producción HDFS se protegen en la conversión a datos NFS.
- La solución posibilita funciones de gestión empresarial mediante las herramientas de NetApp.

La desventaja de esta solución es que se implementa en el clúster de producción, que puede agregar tareas adicionales de administrador en el clúster de producción.

Solución C

En la solución C, los volúmenes SAN de NetApp se aprovisionan directamente en el clúster de producción de Hadoop para el almacenamiento HDFS, tal y como se muestra en la siguiente figura.



Los pasos detallados de la solución C incluyen:

- El almacenamiento SAN ONTAP de NetApp se aprovisiona en el clúster Hadoop de producción para el almacenamiento de datos HDFS.
- Las tecnologías Snapshot y SnapMirror de NetApp se usan para realizar backups de los datos HDFS desde el clúster de producción de Hadoop.
- No existe ningún efecto sobre el rendimiento en la producción del clúster Hadoop/Spark durante el proceso de backup de copias de Snapshot debido a que el backup se encuentra en la capa de almacenamiento.

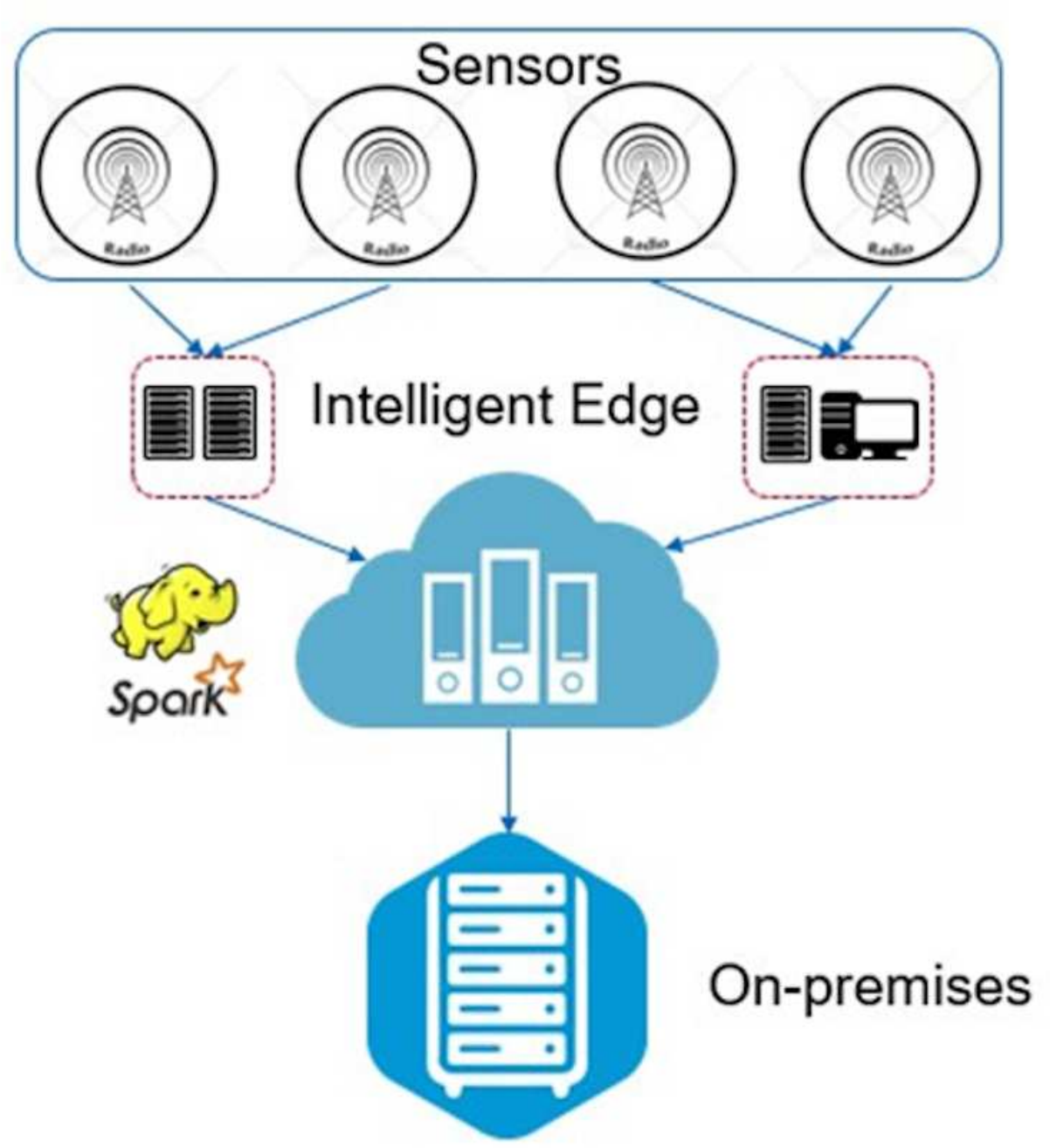
> La tecnología Snapshot ofrece backups que se realizan en cuestión de segundos independientemente del tamaño de los datos.

