MapReduce Service

Pasos iniciales

Edición 01

Fecha 2023-11-18

HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD.

Marcas registradas y permisos

El logotipo HUAWEI y otras marcas registradas de Huawei pertenecen a Huawei Technologies Co., Ltd.
Todas las demás marcas registradas y los otros nombres comerciales mencionados en este documento son propiedad de sus respectivos titulares.

Aviso

Es posible que la totalidad o parte de los productos, las funcionalidades y/o los servicios que figuran en el presente documento no se encuentren dentro del alcance de un contrato vigente entre Huawei Cloud y el cliente. Las funcionalidades, los productos y los servicios adquiridos se limitan a los estipulados en el respectivo contrato. A menos que un contrato especifique lo contrario, ninguna de las afirmaciones, informaciones ni recomendaciones contenidas en el presente documento constituye garantía alguna, ni expresa ni implícita.

Huawei está permanentemente preocupada por la calidad de los contenidos de este documento; sin embargo, ninguna declaración, información ni recomendación aquí contenida constituye garantía alguna, ni expresa ni implícita. La información contenida en este documento se encuentra sujeta a cambios sin previo aviso.

Huawei Cloud Computing Technologies Co., Ltd.

Dirección: Huawei Cloud Data Center Jiaoxinggong Road
Avenida Qianzhong
Nuevo distrito de Gui'an
Gui Zhou, 550029
República Popular China

Sitio web: https://www.huaweicloud.com/intl/es-us/

Índice

1 Compra y uso de un clúster MRS....1
1.1 Pasos iniciales con MapReduce Service....1
1.2 Compra de un clúster....2
1.3 Carga de datos....5
1.4 Creación de un trabajo....8
1.5 Terminación de un clúster....11
2 Instalación y uso del cliente de clúster....12
3 Uso de clústeres con autenticación Kerberos habilitada....17
4 Uso de Hadoop desde el principio....27
5 Uso de Kafka desde principio....31
6 Uso de HBase desde principio....36
7 Modificación de configuraciones de MRS....44
8 Configuración del escalado automático para un clúster MRS....49
9 Configuración de Hive con almacenamiento y cómputo desacoplado....56
10 Envío de tareas de Spark a nuevos nodos de Task....61
11 Configuración de Umbrales para Alarmas....66
12 Desarrollo de aplicaciones de componentes MRS....97
12.1 Desarrollo de aplicaciones de HBase....97
12.2 Desarrollo de aplicaciones de HDFS....106
12.3 Desarrollo de aplicaciones de Hive JDBC....112
12.4 Desarrollo de aplicaciones de Hive HCatalog....115
12.5 Desarrollo de aplicaciones de Kafka....119
12.6 Desarrollo de aplicaciones de Flink....124
12.7 Desarrollo de aplicaciones de ClickHouse....132
12.8 Desarrollo de aplicaciones de Spark....140
13 Prácticas....147

1 Compra y uso de un clúster MRS

1.1 Pasos iniciales con MapReduce Service

MapReducce Service (MRS) es un servicio de Huawei Cloud que se utiliza para desplegar y gestionar clústeres de Hadoop. MRS proporciona clústeres de big data de clase empresarial en la nube. Los tenants pueden controlar completamente estos clústeres y ejecutar fácilmente componentes de big data como Hadoop, Spark, HBase y Kafka en ellos.

MRS es fácil de usar. Puede ejecutar varias tareas y procesar o almacenar datos a nivel de PB mediante equipos conectados en un clúster.

El procedimiento de uso de MRS es el siguiente:

1. Comprar un clúster en la consola MRS. Durante este período, puede especificar el tipo de clúster, las especificaciones y el recuento de nodos, el tipo de disco de datos (**High I/O** o **Ultra-high I/O**), y los componentes que se van a instalar.
2. Desarrolle un programa de procesamiento de datos. Para obtener detalles sobre cómo desarrollar rápidamente un programa de este tipo y ejecutarlo correctamente, consulte el código de ejemplo y los tutoriales proporcionados en **Método de construcción de un proyecto de ejemplo de MRS**.
3. Cargue el programa y los archivos de datos preparados al Object Storage Service (OBS) o el HDFS en el clúster.
4. Después de crear un clúster, puede agregar directamente trabajos y ejecutar programas o sentencias SQL para procesar y analizar datos.
5. MRS le proporciona MRS Manager, una plataforma de gestión unificada de clase empresarial de clústeres de big data, que le ayuda a conocer rápidamente el estado de salud de los servicios y hosts. A través del monitoreo y la personalización de métricas gráficas, puede obtener información crítica del sistema de manera oportuna. Además, puede modificar las configuraciones de atributos de servicio en función de los requisitos de rendimiento del servicio e iniciar o detener clústeres, servicios e instancias de rol con un solo clic.
6. Termine el clúster si ya no es necesario después de la ejecución del trabajo. El clúster terminado ya no se factura.

1.2 Compra de un clúster

Para usar MRS, compre un clúster en la consola de MRS. El siguiente procedimiento toma MRS 3.2.0-LTS.1 como ejemplo para describir cómo crear un clúster en la consola de gestión de MRS. Las operaciones para otra versión son subjetivas a la interfaz de usuario.

Procedimiento

Paso 1 Vaya a la página **Comprar clúster**.

Paso 2 En la página para comprar un clúster, haga clic en la pestaña **Custom Config**.

NOTA

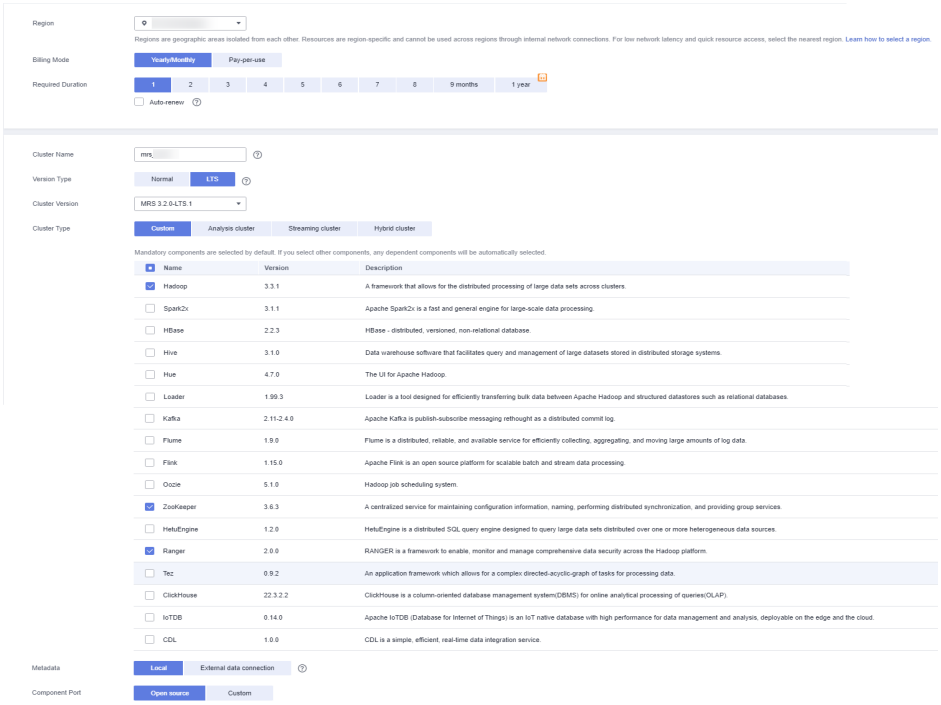
Al crear un clúster, preste atención a la notificación de cuota. Si una cuota de recursos es insuficiente, aumente la cuota de recursos según se le solicite y cree un clúster.

Paso 3 Configurar la información del software del clúster.

- **Region**: Conserve el valor predeterminado.
- **Billing Mode**: Conserve el valor predeterminado.
- **Required Duration**: Seleccione una duración según sea necesario.
- **Cluster Name**: Puede utilizar el nombre predeterminado. Sin embargo, se recomienda incluir una abreviatura de nombre de proyecto o fecha para la memoria consolidada y fácil de distinguir, por ejemplo, **mrs_20180321**.
- **Version Type**: **Normal** (predeterminado) o **LTS**
- **Cluster Version**: Seleccione la última versión, que es el valor predeterminado.
- **Cluster Type**: Conserve el valor **Analysis cluster**.
- **Component**: Seleccione componentes como Spark2x, HBase y Hive para el clúster de análisis. Para un clúster de streaming, seleccione componentes como Kafka y Storm. Para un clúster híbrido, puede seleccionar los componentes del clúster de análisis y del clúster de streaming en función de los requisitos de servicio.

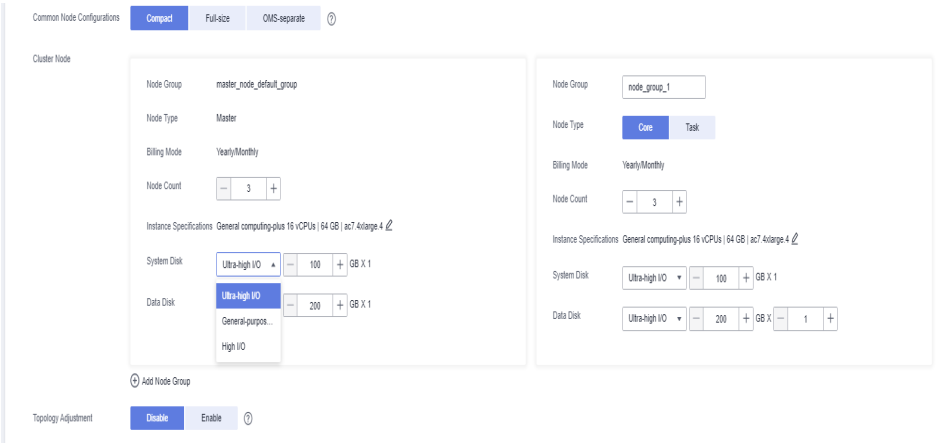
  NOTA

  Para versiones anteriores a MRS 3.*x*, seleccione componentes como Spark, HBase y Hive para un clúster de análisis.
- **Metadata**: Conserve el valor predeterminado. Este parámetro solo es compatible con MRS 3.*x*.
- Puerto de componente: Política para establecer el puerto de comunicación predeterminado de cada componente del clúster.



Paso 4 Haga clic en **Next**.

- **AZ**: Conservar el valor predeterminado.
- **Enterprise Project**: Seleccione **default**.
- **VPC**: Utilice el valor predeterminado. Si no hay una VPC disponible, haga clic en **View VPC** para acceder a la consola de VPC y crear una nueva VPC.
- **Subnet**: Conservar el valor predeterminado.
- **Security Group**: Conserve el valor predeterminado.
- **EIP**: Conserve el valor predeterminado.
- **Cluster Node**
  - **Node Count**: el número de nodos que desea comprar. Para los clústeres de MRS 3.x, el valor predeterminado es **3**. Puede establecer el valor según lo necesite.
  - **Instance Specifications**: Conservar la configuración predeterminada para los nodos principal y principal o seleccionar las especificaciones adecuadas en función de los requisitos de servicio.
  - **System Disk**: Conserve el **Ultra-high I/O** y la capacidad de almacenamiento predeterminados.
  - **Data Disk**: Conservar el **Ultra-high I/O**, la capacidad de almacenamiento y la cantidad predeterminados.



Paso 5 Haga clic en **Next**. Se muestra la página **Set Advanced Options**. Configure los siguientes parámetros. Conservar la configuración predeterminada para los demás parámetros.

- Autenticación de Kerberos:
  - **Kerberos Authentication**: Deshabilitar la autenticación de Kerberos.
  - **Username**: nombre del administrador del Manager. **admin** se utiliza de forma predeterminada.
  - **Password**: contraseña del administrador de Manager.
  - **Confirm Password**: Ingrese la contraseña de nuevo.
- **Login Mode**: Seleccionar un modo para iniciar sesión en un ECS.
  - **Password**: Establecer una contraseña para iniciar sesión en un ECS.
  - **Key Pair**: Seleccionar un par de claves de la lista desplegable. Seleccione **"I acknowledge that I have obtained private key file** *SSHkey-xxx* **and that without this file I will not be able to log in to my ECS."** Si nunca ha creado un par de claves, haga clic en **View Key Pair** para crear o importar un par de claves. Y luego, obtener un archivo de clave privada.
- **Hostname Prefix**: prefijo para el nombre de un ECS o BMS en el clúster.

  Introduzca un máximo de 20 caracteres que no comiencen o terminen con un guion (-). Solo se permiten letras, números y guiones (-).

  Cuando se crea un clúster, se registra un nombre de dominio DNS para los nodos del clúster. El nombre de dominio completo está en el siguiente formato: **[prefix]-hostname.mrs-{***XXXX***}.com**. (*XXXX* es una cadena de cuatro caracteres generada basada en el UUID.)
- **Set Advanced Options**: Para configurar algunos parámetros avanzados, seleccione **Configure**.

Paso 6 Haga clic en **Next**.

- **Configure**: Confirme los parámetros configurados en las áreas **Configure Software**, **Configure Hardware** y **Set Advanced Options**.
- **Secure Communications**: Seleccione **Enable**.

Paso 7 Haga clic en **Buy Now**.

Si la autenticación de Kerberos está habilitada para un clúster, compruebe si es necesaria la autenticación de Kerberos. En caso afirmativo, haga clic en **Continue**. Si no, haga clic en **Back** para deshabilitar la autenticación de Kerberos y, a continuación, cree un clúster.

Paso 8 Haga clic en **Back to Cluster List** para ver el estado del clúster.

Se necesita algún tiempo para crear un clúster. El estado inicial del clúster es **Starting**. Una vez que el clúster se ha creado correctamente, el estado del clúster pasa a ser **Running**.

----Fin

1.3 Carga de datos

En la página **Files**, puede crear y eliminar directorios HDFS, así como importar, exportar y eliminar archivos de un clúster de análisis.

Para los clústeres con autenticación Kerberos habilitada, sincronice a los usuarios de IAM antes de realizar operaciones en la página **Files**. En la página de detalles del clúster, haga clic en **Dashboard** y haga clic en **Synchronize** a la derecha de **IAM User Sync** para sincronizar usuarios de IAM.

Contexto

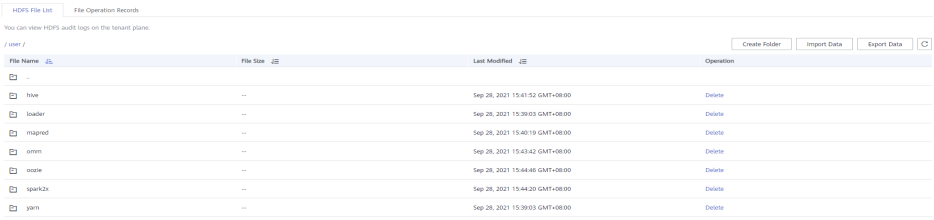
Los clústeres MRS generalmente procesan datos de OBS o HDFS. OBS le proporciona las capacidades de almacenamiento de datos que son masivas, seguras, confiables y rentables. MRS puede procesar datos directamente en OBS. Puede examinar, gestionar y usar datos tanto en la consola de gestión como en el cliente OBS. Si necesita importar datos OBS en el sistema HDFS del clúster para su procesamiento, realice los pasos de esta sección.

Importación de datos de OBS a HDFS

Actualmente, MRS solo admite la importación de datos de OBS a HDFS. La tasa de carga de archivos disminuye con el aumento del tamaño del archivo. Este modo se aplica a escenarios en los que el volumen de datos es pequeño.

Puede realizar los siguientes pasos para importar archivos y directorios:

1. Inicie sesión en la consola de MRS.
2. Seleccione **Clusters** > **Active Clusters** y haga clic en el nombre del clúster de destino para ingresar a la página de detalles del clúster.
3. Haga clic en **Files** para ir a la página de gestión de archivos.
4. Seleccione **HDFS File List**.



5. Vaya al directorio de almacenamiento de datos, por ejemplo, **bd_app1**.

   El directorio **bd_app1** es solo un ejemplo. Puede utilizar cualquier directorio de la página o crear uno nuevo.

   Los requisitos para crear una carpeta son los siguientes:

   - El nombre de la carpeta contiene un máximo de 255 caracteres.
   - El nombre de la carpeta no puede estar vacío.

- El nombre de la carpeta no puede contener las siguientes characters especiales: /:*?"<>|\;&,`'!{}[]$%+
- El valor no puede comenzar ni finalizar con un período (.).
- Los espacios al principio y al final se ignoran.

6. Haga clic en **Import Data** y configure las rutas de HDFS y OBS correctamente. Cuando configure la ruta de acceso OBS o HDFS, haga clic en **Browse**, seleccione un directorio de archivo y haga clic en **Yes**.

   Figura 1-1 Importación de datos de OBS a HDFS

   - **Ruta de OBS**
     - La ruta debe comenzar con **obs://**.
     - Los archivos o programas cifrados por KMS no se pueden importar.
     - No se puede importar una carpeta vacía.
     - El directorio y el nombre del archivo pueden contener letras, dígitos, guiones (-) y guiones bajos (_), pero no pueden contener caracteres especiales ;|&>,<'$*?\
     - El directorio y el nombre de archivo no pueden comenzar o terminar con un espacio, pero pueden contener espacios entre ellos.
     - La ruta de acceso completa de OBS contiene un máximo de 255 caracteres.
   - **Ruta de HDFS**
     - La ruta comienza por **/user** de forma predeterminada.
     - El directorio y el nombre del archivo pueden contener letras, dígitos, guiones (-) y guiones bajos (_), pero no pueden contener los siguientes caracteres especiales: ;|&>,<'$*?\:
     - El directorio y el nombre de archivo no pueden comenzar o terminar con un espacio, pero pueden contener espacios entre ellos.
     - La ruta de acceso completa de HDFS contiene un máximo de 255 caracteres.

7. Haga clic en **OK**.

   Puede ver el progreso de la carga de archivos en la página **File Operation Records**. MRS procesa la operación de importación de datos como un trabajo de DistCp. También puede comprobar si el trabajo DistCp se ejecuta correctamente en la página **Jobs**.

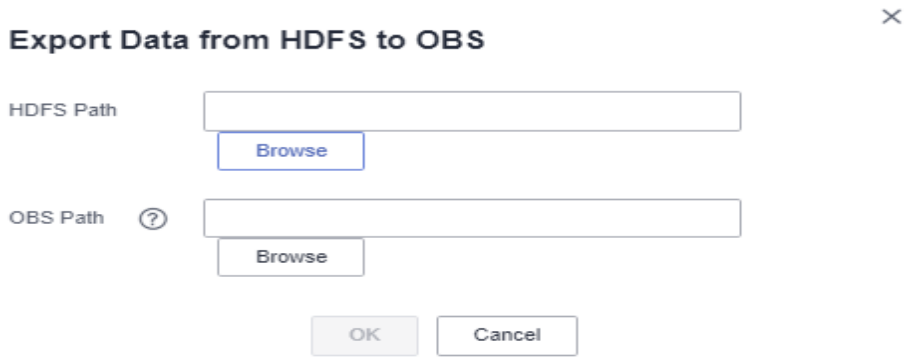
Exportación de datos de HDFS a OBS

Después de completar el análisis de datos y la computación, puede almacenar los datos en el HDFS o exportarlos a OBS.

Puede realizar los siguientes pasos para exportar archivos y directorios:

1. Inicie sesión en la consola de MRS.
2. Seleccione **Clusters** > **Active Clusters** y haga clic en el nombre del clúster de destino para ingresar a la página de detalles del clúster.
3. Haga clic en **Files** para ir a la página de gestión de archivos.
4. Seleccione **HDFS File List**.
5. Vaya al directorio de almacenamiento de datos, por ejemplo, **bd_app1**.
6. Haga clic en **Export Data** y configure las rutas OBS y HDFS. Cuando configure la ruta de acceso OBS o HDFS, haga clic en **Browse**, seleccione un directorio de archivo y haga clic en **Yes**.

Figura 1-2 Exportación de datos de HDFS a OBS



- **Ruta de OBS**
  - La ruta debe comenzar con **obs://**.
  - El directorio y el nombre del archivo pueden contener letras, dígitos, guiones (-) y guiones bajos (_), pero no pueden contener caracteres especiales ;| &>,<'$*?\
  - El directorio y el nombre de archivo no pueden comenzar o terminar con un espacio, pero pueden contener espacios entre ellos.
  - La ruta de acceso completa de OBS contiene un máximo de 255 caracteres.
- **Ruta de HDFS**
  - La ruta comienza por **/user** de forma predeterminada.
  - El directorio y el nombre del archivo pueden contener letras, dígitos, guiones (-) y guiones bajos (_), pero no pueden contener los siguientes caracteres especiales: ;|&>,<'$*?\:
  - El directorio y el nombre de archivo no pueden comenzar o terminar con un espacio, pero pueden contener espacios entre ellos.
  - La ruta de acceso completa de HDFS contiene un máximo de 255 caracteres.

NOTA

Cuando se exporta una carpeta a OBS, se agrega un archivo de etiquetas denominado **folder name_$folder$** a la ruta de acceso de OBS. Asegúrese de que la carpeta exportada no está vacía. Si la carpeta exportada está vacía, OBS no puede mostrarla y solo genera un archivo denominado **folder name_$folder$**.

7. Haga clic en **OK**.

   Puede ver el progreso de la carga de archivos en la página **File Operation Records**. MRS procesa la operación de exportación de datos como un trabajo DistCp. También puede comprobar si el trabajo DistCp se ejecuta correctamente en la página **Jobs**.

1.4 Creación de un trabajo

Puede enviar programas desarrollados por usted mismo a MRS para ejecutarlos y obtener los resultados.

Esta sección describe cómo enviar un trabajo (tome un trabajo MapReduce como ejemplo) en la consola MRS. Los trabajos de MapReduce se utilizan para enviar programas JAR para procesar rápidamente cantidades masivas de datos en paralelo y crear un entorno de procesamiento y ejecución de datos distribuidos.

Si las funciones de gestión de archivos y trabajos no se admiten en la página de detalles del clúster, envíe los trabajos en segundo plano.

Antes de crear un trabajo, debe cargar datos locales en OBS para la computación y análisis de datos. MRS permite exportar datos de OBS a HDFS para computación y análisis. Después de completar el análisis de datos y la computación, puede almacenar los datos en HDFS o exportarlos a OBS. HDFS y OBS también pueden almacenar los datos comprimidos en el formato **bz2** o **gz**.

NOTA

Si el nombre de usuario de IAM contiene espacios (por ejemplo, **admin 01**), no se puede crear un trabajo.

Enviar un trabajo en la GUI

Paso 1 Inicie sesión en la consola de MRS.

Paso 2 Seleccione **Clusters** > **Active Clusters**, seleccione un clúster en ejecución y haga clic en su nombre para acceder a la página de detalles del clúster.

Paso 3 Si la autenticación de Kerberos está habilitada para el clúster, realice los siguientes pasos. Si la autenticación de Kerberos no está habilitada para el clúster, omita este paso.

En el área **Basic Information** de la página **Dashboard**, haga clic en **Synchronize** en el lado derecho de **IAM User Sync** para sincronizar usuarios de IAM.

Paso 4 Haga clic en la pestaña **Jobs**.

Paso 5 Haga clic en **Create**. Aparece el cuadro de diálogo **Create Job**.

Paso 6 En **Type**, seleccione **MapReduce**. Configurar otra información del trabajo.

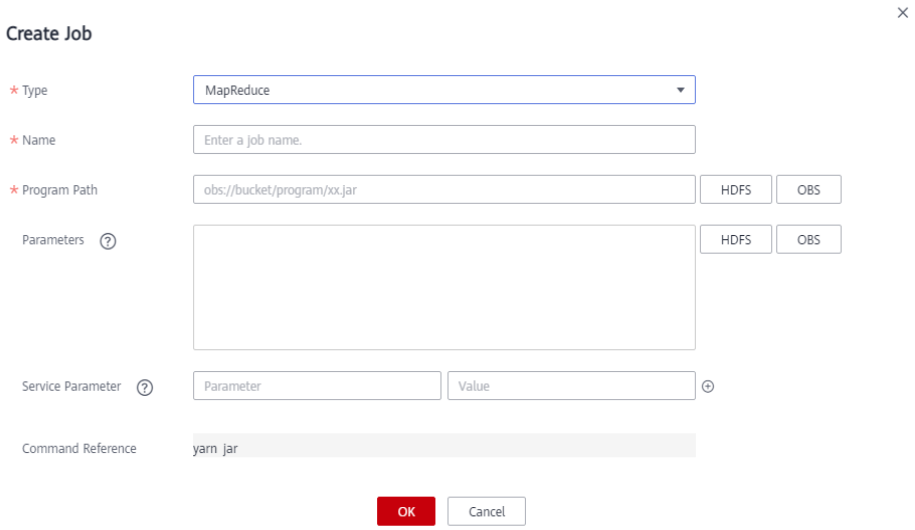


Tabla 1-1 Parámetros de trabajo

Parámetro	Descripción
Name	Nombre del trabajo. Contiene de 1 a 64 caracteres. Solo se permiten letras, dígitos, guiones medios (-) y guiones bajos (_). **NOTA** Se recomienda establecer diferentes nombres para diferentes trabajos.
Program Path	Ruta del paquete de programa que se va a ejecutar. Se deben cumplir los siguientes requisitos: ● Contiene un máximo de 1,023 caracteres, excluidos caracteres especiales como **;|&><'$**. El valor del parámetro no puede estar vacío ni lleno de espacios. ● La ruta del programa a ejecutar se puede almacenar en HDFS u OBS. La ruta de acceso varía según el sistema de archivos. – OBS: La ruta comienza con **obs://**. Ejemplo: **obs://wordcount/program/*xxx*.jar** – HDFS: La ruta debe comenzar con **/user**. ● Para SparkScript y HiveScript, el camino debe terminar con **.sql**. En el caso de MapReduce, la ruta debe terminar con **.jar**. Para Flink y SparkSubmit la ruta debe terminar con **.jar** o **.py**. El **.sql**, **.jar** y el **.py** no distinguen entre mayúsculas y minúsculas.

Parámetro	Descripción
Parameters	(Opcional) Es el parámetro clave para la ejecución del programa. Separe múltiples parámetros con espacio. Método de configuración: *Program class name Data input path Data output path* ● Nombre de la clase del programa: Es especificada por una función en su programa. MRS es responsable de la transferencia de parámetros solamente. ● Ruta de entrada de datos: Haga clic en **HDFS** o **OBS** para seleccionar una ruta o introduzca manualmente una ruta correcta. ● Ruta de salida de datos: Ingrese un directorio que no existe. El valor puede contener un máximo de 150,000 caracteres, incluidos caracteres especiales (;|&'$), pero no puede contener > ni <. Este parámetro también se puede dejar en blanco. **ATENCIÓN** Si introduce un parámetro con información confidencial (como la contraseña de inicio de sesión), el parámetro puede estar expuesto en la pantalla de detalles del trabajo y en la impresión del registro. Tenga cuidado al realizar esta operación.
Service Parameters	(Opcional) Se utiliza para modificar los parámetros de configuración del servicio para el trabajo que se va a ejecutar. La modificación del parámetro solo se aplica al trabajo que se va a ejecutar. Para agregar varios parámetros, haga clic en ⊕ a la derecha. Para eliminar un parámetro, haga clic en **Delete** a la derecha. **Tabla 1-2** describe los parámetros comunes de un servicio.
Command Reference	Comando enviado en segundo plano para su ejecución cuando se envía un trabajo.

Tabla 1-2 Parámetros de configuración del servicio

Parámetro	Descripción	Valor de ejemplo
fs.obs.access.key	ID de clave para acceder a OBS.	-
fs.obs.secret.key	Clave correspondiente al ID de clave para acceder a OBS.	-

Paso 7 Confirme la información de configuración del trabajo y haga clic en **OK**.

Después de crear el trabajo, puede gestionarlo.

----Fin

1.5 Terminación de un clúster

Puede terminar un clúster MRS que ya no se utiliza una vez completada la ejecución del trabajo. El clúster terminado o cancelado ya no se factura.

Contexto

Por lo general, después de analizar y almacenar los datos, o cuando el clúster encuentra una excepción y no puede funcionar, puede terminar un clúster. Un clúster que no se despliega se terminará automáticamente.

Procedimiento

Paso 1 Inicie sesión en la consola de gestión de MRS.

Paso 2 En el panel de navegación de la izquierda, elija **Clusters** > **Active Clusters**.

Paso 3 En la lista de clústeres, busque la fila que contiene el clúster que se va a terminar y haga clic en **Terminate** en la columna **Operation**.

El estado del clúster cambia de **Running** a **Terminating** y finalmente a **Terminated**. Puede ver el clúster terminado de **Cluster History**. El clúster terminado ya no se factura.

----Fin

2 Instalación y uso del cliente de clúster

Instale y use rápidamente los clientes de todos los servicios en un clúster MRS 3.*x* o posterior.

Los clientes se pueden instalar en los nodos dentro o fuera del clúster. A continuación se proporciona un ejemplo de cómo instalar y usar un cliente en un clúster.

NOTA

Si Flume se ha instalado en el clúster, el cliente de Flume debe instalarse de forma independiente. Para obtener detalles sobre cómo instalar el cliente Flume, consulte **Instalación de cliente de Flume**.

Puede comenzar leyendo los siguientes temas:

1. **Descargar un cliente**
2. **Instalación de un cliente**
3. **Uso de un cliente**

Vídeo Tutorial

Este vídeo utiliza un clúster MRS 3.1.0 como ejemplo para describir cómo instalar y usar el cliente de clúster después de crear un clúster. Para obtener más información, consulte **Instalación y uso de cliente de MRS**.

NOTA

La interfaz de usuario puede variar dependiendo de la versión. El video tutorial es solo para referencia.

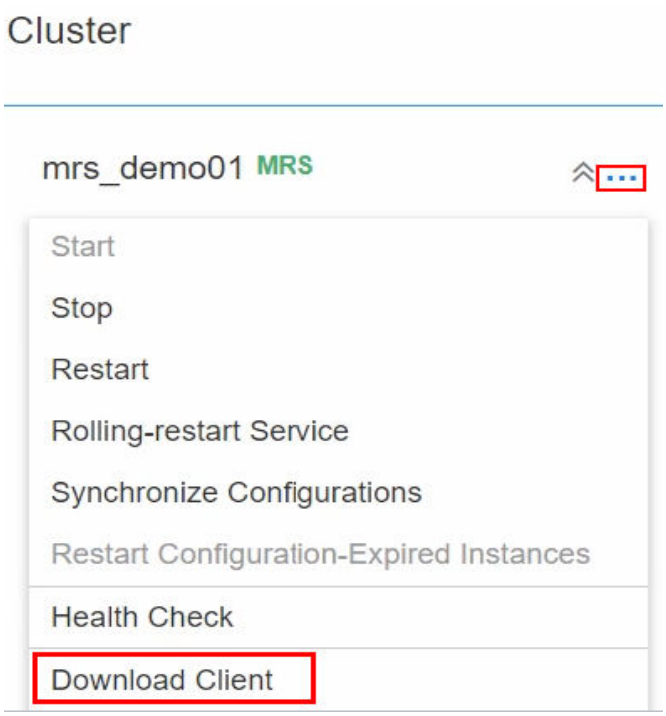
Descargar un cliente

Paso 1 Inicie sesión en FusionInsight Manager del clúster haciendo referencia a **Acceder a FusionInsight Manager (MRS 3.x o posterior)**.

Paso 2 Descargue el paquete de software del cliente de clúster en el nodo de destino.

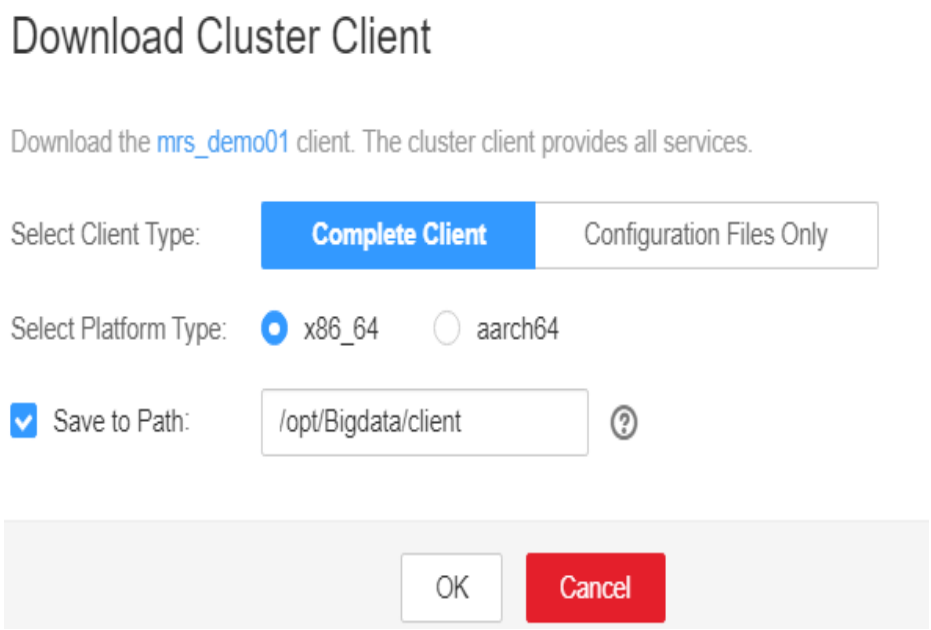
En la página principal, haga clic en ··· junto al nombre del clúster y haga clic en **Download Client** para descargar el cliente del clúster.

Figura 2-1 Descargar un cliente



Paso 3 En la página **Download Cluster Client**, ingrese la información de descarga del cliente de clúster.

Figura 2-2 Descargar el cliente de clúster



- Establezca **Select Client Type** en **Complete Client**.
- Establezca **Select Platform Type** en la arquitectura del nodo para instalar el cliente. **x86_64** se utiliza como ejemplo.

- Seleccione **Save to Path** e introduzca la ruta de descarga, por ejemplo, **/opt/Bigdata/client**. Asegúrese de que el usuario **omm** tiene el permiso de operación en la ruta de acceso.

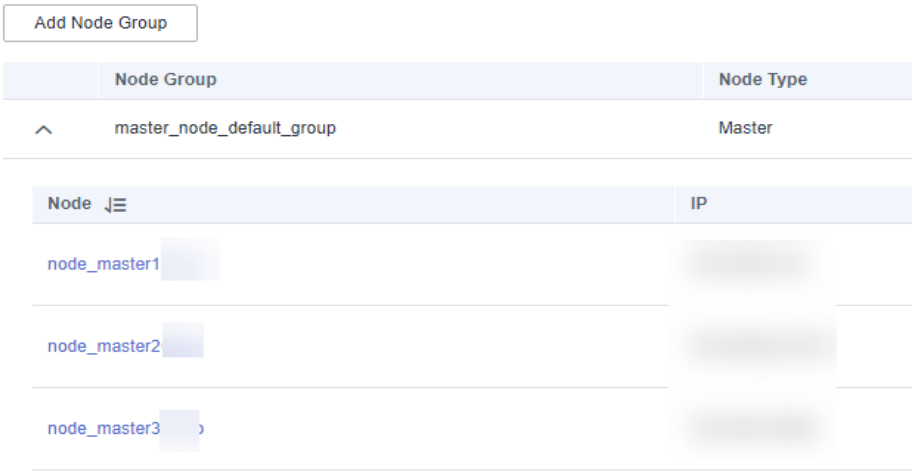
  NOTA

  El clúster admite dos tipos de clientes: **x86_64** y **aarch64**. El tipo de cliente debe coincidir con la arquitectura del nodo para instalar el cliente. De lo contrario, la instalación del cliente fallará.

Paso 4 Una vez descargado el paquete de software cliente, inicie sesión en el nodo OMS activo del clúster como usuario **root**.

De forma predeterminada, el paquete de software cliente se descarga en el nodo OMS activo del clúster. Puede ver el nodo marcado con ★ en la página host del FusionInsight Manager. Si necesita instalar el paquete de software cliente en otro nodo del clúster, ejecute el siguiente comando para transferir el paquete de software al nodo de destino.

En la lista de clústeres de la consola MRS, haga clic en el nombre del clúster. En la página **Nodes**, haga clic en el nombre del nodo de destino. En la página de detalles de ECS, puede iniciar sesión de forma remota en este nodo.



scp -p /opt/Bigdata/client/FusionInsight_Cluster_1_Services_Client.tar *IP address of the node where the client is to be installed:/opt/Bigdata/client*

----Fin

Instalación de un cliente

Paso 1 Inicie sesión en el nodo donde está instalado el paquete de software cliente como usuario cliente (por ejemplo, usuario **root**) y ejecute los siguientes comandos para descomprimir el paquete de software:

cd /opt/Bigdata/client

tar -xvf FusionInsight_Cluster_1_Services_Client.tar

Paso 2 Ejecute el comando **sha256sum** para verificar el archivo descomprimido.

sha256sum -c FusionInsight_Cluster_1_Services_ClientConfig.tar.sha256

```
FusionInsight_Cluster_1_Services_Client.tar: OK
```


Paso 3 Descomprima el archivo de instalación obtenido.

tar -xvf FusionInsight_Cluster_1_Services_ClientConfig.tar

Paso 4 Vaya al directorio donde está almacenado el paquete de instalación e instale el cliente.

cd /opt/Bigdata/client/FusionInsight_Cluster_1_Services_ClientConfig

Ejecute el siguiente comando para instalar el cliente en un directorio especificado (una ruta absoluta), por ejemplo, **/opt/hadoopclient**.

./install.sh /opt/hadoopclient

```
...
The component client is installed successfully
```

NOTA

- Si los clientes de servicio existentes han utilizado el directorio **/opt/hadoopclient**, debe utilizar otro directorio en este paso al instalar otros clientes de servicio.
- Debe eliminar el directorio de instalación del cliente al desinstalar un cliente.
- Si desea evitar que otros usuarios accedan a este cliente, agregue el parámetro **-o** durante la instalación. Es decir, ejecute el comando **./install.sh /opt/hadoopclient -o** para instalar el cliente.
- Si se instala un cliente HBase, se recomienda que el directorio de instalación del cliente contenga solo letras mayúsculas y minúsculas, dígitos y caracteres especiales (_-?.@+=) debido a la limitación de la sintaxis Ruby utilizada por HBase.
- Si el servidor NTP se va a instalar en modo **chrony**, asegúrese de que el parámetro **chrony** se agrega durante la instalación, es decir, ejecute el comando **./install.sh /opt/hadoopclient -o chrony** para instalar el cliente.

----Fin

Uso de un cliente

Paso 1 Inicie sesión en el nodo donde está instalado el cliente como usuario de instalación del cliente y ejecute el siguiente comando para cambiar al directorio del cliente:

cd /opt/hadoopclient

Paso 2 Ejecute el siguiente comando para cargar variables de entorno:

source bigdata_env

Paso 3 Si la autenticación de Kerberos está habilitada para el clúster actual, ejecute el siguiente comando para autenticar al usuario. Si la autenticación de Kerberos está deshabilitada para el clúster actual, la autenticación no es necesaria.

kinit *MRS cluster user*

Por ejemplo:

kinit admin

Paso 4 Ejecute el comando de cliente de un componente directamente.

Por ejemplo:

Ejecute el siguiente comando para ver los archivos en el directorio raíz de HDFS:

hdfs dfs -ls /


```
Found 15 items
drwxrwx--x   - hive       hive                0 2021-10-26 16:30 /apps
drwxr-xr-x   - hdfs       hadoop              0 2021-10-18 20:54 /datasets
drwxr-xr-x   - hdfs       hadoop              0 2021-10-18 20:54 /datastore
drwxrwx---+  - flink      hadoop              0 2021-10-18 21:10 /flink
drwxr-x---   - flume      hadoop              0 2021-10-18 20:54 /flume
drwxrwx--x   - hbase      hadoop              0 2021-10-30 07:31 /hbase
...
```

----Fin

3 Uso de clústeres con autenticación Kerberos habilitada

Utilice clústeres de seguridad y ejecute los programas MapReduce, Spark y Hive.

En MRS 3.*x*, Presto no admite la autenticación Kerberos.

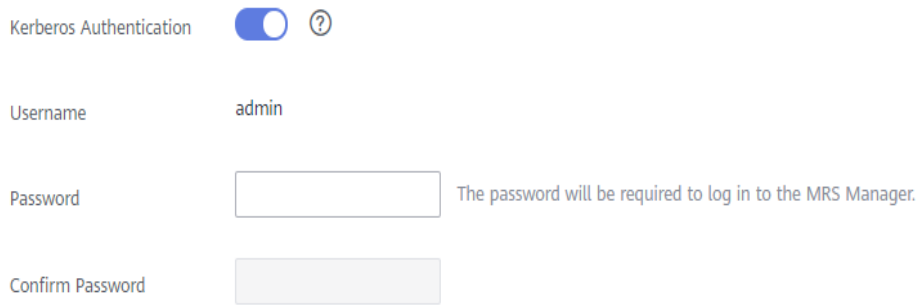
Puede comenzar leyendo los siguientes temas:

1. **Creación de un clúster de seguridad e inicio de sesión en Manager**
2. **Creación de un rol y un usuario**
3. **Ejecución de un programa de MapReduce**
4. **Ejecución de un programa Spark**
5. **Ejecución de un programa Hive**

Creación de un clúster de seguridad e inicio de sesión en Manager

Paso 1 Cree un clúster de seguridad. Para obtener más información, consulte **Compra de un clúster personalizado**. Habilite **Kerberos Authentication** y configure **Password** y confirme la contraseña. Esta contraseña se utiliza para iniciar sesión en Manager. Manténgalo seguro.

Figura 3-1 Configuración de los parámetros del clúster de seguridad



Paso 2 Inicie sesión en la consola de MRS.

Paso 3 En el panel de navegación de la izquierda, elija **Active Clusters** y haga clic en el nombre del clúster de destino de la derecha para acceder a la página de detalles del clúster.

Paso 4 Haga clic en **Access Manager** a la derecha de **MRS Manager** para iniciar sesión en Manager.

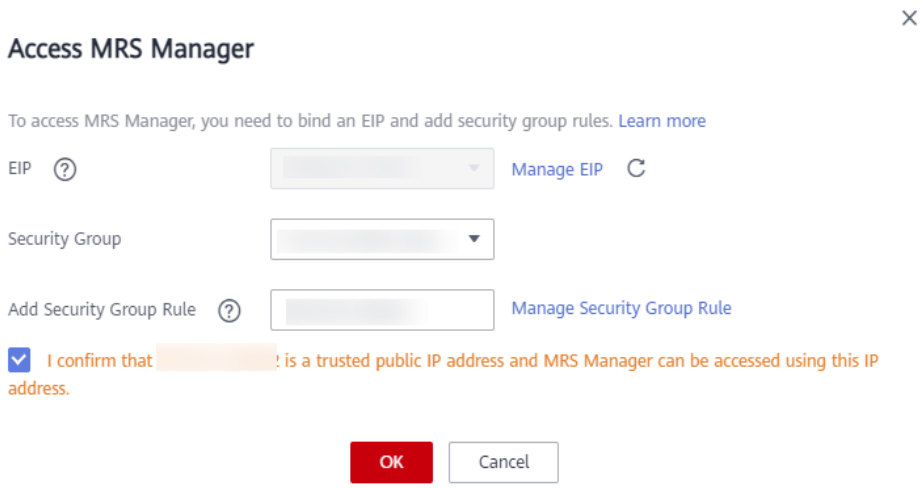
- Si ha enlazado una EIP al crear el clúster, realice las siguientes operaciones:
  a. Agregue una regla de grupo de seguridad. De forma predeterminada, la dirección IP pública utilizada para acceder al puerto 9022 se completa en la regla. Si desea ver, modificar o eliminar una regla de grupo de seguridad, haga clic en **Manage Security Group Rule**.

     NOTA

     - Es normal que la dirección IP pública generada automáticamente sea diferente de su dirección IP local y no se requiere ninguna acción.
     - Si el puerto 9022 es un puerto Knox, debe habilitar el permiso para acceder al puerto 9022 de Knox para acceder al Manager.

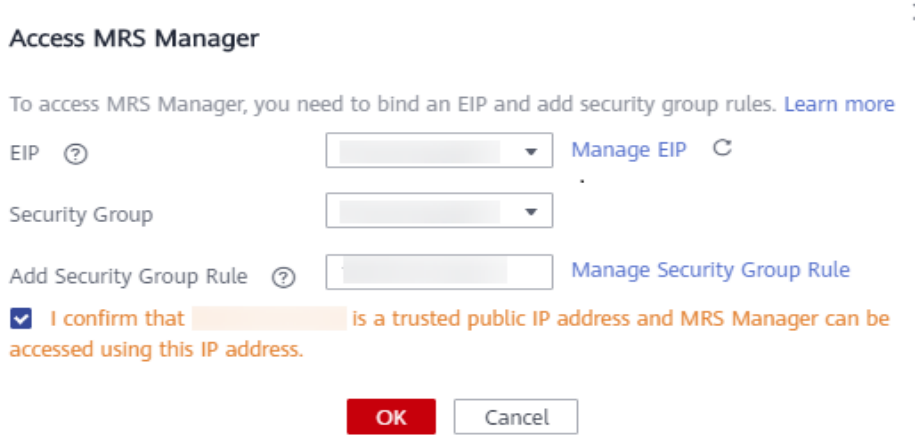
  b. Seleccione **I confirm that xx.xx.xx.xx is a trusted public IP address and MRS Manager can be accessed using this IP address.**

  Figura 3-2 Acceder a Manager



- Si no ha enlazado una EIP al crear el clúster, realice las siguientes operaciones:
  a. Seleccione un EIP en la lista desplegable o haga clic en **Manage EIP** para comprar una.
  b. Agregue una regla de grupo de seguridad. De forma predeterminada, la dirección IP pública utilizada para acceder al puerto 9022 se completa en la regla. Si desea ver, modificar o eliminar una regla de grupo de seguridad, haga clic en **Manage Security Group Rule**.
  c. Seleccione **I confirm that xx.xx.xx.xx is a trusted public IP address and MRS Manager can be accessed using this IP address.**

Figura 3-3 Acceder a Manager



Paso 5 Haga clic en **OK**. Se muestra la página de inicio de sesión de Manager. Para asignar a otros usuarios el permiso de acceso al Manager, agregue las direcciones IP como de confianza haciendo referencia a **Acceder a Manager**.

Paso 6 Ingrese el nombre de usuario predeterminado **admin** y la contraseña que configuró al crear el clúster, y haga clic en **Log In**.

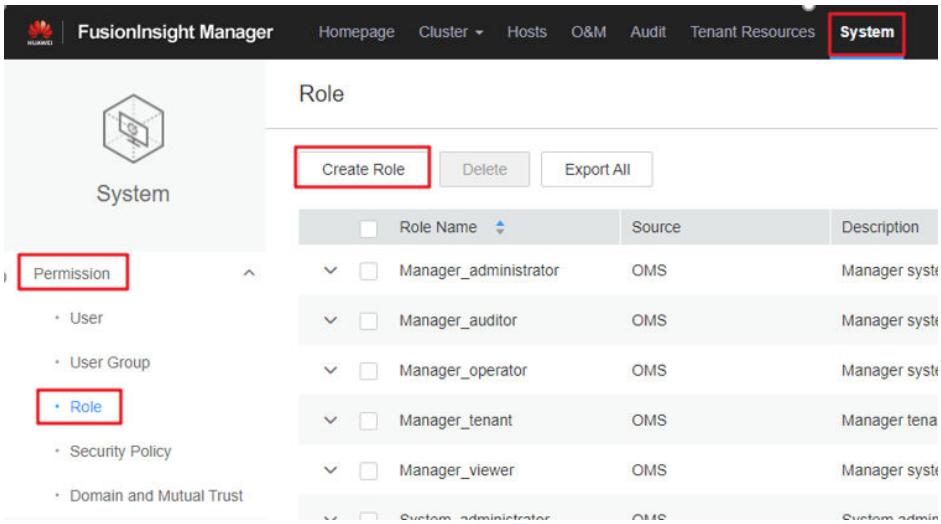
----Fin

Creación de un rol y un usuario

Para clústeres con autenticación Kerberos habilitada, realice los siguientes pasos para crear un usuario y asignar permisos al usuario para ejecutar programas.