Las ventajas de la solución C incluyen:

- Los backups con gestión eficiente del espacio pueden crearse utilizando la tecnología Snapshot.
- Permite funciones de gestión empresarial mediante herramientas de NetApp.

Caso de uso 2: Backup y recuperación ante desastres del cloud a las instalaciones

Este caso de uso se basa en un cliente de retransmisión que necesita realizar backups de los datos de análisis basados en cloud en su centro de datos en las instalaciones, como se muestra en la siguiente figura.



Situación

En este escenario, los datos del sensor del IoT se ingieren en el cloud y se analizan mediante un clúster de código abierto Apache Spark dentro de AWS. El requisito es realizar backups de los datos procesados desde el cloud a las instalaciones.

Requisitos y retos

Los principales requisitos y retos de este caso de uso son:

- La habilitación de la protección de datos no debe provocar ningún efecto sobre el rendimiento en el clúster de producción de Spark/Hadoop en el cloud.
- Los datos de los sensores del cloud necesitan moverse y protegerse a las instalaciones de una forma eficiente y segura.
- Flexibilidad para transferir datos del cloud a las instalaciones en distintas condiciones, como bajo demanda, instantáneas y durante tiempos de carga en un clúster bajo.

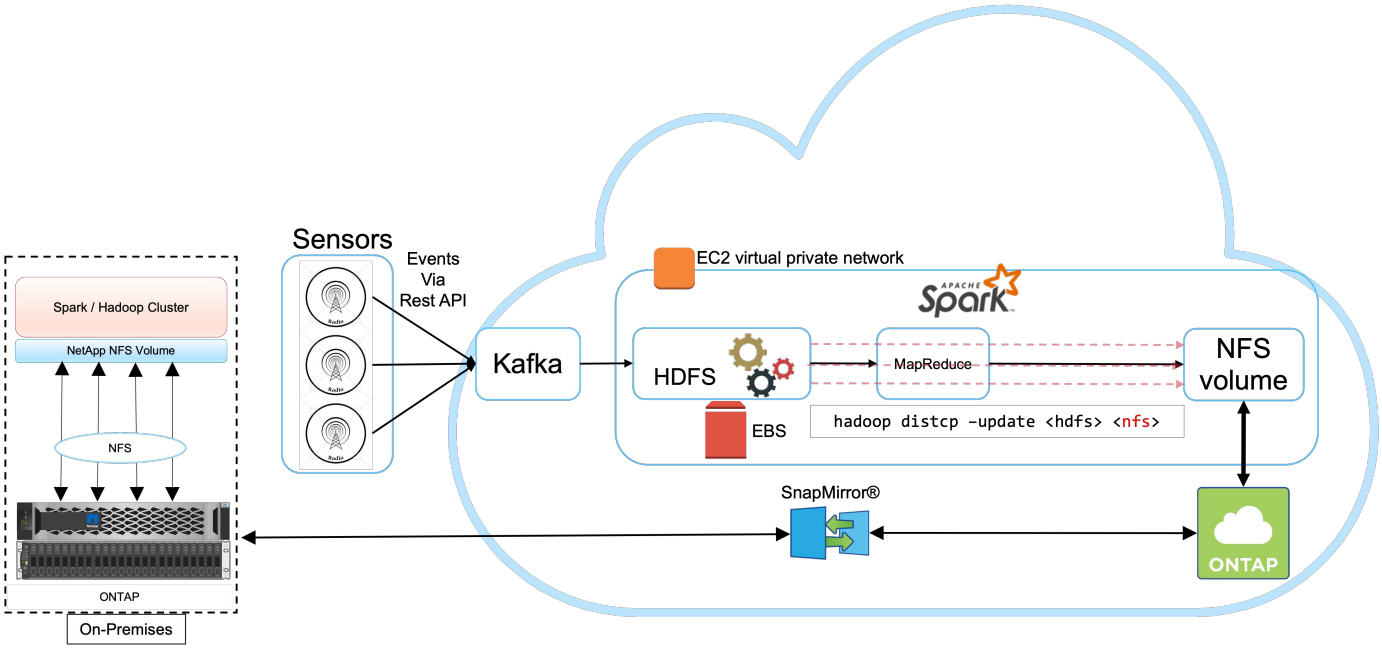
Solución

El cliente utiliza Elastic Block Store (EBS) de AWS para su almacenamiento HDFS con clúster Spark y recibir datos de sensores remotos a través de Kafka. Por lo tanto, el almacenamiento HDFS actúa como el origen de los datos de backup.

Para cumplir estos requisitos, ONTAP Cloud de NetApp se pone en marcha en AWS y se crea un recurso compartido de NFS para que actúe como objetivo de backup para el clúster de Spark/Hadoop.

Una vez creado el recurso compartido NFS, copie los datos del almacenamiento HDFS EBS en el recurso compartido NFS de ONTAP. Después de que los datos residen en NFS en ONTAP Cloud, la tecnología SnapMirror puede utilizarse para reflejar los datos del cloud en un almacenamiento en las instalaciones según sea necesario de una forma segura y eficiente.

Esta imagen muestra el backup y la recuperación ante desastres del cloud a la solución en las instalaciones.



Caso de uso 3: Activación de DevTest en datos de Hadoop existentes

En este caso de uso, el requisito del cliente es crear de forma rápida y eficiente clústeres Hadoop/Spark basados en un clúster de Hadoop existente que contenga una gran

cantidad de datos de análisis para DevTest y fines de generación de informes en el mismo centro de datos y en ubicaciones remotas.

Situación

En este escenario, se han creado varios clústeres de Spark/Hadoop a partir de una implementación de lagos de datos de Hadoop en las instalaciones y en ubicaciones de recuperación ante desastres.

Requisitos y retos

Los principales requisitos y retos de este caso de uso son:

- Cree varios clústeres de Hadoop para DevTest, QA o cualquier otro propósito que requiera el acceso a los mismos datos de producción. El reto que se plantea es clonar un clúster de Hadoop de gran tamaño varias veces al instante y con una gestión muy eficiente del espacio.
- Sincronice los datos de Hadoop con los equipos de generación de informes y DevTest para obtener eficiencia operativa.
- Distribuya los datos de Hadoop mediante las mismas credenciales en los clústeres de producción y nuevos.
- Use directivas programadas para crear de forma eficaz clústeres de control de calidad sin que ello afecte al clúster de producción.