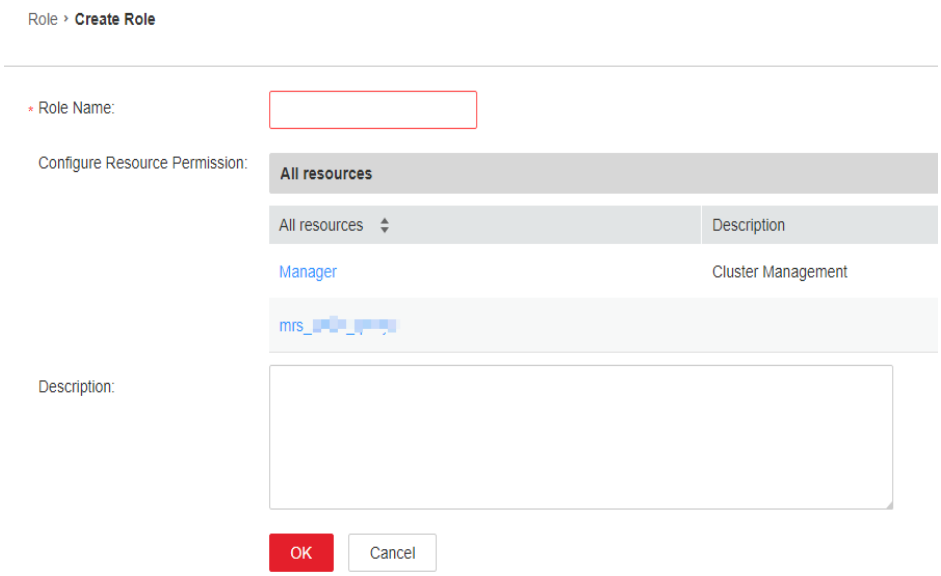
Paso 1 En Manager, elija **System** > **Permission** > **Role**.

Figura 3-4 Rol



Paso 2 Haga clic en **Create Role**. Para obtener más información, consulte **Creación de un rol**.

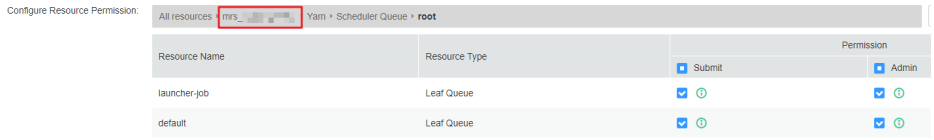
Figura 3-5 Creación de un rol



Especifique la siguiente información:

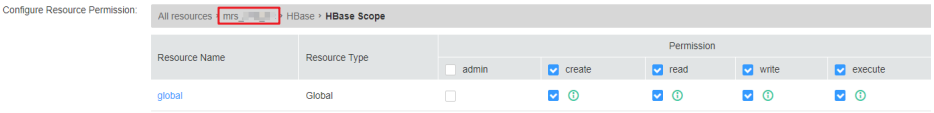
- Introduzca un nombre de rol, por ejemplo, **mrrole**.
- En **Configure Resource Permission**, seleccione el clúster que se va a operar, elija **Yarn** > **Scheduler Queue** > **root** y seleccione **Submit** y **Admin** en la columna **Permission**. Después de finalizar la configuración, no haga clic en **OK** sino en el nombre del clúster de destino que se muestra en la siguiente figura y, a continuación, configure otros permisos.

  Figura 3-6 Configuración de permisos de recursos para Yarn

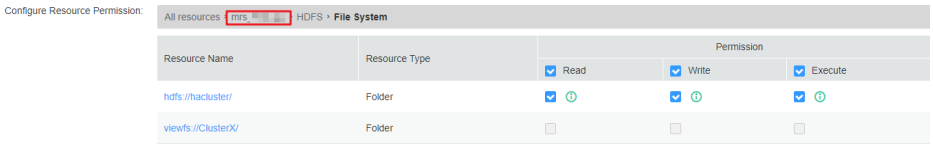
- Elija **HBase** > **HBase Scope**. Busque la fila que contiene **global** y seleccione **create**, **read**, **write** y **execute** en la columna **Permission**. Después de finalizar la configuración, no haga clic en **OK** sino en el nombre del clúster de destino que se muestra en la siguiente figura y, a continuación, configure otros permisos.

  Figura 3-7 Configuración de permisos de recursos para HBase

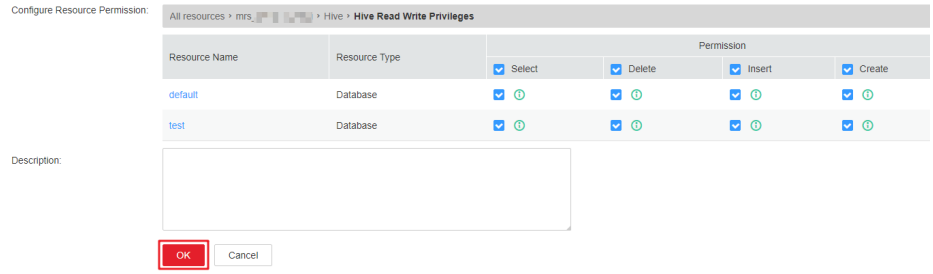
- Elija **HDFS** > **File System** > **hdfs://hacluster/** y seleccione **Read**, **Write** y **Execute** en la columna **Permission**. Después de finalizar la configuración, no haga clic en **OK** sino en el nombre del clúster de destino que se muestra en la siguiente figura y, a continuación, configure otros permisos.

Figura 3-8 Configuración de permisos de recursos para HDFS



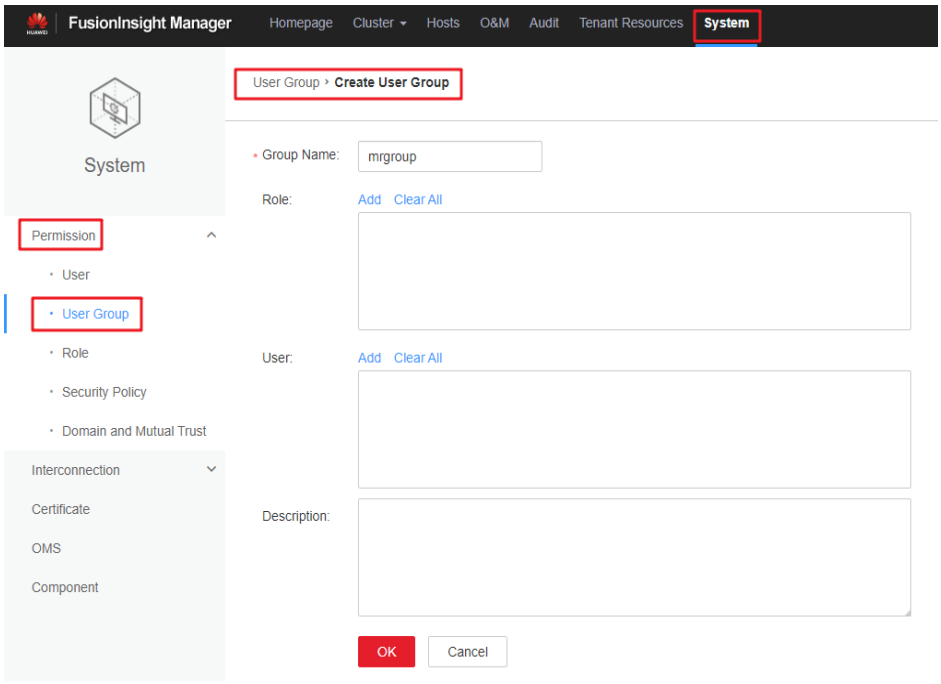
- Elija **Hive** > **Hive Read Write Privileges**, seleccione **Select**, **Delete**, **Insert**, y **Create** en la columna **Permission**, y haga cic en **OK**.

Figura 3-9 Configuración de permisos de recursos para Hive



Paso 3 Elija **System**. En el panel de navegación de la izquierda, elija **Permission** > **User Group** > **Create User Group** para crear un grupo de usuarios para el proyecto de ejemplo, por ejemplo, **mrgroup**. Para obtener más información, consulte **Creación de un grupo de usuario**.

Figura 3-10 Crear un grupo de usuarios



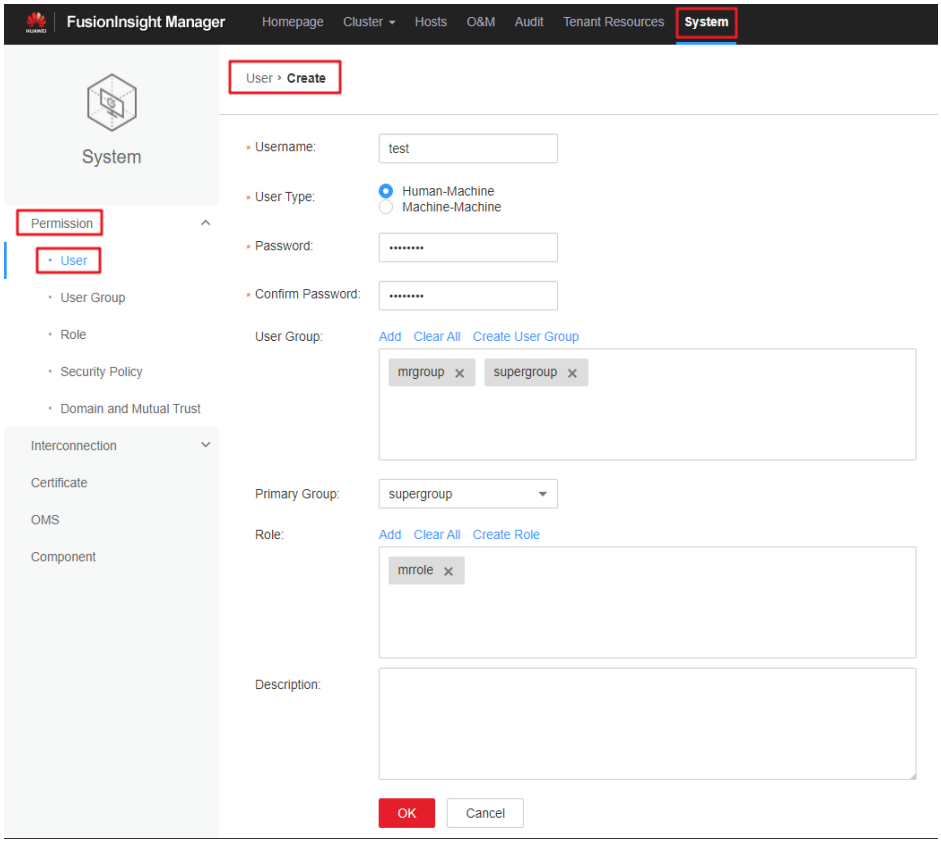
Paso 4 Elija **System**. En el panel de navegación de la izquierda, elija **Permission** > **User** > **Create** para crear un usuario para el proyecto de ejemplo. Para obtener más información, consulte **Creación de un usuario**.

- Introduzca un nombre de usuario, por ejemplo, **test**. Si desea ejecutar un programa Hive, escriba **hiveuser** en el archivo **Username**.

- Ajusta **User Type** a **Human-Machine**.
- Ingrese una contraseña. Esta contraseña se usará cuando ejecute el programa.
- En **User Group**, agregue **mrgroup** y **supergroup**.
- Establezca **Primary Group** en **supergroup** y vincule el rol **mrrole** para obtener el permiso.

  Haga clic en **OK**.

Figura 3-11 Creación de un usuario



Paso 5 Elija **System**. En el panel de navegación de la izquierda, elija **Permission** > **User**, busque la fila donde se encuentra el usuario **test** y seleccione **Download Authentication Credential** en la lista desplegable **More**. Guarde el paquete descargado y descomprima para obtener los archivos **keytab** y **krb5.conf**.

Figura 3-12 Descargar la credencial de autenticación

----Fin

Ejecución de un programa de MapReduce

Esta sección describe cómo ejecutar un programa de MapReduce en modo de clúster de seguridad.

Prerrequisitos

Ha compilado el programa y preparado archivos de datos, por ejemplo, **mapreduce-examples-1.0.jar**, **input_data1.txt**, y **input_data2.txt**. Para obtener más información sobre el desarrollo del programa MapReduce y la preparación de datos, consulte **Introducción de MapReduce**.

Procedimiento

Paso 1 Utilice un software de inicio de sesión remoto (por ejemplo, MobaXterm) para iniciar sesión en el nodo master del clúster de seguridad mediante SSH (usando la EIP).

Paso 2 Después de iniciar sesión correctamente, ejecute los siguientes comandos para crear la carpeta **test** en el directorio **/opt/Bigdata/client** y crear la carpeta **conf** en el directorio **test**:

```
cd /opt/Bigdata/client
mkdir test
cd test
mkdir conf
```

Paso 3 Utilice una herramienta de carga (por ejemplo, WinSCP) para copiar **mapreduce-examples-1.0.jar**, **input_data1.txt**, y **input_data2.txt** al directorio **test**, y copie los archivos **keytab** y **krb5.conf** obtenidos en **Paso 5** en **Creating Roles and Users** al directorio **conf**.

Paso 4 Ejecute los siguientes comandos para configurar variables de entorno y autenticar el usuario creado, por ejemplo **test**:

```
cd /opt/Bigdata/client
source bigdata_env
export YARN_USER_CLASSPATH=/opt/Bigdata/client/test/conf/
kinit test
```

Ingrese la contraseña como se le solicite. Si no se muestra ningún mensaje de error (necesita cambiar la contraseña como se le solicite en el primer inicio de sesión) , la autenticación Kerberos se ha completado.

Paso 5 Ejecute los siguientes comandos para importar datos a HDFS:

```
cd test
hdfs dfs -mkdir /tmp/input
hdfs dfs -put input_data* /tmp/input
```

Paso 6 Ejecute los siguientes comandos para ejecutar el programa:

```
yarn jar mapreduce-examples-1.0.jar
com.huawei.bigdata.mapreduce.examples.FemaleInfoCollector /tmp/input /tmp/
mapreduce_output
```

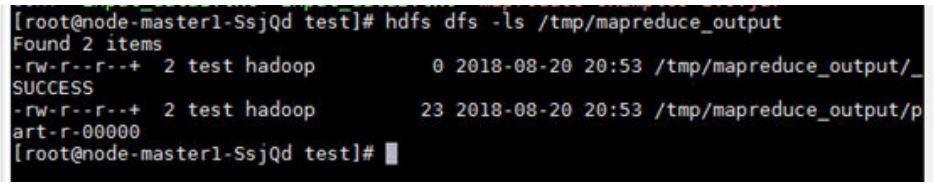
En los comandos anteriores:

/tmp/input indica la ruta de entrada en el HDFS.

/tmp/mapreduce_output indica la ruta de salida en el HDFS. Este directorio no debe existir. De lo contrario, se informará de un error.

Paso 7 Una vez que el programa se ejecute correctamente, ejecute el comando **hdfs dfs -ls /tmp/mapreduce_output**. Se muestra el siguiente resultado del comando.

Figura 3-13 Resultado de ejecución del programa



----Fin

Ejecución de un programa Spark

En esta sección se describe cómo ejecutar un programa Spark en modo de clúster de seguridad.

Prerrequisitos

Ha compilado el programa y preparado archivos de datos, por ejemplo, **FemaleInfoCollection.jar**, **input_data1.txt**, y **input_data2.txt**. Para obtener detalles sobre el desarrollo del programa de Spark y la preparación de datos, consulte **Descripción de desarrollo de aplicación de Spark**.

Procedimiento

Paso 1 Utilice un software de inicio de sesión remoto (por ejemplo, MobaXterm) para iniciar sesión en el nodo master del clúster de seguridad mediante SSH (usando la EIP).

Paso 2 Después de iniciar sesión correctamente, ejecute los siguientes comandos para crear la carpeta **test** en el directorio **/opt/Bigdata/client** y crear la carpeta **conf** en el directorio **test**:

```
cd /opt/Bigdata/client
mkdir test
cd test
mkdir conf
```

Paso 3 Utilice una herramienta de carga (por ejemplo, WinSCP) para copiar **FemaleInfoCollection.jar**, **input_data1.txt**, y **input_data2.txt** al directorio **test**, y copie los archivos **keytab** y **krb5.conf** obtenidos en **Paso 5** en la sección **Creating Roles and Users** al directorio **conf**.

Paso 4 Ejecute los siguientes comandos para configurar variables de entorno y autenticar el usuario creado, por ejemplo **test**:

```
cd /opt/Bigdata/client
source bigdata_env
export YARN_USER_CLASSPATH=/opt/Bigdata/client/test/conf/
kinit test
```

Ingrese la contraseña como se le solicite. Si no se muestra ningún mensaje de error, se completa la autenticación Kerberos.

Paso 5 Ejecute los siguientes comandos para importar datos a HDFS:

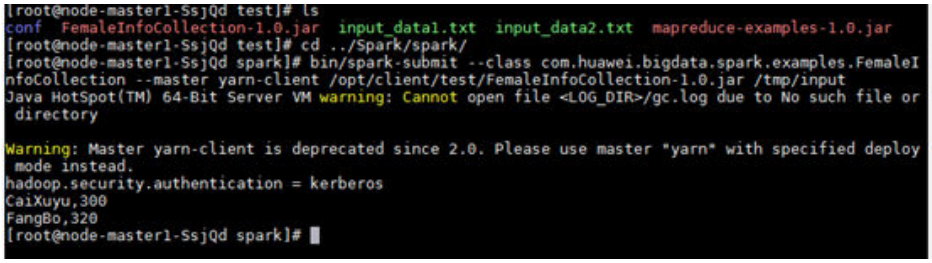
```
cd test
hdfs dfs -mkdir /tmp/input
hdfs dfs -put input_data* /tmp/input
```

Paso 6 Ejecute los siguientes comandos para ejecutar el programa:

```
cd /opt/Bigdata/client/Spark/spark
bin/spark-submit --class com.huawei.bigdata.spark.examples.FemaleInfoCollection --
master yarn-client /opt/Bigdata/client/test/FemaleInfoCollection-1.0.jar /tmp/
input
```


Paso 7 Una vez que el programa se ejecuta correctamente, se muestra la siguiente información.

Figura 3-14 Resultado de ejecución del programa



----Fin

Ejecución de un programa Hive

En esta sección se describe cómo ejecutar un programa Hive en modo de clúster de seguridad.

Prerrequisitos

Ha compilado el programa y preparado archivos de datos, por ejemplo, **hive-examples-1.0.jar**, **input_data1.txt**, y **input_data2.txt**. Para obtener detalles sobre el desarrollo del programa Hive y la preparación de datos, consulte **Descripción de desarrollo de aplicación de Hive**.

Procedimiento

Paso 1 Utilice un software de inicio de sesión remoto (por ejemplo, MobaXterm) para iniciar sesión en el nodo master del clúster de seguridad mediante SSH (usando la EIP).

Paso 2 Después de iniciar sesión correctamente, ejecute los siguientes comandos para crear la carpeta **test** en el directorio **/opt/Bigdata/client** y crear la carpeta **conf** en el directorio **test**:

```
cd /opt/Bigdata/client
mkdir test
cd test
mkdir conf
```

Paso 3 Utilice una herramienta de carga (por ejemplo, WinSCP) para copiar **FemaleInfoCollection.jar**, **input_data1.txt**, y **input_data2.txt** al directorio **test**, y copie los archivos **keytab** y **krb5.conf** obtenidos en **Paso 5** en la sección **Creating Roles and Users** al directorio **conf**.

Paso 4 Ejecute los siguientes comandos para configurar variables de entorno y autenticar el usuario creado, por ejemplo **test**:

```
cd /opt/Bigdata/client
source bigdata_env
export YARN_USER_CLASSPATH=/opt/Bigdata/client/test/conf/
kinit test
```

Ingrese la contraseña como se le solicite. Si no se muestra ningún mensaje de error, se completa la autenticación Kerberos.

Paso 5 Ejecute el siguiente comando para ejecutar el programa:

```
chmod +x /opt/hive_examples -R   cd /opt/hive_examples   java -cp .:hive-
examples-1.0.jar:/opt/hive_examples/conf:/opt/Bigdata/client/Hive/
Beeline/lib/*:/opt/Bigdata/client/HDFS/hadoop/lib/*
com.huawei.bigdata.hive.example.ExampleMain
```


Paso 6 Una vez que el programa se ejecuta correctamente, se muestra la siguiente información.

Figura 3-15 Resultado de ejecución del programa

```
[root@node-master1-iYpxp hive_examples]# java -cp .:hive-examples-mrs-1.7.0.jar:
/opt/hive_examples/conf:/opt/client/Hive/Beeline/lib/*:/opt/client/HDFS/hadoop/l
ib/* com.huawei.bigdata.hive.example.ExampleMain
log4j:WARN No appenders could be found for logger (com.huawei.bigdata.security.L
oginUtil).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.
Create table success!
_c0
0
Delete table success!
[root@node-master1-iYpxp hive_examples]#
```

----Fin

4 Uso de Hadoop desde el principio

- MRS proporciona componentes de big data de alto rendimiento basados en Hadoop, como Spark, HBase, Kafka y Storm.
- Esta sección describe cómo usar Hadoop para enviar trabajos de wordcount a través de la GUI y los nodos del clúster. El trabajo de recuento de palabras es el trabajo más clásico de Hadoop que cuenta las palabras en grandes cantidades de texto.
- Adquiera un clúster; prepare el programa de muestra Hadoop y los archivos de datos; cargue datos a OBS; cree un trabajo y vea los resultados de la ejecución del trabajo.

  Puede comenzar leyendo los siguientes pasos:

  a. **Comprar un clúster MRS.**
  b. **Configurar software.**
  c. **Configurar hardware.**
  d. **Establecer opciones avanzadas.**
  e. **Confirmar la configuración.**
  f. **Preparar el programa de muestra Hadoop y los archivos de datos.**
  g. **Subir datos a OBS.**
  h. **Enviar un trabajo en la GUI.**
  i. **Enviar un trabajo a través de un nodo de clúster.**
  j. **Consultar resultados de ejecución de trabajos.**

Procedimiento

Paso 1 Comprar un clúster MRS.

1. Inicie sesión en la consola Huawei Cloud.
2. Elija **Service List** >**Analytics** >**MapReduce Service**.
3. En la página **Active Clusters** que se muestra, haga clic en **Buy Cluster**.
4. Haga clic en la pestaña **Custom Config**.

Paso 2 Configurar software.

1. **Region**: Seleccione una región según sea necesario.
2. **Billing Mode**: Seleccione **Pay-per-use**.

3. **Cluster Name**: Introduzca **mrs_demo** o especifique un nombre de acuerdo con las reglas de nomenclatura.
4. **Cluster Version**: Seleccione **MRS 3.1.0**.
5. **Cluster Type**: Seleccione **Analysis Cluster**.
6. Seleccione todos los componentes del clúster de análisis.
7. Haga clic en **Next**.

Paso 3 Configurar hardware.

1. **AZ**: Seleccione **AZ2**.
2. **Enterprise Project**: Seleccione **default**.
3. **VPC** y **Subnet**: Conserve los valores predeterminados o haga clic en **View VPC** y **View Subnet** para crearlos.
4. **Security Group** - Utilice el valor predeterminado **Security Group**.
5. **EIP**: **Bind later** está seleccionado de forma predeterminada.
6. **Cluster Node**: Conserve los valores predeterminados. No agregue nodos de tarea.
7. Haga clic en **Next**.

Paso 4 Establecer opciones avanzadas.

1. **Kerberos Authentication**: **Disabled**
2. **Username**: **admin** se utiliza de forma predeterminada.
3. **Password** y **Confirm Password**: Configúrelos con la contraseña del administrador del FusionInsight Manager.
4. **Login Mode**: Seleccione **Password**. Ingrese una contraseña y confirme la contraseña para usuario **root**.
5. **Host Name Prefix**: Conserve el valor predeterminado.
6. Seleccione **Advanced Settings** y establezca **Agency** en **MRS_ECS_DEFAULT_AGENCY**.
7. Haga clic en **Next**.

Paso 5 Confirmar la configuración.

1. **Configure**: Confirme los parámetros configurados en las áreas **Configure Software**, **Configure Hardware** y **Set Advanced Options**.
2. **Secure Communications**: Seleccione **Enable**.
3. Haga clic en **Buy Now**. Se muestra la página que muestra que la tarea se ha enviado.
4. Haga clic en **Back to Cluster List**. Puede ver el estado del clúster en la página **Active Clusters**. Espere a que se complete la creación del clúster. El estado inicial del clúster es **Starting**. Una vez creado el clúster, el estado del clúster pasa a ser **Running**.

Paso 6 Preparar el programa de muestra de Hadoop y archivos de datos.