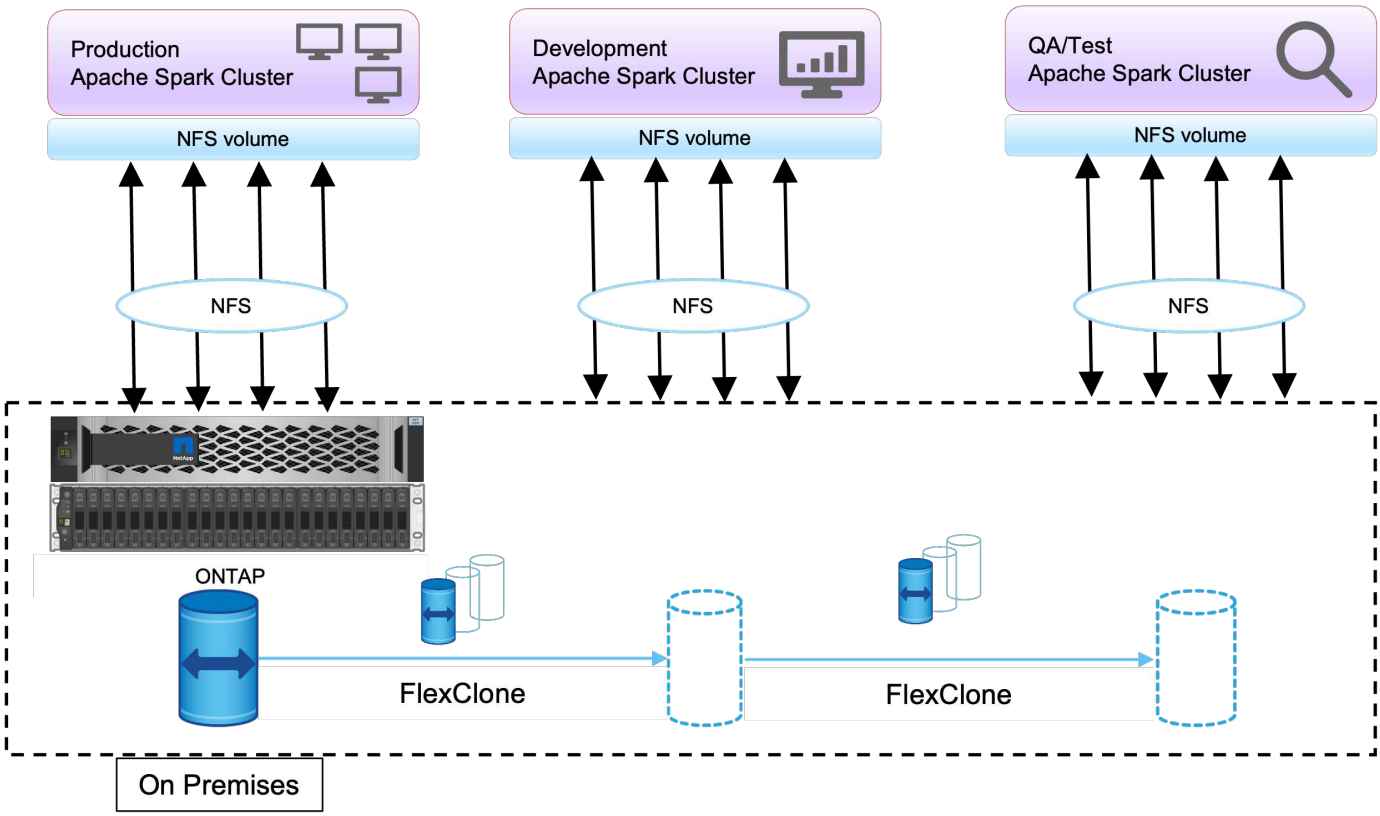
Solución

La tecnología FlexClone se utiliza para responder a los requisitos que acabamos de describir. La tecnología FlexClone es la copia de lectura/escritura de una copia Snapshot. Lee los datos de la copia snapshot principal y consume únicamente espacio adicional para los bloques nuevos o modificados. Es rápida y permite ahorrar espacio.

En primer lugar, se creó una copia Snapshot del clúster existente mediante un grupo de consistencia de NetApp.

Las copias de Snapshot dentro de NetApp System Manager o del símbolo del sistema del administrador del almacenamiento. Las copias Snapshot del grupo de consistencia son copias Snapshot de grupos coherentes con las aplicaciones y el volumen FlexClone se crea en función de las copias Snapshot del grupo de consistencia. Vale la pena mencionar que un volumen FlexClone hereda la política de exportación NFS del volumen principal. Una vez creada la copia snapshot, debe instalarse un nuevo clúster de Hadoop con fines de DevTest y generación de informes, como se muestra en la siguiente figura. El volumen NFS clonado del nuevo clúster de Hadoop accede a los datos de NFS.

Esta imagen muestra el clúster de Hadoop para DevTest.



Caso de uso 4: Protección de datos y conectividad multicloud

Este caso de uso es relevante para un partner de servicio cloud que se encarga de proporcionar conectividad multicloud para los datos de análisis de Big Data de los clientes.

Situación

En esta situación, los datos de IoT que se reciben en AWS de diferentes orígenes se almacenan en una ubicación central en NPS. El almacenamiento NPS está conectado a clústeres de Spark/Hadoop ubicados en AWS y Azure y permite que las aplicaciones de análisis de Big Data se ejecuten en varios clouds accedan a los mismos datos.

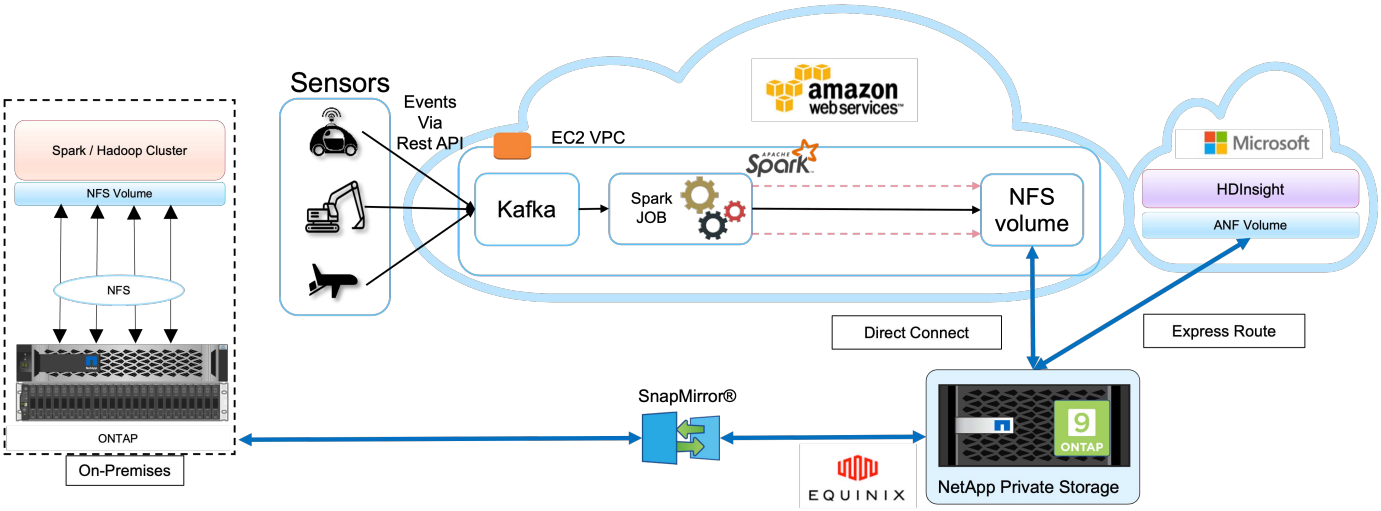
Requisitos y retos

Los principales requisitos y retos de este caso de uso son:

- Los clientes quieren ejecutar trabajos de análisis en los mismos datos utilizando varios clouds.
- Los datos deben recibirse de diferentes fuentes, como on-premises y en el cloud, mediante distintos sensores y concentradores.
- La solución debe ser eficiente y rentable.
- El principal reto es crear una solución rentable y eficiente que ofrezca servicios de análisis híbridos entre sus instalaciones y clouds diferentes.

Solución

Esta imagen ilustra la solución de conectividad de múltiples clouds y protección de datos.



Como se muestra en la figura anterior, los datos de los sensores se transmiten y ingieren en el clúster de AWS Spark a través de Kafka. Los datos se almacenan en un recurso compartido de NFS que reside en NPS, que se encuentra fuera del proveedor de cloud en un centro de datos Equinix. Dado que el almacenamiento privado de NetApp de NetApp está conectado a Amazon AWS y Microsoft Azure a través de conexiones Direct Connect y Express Route, respectivamente, los clientes pueden acceder a los datos NFS desde los clústeres de análisis de Amazon y AWS. Este enfoque soluciona el tener que hacer análisis en cloud en varios proveedores a hiperescala.

Como consecuencia, tanto el almacenamiento en las instalaciones como el almacenamiento privado de NetApp ejecutan el software ONTAP, SnapMirror puede reflejar los datos del almacenamiento privado de NetApp en el clúster en las instalaciones, lo cual proporciona análisis del cloud híbrido entre las instalaciones y varios clouds.

Para obtener el mejor rendimiento, NetApp suele recomendar el uso de varias interfaces de red, así como rutas de conexión directa/exprés para acceder a los datos desde instancias del cloud.

Caso de uso 5: Acelerar las cargas de trabajo analíticas

En este escenario, se modernizó la plataforma de análisis de un gran banco de inversiones y servicios financieros mediante la solución de almacenamiento NFS de NetApp para lograr mejoras significativas en el análisis de riesgos de inversión y derivados para su unidad de negocio cuantitativa y gestión de activos.