1. Prepare el programa de recuento de palabras.

   Descargue el **programa de muestra de Hadoop** (incluido wordcount). **hadoop-3.3.1.tar.gz** se utiliza como ejemplo. Utilice la versión real del programa proporcionada en el enlace. Por ejemplo, elija **hadoop-3.3.1**. En la página que se muestra, haga clic en **hadoop-3.3.1.tar.gz** para descargarla. A continuación, descomprima para obtener **hadoop-mapreduce-examples-3.3.1.jar** (el programa de muestra de Hadoop) de **hadoop-3.3.1\share\hadoop\mapreduce**.

2. Prepare los archivos de datos.

   No se requiere el formato de los archivos de datos. Prepare dos archivos **.txt**. En este ejemplo, se utilizan los archivos **wordcount1.txt** y **wordcount2.txt**.

Paso 7 Subir datos a OBS.

1. Inicie sesión en la consola de OBS y elija **Parallel File Systems**. En la página **Parallel File Systems**, haga clic en **Create Parallel File System**. En la página **Create Parallel File System** que se muestra, configure los parámetros para crear un sistema de archivos denominado **mrs-word01**.
2. Haga clic en el nombre del sistema de archivos **mrs-word01**. En el panel de navegación de la izquierda, elija **Files**. En la página que se muestra, haga clic en **Create Folder** para crear las carpetas **program** e **input**.
3. Vaya a la carpeta **program** y cargue el programa de muestra de Hadoop descargado en el programa de **Paso 6**.
4. Vaya a la carpeta **input** y cargue los archivos de datos **wordcount1.txt** y **wordcount2.txt** preparados en **Paso 6**.
5. Para enviar un trabajo en la interfaz gráfica de usuario, vaya a **Paso 8**.

   Para enviar un trabajo a través de un nodo de clúster, vaya a **Paso 9**.

Paso 8 Enviar un trabajo en la GUI.

1. En el panel de navegación de la consola de MRS, elija **Clusters** > **Active Clusters**. En la página **Active Clusters**, haga clic en el clúster **mrs_demo**.
2. En la página de información del clúster, haga clic en la pestaña **Jobs** y luego en **Create** para crear un trabajo. Para enviar un trabajo a través de un nodo de clúster, vaya a **Paso 9**.
3. **Type**: **MapReduce**
4. **Job Name**: Ingrese **wordcount**.
5. **Program Path**: Haga clic en **OBS** y seleccione el programa de muestra de Hadoop subido a **Paso 7**.
6. **Parameters**: Ingrese **wordcount obs://mrs-word01/input/ obs://mrs-word01/output/**. **output** indica la ruta de salida. Introduzca un directorio que no existe.
7. **Service Parameters**: Déjalo en blanco.
8. Haga clic en **OK** para enviar el trabajo. Después de enviar un trabajo, se encuentra en el estado **Accepted** de forma predeterminada. No necesita ejecutar el trabajo de forma manual.
9. Vaya a la página de pestaña **Jobs**, vea el estado y los registros del trabajo y vaya a **Paso 10** para ver el resultado de la ejecución del trabajo.

Paso 9 Enviar un trabajo a través de un nodo de clúster.

1. Inicie sesión en la consola MRS y haga clic en el clúster denominado **mrs_demo** para ir a su página de detalles.
2. Haga clic en la pestaña **Nodes**. En esta página de pestaña, haga clic en el nombre de un nodo master para ir a la consola de gestión de ECS.
3. Haga clic en **Remote Login** en la esquina superior derecha de la página.
4. Introduzca el nombre de usuario y la contraseña del nodo de Master como se le solicite. El nombre de usuario es **root** y la contraseña es la configurada durante la creación del clúster.

5. Ejecute el comando **source /opt/Bigdata/client/bigdata_env** para configurar las variables de entorno.
6. Si se ha habilitado la autenticación Kerberos, ejecute el comando **kinit** *MRS cluster user*, por ejemplo, **kinit admin**, para autenticar al usuario del clúster actual. Omita este paso si la autenticación Kerberos no está habilitada.
7. Ejecute el siguiente comando para copiar el programa de ejemplo en el bucket OBS al nodo master del clúster:

   hadoop fs -Dfs.obs.access.key=AK -Dfs.obs.secret.key=SK -copyToLocal source_path.jar target_path.jar Example: **hadoop fs -Dfs.obs.access.key=XXXX -Dfs.obs.secret.key=XXXX -copyToLocal "obs://mrs-word01/program/hadoop-mapreduce-examples-XXX.jar" "/home/omm/hadoop-mapreduce-examples-XXX.jar"** Para obtener el par AK/SK para iniciar sesión en la consola OBS, coloque el cursor sobre el nombre de usuario en la esquina superior derecha de la consola de gestión, y elija **My Credentials** > **Access Keys**, o haga clic en **Create Access Key** para crear una.
8. Ejecute el siguiente comando para enviar un trabajo de wordcount. Para leer o escribir datos en OBS, agregue parámetros AK/SK. **source /opt/Bigdata/client/bigdata_env;hadoop jar execute_jar wordcount input_path output_path** Example: **source /opt/Bigdata/client/bigdata_env;hadoop jar /home/omm/hadoop-mapreduce-examples-XXX.jar wordcount -Dfs.obs.access.key=XXXX -Dfs.obs.secret.key=XXXX "obs://mrs-word01/input/*" "obs://mrs-word01/output/"** En este comando, **input_path** indica una ruta para almacenar archivos de entrada de trabajo en OBS. **output_path** indica una ruta para almacenar archivos de salida de trabajo en OBS y debe establecerse en un directorio que no existe

Paso 10 Consultar resultados de ejecución de trabajos.

1. Inicie sesión en la consola OBS y haga clic en el nombre del sistema de archivos paralelo **mrs-word01**.
2. En la página que se muestra, elija **Files** en el panel de navegación de la izquierda. Vaya a la ruta de salida en el bucket **mrs-word01** especificado durante el envío del trabajo y vea el archivo de salida del trabajo. Necesita descargar el archivo en el host local y abrirlo en formato **.txt**.

----Fin

5 Uso de Kafka desde principio

MRS proporciona componentes de big data de alto rendimiento basados en Hadoop, como Spark, HBase, Kafka y Storm.

En esta sección se utiliza un clúster con la autenticación Kerberos deshabilitada como ejemplo para describir cómo generar y consumir mensajes en un topic de Kafka.

Puede comenzar leyendo los siguientes pasos:

1. **Compra de un clúster**
2. **Instalar el cliente Kafka**
3. **Iniciar sesión en un nodo master usando VNC**
4. **Creación de un topic mediante el cliente Kafka**
5. **Gestión de mensajes en Kafka Topics**

Vídeo Tutorial

Este vídeo utiliza un clúster MRS 3.1.0 (con la autenticación Kerberos deshabilitada) como ejemplo para describir cómo utilizar un cliente Kafka para crear, consultar y eliminar un topic. Para obtener más información sobre cómo crear un topic, vea **Creación de un topic con el cliente Kafka**.

NOTA

La interfaz de usuario puede variar dependiendo de la versión. El video tutorial es solo para referencia.

Compra de un clúster

Paso 1 Comprar un clúster MRS.

1. Inicie sesión en la consola Huawei Cloud.
2. Elija **Service List** >**Analytics** >**MapReduce Service**.
3. En la página **Active Clusters** que se muestra, haga clic en **Buy Cluster**.
4. Haga clic en la pestaña **Custom Config**.

Paso 2 Configurar software.

1. **Region**: Seleccione una región según sea necesario.
2. **Billing Mode**: Seleccione **Pay-per-use**.
3. **Cluster Name**: Introduzca **mrs_demo** o especifique un nombre de acuerdo con las reglas de nomenclatura.
4. **Cluster Version**: Seleccione **MRS 3.1.0**.
5. **Cluster Type**: Seleccione **Streaming cluster**.
6. Seleccione todos los componentes del clúster de streaming.
7. Haga clic en **Next**.

Paso 3 Configurar hardware.

1. **AZ**: Seleccione **AZ2**.
2. **Enterprise Project**: Seleccione **default**.
3. **VPC** y **Subnet**: Conserve los valores predeterminados o haga clic en **View VPC** y **View Subnet** para crearlos.
4. **Security Group** - Utilice el valor predeterminado **Security Group**.
5. **EIP**: **Bind later** está seleccionado de forma predeterminada.
6. **Cluster Node**: Conserve los valores predeterminados. No agregue nodos de tarea.
7. Haga clic en **Next**.

Paso 4 Establecer opciones avanzadas.

1. **Kerberos Authentication**: **Disabled**
2. **Username**: **admin** se utiliza de forma predeterminada.
3. **Password** y **Confirm Password**: Configúrelos con la contraseña del administrador del FusionInsight Manager.
4. **Login Mode**: Seleccione **Password**. Ingrese una contraseña y confirme la contraseña para usuario **root**.
5. **Host Name Prefix**: Conserve el valor predeterminado.
6. Seleccione **Advanced Settings** y establezca **Agency** en **MRS_ECS_DEFAULT_AGENCY**.
7. Haga clic en **Next**.

Paso 5 Confirmar la configuración.

1. **Configure**: Confirme los parámetros configurados en las áreas **Configure Software**, **Configure Hardware** y **Set Advanced Options**.
2. **Secure Communications**: Seleccione **Enable**.
3. Haga clic en **Buy Now**. Se muestra la página que muestra que la tarea se ha enviado.
4. Haga clic en **Back to Cluster List**. Puede ver el estado del clúster en la página **Active Clusters**. Espere a que se complete la creación del clúster. El estado inicial del clúster es **Starting**. Una vez creado el clúster, el estado del clúster pasa a ser **Running**.

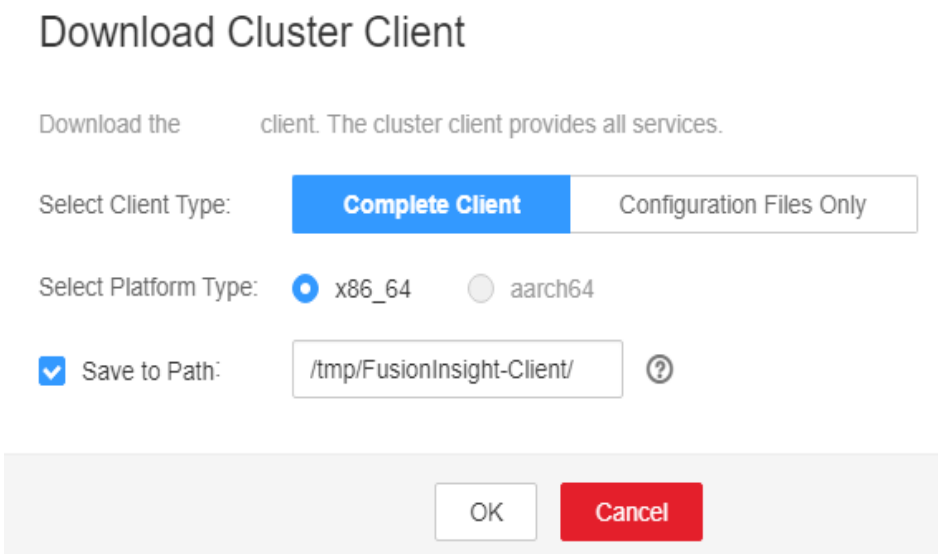
----Fin

Instalación del cliente Kafka

Paso 1 Elija **Clusters** > **Active Clusters**. En la página **Active Clusters**, haga clic en el clúster denominado **mrs_demo** para ir a su página de detalles.

Paso 2 Haga clic en **Access Manager** junto a **MRS Manager**. En la página que se muestra, configure la información de EIP y haga clic en **OK**. Ingrese el nombre de usuario y la contraseña para acceder al FusionInsight Manager.

Paso 3 Elija **Cluster** >**Services** >**HBase**. En la página mostrada, elija **More** >**Download Client**. En el cuadro de diálogo **Download Cluster Client**, seleccione **Complete Client** para **Select Client Type** y seleccione un tipo de plataforma, seleccione **Save to Path** y haga clic en **OK**. Se descarga el paquete de software cliente Kafka, por ejemplo **FusionInsight_Cluster_1_Kafka_Client.tar**.



Paso 4 Inicie sesión en el nodo activo como usuario **root**.

Paso 5 Vaya al directorio donde está almacenado el paquete de software y ejecute los siguientes comandos para descomprimir y verificar el paquete de software, y descomprimir el archivo de instalación obtenido:

cd /tmp/FusionInsight-Client

tar -xvf FusionInsight_Cluster_1_Kafka_Client.tar

sha256sum -c FusionInsight_Cluster_1_Kafka_ClientConfig.tar.sha256

tar -xvf FusionInsight_Cluster_1_Kafka_ClientConfig.tar

Paso 6 Vaya al directorio donde está almacenado el paquete de instalación y ejecute el siguiente comando para instalar el cliente en un directorio especificado (ruta absoluta), por ejemplo **/opt/hadoopclient**:

cd /tmp/FusionInsight-Client/FusionInsight_Cluster_1_Kafka_ClientConfig

Ejecute el comando **./install.sh /opt/hadoopclient** y espere hasta que se complete la instalación del cliente.

Paso 7 Compruebe si el cliente está instalado.

cd /opt/hadoopclient

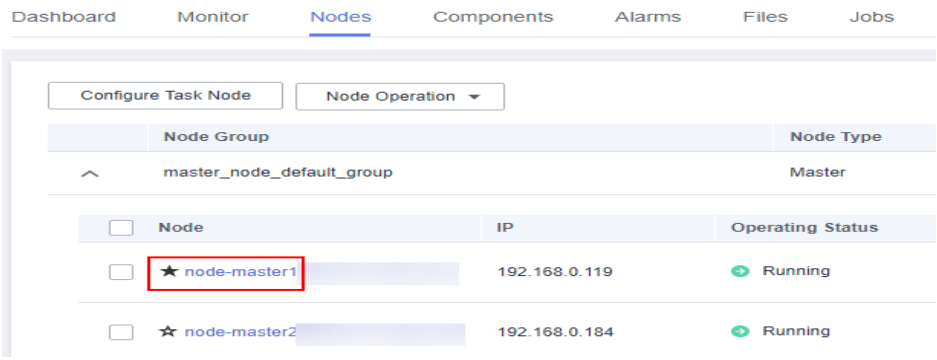
source bigdata_env

Ejecute el comando **klist** para consultar y confirmar los detalles de autenticación. Si se ejecuta el comando, se instala el cliente Kafka.

----Fin

Inicio de sesión en un nodo master mediante VNC

Paso 1 Elija **Clusters** > **Active Clusters**. En la página **Active Clusters** que se muestra, haga clic en el clúster denominado **mrs_demo**. En la página de detalles del clúster que se muestra, haga clic en la pestaña **Nodes**. En esta página de pestaña, busque el nodo cuyo tipo es **Master1** y haga clic en el nombre del nodo para ir a la página de detalles de ECS.



Paso 2 Haga clic en **Remote Login** en la esquina superior derecha de la página para iniciar sesión de forma remota en el nodo master. Inicie sesión con el nombre de usuario **root** y la contraseña configurada durante la compra del clúster.

----Fin

Creación de un topic mediante el cliente Kafka

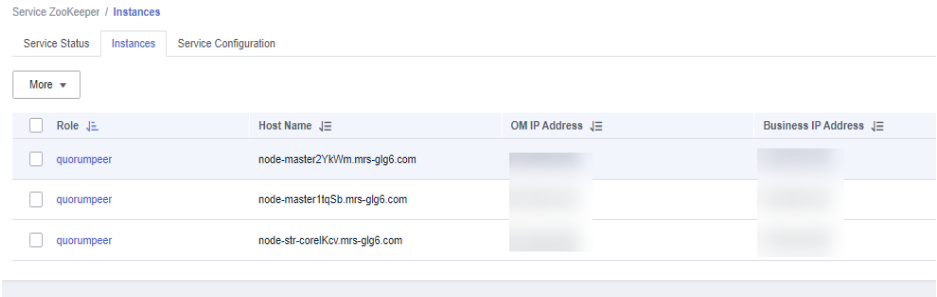
Paso 1 Configure variables de entorno. Por ejemplo, si el directorio de instalación del cliente de Kafka es **/opt/hadoopclient**, ejecute el siguiente comando:

source /opt/hadoopclient/bigdata_env

Paso 2 Elija **Clusters** > **Active Clusters**. En la página **Active Clusters**, haga clic en el clúster denominado **mrs_demo** para ir a la página de pestaña **Dashboard**. En esta página, haga clic en **Synchronize** junto a **IAM User Sync**.

Paso 3 Una vez completada la sincronización, haga clic en la pestaña **Components**. En esta página de pestaña, seleccione **ZooKeeper**. En la página que se muestra, haga clic en la pestaña **Instances**. Registre la dirección IP de cualquier instancia de ZooKeeper, por ejemplo, **192.168.7.35**.

Figura 5-1 Direcciones IP de instancias de rol de ZooKeeper



Paso 4 Ejecute el siguiente comando para crear un topic de Kafka:

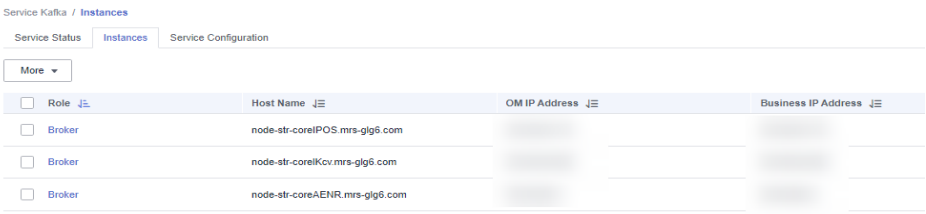
kafka-topics.sh --create --zookeeper <*IP address of the node where the ZooKeeper instance resides***:2181/kafka> --partitions 2 --replication-factor 2 --topic <***Topic name***>**

----Fin

Gestión de mensajes en Kafka Topics

Paso 1 Haga clic en la pestaña **Components**. En esta página de pestaña, seleccione **Kafka**. En la página que se muestra, haga clic en la pestaña **Instances**. En la página de pestaña **Instances**, vea las direcciones IP de las instancias de Kafka. Registre la dirección IP de cualquier instancia de Kafka, por ejemplo **192.168.7.15**.

Figura 5-2 Direcciones IP de instancias de rol de Kafka



Paso 2 Inicie sesión en el nodo master y ejecute el siguiente comando para generar mensajes en una prueba de topic:

kafka-console-producer.sh --broker-list <*IP address of the node where the Kafka instance resides***:9092> --topic <***Topic name***> --producer.config /opt/hadoopclient/Kafka/kafka/config/producer.properties**

Introduzca el contenido especificado como los mensajes generados por el productor y pulse **Enter** para enviar los mensajes. Para dejar de generar mensajes, presione **Ctrl+C** para salir.

Paso 3 Consumir mensajes en la prueba de topic.

kafka-console-consumer.sh --topic <*Topic name***> --bootstrap-server <***IP address of the node where the Kafka instance resides***:9092> --consumer.config /opt/hadoopclient/Kafka/kafka/config/consumer.properties**

----Fin

6 Uso de HBase desde principio

MRS proporciona componentes de big data de alto rendimiento basados en Hadoop, como Spark, HBase y Kafka.

Esta sección utiliza un clúster con la autenticación Kerberos deshabilitada como ejemplo para describir cómo iniciar sesión en el cliente HBase, crear una tabla, insertar datos en la tabla y modificar la tabla.

Puede comenzar leyendo los siguientes temas:

1. **Preparación de un clúster MRS**
2. **Instalación del cliente HBase**
3. **Creación de una tabla con el cliente HBase**

Vídeo Tutorial

Este vídeo utiliza un clúster MRS 3.1.0 (con la autenticación Kerberos deshabilitada) como ejemplo para describir cómo utilizar un cliente HBase para crear una tabla, insertar datos en la tabla y modificar los datos de la tabla. Para obtener más información sobre cómo utilizar un cliente HBase para crear una tabla, consulte **Uso de un cliente HBase**.

NOTA

La interfaz de usuario puede variar dependiendo de la versión. El video tutorial es solo para referencia.

Preparación de un clúster MRS

Paso 1 Adquiera un clúster MRS.

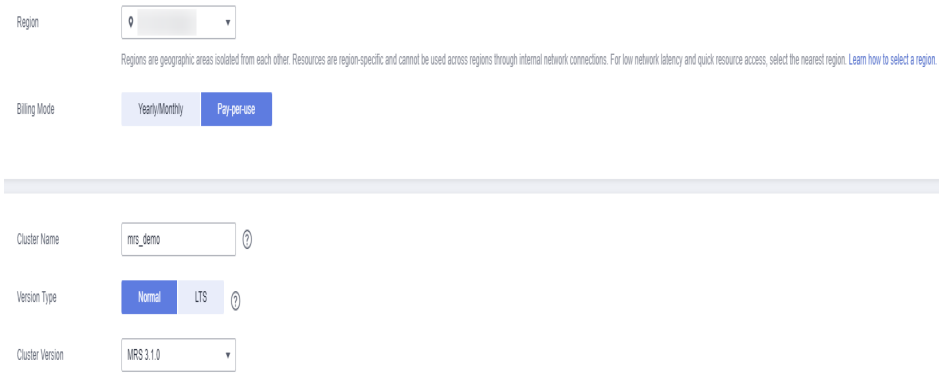
1. Vaya a la página **Comprar clúster**.
2. Haga clic en la pestaña **Custom Config**.

Paso 2 Defina los siguientes parámetros y haga clic en **Next**.

- **Region**: Seleccione una región según sea necesario.
- **Billing Mode**: Seleccione **Pay-per-use**.
- **Cluster Name**: Introduzca **mrs_demo** o especifique un nombre de acuerdo con las reglas de nomenclatura.
- **Version Type**: Seleccione **Normal**.

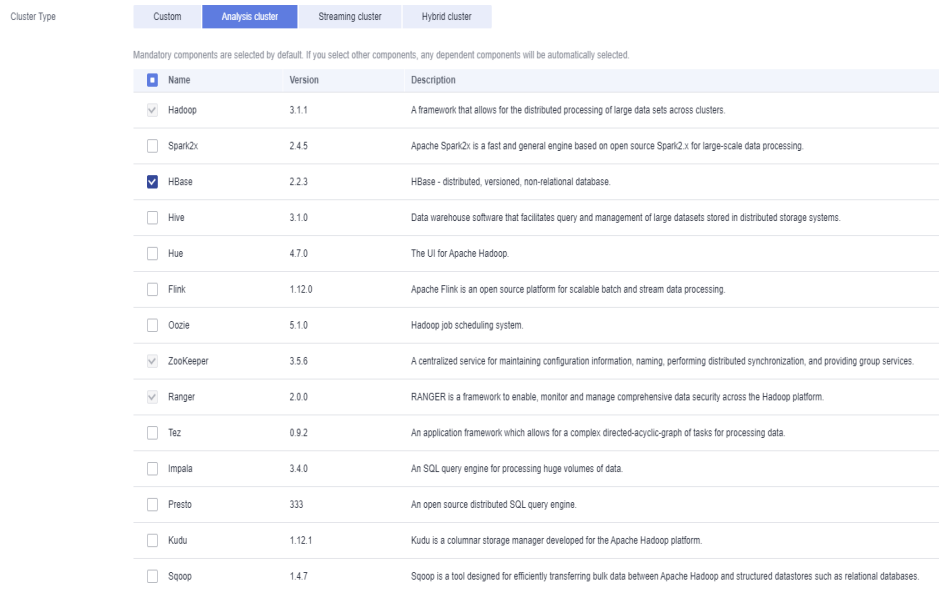
- **Cluster Version**: Seleccione **MRS 3.1.0**.

Figura 6-1 Página de Configure Software



- **Cluster Type**: Seleccione **Analysis Cluster** y seleccione HBase.