Situación

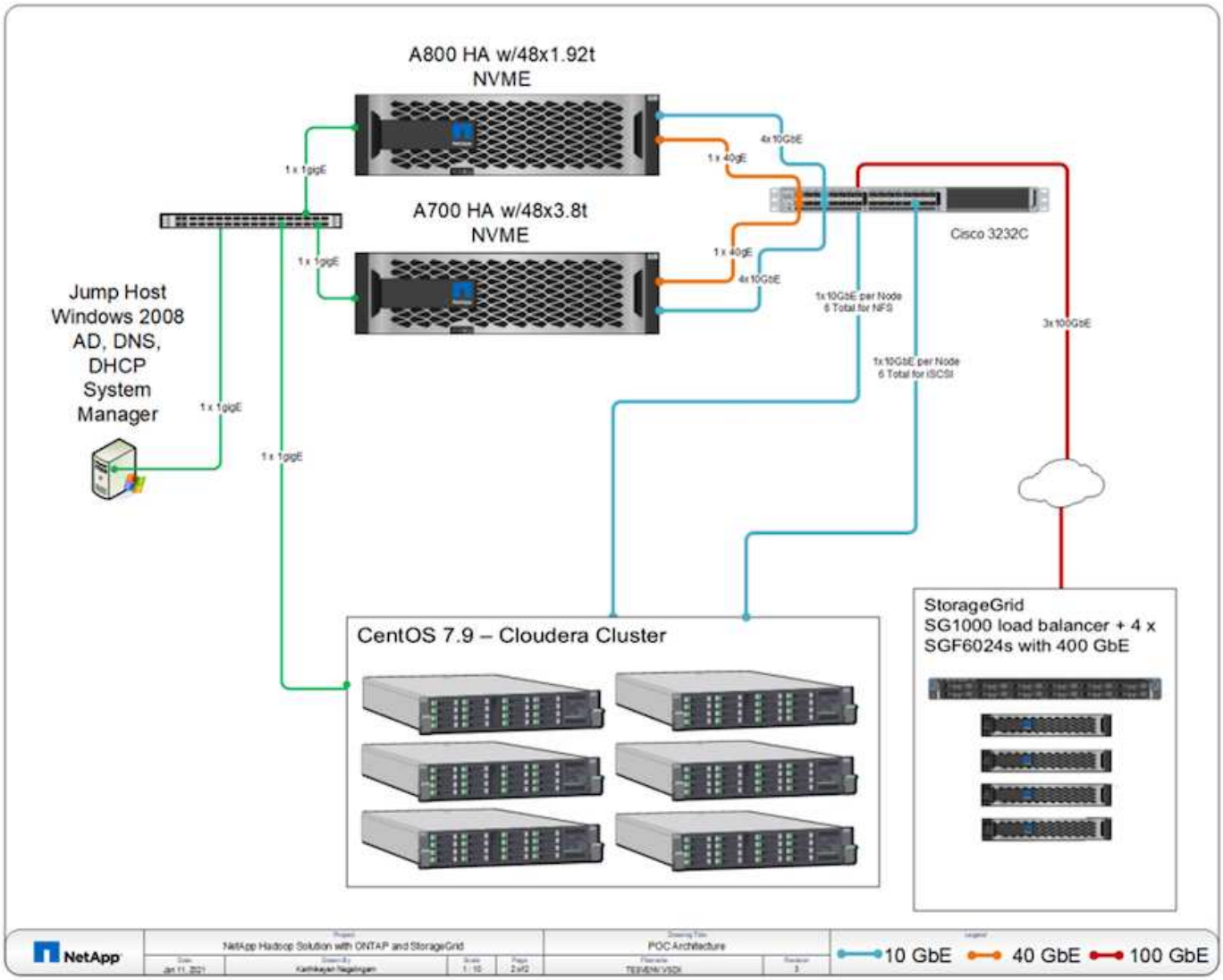
En el entorno existente del cliente, la infraestructura Hadoop utilizada para la plataforma de análisis aprovechó el almacenamiento interno de los servidores Hadoop. Debido a la naturaleza propia del entorno JBOD, muchos clientes internos de la organización no pudieron aprovechar su modelo cuantitativo de Monte Carlo, una simulación que se basa en muestras recurrentes de datos en tiempo real. La capacidad subóptima para comprender los efectos de la incertidumbre en los movimientos del mercado no era favorable para la unidad de negocio de gestión de activos cuantitativos.

Requisitos y retos

La unidad de negocios cuantitativos del banco quería un método eficiente de pronóstico para lograr predicciones precisas y oportunas. Para ello, el equipo reconoció la necesidad de modernizar la infraestructura, reducir el tiempo de espera de I/o existente y mejorar el rendimiento en las aplicaciones analíticas como Hadoop y Spark para simular de manera eficiente modelos de inversiones, medir posibles ganancias y analizar riesgos.

Solución

El cliente tenía JBOD para su solución de Spark existente. Entonces, se aprovechó ONTAP de NetApp, StorageGRID de NetApp y Minio Gateway a NFS para reducir el tiempo de espera de I/o para el grupo financiero cuantitativo del banco que ejecuta simulación y análisis en modelos de inversión que evalúan posibles ganancias y riesgos. Esta imagen muestra la solución Spark con el almacenamiento de NetApp.



Como se muestra en la figura anterior, los sistemas AFF A800, A700 y StorageGRID se pusieron en marcha para acceder a archivos de parqué a través de protocolos NFS y S3 en un clúster Hadoop de seis nodos con Spark y los servicios de metadatos DE YARN y Hive para operaciones de análisis de datos.

Una solución de almacenamiento de conexión directa (DAS) en el antiguo entorno del cliente tenía la desventaja de escalar los recursos informáticos y el almacenamiento de forma independiente. Con la solución

ONTAP de NetApp para Spark, la unidad de negocio de análisis financieros del banco pudo desvincular el almacenamiento de la computación y ofrecer sin problemas recursos de infraestructura de una manera más eficiente posible según lo necesite.

Al utilizar ONTAP con NFS, las CPU de los servidores de computación se utilizaron casi por completo para tareas de Spark SQL y el tiempo de espera de I/o se redujo en casi un 70 % y, así, se obtuvo una mejor potencia informática y un mejor impulso de rendimiento para las cargas de trabajo de Spark. Posteriormente, el aumento del uso de la CPU también permitió al cliente aprovechar las GPU, como GPUDirect, para una mayor modernización de la plataforma. Además, StorageGRID proporciona una opción de almacenamiento de bajo coste para las cargas de trabajo de Spark y Minio Gateway proporciona acceso seguro a los datos NFS mediante el protocolo S3. Para los datos en el cloud, NetApp recomienda Cloud Volumes ONTAP, Azure NetApp Files y Cloud Volumes Service de NetApp.

Conclusión

En esta sección se proporciona un resumen de los casos de uso y las soluciones que ofrece NetApp para satisfacer diversos requisitos de protección de datos de Hadoop. Gracias al Data Fabric con tecnología de NetApp, los clientes pueden:

- Disponga de la flexibilidad necesaria para elegir las soluciones de protección de datos adecuadas aprovechando las completas funcionalidades de gestión de datos de NetApp y su integración con los flujos de trabajo nativos de Hadoop.
- Reduzca el tiempo necesario para realizar backups en clústeres de Hadoop en casi un 70 %.
- Elimine cualquier efecto en el rendimiento que se pueda obtener gracias a los backups de clústeres de Hadoop.
- Ofrezca protección de datos multicloud y acceso a los datos desde distintos proveedores de cloud simultáneamente a una única fuente de datos de análisis.
- Cree copias de clúster de Hadoop rápidas y con gestión eficiente del espacio mediante la tecnología FlexClone.

Dónde encontrar información adicional

Para obtener más información sobre la información descrita en este documento, consulte los siguientes documentos y/o sitios web:

- Soluciones de análisis de Big Data de NetApp

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

- Carga de trabajo de Apache Spark con el almacenamiento de NetApp

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

- Soluciones de almacenamiento de NetApp para Apache Spark

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

- Apache Hadoop en Data Fabric habilitado por NetApp

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

Reconocimientos

- Paul Burland, representante de ventas, ANZ Victoria District Sales, NetApp
- Hoseb Dermanilian, Director de Desarrollo Empresarial de NetApp
- Lee Dorrier, director de MPSG, NetApp
- David Thiessen, ingeniero de sistemas, ANZ Victoria District se, NetApp

Historial de versiones

Versión	Fecha	Historial de versiones del documento
Versión 1.0	Enero de 2018	Versión inicial
Versión 2.0	Octubre de 2021	Actualizado con el caso de uso n.o 5: Acelere la carga de trabajo analítica
Versión 3.0	Noviembre de 2023	Se han eliminado los detalles de NIPAM

Información de copyright

Copyright © 2024 NetApp, Inc. Todos los derechos reservados. Imprimido en EE. UU. No se puede reproducir este documento protegido por copyright ni parte del mismo de ninguna forma ni por ningún medio (gráfico, electrónico o mecánico, incluidas fotocopias, grabaciones o almacenamiento en un sistema de recuperación electrónico) sin la autorización previa y por escrito del propietario del copyright.