Figura 6-2 Selección del tipo de clúster y los componentes



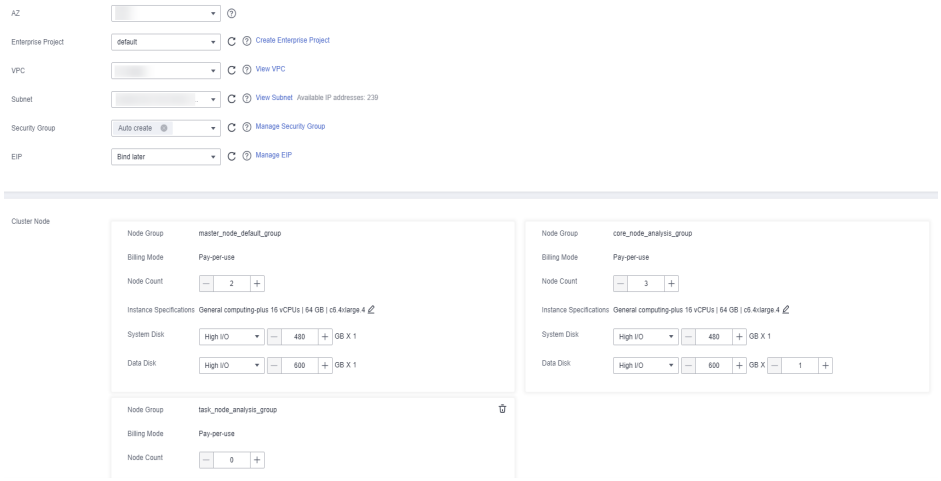
Paso 3 En la página **Configure Hardware**, establezca los parámetros haciendo referencia a **Tabla 6-1** y haga clic en **Next**.

Tabla 6-1 Configuración de hardware de clúster de MRS

Parámetro	Valor de ejemplo
AZ	**AZ2**
Enterprise Project	**default**
VPC	Conserve el valor predeterminado. También puede hacer clic en **View VPC** para crear una VPC.

Parámetro	Valor de ejemplo
EIP	Puede seleccionar una EIP existente en la lista desplegable. Si no hay ningún EIP disponible en la lista desplegable, haga clic en **Manage EIP** para acceder a la página **EIPs** y crear una.

Figura 6-3 Configuraciones de hardware



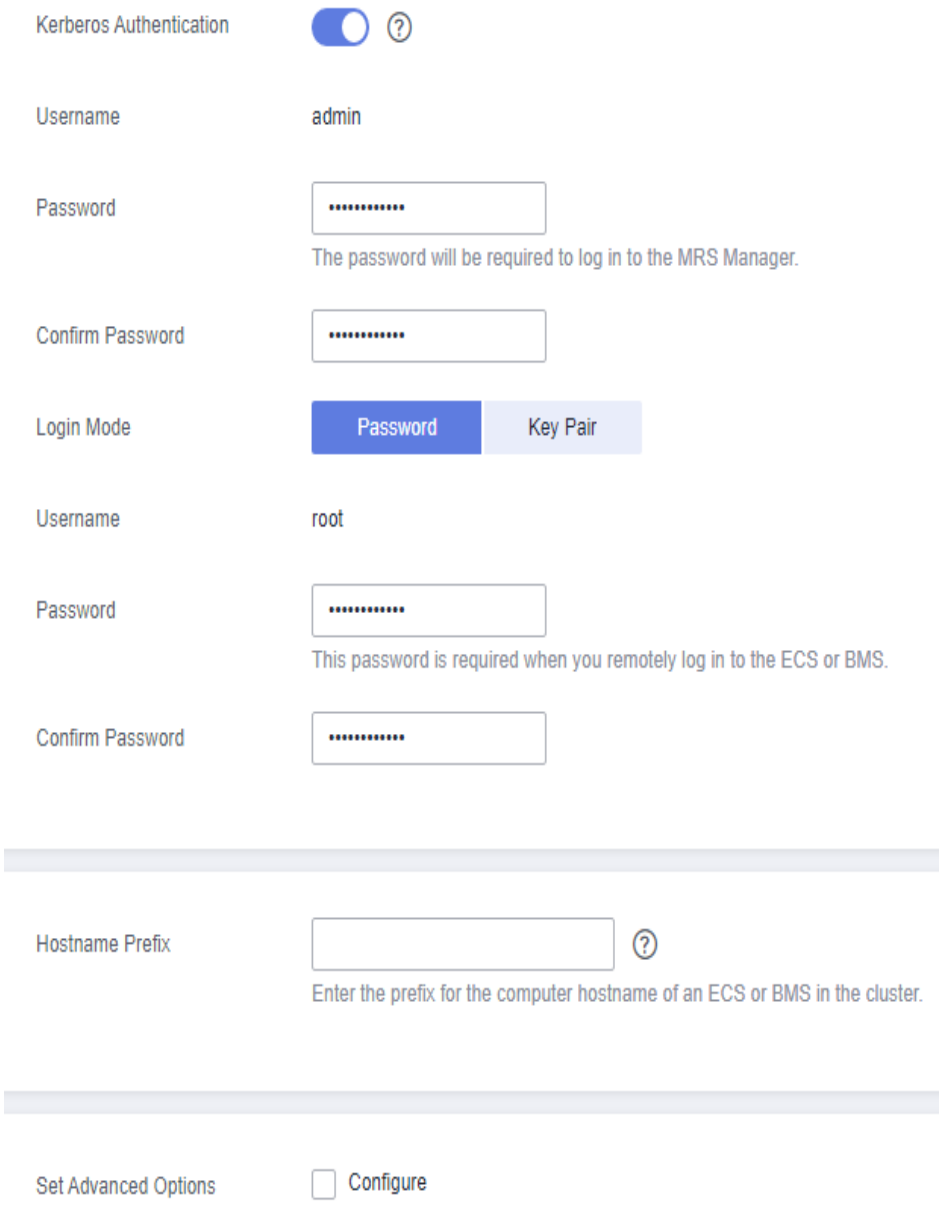
Paso 4 Configurar opciones avanzadas.

1. En la página **Set Advanced Options**, configure los parámetros de acuerdo con **Tabla 6-2** y haga clic en **Next**.

   Tabla 6-2 Opciones avanzadas del clúster MRS

Parámetro	Valor de ejemplo
Kerberos Authentication	Deshabilite esta función.
Password	**Test@!123456**
Confirm Password	**Test@!123456**
Login Mode	**Password**
Password	**Test@#123456**
Confirm Password	**Test@#123456**

Figura 6-4 Configurar opciones avanzadas



2. En la página **Confirm Configuration**, compruebe la información de configuración del clúster. Si necesita ajustar la configuración, haga clic en para ir a la pestaña correspondiente y configurar los parámetros de nuevo.
3. Seleccione **Enable** para **Secure Communications**. Haga clic en **Buy Now**. Se muestra una página que indica que la tarea se ha enviado correctamente.
4. Haga clic en **Back to Cluster List**. Puede ver el estado del clúster en la página **Active Clusters**.
5. Espere a que se complete la creación del clúster. El estado inicial del clúster es **Starting**. Una vez que el clúster se ha creado correctamente, el estado del clúster pasa a ser **Running**.

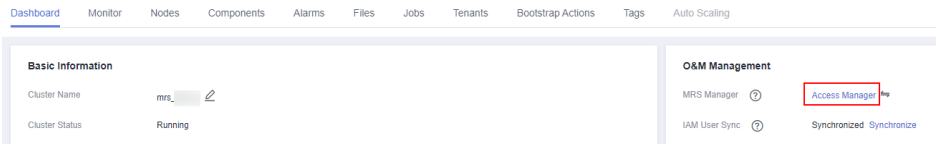
----Fin

Instalación del cliente HBase

Paso 1 Elija **Clusters** > **Active Clusters** y haga clic en **mrs_demo**. Se muestra la página de información del clúster.

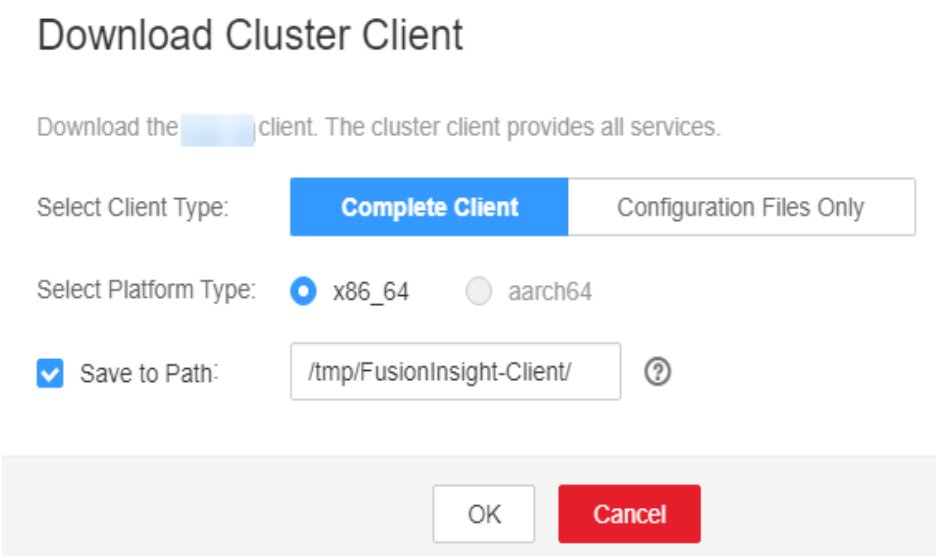
Paso 2 Haga clic en **Access Manager** junto a **MRS Manager**. En la página que se muestra, configure la información de EIP y haga clic en **OK**. Ingrese el nombre de usuario y la contraseña para acceder al FusionInsight Manager.

Figura 6-5 Iniciar sesión en FusionInsight Manager desde la consola de gestión



Paso 3 Elija **Cluster** > **Services** > **HBase** y seleccione **Download Client** en la lista desplegable **More**. Seleccione **Complete Client**, el tipo de plataforma correspondiente, y **Save to path** y haga clic en OK.

Figura 6-6 Descargar el cliente de clúster



Paso 4 Inicie sesión en el nodo de gestión activo como usuario **root**.

NOTA

Para identificar los nodos de gestión activos y en espera, consulte **Determinación de nodos de gestión activos y en espera del Manager**.

Paso 5 Vaya al directorio donde está almacenado el paquete de instalación y ejecute los siguientes comandos para descomprimir y verificar el paquete de instalación, y descomprimir el archivo de instalación obtenido:

cd /tmp/FusionInsight-Client

tar -xvf FusionInsight_Cluster_1_HBase_Client.tar

sha256sum -c FusionInsight_Cluster_1_HBase_ClientConfig.tar.sha256

tar -xvf FusionInsight_Cluster_1_HBase_ClientConfig.tar

Paso 6 Vaya al directorio donde está almacenado el paquete de instalación y ejecute el siguiente comando para instalar el cliente en un directorio especificado (una ruta absoluta), por ejemplo, **/opt/hbaseclient**:

cd /tmp/FusionInsight-Client/FusionInsight_Cluster_1_HBase_ClientConfig

Ejecute el comando **./install.sh /opt/hbaseclient** y espere hasta que se complete la instalación del cliente.

Paso 7 Compruebe si el cliente se ha instalado correctamente.

cd /opt/hbaseclient

source bigdata_env

hbase shell

Si el comando se ejecuta correctamente, el cliente HBase se instala correctamente.

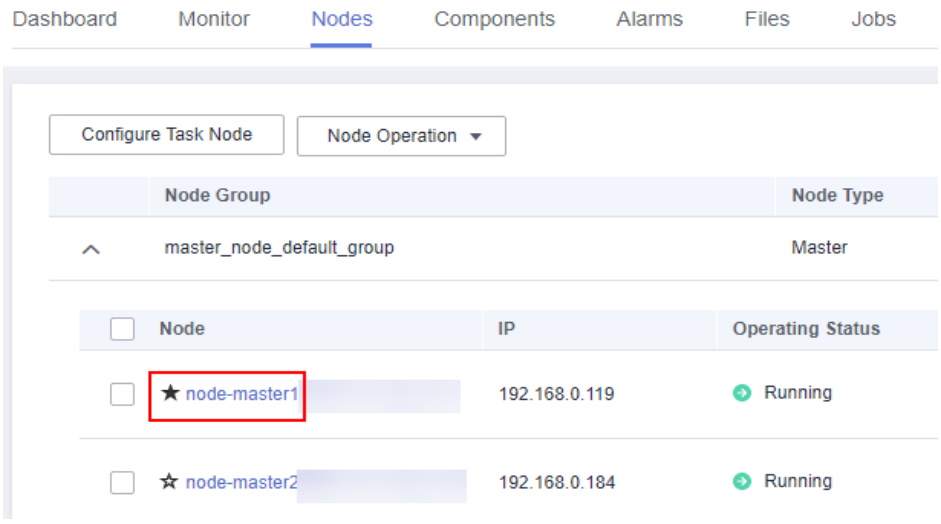
----Fin

Creación de una tabla con el cliente HBase

Paso 1 Inicie sesión en el nodo master usando VNC.

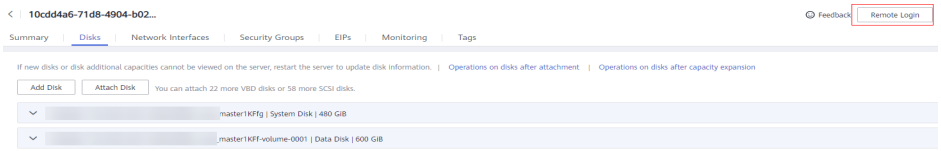
1. En la consola MRS, elija **Clusters** > **Active Clusters** y seleccione **mrs_demo** en la lista de clústeres. Haga clic en **Nodes** y haga clic en el nodo cuyo nombre contiene **master1** para acceder a su página de detalles de ECS.

   Figura 6-7 Página de pestaña Nodes donde se encuentra el nodo Master1

2. Haga clic en **Remote Login** en la esquina superior derecha de la página para iniciar sesión en el nodo master como usuario **root**. La contraseña es la que se establece cuando se compra el clúster.

   Figura 6-8 Iniciar sesión de forma remota en el nodo Master1

Paso 2 Ejecute el siguiente comando para ir al directorio del cliente:

cd /opt/hbaseclient

Paso 3 Ejecute el siguiente comando para configurar las variables de entorno:

source bigdata_env

NOTA

Si la autenticación Kerberos está habilitada para el clúster, ejecute el siguiente comando para autenticar al usuario actual. El usuario actual debe tener el permiso para crear tablas HBase.

Por ejemplo:

kinit *hbaseuser*

Paso 4 Ejecute el siguiente comando para acceder a la HBase shell CLI:

hbase shell

Paso 5 Ejecute el comando HBase client para crear la tabla **user_info**.

1. Cree la tabla **user_info** y agregue datos relacionados.

   create '*user_info*',{**NAME** => 'i'}

   put '*user_info*','*12005000201*','***i:name***','*A*'

   put '*user_info*','*12005000201*','***i:gender***','*Male*'

   put '*user_info*','*12005000201*','***i:age***','*19*'

   put '*user_info*','*12005000201*','***i:address***','*City A*'

2. Agregue los antecedentes educativos y los títulos profesionales de los usuarios a la tabla **user_info**.

   put '*user_info*','*12005000201*','**i:degree**','*master*'

   put '*user_info*','*12005000201*','**i:pose**','*manager*'

3. Consultar nombres de usuario y direcciones por ID de usuario.

   scan'*user_info*',
   {**STARTROW**=>'*12005000201*',**STOPROW**=>'*12005000201*',**COLUMNS**=>['**i:name**','**i:address**']}

   ```
   ROW                                    COLUMN
   +CELL

    12005000201                           column=i:address,
   timestamp=2021-10-30T10:21:42.196, value=City
   A
    12005000201                           column=i:name,
   timestamp=2021-10-30T10:21:18.594,
   value=A
   1 row(s)
   Took 0.0996 seconds
   ```

4. Consultar información por nombre de usuario.

   scan'*user_info*',{**FILTER**=>"SingleColumnValueFilter('i','name',=,'binary:*A*')"}

   ```
   ROW                                    COLUMN
   +CELL

    12005000201                           column=i:address,
   timestamp=2021-10-30T10:21:42.196, value=City
   A
    12005000201                           column=i:age,
   timestamp=2021-10-30T10:21:30.777,
   ```



```
value=19
 12005000201                              column=i:degree,
timestamp=2021-10-30T10:21:53.284,
value=master
 12005000201                              column=i:gender,
timestamp=2021-10-30T10:21:18.711,
value=Male
 12005000201                              column=i:name,
timestamp=2021-10-30T10:21:18.594,
value=A
 12005000201                              column=i:pose,
timestamp=2021-10-30T10:22:07.152,
value=manager
1 row(s)
Took 0.2158 seconds
```

5. Eliminar datos de usuario de la tabla de información de usuario.

   delete'*user_info*','*12005000201*','i'

6. Eliminar la tabla de información del usuario.

   disable *'user_info'*

   drop *'user_info'*

----Fin

7 Modificación de configuraciones de MRS

Después de crear un clúster MRS, puede modificar los parámetros de configuración de los servicios en el clúster en la consola MRS o en Manager.

Esta sección utiliza el parámetro **hbase.log.maxbackupindex** del servicio HBase como ejemplo para describir cómo modificar los parámetros de configuración MRS.

Puede comenzar leyendo los siguientes temas:

1. **Modificación de parámetros de servicio en la consola MRS**
2. **Modificación de parámetros de servicio en FusionInsight Manager**

Vídeo Tutorial

Este vídeo utiliza un clúster MRS 3.1.0 como ejemplo para describir cómo modificar los parámetros de servicio en la consola de gestión y FusionInsight Manager. Para obtener más información, consulte **Modificación de la configuración del servicio de clúster**.

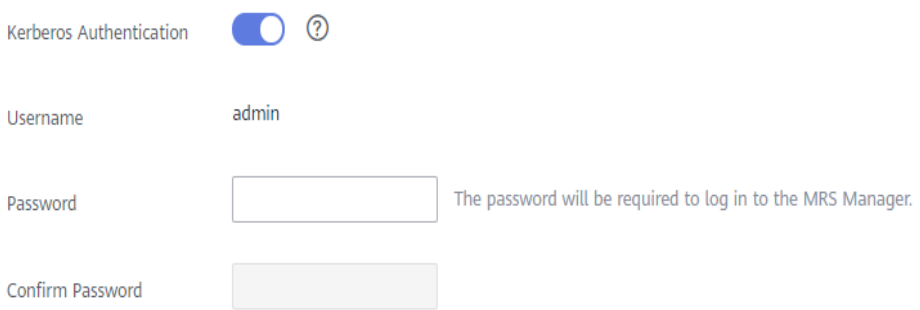
NOTA

La interfaz de usuario puede variar dependiendo de la versión. El video tutorial es solo para referencia.

Modificación de parámetros de servicio en la consola MRS

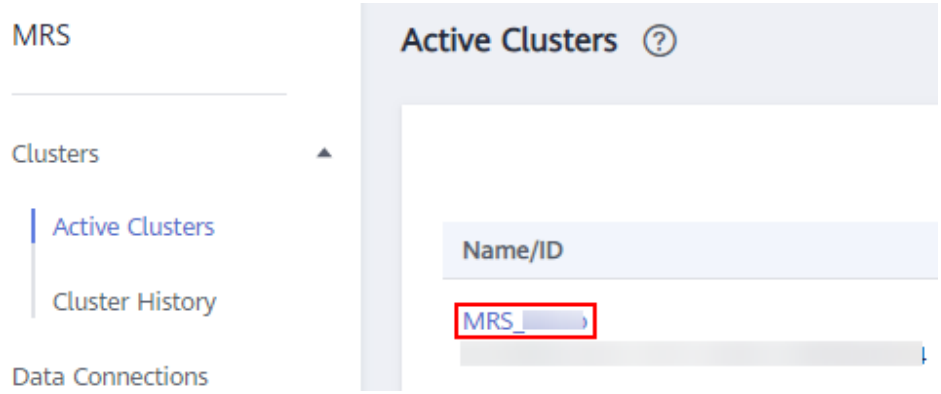
Paso 1 Cree un clúster de seguridad. Para obtener más información, consulte **Comprar un clúster personalizado**. Habilite **Kerberos Authentication** y configure **Password** y confirme la contraseña. Esta contraseña se utiliza para iniciar sesión en Manager. Manténgalo seguro.

Figura 7-1 Configuración de los parámetros del clúster de seguridad



Paso 2 Inicie sesión en la consola de MRS. En el panel de navegación de la izquierda, elija **Clusters** > **Active Clusters** y haga clic en un nombre de clúster.

Figura 7-2 Hacer clic en un nombre de clúster

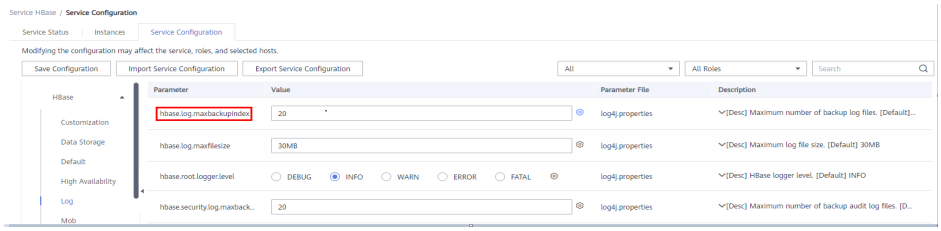


Paso 3 Elija **Components** > **HBase**, haga clic en **Service Configuration**, y elija **All** en la esquina superior derecha de la página.

Paso 4 En el árbol de navegación de la izquierda, elija **HBase** >**Log**.

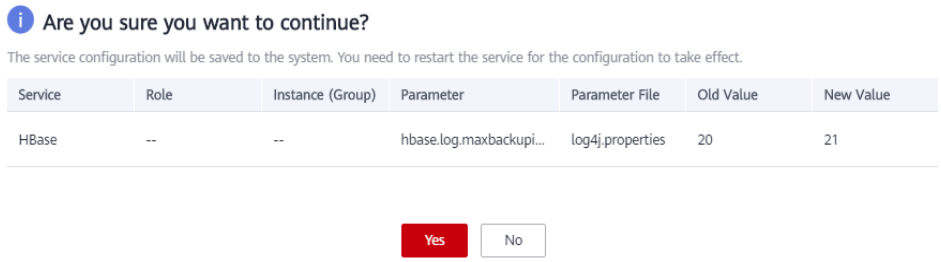
Paso 5 Busque el parámetro **hbase.log.maxbackupindex** y cambie su valor en función de los requisitos de servicio.

Figura 7-3 Cambio del valor del parámetro



Paso 6 Haga clic en **Save Configuration**. En el cuadro de diálogo mostrado, confirme el valor del parámetro cambiado y haga clic en **Yes**. Espere a que el sistema guarde y actualice la configuración y haga clic en **Finish**.

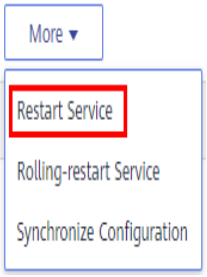
Figura 7-4 Confirmación de la modificación



Paso 7 Compruebe el estado actual de la configuración del servicio.

Haga clic en **Service Status** para ver el estado actual de la configuración del servicio. Si la configuración de un servicio ha caducado, haga clic en **More** y seleccione **Restart Service** para reiniciar el servicio. En el cuadro de diálogo que se muestra, haga clic en **Yes**. A continuación, espere hasta que se reinicie el servicio.

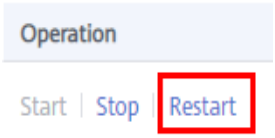
Figura 7-5 Reinicio de un servicio



Paso 8 Compruebe el estado de configuración del servicio de los servicios relacionados.

Vuelva a la página **Components** para comprobar el estado de configuración de los servicios relacionados. Si la configuración de un servicio ha caducado, haga clic en **Restart** en la columna **Operation** del servicio. En el cuadro de diálogo mostrado, haga clic en **Yes** para reiniciarlo.

Figura 7-6 Reinicio de un servicio



----Fin

Modificación de parámetros de servicio en FusionInsight Manager

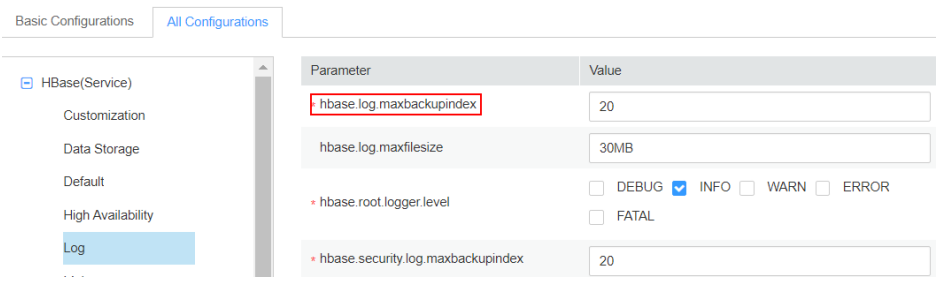
Paso 1 Cree un clúster e inicie sesión en FusionInsight Manager. Para obtener más información, consulte **Creación de un clúster de seguridad e inicio de sesión en Manager**.