El software derivado del material de NetApp con copyright está sujeto a la siguiente licencia y exención de responsabilidad:

ESTE SOFTWARE LO PROPORCIONA NETAPP «TAL CUAL» Y SIN NINGUNA GARANTÍA EXPRESA O IMPLÍCITA, INCLUYENDO, SIN LIMITAR, LAS GARANTÍAS IMPLÍCITAS DE COMERCIALIZACIÓN O IDONEIDAD PARA UN FIN CONCRETO, CUYA RESPONSABILIDAD QUEDA EXIMIDA POR EL PRESENTE DOCUMENTO. EN NINGÚN CASO NETAPP SERÁ RESPONSABLE DE NINGÚN DAÑO DIRECTO, INDIRECTO, ESPECIAL, EJEMPLAR O RESULTANTE (INCLUYENDO, ENTRE OTROS, LA OBTENCIÓN DE BIENES O SERVICIOS SUSTITUTIVOS, PÉRDIDA DE USO, DE DATOS O DE BENEFICIOS, O INTERRUPCIÓN DE LA ACTIVIDAD EMPRESARIAL) CUALQUIERA SEA EL MODO EN EL QUE SE PRODUJERON Y LA TEORÍA DE RESPONSABILIDAD QUE SE APLIQUE, YA SEA EN CONTRATO, RESPONSABILIDAD OBJETIVA O AGRAVIO (INCLUIDA LA NEGLIGENCIA U OTRO TIPO), QUE SURJAN DE ALGÚN MODO DEL USO DE ESTE SOFTWARE, INCLUSO SI HUBIEREN SIDO ADVERTIDOS DE LA POSIBILIDAD DE TALES DAÑOS.

NetApp se reserva el derecho de modificar cualquiera de los productos aquí descritos en cualquier momento y sin aviso previo. NetApp no asume ningún tipo de responsabilidad que surja del uso de los productos aquí descritos, excepto aquello expresamente acordado por escrito por parte de NetApp. El uso o adquisición de este producto no lleva implícita ninguna licencia con derechos de patente, de marcas comerciales o cualquier otro derecho de propiedad intelectual de NetApp.

Es posible que el producto que se describe en este manual esté protegido por una o más patentes de EE. UU., patentes extranjeras o solicitudes pendientes.

LEYENDA DE DERECHOS LIMITADOS: el uso, la copia o la divulgación por parte del gobierno están sujetos a las restricciones establecidas en el subpárrafo (b)(3) de los derechos de datos técnicos y productos no comerciales de DFARS 252.227-7013 (FEB de 2014) y FAR 52.227-19 (DIC de 2007).

Los datos aquí contenidos pertenecen a un producto comercial o servicio comercial (como se define en FAR 2.101) y son propiedad de NetApp, Inc. Todos los datos técnicos y el software informático de NetApp que se proporcionan en este Acuerdo tienen una naturaleza comercial y se han desarrollado exclusivamente con fondos privados. El Gobierno de EE. UU. tiene una licencia limitada, irrevocable, no exclusiva, no transferible, no sublicenciable y de alcance mundial para utilizar los Datos en relación con el contrato del Gobierno de los Estados Unidos bajo el cual se proporcionaron los Datos. Excepto que aquí se disponga lo contrario, los Datos no se pueden utilizar, desvelar, reproducir, modificar, interpretar o mostrar sin la previa aprobación por escrito de NetApp, Inc. Los derechos de licencia del Gobierno de los Estados Unidos de América y su Departamento de Defensa se limitan a los derechos identificados en la cláusula 252.227-7015(b) de la sección DFARS (FEB de 2014).

Información de la marca comercial

NETAPP, el logotipo de NETAPP y las marcas que constan en http://www.netapp.com/TM son marcas comerciales de NetApp, Inc. El resto de nombres de empresa y de producto pueden ser marcas comerciales de sus respectivos propietarios.