Paso 2 Elija **Cluster** > **Services** > **HBase**, elija **Configurations**, y haga clic en **All Configurations**.

Paso 3 Elija **HBase(Service)** > **Log**.

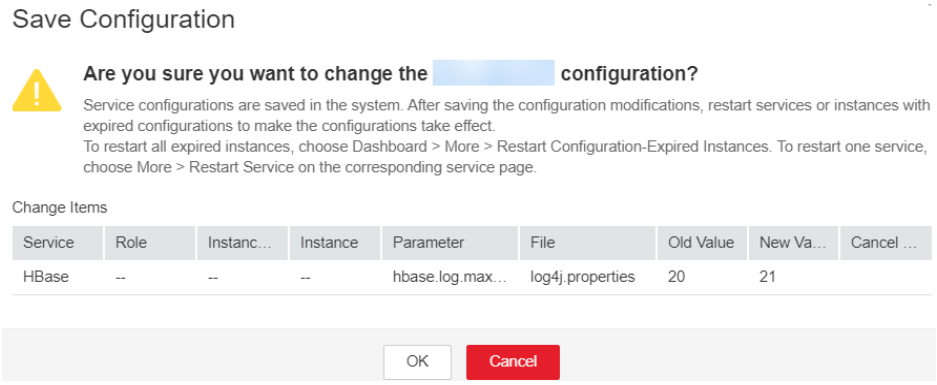
Paso 4 Busque el parámetro **hbase.log.maxbackupindex** y cambie su valor en función de los requisitos de servicio.

Figura 7-7 Cambio del valor del parámetro



Paso 5 Haga clic en **Save**. En el cuadro de diálogo que se muestra, confirme el valor del parámetro cambiado y haga clic en **OK**. Espere a que el sistema guarde y actualice la configuración y haga clic en **Finish**.

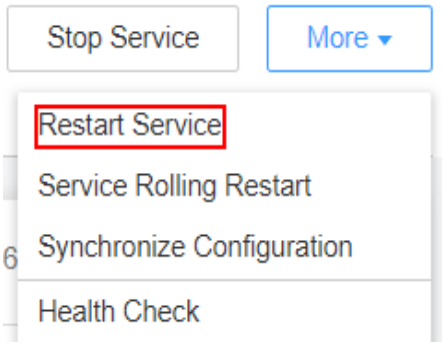
Figura 7-8 Confirmación de la modificación



Paso 6 Compruebe el estado actual de la configuración del servicio.

Haga clic en **Dashboard** para ver el estado actual de la configuración del servicio. Si la configuración de un servicio ha caducado, haga clic en **More** y seleccione **Restart Service**. A continuación, introduzca la contraseña y haga clic en **OK** para reiniciar el servicio. Espere hasta que se reinicie el servicio.

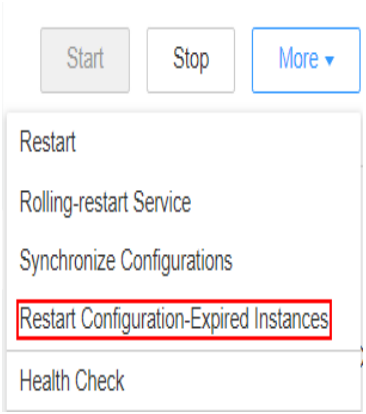
Figura 7-9 Reinicio de un servicio



Paso 7 Compruebe el estado de configuración del servicio de los servicios relacionados.

Seleccione **Cluster** > **Service** para ver el estado de configuración de otros servicios relacionados. Si la configuración de un servicio ha caducado, elija **Cluster** > **Dashboard**, seleccione **Restart Configuration-Expired Instances** en la lista desplegable **More**, escriba la contraseña y haga clic en **OK** para reiniciarla.

Figura 7-10 Reiniciar las instancias de configuración caducadas



----Fin

8 Configuración del escalado automático para un clúster MRS

En escenarios de aplicaciones de big data, especialmente análisis y procesamiento de datos en tiempo real, el número de nodos de clúster debe ajustarse dinámicamente de acuerdo con los cambios en el volumen de datos para proporcionar recursos adecuados. La función de escalado automático de MRS permite que los clústeres sean escalados automáticamente o en función de la carga del clúster.

- Reglas de escalado automático: puede aumentar o disminuir los nodos de tarea en función de las cargas de clúster en tiempo real. El escalado automático se activará cuando el volumen de datos cambie, pero puede haber algunos retrasos.
- Plan de recursos (configuración de la cantidad de nodo de tarea en función del intervalo de tiempo): Si el volumen de datos cambia periódicamente, puede crear planes de recursos para cambiar el tamaño del clúster antes de que cambie el volumen de datos, evitando así retrasos en el aumento o la disminución de recursos.

Puede configurar reglas de escalado automático o planes de recursos o ambos para activar el escalado automático.

Escenario

En el ejemplo siguiente se describe cómo utilizar las reglas de escalado automático y los planes de recursos:

Un servicio de procesamiento en tiempo real observa un aumento inestable en el volumen de datos de 7:00 a 13:00 los lunes, martes y sábados. Por ejemplo, se requieren de 5 a 8 nodos de tarea de 7:00 a 13:00 los lunes, martes y sábado, y de 2 a 4 más allá de este período.

Puede establecer una regla de escalado automático basada en un plan de recursos. Cuando el volumen de datos excede el valor esperado, el número de nodos Task cambia con las cargas de recursos, sin exceder el rango de nodos especificado en el plan de recursos. Cuando se activa un plan de recursos, el número de nodos cambia dentro del rango especificado con un efecto mínimo. Es decir, aumentar los nodos hasta el límite superior y disminuir los nodos hasta el límite inferior.

Vídeo Tutorial

Este vídeo utiliza un clúster MRS 3.1.0 como ejemplo para describir cómo configurar una política de escalado automático cuando compra un clúster y cómo agregar una política de

escalado automático a un clúster existente. Para obtener más información, consulte **Configuración de escalado automático para un clúster MRS**.

NOTA

La interfaz de usuario puede variar según la versión. El video tutorial es solo para referencia.

Adición de un nodo de Task

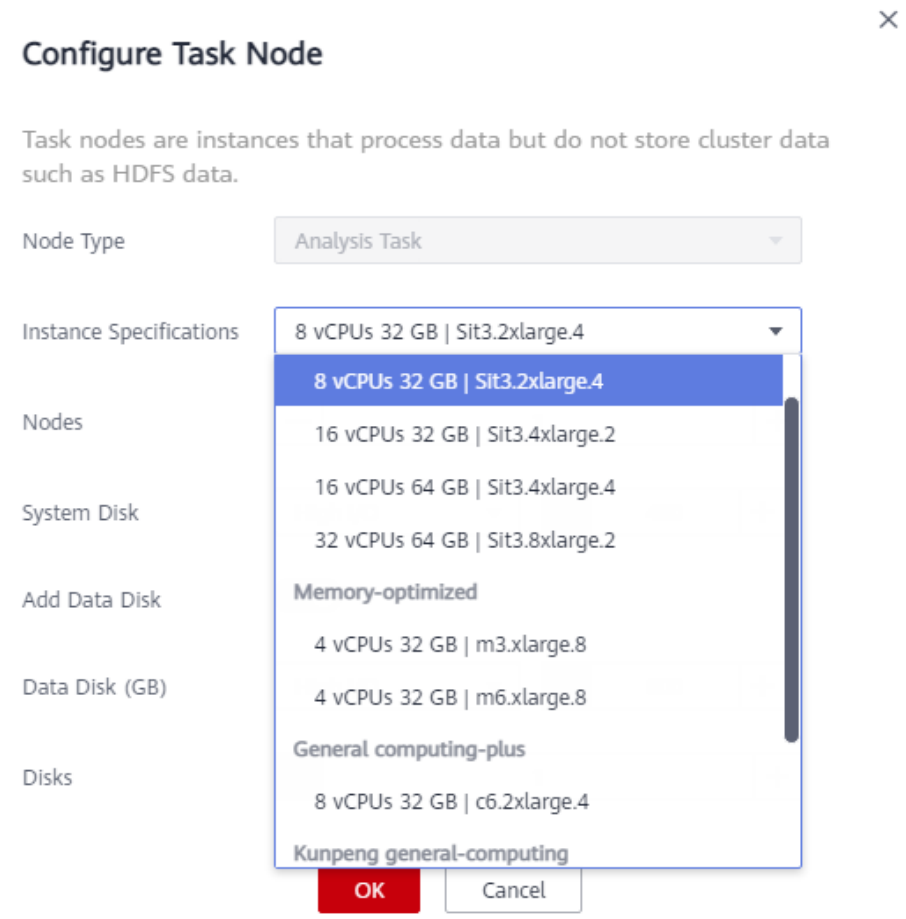
Puede escalar un clúster de MRS agregando manualmente nodos de tarea.

Para agregar un nodo de tarea a un clúster personalizado, realice los siguientes pasos:

1. En la página de detalles del clúster, haga clic en la pestaña **Nodes** y haga clic en **Add Node Group**. Se muestra la página **Add Node Group**.
2. Seleccione **NM** para **Deploy Roles** y establezca otros parámetros según sea necesario.

Para agregar un nodo de task a un clúster no personalizado, realice los siguientes pasos:

1. En la página de detalles del clúster, haga clic en la pestaña **Nodes** y haga clic en **Configure Task Node**. Se muestra la página **Configure Task Node**.
2. En la página **Configure Task Node**, establezca **Node Type**, **Instance Specifications**, **Nodes**, **System Disk**. Además, si **Add Data Disk** está habilitado, configure el tipo de almacenamiento, el tamaño y el número de discos de datos.



3. Haga clic en **OK**.

Uso de reglas de escalado automático y planes de recursos juntos

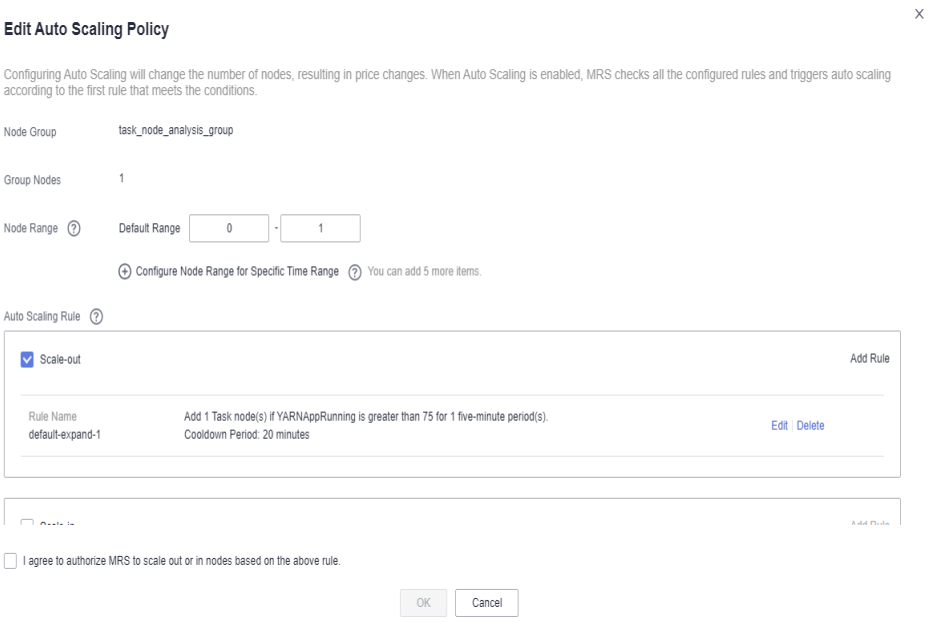
Paso 1 Inicie sesión en la consola de gestión de MRS.

Paso 2 Seleccione **Clusters** > **Active Clusters**, y haga clic en el nombre del clúster de destino. Se muestra la página de detalles del clúster.

Paso 3 En la página que se muestra, haga clic en la pestaña **Auto Scaling**.

Paso 4 Haga clic en **Add Auto Scaling Policy** y establezca **Node Range** en **2-4**.

Figura 8-1 Configuración del escalado automático



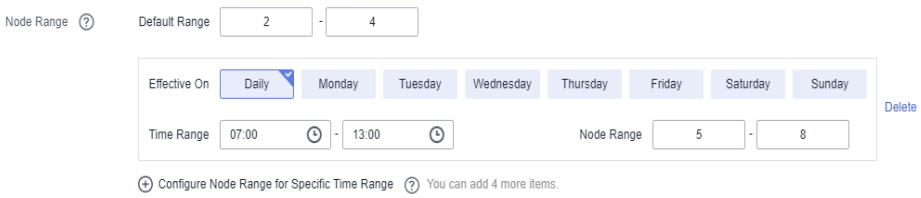
Paso 5 Configurar un plan de recursos.

1. Haga clic en **Configure Node Range for Specific Time Range** en **Default Range**.
2. Configure los parámetros **Time Range** y **Node Range**.

   Time Range - Póngalo en **07:00-13:00**.

   Node Range - Póngalo en **5-8**.

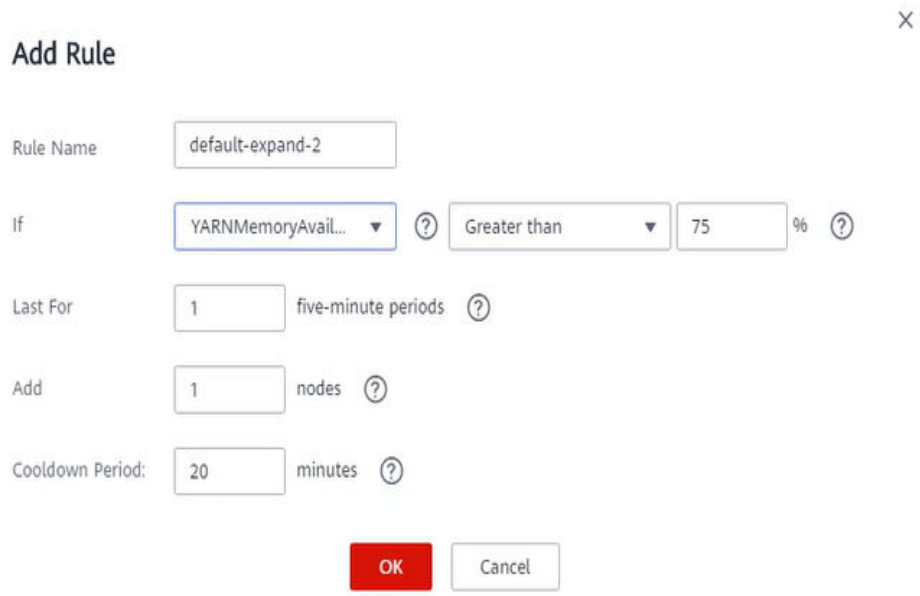
Figura 8-2 Escalamiento automático



Paso 6 Configure una regla de escalado automático.

1. Seleccione **Scale-out**.
2. Haga clic en **Add Rule** a la derecha.

Figura 8-3 Adición de una regla



Rule Name: **default-expand-2**.

If: Seleccione los objetos de regla y las restricciones en los cuadros de lista desplegables; por ejemplo, YARNAppRunning es mayor que 75.

Last For: Póngalo en **1 five-minute periods**.

Add: Póngalo en **1 node**.

Cooldown Period: Póngalo en **20 minutes**.

3. Haga clic en **OK**.

Paso 7 Seleccione **I agree to authorize MRS to scale out or in nodes based on the above rule**.

Paso 8 Haga clic en **OK**.

----Fin

Información de referencia

Al agregar una regla, puede consultar **Tabla 8-1** para configurar las métricas correspondientes.

NOTA

- Los clústeres híbridos admiten todas las métricas de los clústeres de análisis y streaming.
- La precisión de los diferentes tipos de valores de **Tabla 8-1** es la siguiente:
  - **Integer**: entero
  - **Percentage**: 0.01
  - **Ratio**: 0.01

Tabla 8-1 Métricas de escalado automático

Tipo de clúster	Métrica	Tipo de valor	Descripción
Clúster de streaming	StormSlotAvailable	Integer	Número de slot de Storm disponibles. Rango de valores: 0 a 2147483646.
	StormSlotAvailable-Percentage	Percentage	Porcentaje de slot de Storm disponibles, es decir, la proporción de slot disponibles respecto al total de slot. Rango de valores: 0 a 100.
	StormSlotUsed	Integer	Número de slot de Storm usadas. Rango de valores: 0 a 2147483646.
	StormSlotUsedPercentage	Percentage	Porcentaje de las slot usadas de Storm, es decir, la proporción de las slot usadas con respecto al total de slot. Rango de valores: 0 a 100.
	StormSupervisor-MemAverageUsage	Integer	Uso medio de memoria del proceso Supervisor de Storm. Rango de valores: 0 a 2147483646.
	StormSupervisor-MemAverageUsage Percentage	Percentage	Porcentaje medio de la memoria utilizada del proceso Supervisor de Storm con respecto a la memoria total del sistema. Rango de valores: 0 a 100.
	StormSupervisorCPUAverageUsagePercentage	Percentage	Porcentaje promedio de las CPU usadas del proceso Supervisor de Storm con respecto al total de CPU. Rango de valores: [0, 6000].
Clúster de análisis	YARNAppPending	Integer	Número de tareas pendientes en Yarn. Rango de valores: 0 a 2147483646.
	YARNAppPending Ratio	Ratio	Relación de tareas pendientes en Yarn, es decir, la relación entre tareas pendientes y tareas en ejecución en Yarn. Rango de valores: 0 a 2147483646.
	YARNAppRunning	Integer	Número de tareas en ejecución en Yarn. Rango de valores: 0 a 2147483646.
	YARNContainerAllocated	Integer	Número de los container asignados a YARN. Rango de valores: 0 a 2147483646.

Tipo de clúster	Métrica	Tipo de valor	Descripción
	YARNContainerPending	Integer	Número de los container pendientes en Yarn. Rango de valores: 0 a 2147483646.
	YARNContainerPendingRatio	Ratio	Relación de los container pendientes en Yarn, es decir, la relación de los container pendientes a los containers en ejecución en Yarn. Rango de valores: 0 a 2147483646.
	YARNCPUAllocated	Integer	Número de CPU virtuales (vCPUs) asignadas a Yarn. Rango de valores: 0 a 2147483646.
	YARNCPUAvailable	Integer	Número de las vCPU disponibles en Yarn. Rango de valores: 0 a 2147483646.
	YARNCPUAvailablePercentage	Percentage	Porcentaje de las vCPU disponibles en Yarn, es decir, la proporción de vCPU disponibles respecto al total de vCPU. Rango de valores: 0 a 100.
	YARNCPUPending	Integer	Número de vCPU pendientes en Yarn. Rango de valores: 0 a 2147483646.
	YARNMemoryAllocated	Integer	Memoria asignada a Yarn. La unidad es MB. Rango de valores: 0 a 2147483646.
	YARNMemoryAvailable	Integer	Memoria disponible en Yarn. La unidad es MB. Rango de valores: 0 a 2147483646.
	YARNMemoryAvailablePercentage	Percentage	Porcentaje de memoria disponible en Yarn, es decir, la proporción de memoria disponible a memoria total en Yarn. Rango de valores: 0 a 100.
	YARNMemoryPending	Integer	Memoria pendiente en Yarn. Rango de valores: 0 a 2147483646.

Al agregar un plan de recursos, puede establecer parámetros haciendo referencia a **Tabla 8-2**.

Tabla 8-2 Conceptos de configuración de un plan de recursos

Parámetro	Descripción
Effective On	La fecha de entrada en vigor de un plan de recursos. **Daily** está seleccionado de forma predeterminada. También puede seleccionar uno o varios días de lunes a domingo.
Time Range	La hora de inicio y la hora de finalización de un plan de recursos son exactas a los minutos, con un valor que oscila entre **00:00** y **23:59**. Por ejemplo, si un plan de recursos comienza a las 8:00 y termina a las 10:00, establezca este parámetro en **8:00-10:00**. La hora de finalización debe ser al menos 30 minutos más tarde que la hora de inicio. Los intervalos de tiempo configurados para diferentes planes de recursos no pueden superponerse.
Node Range	El número de nodos de un plan de recursos varía entre **0** y **500**. En el intervalo de tiempo especificado en el plan de recursos, si el número de nodos de tarea es menor que el número mínimo especificado de nodos, se incrementará al valor mínimo especificado del intervalo de nodos a la vez. Si el número de nodos de tarea es mayor que el número máximo de nodos especificado en el plan de recursos, la función de escalado automático reduce el número de nodos de tarea al valor máximo del intervalo de nodos a la vez. La cantidad mínima de nodos debe ser inferior o igual a la cantidad máxima de estos.

9 Configuración de Hive con almacenamiento y cómputo desacoplado

El MRS permite almacenar datos en el OBS y utilizar un clúster de MRS solo para el cómputo de datos. De esta manera, el almacenamiento y los cálculos se desacoplan. Puede utilizar el servicio IAM para realizar configuraciones sencillas para acceder a OBS.

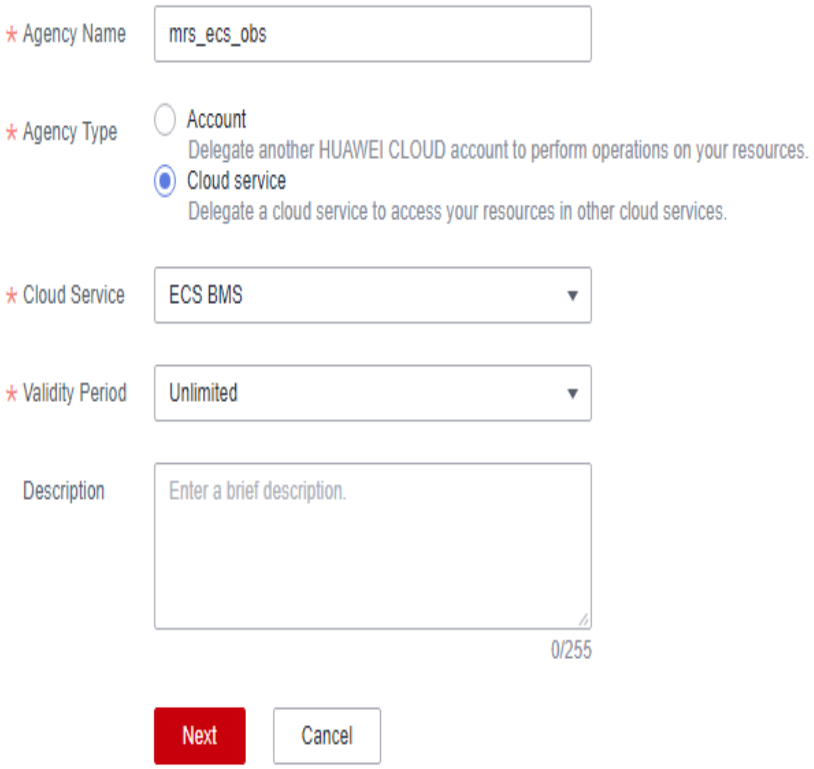
En esta sección se describe cómo crear una tabla Hive para almacenar datos en OBS.

1. **Creación de una agencia ECS**
2. **Configuración de una agencia para un clúster MRS**
3. **Creación de un sistema de archivos OBS**
4. **Acceso al sistema de archivos OBS a través de Hive**

Creación de una agencia ECS

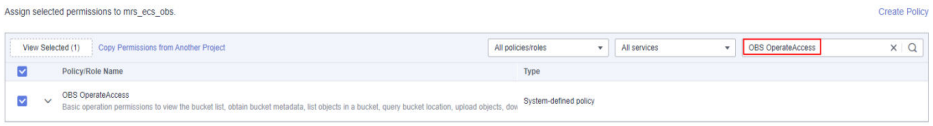
1. Inicie sesión en la consola de gestión de Huawei Cloud.
2. Elija **Service List** >**Management & Governance** >**Identity and Access Management**.
3. Haga clic en **Agencies**. En la página mostrada, haga clic en **Create Agency**.
4. Introduzca el nombre de una agencia, por ejemplo, **mrs_ecs_obs**.
5. Establezca **Agency Type** en **Cloud service** y seleccione **ECS BMS** para autorizar a ECS o BMS a invocar OBS.
6. Establezca **Validity Period** en **Unlimited** y haga clic en **Next**.

Figura 9-1 Creación de una agencia



7. En la página que se muestra, busque **OBS OperateAccess** en el cuadro de búsqueda y selecciónelo en la lista de resultados.

Figura 9-2 Asignación de permisos



8. Haga clic en **Next**. En la página que se muestra, seleccione el ámbito deseado para los permisos seleccionados. De forma predeterminada, se selecciona **All resources**. Haga clic en **Show More**, seleccione **Global resources** y haga clic en **OK**.
9. En el cuadro de diálogo que se muestra, haga clic en **OK** para iniciar la autorización. Después de que aparezca el mensaje "Authorization successful.", haga clic en **Finish**. La agencia se crea con éxito.

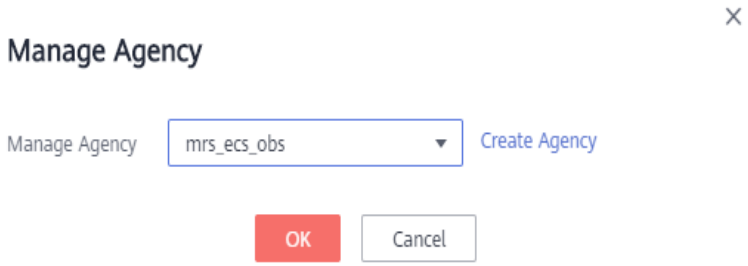
Configuración de una agencia para un clúster MRS

Puede configurar una agencia al crear un clúster o vincular una agencia a un clúster existente para desacoplar almacenamiento y procesamiento. En esta sección se utiliza un clúster existente como ejemplo para describir cómo configurar una agencia.

1. Inicie sesión en la consola de MRS. En el panel de navegación de la izquierda, elija **Clusters** >**Active Clusters**.
2. Haga clic en el nombre de un clúster para ir a la página de detalles del clúster.
3. En la página **Dashboard**, haga clic en **Synchronize** en el lado derecho de **IAM User Sync** para sincronizar usuarios de IAM.

4. En la página **Dashboard**, haga clic en **Manage Agency** en el lado derecho de **Agency** para seleccionar la agencia creada en **Creación de una agencia ECS**, y haga clic en **OK** para vincularla al clúster. Alternativamente, haga clic en **Create Agency** para ir a la consola de IAM para crear una agencia y vincularla al clúster.

Figura 9-3 Vinculación de una agencia

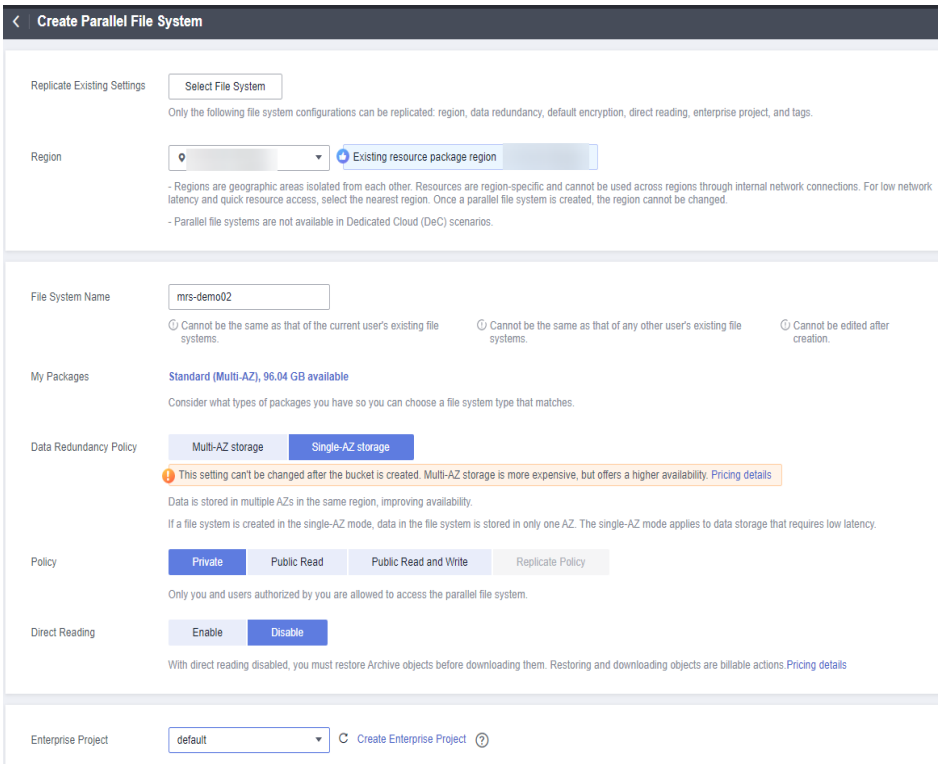


Creación de un sistema de archivos OBS

1. Inicie sesión en la consola OBS.
2. Elija **Parallel File System** > **Create Parallel File System**.
3. Introduzca el nombre del sistema de archivos, por ejemplo, **mrs-demo01**.

   Establezca otros parámetros según sea necesario.

Figura 9-4 Crear un sistema de archivos paralelo



4. Haga clic en **Create Now**.
5. En la lista del sistema de archivos paralelo de la consola OBS, haga clic en el nombre del sistema de archivos para ir a la página de detalles.

6. En el panel de navegación, elija **Files** y cree carpetas **program** y **input**.
   - **program**: Subir el paquete de programa a esta carpeta.
   - **input**: Subir los datos de entrada a esta carpeta.

Acceso al sistema de archivos OBS a través de Hive

1. Inicie sesión en un nodo master como usuario **root**. Para obtener más información, consulte **Iniciar sesión en un ECS**.
2. Verifique que Hive pueda acceder a OBS.
   a. Inicie sesión en el nodo master del clúster como usuario **root** y ejecute los siguientes comandos:

      cd /opt/Bigdata/client

      source bigdata_env

      source Hive/component_env

   b. Vea la lista de archivos en el sistema de archivos **mrs-demo01**.

      hadoop fs -ls obs://mrs-demo01/

   c. Compruebe si se devuelve la lista de archivos. Si se devuelve, el acceso a OBS se realiza correctamente.

      Figura 9-5 Consulta de la lista de archivos en mrs-demo01

      ```
      Found 2 items
      drwxrwxrwx   - hive hive          0 2021-10-22 10:08 obs://mrs-demo01/input
      drwxrwxrwx   - hive hive          0 2021-10-22 10:08 obs://mrs-demo01/program
      ```

   d. Ejecute el siguiente comando para autenticar al usuario (Sáltese este paso para un clúster normal, es decir, con la autenticación Kerberos deshabilitada):

      kinit hive

      Introduzca la contraseña del usuario **hive**. La contraseña predeterminada es **Hive@123**. Cambie la contraseña cuando inicie sesión por primera vez.

   e. Ejecute el comando de cliente de Hive.

      beeline

   f. Acceda al directorio OBS en Beeline. Por ejemplo, ejecute el siguiente comando para crear una tabla Hive y especifique que los datos se almacenan en la tabla **test_demo01** del sistema de archivos **mrs-demo01**:

      create table test_demo01(name string) location "obs://mrs-demo01/test_demo01";

   g. Ejecute el siguiente comando para consultar todas las tablas. Si se muestra la tabla **test_demo01** en la salida del comando, el acceso a OBS se realiza correctamente.

      show tables;

      Figura 9-6 Comprobar si existe la tabla test_demo01

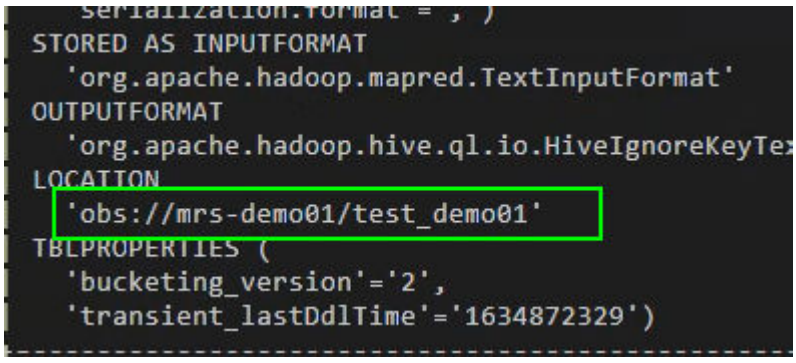
      ```
      +--------------+
      |   tab_name   |
      +--------------+
      | test_demo01  |
      +--------------+
      1 row selected (0.301 seconds)
      ```


h. Ejecute el siguiente comando para comprobar la ubicación de la tabla.

show create table test_demo01;

Compruebe si la ubicación de la tabla comienza con **obs:**//*OBS bucket name*/.

Figura 9-7 Comprobación de la ubicación de la tabla test_demo01

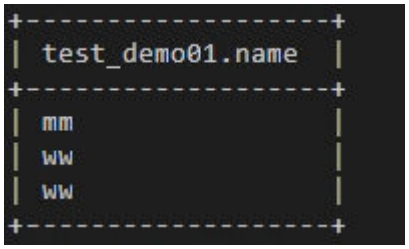


i. Ejecute el siguiente comando para escribir datos en la tabla.

insert into test_demo01 values('mm'),('ww'),('ww');

Ejecute el comando **select * from test_demo01;** para comprobar si los datos se escriben correctamente.

Figura 9-8 Consulta de datos en la tabla test_demo01



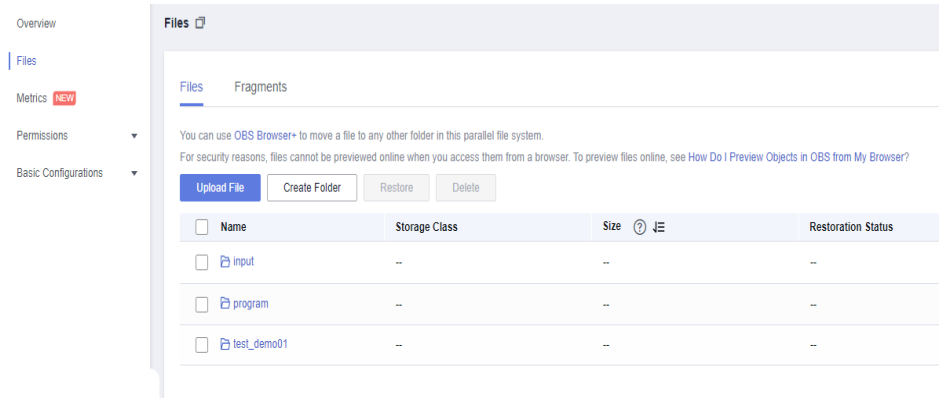
j. Ejecute el comando **!q** para salir del cliente Beeline.

k. Inicie sesión de nuevo en la consola OBS.

l. Haga clic en **Parallel File System** y seleccione el sistema de archivos creado.

m. Haga clic en **Files** para comprobar si los datos existen en la tabla creada.

Figura 9-9 Consulta de datos



10 Envío de tareas de Spark a nuevos nodos de Task

Agregue nodos de task a un clúster MRS personalizado para aumentar la capacidad informática. Los nodos de task se utilizan principalmente para procesar datos en lugar de almacenarlos permanentemente.

NOTA

Actualmente, los nodos de task solo se pueden agregar a clústeres MRS personalizados.

En esta sección se describe cómo enlazar un nuevo nodo de task con recursos del tenant y enviar tareas de Spark al nuevo nodo de task. Puede comenzar leyendo los siguientes temas:

1. **Adición de nodos de Task**
2. **Creación de un grupo de recursos**
3. **Creación de un tenant**
4. **Configuración de colas**
5. **Configuración de políticas de distribución de recursos**
6. **Creación de un usuario**
7. **Uso de spark-submit para enviar una tarea**
8. **Eliminación de nodos de Task**

Adición de nodos de Task

1. En la página de detalles de un clúster MRS personalizado, haga clic en la pestaña **Nodes**. En esta página de fichas, haga clic en **Add Node Group**.
2. En la página **Add Node Group** que se muestra, establezca los parámetros según sea necesario.

Tabla 10-1 Parámetros para agregar un grupo de nodos

Parámetro	Descripción
Instance Specifications	Seleccione el tipo de variante de los hosts en el grupo de nodos.

Parámetro	Descripción
Nodes	Configure el número de nodos en el grupo de nodos.
System Disk	Configure las especificaciones y la capacidad de los discos del sistema en los nuevos nodos.
Data Disk (GB)/Disks	Establezca las especificaciones, la capacidad y el número de discos de datos de los nuevos nodos.
Deploy Roles	Seleccione **NM** para agregar un rol de NodeManager.

3. Haga clic en **OK**.

Creación de un grupo de recursos

Paso 1 En la página de detalles del clúster, haga clic en **Tenants**.

Paso 2 Haga clic en **Resource Pools**.

Paso 3 Haga clic en **Create Resource Pool**.

Paso 4 En la página **Create Resource Pool**, establezca las propiedades del grupo de recursos.

- **Name**: Introduzca el nombre del grupo de recursos, por ejemplo, **test1**.
- **Resource Label**: Introduzca la etiqueta del grupo de recursos, por ejemplo, **1**.
- **Available Hosts**: Ingrese el nodo agregado a **Adición de nodos de Task**.

Paso 5 Haga clic en **OK**.

----Fin

Creación de un tenant

Paso 1 En la página de detalles del clúster, haga clic en **Tenants**.

Paso 2 Haga clic en **Create Tenant**. En la página mostrada, configure las propiedades de tenant.

Tabla 10-2 Parámetros del tenant

Parámetro	Descripción
Name	Establezca el nombre del tenant, por ejemplo, **tenant_spark**.
Tenant Type	Seleccione **Leaf**. Si se selecciona **Leaf**, el tenant actual es un tenant hoja y no se puede agregar ningún subtenant. Si se selecciona **Non-leaf**, se pueden agregar subtenants al tenant actual.
Dynamic Resource	Si se selecciona **Yarn**, el sistema crea automáticamente una cola de tareas con el nombre del tenant en Yarn. Si **Yarn** no está seleccionado, el sistema no crea automáticamente una cola de tareas.

Parámetro	Descripción
Default Resource Pool Capacity (%)	Establezca el porcentaje de recursos informáticos utilizados por el tenant actual en el grupo de recursos **default** por ejemplo, **20%**.
Default Resource Pool Max. Capacity (%)	Establezca el porcentaje máximo de recursos informáticos utilizados por el tenant actual en el grupo de recursos **default** por ejemplo, **80%**.
Storage Resource	Si se selecciona HDFS, el sistema crea automáticamente el directorio **/tenant** bajo el directorio raíz del HDFS cuando se crea un tenant por primera vez. Si **HDFS** no está seleccionado, el sistema no crea un directorio de almacenamiento bajo el directorio raíz del HDFS.
Maximum Number of Files/ Directories	Establezca el número máximo de archivos o directorios, por ejemplo, **100000000000**.
Storage Space Quota (MB)	Establezca la cuota para usar el espacio de almacenamiento, por ejemplo, 50000 MB. Este parámetro indica el espacio de almacenamiento HDFS máximo que puede utilizar un tenant, pero no el espacio real utilizado. Si su valor es mayor que el tamaño del disco físico HDFS, el espacio máximo disponible es el espacio completo del disco físico HDFS. **NOTA** Para garantizar la confiabilidad de los datos, el sistema genera automáticamente un archivo de copia de respaldo cuando se almacena un archivo en el HDFS. Es decir, dos réplicas del mismo archivo se almacenan de forma predeterminada. El espacio de almacenamiento HDFS indica el espacio total en disco ocupado por todas estas réplicas. Por ejemplo, si el valor de **Storage Space Quota** se establece en **500**, el espacio real para almacenar archivos es de aproximadamente 250 MB (500/2 = 250).
Storage Path	Establezca la ruta de almacenamiento, por ejemplo, **tenant/spark_test**. El sistema crea automáticamente una carpeta con el nombre del tenant en el directorio **/tenant** de forma predeterminada, por ejemplo, **spark_test**. El directorio de almacenamiento HDFS predeterminado para tenant **spark_test** es **tenant/spark_test**. Cuando se crea un tenant por primera vez, el sistema crea el directorio **/tenant** en el directorio raíz HDFS. La ruta de almacenamiento es personalizable.
Services	Establezca otros recursos de servicio asociados con el tenant actual. HBase es compatible. Para configurar este parámetro, haga clic en **Associate Services**. En el cuadro de diálogo que se muestra, establezca **Service** en **HBase**. Si **Association Mode** se establece en **Exclusive**, los recursos de servicio se ocupan exclusivamente. Si se selecciona **share**, se comparten los recursos de servicio.
Description	Introduzca la descripción del tenant actual.

Paso 3 Haga clic en **OK** para guardar la configuración.

Se tarda unos minutos en guardar la configuración. Si el **Tenant created successfully** se muestra en la esquina superior derecha, el tenant se agrega correctamente.

NOTA

- Los roles, los recursos informáticos y los recursos de almacenamiento se crean automáticamente cuando se crean los tenants.
- El nuevo rol tiene permisos sobre los recursos informáticos y de almacenamiento. El rol y sus permisos son controlados por el sistema automáticamente y no pueden ser controlados manualmente en **Manage Role**.
- Si desea utilizar el tenant, cree un usuario del sistema y asigne al usuario el rol Manager_tenant y el rol correspondiente al tenant.

----Fin

Configuración de colas

Paso 1 En la página de detalles del clúster, haga clic en **Tenants**.

Paso 2 Haga clic en la pestaña **Queue Configuration**.

Paso 3 En la tabla de colas de tenant, haga clic en **Modify** en la columna **Operation** de la cola de tenant especificada.

NOTA

- En la lista de tenants a la izquierda de la página **Tenant Management**, haga clic en el tenant de destino. En la ventana que se muestra, elija **Resource**. En la página mostrada, haga clic en para abrir la página de modificación de cola.
- Una cola puede estar enlazada a un solo grupo de recursos no predeterminado.

De forma predeterminada, la etiqueta de recurso es la especificada en **Creación de un grupo de recursos**. Establezca otros parámetros en función de los requisitos del sitio.

Paso 4 Haga clic en **OK**.

----Fin

Configuración de políticas de distribución de recursos

Paso 1 En la página de detalles del clúster, haga clic en **Tenants**.

Paso 2 Haga clic en **Resource Distribution Policies** y seleccione el grupo de recursos creado en **Creación de un grupo de recursos**.

Paso 3 Busque la fila que contiene **tenant_spark** y haga clic en **Modify** en la columna **Operation**.

- **Weight**: **20**
- **Minimum Resource**: **20**
- **Maximum Resource**: **80**
- **Reserved Resource**: **10**

Paso 4 Haga clic en **OK**.

----Fin

Creación de un usuario

Paso 1 Inicie sesión en FusionInsight Manager. Para obtener más información, consulte **Acceso a FusionInsight Manager**.

Paso 2 Elija **System** > **Permission** > **User**. En la página mostrada, haga clic en **Create User**.

- **Username**: **spark_test**
- **User Type**: **Human-Machine**
- **User Group**: hadoop and hive
- **Primary Group**: **hadoop**
- **Role**: **tenant_spark**

Paso 3 Haga clic en **OK** para agregar el usuario.

----Fin

Uso de spark-submit para enviar una tarea

1. Inicie sesión en el nodo cliente como usuario **root** y ejecute los siguientes comandos:

   cd *Client installation directory*

   source bigdata_env

   source Spark2x/component_env

   Para un clúster con autenticación Kerberos habilitada, ejecute el comando **kinit spark_test**. Para un clúster con la autenticación Kerberos deshabilitada, omita este paso.

   Introduzca la contraseña para la autenticación. Cambie la contraseña cuando inicie sesión por primera vez.

   cd Spark2x/spark/bin

   sh spark-submit --queue tenant_spark --class org.apache.spark.examples.SparkPi --master yarn-client ../examples/jars/spark-examples_*.jar

Eliminación de nodos de Task

1. En la página de detalles del clúster, haga clic en **Nodes**.
2. Busque la fila que contiene el grupo de nodos de tarea de destino y haga clic en **Scale In** en la columna **Operation**.
3. Establezca el **Scale-In Type** en **Specific node** y seleccione los nodos de destino.

   NOTA

   Los nodos de destino necesitan ser apagados.

4. Seleccione **I understand the consequences of performing the scale-in operation** y haga clic en **OK**.

11 Configuración de Umbrales para Alarmas

Los clústeres MRS proporcionan funciones alarmantes fáciles de usar con vistas métricas de monitorización intuitivas. Puede ver rápidamente estadísticas sobre las métricas clave de rendimiento (KPI) de un clúster y evaluar el estado del clúster. MRS le permite configurar umbrales de métricas para mantenerse informado del estado del clúster. Si se cumple un valor umbral, el sistema genera y muestra una alarma en el panel de métricas.

Si se **comprueba** que el impacto de algunas alarmas en los servicios puede ignorarse o que es necesario ajustar los umbrales de alarma, puede personalizar las métricas del clúster o enmascarar algunas alarmas según sea necesario.

Puede establecer umbrales para alarmas de métricas de información de nodo y métricas de servicio de clúster. Para obtener más información sobre estas métricas, sus impactos en el sistema y los umbrales predeterminados, consulte **Referencia de métrica de monitoreo**.

AVISO

Estas alarmas pueden afectar a las funciones del clúster o a la ejecución del trabajo. Si desea enmascarar o modificar las reglas de alarma, evalúe los riesgos de operación con anticipación.

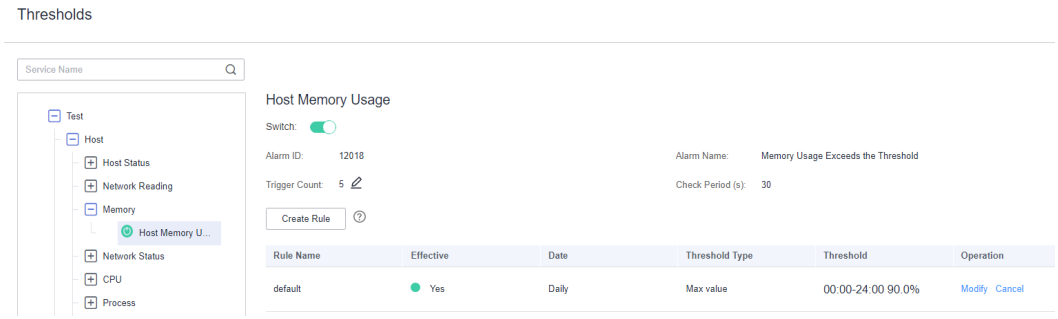
Modificación de reglas para alarmas con umbrales personalizados

Paso 1 Inicie sesión en FusionInsight Manager del clúster MRS de destino haciendo referencia a **Acceso a FusionInsight Manager (MRS 3.x o posterior)** (MRS 3.x o posterior).

Paso 2 Elija **O&M** > **Alarm** > **Thresholds**.

Paso 3 Seleccione una métrica para un host o servicio en el clúster. Por ejemplo, seleccione **Host Memory Usage**.

Figura 11-1 Consulta de un umbral de alarma



- **Switch**: Si este interruptor está encendido, se activará una alarma cuando la métrica incumpla este umbral.
- **Trigger Count**: Manager comprueba si la métrica cumple con el valor umbral. Si el número de comprobaciones consecutivas donde falla la métrica es igual al valor de **Trigger Count**, se genera una alarma. El valor se puede personalizar. **Si una alarma se notifica con frecuencia, puede ajustar Trigger Count a un valor mayor para reducir la frecuencia de las alarmas.**
- **Check Period (s)**: Intervalo entre cada dos comprobaciones
- Las reglas para activar alarmas se enumeran en la página.

Paso 4 Modificar una regla de alarma.

- Agregar una nueva regla.
  a. Haga clic en **Create Rule** para agregar una regla que define cómo se activará una alarma. Para obtener más información, consulte **Tabla 11-1**.
  b. Haga clic en **OK** para guardar la regla.
  c. Busque la fila que contiene una regla que está en uso y haga clic en **Cancel** en la columna **Operation**. Si no hay ninguna regla en uso, omita este paso.
  d. Busque la fila que contiene la nueva regla y haga clic en **Apply** en la columna **Operation**. El valor de **Effective** para esta regla cambia a **Yes**.
- Modificar una regla existente.
  a. Haga clic en **Modify** en la columna **Operation** de la fila que contiene la regla de destino.
  b. Modifique los parámetros de regla haciendo referencia a **Tabla 11-1**.
  c. Haga clic en **OK**.

En la siguiente tabla se enumeran los parámetros de regla que debe establecer para activar una alarma de **Host Memory Usage**.

Tabla 11-1 Parámetros de reglas de alarma

Parámetro	Descripción	Valor de ejemplo
Rule Name	Nombre de la regla	mrs_test

Parámetro	Descripción	Valor de ejemplo
Severity	Gravedad de alarma. Las opciones son las siguientes: ● **Critical** ● **Major** ● **Minor** ● **Warning**	Major
Threshold Type	Valor máximo o mínimo de una métrica ● **Max value**: Se generará una alarma cuando el valor de la métrica sea mayor que este valor. ● **Min value**: Se generará una alarma cuando el valor de la métrica sea menor que este valor.	Max. Value
Date	Con qué frecuencia entra en vigor la regla ● Daily ● Weekly ● Others	Daily
Add Date	Fecha en que la regla entra en vigor. Este parámetro solo está disponible cuando **Date** está establecido en **Others**. Puede establecer varias fechas.	-
Thresholds	**Start and End Time**: Período en el que la regla entra en vigor.	00:00 - 23:59
	Threshold: Valor umbral de alarma	85

----Fin

Alarmas especificadas de enmascaramiento

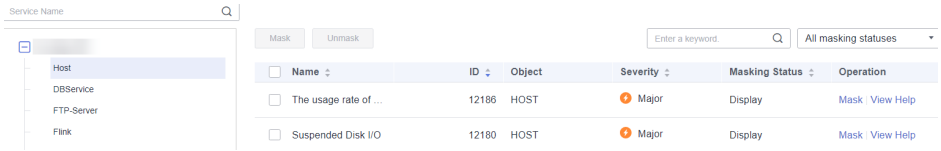
Paso 1 Inicie sesión en FusionInsight Manager del clúster MRS de destino haciendo referencia a **Acceso a FusionInsight Manager (MRS 3.x o posterior)** (MRS 3.x o posterior).

Paso 2 Elija **O&M** > **Alarm** > **Masking**.

Paso 3 En la lista a la izquierda de la página mostrada, seleccione el servicio o módulo de destino.

Paso 4 Haga clic en **Mask** en la columna **Operation** de la alarma que desea enmascarar. En el cuadro de diálogo que se muestra, haga clic en **OK** para cambiar el estado de enmascaramiento de la alarma a **Mask**.

Figura 11-2 Enmascarar una alarma



NOTA

- Puede buscar las alarmas especificadas en la lista.
- Para cancelar el enmascaramiento de alarma, haga clic en **Unmask** en la fila de la alarma de destino. En el cuadro de diálogo que se muestra, haga clic en **OK** para cambiar el estado de enmascaramiento de alarma a **Display**.
- Si necesita realizar operaciones con varias alarmas a la vez, seleccione las alarmas y haga clic en **Mask** o **Unmask** en la parte superior de la lista.

----Fin

Preguntas frecuentes

- **¿Cómo puedo ver las alarmas no confirmadas de un cluster?**

  a. Inicie sesión en la consola de gestión de MRS.

  b. Haga clic en el nombre del clúster de destino y haga clic en la pestaña **Alarms**.

  c. Haga clic en **Advanced Search**, establezca **Alarm Status** en **Uncleared**, y haga clic en **Search**.

  d. Se muestran las alarmas borradas del clúster actual.

- **¿Cómo borro una alarma de clúster?**

  Puede manejar las alarmas haciendo referencia a la ayuda de alarma. Para ver el documento de ayuda, realice los siguientes pasos:

  - Consola: Inicie sesión en la consola de gestión de MRS, haga clic en el nombre del clúster de destino, haga clic en la pestaña **Alarms** y haga clic en **View Help** en la columna **Operation** de la lista de alarmas. A continuación, borre la alarma haciendo referencia al procedimiento de manejo de alarmas.
  - Manager: Inicie sesión en FusionInsight Manager, seleccione **O&M** > **Alarm** > **Alarms**, y haga clic en **View Help** en la columna **Operation**. A continuación, borre la alarma haciendo referencia al procedimiento de manejo de alarmas.

Referencia de métrica de monitoreo

Las métricas de monitoreo de FusionInsight Manager se clasifican como métricas de información de nodo y métricas de servicio de clúster. **Tabla 11-2** enumera las métricas cuyos umbrales se pueden configurar en un nodo, y **Tabla 11-3** enumera las métricas cuyos umbrales se pueden configurar para un componente.

Tabla 11-2 Métricas de monitorización de nodos

Grupo métrico	Métrica	ID	Alarma	Impacto en el sistema	Umbral predeterminado
CPU	Uso de la CPU del host	12016	El uso de la CPU excede el umbral	Los procesos de servicio responden lentamente o no están disponibles.	90.0%

Grupo métrico	Métrica	ID	Alarma	Impacto en el sistema	Umbral predeterminado
Disco	Uso de disco	12017	Capacidad de disco insuficiente	Los procesos de servicio no están disponibles.	90.0%
	Uso de Inode de disco	12051	El uso del Inode de Disco Supera el Umbral	Los datos no se pueden escribir correctamente en el sistema de archivos.	80.0%
Memoria	Uso de memoria de host	12018	El uso de la memoria excede el umbral.	Los procesos de servicio responden lentamente o no están disponibles.	90.0%
Estado de host	Uso del identificador de archivo de host	12053	El uso del identificador de archivo de host supera el umbral	Las operaciones de E/S, como abrir un archivo o conectarse a la red, no se pueden realizar y los programas son anormales.	80.0%
	Uso de PID de host	12027	El uso de PID del host supera el umbral	No hay ningún PID disponible para los nuevos procesos y los procesos de servicio no están disponibles.	90%
Estado de la red	Uso del puerto temporal TCP	12052	El uso de puerto temporal TCP supera el umbral	Los servicios en el host no pueden establecer conexiones con el externo y los servicios se interrumpen.	80.0%
Lectura de red	Tasa de error de paquete de lectura	12047	La tasa de error de paquetes de lectura supera el umbral	La comunicación se interrumpe intermitentemente y los servicios expiran.	0.5%
	Tasa de paquetes perdidos de lectura	12045	La tasa de caída de paquetes de lectura supera el umbral	El rendimiento del servicio se deteriora o el tiempo de espera de algunos servicios.	0.5%

Grupo métrico	Métrica	ID	Alarma	Impacto en el sistema	Umbral predeterminado
	Tasa de rendimiento de lectura	12049	La tasa de rendimiento de lectura supera el umbral	El sistema de servicio se ejecuta de forma anormal o no está disponible.	80%
Escritura de red	Tasa de errores de paquetes de escritura	12048	La tasa de error de paquete de escritura supera el umbral	La comunicación se interrumpe intermitentemente y los servicios expiran.	0.5%
	Tasa de paquetes perdidos de escritura	12046	La tasa de escritura de paquetes caídos supera el umbral	El rendimiento del servicio se deteriora o el tiempo de espera de algunos servicios.	0.5%
	Tasa de rendimiento de escritura	12050	La tasa de rendimiento de escritura supera el umbral	El sistema de servicio se ejecuta de forma anormal o no está disponible.	80%
Proceso	Número total de procesos en los estados D y Z	12028	Número de procesos en el Estado D y Z en un host supera el umbral	Se utilizan recursos excesivos del sistema y los procesos de servicio responden lentamente.	0
	Uso del proceso omm	12061	El uso del proceso supera el umbral	El cambio al **omm** de usuario falla. No se puede crear un nuevo proceso de omm.	90

Tabla 11-3 Métricas de monitoreo de clústeres

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
DBService	Uso del número de conexiones de base de datos	27005	El uso de conexión a base de datos supera el umbral	Los servicios de capa superior pueden no conectarse a la base de datos de DBService, lo que afecta a los servicios.	90%
	Uso del espacio en disco del directorio de datos	27006	El uso de espacio en disco del directorio de datos supera el umbral	Los procesos de servicio no están disponibles. Cuando el uso de espacio en disco del directorio de datos supera el 90%, la base de datos entra en el modo de solo lectura y se genera **Database Enters the Read-Only Mode**. Como resultado, se pierden datos de servicio.	80%
Flume	Porcentaje de recursos de memoria heap	24006	El uso de memoria heap del Flume Server supera el umbral	El desbordamiento de la memoria heap puede causar una falla en el servicio.	95.0%
	Estadísticas de uso de memoria directa	24007	El uso de memoria directa del Flume Server supera el umbral	El desbordamiento de la memoria directa puede provocar una falla en el servicio.	80.0%
	Uso de memoria no heap	24008	El uso de memoria no heap de Flume Server supera el umbral	El desbordamiento de la memoria no heap puede provocar una falla en el servicio.	80.0%
	Duración total del GC	24009	La duración de Flume Server GC supera el umbral	La eficiencia de transmisión de datos de Flume disminuye.	12000ms

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
HBase	Duración de GC de generación Old	19007	La duración de HBase GC supera el umbral	Si la duración de GC de generación anterior excede el umbral, la lectura y escritura de datos de HBase se ven afectadas.	5000ms
	Estadísticas de uso de memoria directa de RegionServer	19009	El uso de memoria directa del proceso HBase supera el umbral	Si la memoria directa HBase disponible es insuficiente, se produce un desbordamiento de memoria y el servicio se interrumpe.	90%
	Estadísticas de uso de memoria heap de RegionServer	19008	El uso de memoria heap del proceso HBase supera el umbral	Si la memoria HBase disponible es insuficiente, se produce un desbordamiento de memoria y el servicio se interrumpe.	90%
	Uso de memoria directa de HMaster	19009	El uso de memoria directa del proceso HBase supera el umbral	Si la memoria directa HBase disponible es insuficiente, se produce un desbordamiento de memoria y el servicio se interrumpe.	90%
	Estadísticas de uso de memoria heap de HMaster.	19008	El uso de memoria heap del proceso HBase supera el umbral	Si la memoria HBase disponible es insuficiente, se produce un desbordamiento de memoria y el servicio se interrumpe.	90%

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
	Número de regiones en línea de un RegionServer	19011	Número de regions de RegionServer supera el umbral	El rendimiento de lectura/escritura de datos de HBase se ve afectado cuando el número de regions en un RegionServer excede el umbral.	2000
	Región en el estado de RIT que alcanza el umbral de duración	19013	La duración de regiones en el estado de RIT supera el umbral	Algunos datos de la tabla se pierden o no están disponibles.	1
	Uso de handler de RegionServer	19021	Número de handlers activos de RegionServer supera el umbral	RegionServers y HBase no pueden proporcionar servicios correctamente.	90%
	Errores de sincronización en la recuperación ante desastres	19006	Error de sincronización de replicación de HBase	Los datos de HBase en un clúster no se sincronizan con el clúster en espera, lo que provoca incoherencia de datos entre los clústeres activos y en espera.	1
	Número de archivos de registro que se sincronizarán en el clúster activo	19020	El número de archivos WAL de HBase a sincronizar supera el umbral	Si el número de archivos WAL a sincronizar por un RegionServer excede el umbral, el número de ZNodes utilizados por HBase excede el umbral, lo que afecta al estado del servicio HBase.	128

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
	Número de HFiles que se van a sincronizar en el clúster activo	19019	El número de HFiles a sincronizar supera el umbral	Si el número de HFiles a sincronizar por un RegionServer excede el umbral, el número de ZNodes utilizados por HBase excede el umbral, afectando el estado del servicio HBase.	128
	Tamaño de la cola de Compaction	19018	El tamaño de la cola de compactación de HBase supera el umbral	El rendimiento del clúster puede deteriorarse, lo que afecta a la lectura y escritura de datos.	100
HDFS	Bloques perdidos	14003	Número de bloques HDFS perdidos supera el umbral	Los datos almacenados en HDFS se pierden. HDFS puede entrar en el modo de seguridad y no puede proporcionar servicios de escritura. Los datos de bloques perdidos no se pueden restaurar.	0
	Bloques bajo replicación	14028	El número de bloques a complementar supera el umbral	Los datos almacenados en HDFS se pierden. HDFS puede entrar en el modo de seguridad y no puede proporcionar servicios de escritura. Los datos de bloques perdidos no se pueden restaurar.	1000

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
	Tiempo promedio de procesamiento de RPC de NameNode activo	14021	El tiempo promedio de procesamiento de RPC de NameNode supera el umbral	NameNode no puede procesar las solicitudes RPC de clientes HDFS, servicios de capa superior que dependen de HDFS y DataNode de manera oportuna. Específicamente, los servicios que acceden a HDFS se ejecutan lentamente o el servicio HDFS no está disponible.	100ms
	Tiempo promedio de la cola NameNode RPC activa	14022	El tiempo promedio de cola de NameNode RPC supera el umbral	NameNode no puede procesar las solicitudes RPC de clientes HDFS, servicios de capa superior que dependen de HDFS y DataNode de manera oportuna. Específicamente, los servicios que acceden a HDFS se ejecutan lentamente o el servicio HDFS no está disponible.	200ms
	Uso de disco HDFS	14001	El uso del disco HDFS supera el umbral	El rendimiento de la escritura de datos en HDFS se ve afectado.	80%
	Uso del disco DataNode	14002	El uso del disco DataNode supera el umbral	La falta de espacio en disco afectará a la escritura de datos en HDFS.	80%

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
	Porcentaje de espacio reservado para réplicas de espacio no utilizado	14023	El porcentaje de espacio total en disco reservado para réplicas supera el umbral	El rendimiento de la escritura de datos en HDFS se ve afectado. Si todo el espacio de DataNode no utilizado está reservado para réplicas, se produce un error al escribir datos de HDFS.	90%
	Total de DataNodes defectuosos	14009	Número de Dead DataNodes supera el umbral	Los DataNodes defectuosos no pueden proporcionar servicios HDFS.	3
	Estadísticas de uso de memoria no heap de NameNode	14018	El uso de memoria no heap de NameNode supera el umbral	Si el uso de memoria no heap del HDFS NameNode es demasiado alto, el rendimiento de lectura/escritura de datos de HDFS se verá afectado.	90%
	Estadísticas de uso de memoria directa de NameNode	14017	El uso de memoria directa de NameNode supera el umbral	Si la memoria directa disponible de las instancias de NameNode es insuficiente, puede producirse un desbordamiento de memoria y el servicio se interrumpe.	90%
	Estadísticas de uso de memoria heap de NameNode	14007	El uso de memoria heap de NameNode supera el umbral	Si el uso de memoria heap de HDFS NameNode es demasiado alto, el rendimiento de lectura/escritura de datos de HDFS se verá afectado.	95%

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
	Estadísticas de uso de memoria directa de DataNode	14016	El uso de memoria directa de DataNode supera el umbral	Si la memoria directa disponible de las instancias de DataNode es insuficiente, puede producirse un desbordamiento de memoria y el servicio se interrumpe.	90%
	Estadísticas de uso de memoria heap de DataNode	14008	El uso de memoria heap de DataNode supera el umbral	El uso de la memoria heap de HDFS DataNode es demasiado alto, lo que afecta al rendimiento de lectura/escritura de datos del HDFS.	95%
	Estadísticas de uso de memoria no heap de DataNode	14019	El uso de memoria no heap de DataNode supera el umbral	Si el uso de memoria no heap del HDFS DataNode es demasiado alto, el rendimiento de lectura/escritura de datos de HDFS se verá afectado.	90%
	Estadísticas de duración de GC de NameNode	14014	La duración de GC de NameNode supera el umbral	Una larga duración de GC del proceso NameNode puede interrumpir los servicios.	12000ms
	Estadísticas de duración de GC de DataNode	14015	La duración de GC de DataNode supera el umbral	Una larga duración GC del proceso DataNode puede interrumpir los servicios.	12000ms
Hive	Tasa de éxito de ejecución de SQL de Hive (porcentaje)	16002	La Tasa de Éxito de Ejecución SQL de Hive es inferior al Umbral	La configuración y el rendimiento del sistema no pueden cumplir los requisitos de procesamiento del servicio.	90.0%

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
	Uso de subprocesos de background	16003	El uso de subprocesos de background supera el umbral	Hay demasiados subprocesos de background, por lo que la tarea recién enviada no puede ejecutarse a tiempo.	90%
	Duración total del GC de MetaStore	16007	La duración de Hive GC supera el umbral	Si la duración de GC excede el umbral, la lectura y escritura de los datos de Hive se verán afectados.	12000ms
	Duración total del GC de HiveServer	16007	La duración de Hive GC supera el umbral	Si la duración de GC excede el umbral, la lectura y escritura de los datos de Hive se verán afectados.	12000ms
	Porcentaje de espacio HDFS utilizado por Hive con respecto al espacio disponible	16001	El uso del espacio en el almacén de Hive supera el umbral	El sistema no puede escribir datos, lo que causa la pérdida de datos.	85.0%
	Estadísticas de uso de memoria directa de MetaStore	16006	El uso de memoria directa del proceso Hive supera el umbral	Cuando el uso de memoria directa de Hive es excesivo, el rendimiento de la operación de tarea Hive se ve afectado. Además, puede producirse un desbordamiento de memoria para que el servicio Hive no esté disponible.	95%

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
	Estadísticas de uso de memoria no heap de MetaStore	16008	El uso de memoria no heap del servicio Hive supera el umbral	Cuando el uso de memoria no heap de Hive es excesivo, el rendimiento de la operación de tarea Hive se ve afectado. Además, puede producirse un desbordamiento de memoria para que el servicio Hive no esté disponible.	95%
	Estadísticas de uso de memoria heap de MetaStore	16005	El uso de memoria heap del proceso Hive supera el umbral	Cuando el uso de memoria heap de Hive es excesivo, el rendimiento de la operación de tarea de Hive se ve afectado. Además, puede producirse un desbordamiento de memoria para que el servicio Hive no esté disponible.	95%
	Estadísticas de uso de memoria directa de HiveServer	16006	El uso de memoria directa del proceso Hive supera el umbral	Cuando el uso de memoria directa de Hive es excesivo, el rendimiento de la operación de tarea Hive se ve afectado. Además, puede producirse un desbordamiento de memoria para que el servicio Hive no esté disponible.	95%

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
	Estadísticas de uso de memoria no heap de HiveServer	16008	El uso de memoria no heap del servicio Hive supera el umbral	Cuando el uso de memoria no heap de Hive es excesivo, el rendimiento de la operación de tarea Hive se ve afectado. Además, puede producirse un desbordamiento de memoria para que el servicio Hive no esté disponible.	95%
	Estadísticas de uso de memoria heap de HiveServer	16005	El uso de memoria heap del proceso Hive supera el umbral	Cuando el uso de memoria heap de Hive es excesivo, el rendimiento de la operación de tarea de Hive se ve afectado. Además, puede producirse un desbordamiento de memoria para que el servicio Hive no esté disponible.	95%
	Porcentaje de Sessions conectadas al HiveServer con respecto al número máximo de Sessions permitidas por el HiveServer	16000	El porcentaje de Sessions conectadas al HiveServer al número máximo permitido supera el umbral	Si se genera una alarma de conexión, se conectan demasiadas sessions al HiveServer y no se pueden crear nuevas conexiones.	90.0%
Kafka	Porcentaje de Partitions que no están completamente sincronizadas	38006	El porcentaje de particiones de Kafka que no están completamente sincronizadas supera el umbral	Demasiadas particiones de Kafka que no están completamente sincronizadas afectan a la confiabilidad del servicio. Además, los datos pueden perderse cuando se conmutan Leader.	50%

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
	Uso de la conexión de usuario en Broker	38011	El uso de la conexión de usuario en el Broker supera el umbral	Si el número de conexiones de un usuario es excesivo, el usuario no puede crear nuevas conexiones al Broker.	80%
	Uso del disco de Broker	38001	Capacidad insuficiente del disco Kafka	Las operaciones de escritura de datos Kafka fallan.	80.0%
	Tasa de E/S de disco de un Broker	38009	E/S de disco de Broker ocupado	La partición de disco tiene E/S frecuentes. Es posible que los datos no se escriban en el topic de Kafka para el que se genera la alarma.	80%
	Duración de GC de Broker por minuto	38005	La duración de GC del proceso de Broker supera el umbral	Una larga duración de GC del proceso de Broker puede interrumpir los servicios.	12000ms
	Uso de memoria heap de Kafka	38002	El uso de memoria heap de Kafka supera el umbral	Si la memoria heap de Kafka disponible es insuficiente, se produce un desbordamiento de memoria y el servicio se interrumpe.	95%
	Uso de memoria directa de Kafka	38004	El uso de memoria directa de Kafka supera el umbral	Si la memoria directa disponible del servicio Kafka es insuficiente, se produce un desbordamiento de memoria y el servicio se interrumpe.	95%

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
Loader	Uso de memoria heap	23004	El uso de memoria heap de Loader supera el umbral	El desbordamiento de la memoria heap puede causar una falla en el servicio.	95%
	Estadísticas de uso de memoria directa	23006	El uso de memoria directa de Loader supera el umbral	El desbordamiento de la memoria directa puede provocar una falla en el servicio.	80.0%
	Uso de memoria no heap	23005	El uso de memoria no heap de Loader supera el umbral	El desbordamiento de la memoria no heap puede provocar una falla en el servicio.	80%
	Duración total del GC	23007	La duración de GC del proceso de Loader supera el umbral	La respuesta del servicio del Loader es lenta.	12000ms
MapReduce	Estadísticas de duración de GC	18012	La duración de GC de JobHistoryServer supera el umbral	Una larga duración GC del proceso de JobHistoryServer puede interrumpir los servicios.	12000ms
	Estadísticas de uso de memoria directa de JobHistoryServer	18015	El uso de memoria directa de JobHistoryServer supera el umbral	Si la memoria directa disponible del servicio MapReduce es insuficiente, se produce un desbordamiento de memoria y el servicio se interrumpe.	90%

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
	Estadísticas de uso de memoria no heap de JobHistoryServer	18019	El uso de memoria no heap de JobHistoryServer supera el umbral	Cuando el uso de memoria no heap de MapReduce JobHistoryServer es excesivo, el rendimiento del envío y la operación de tareas de MapReduce se ve afectado. Además, puede producirse un desbordamiento de memoria para que el servicio MapReduce no esté disponible.	90%
	Estadísticas de uso de memoria heap de JobHistoryServer	18009	El uso de memoria heap de JobHistoryServer supera el umbral	Cuando el uso de memoria heap de JobHistoryServer de MapReduce es excesivo, el rendimiento del archivo de registros de MapReduce se ve afectado. Además, puede producirse un desbordamiento de memoria para que el servicio Yarn no esté disponible.	95%
Oozie	Uso de memoria heap	17004	El uso de memoria heap de Oozie supera el umbral	El desbordamiento de la memoria heap puede causar una falla en el servicio.	95.0%
	Uso de memoria directa	17006	El uso de memoria directa de Oozie supera el umbral	El desbordamiento de la memoria directa puede provocar una falla en el servicio.	80.0%
	Uso de memoria no heap	17005	El uso de memoria no heap de Oozie supera el umbral	El desbordamiento de la memoria no heap puede provocar una falla en el servicio.	80%

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
	Duración total de GC	17007	La duración de GC del proceso Oozie supera el umbral	Oozie responde lentamente cuando se utiliza para enviar tareas.	12000ms
Spark2x	Estadísticas de uso de memoria heap de JDBCServer2x	43010	El uso de memoria heap del proceso JDBCServer2x supera el umbral	Si la memoria heap de procesos JDBCServe2x disponible es insuficiente, se produce un desbordamiento de memoria y el servicio se rompe	95%
	Estadísticas de uso de memoria directa de JDBCServer2x	43012	El uso memoria heap directa del proceso JDBCServer2x supera el umbral	Si la memoria heap directa del proceso de JDBCServer2x disponible es insuficiente, se produce un desbordamiento de memoria y el servicio se interrumpe.	95%
	Estadísticas de uso de memoria no heap de JDBCServer2x	43011	El uso de memoria heap del proceso JDBCServer2x supera el umbral	Si la memoria no heap de proceso de JDBCServer2x disponible es insuficiente, se produce un desbordamiento de memoria y el servicio se interrumpe.	95%
	Estadísticas de uso de memoria directa de JobHistory2x	43008	El uso de memoria directa del proceso JobHistory2x supera el umbral	Si la memoria directa disponible del proceso JobHistory2x es insuficiente, se produce un desbordamiento de memoria y el servicio se rompe.	95%

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
	Estadísticas de uso de memoria no heap de JobHistory2x	43007	El uso de memoria no heap del proceso JobHistory2x supera el umbral	Si la memoria no heap de JobHistory2x Process disponible es insuficiente, se produce un desbordamiento de memoria y el servicio se rompe.	95%
	Estadísticas de uso de memoria heap de JobHistory2x	43006	El uso de memoria heap del proceso JobHistory2x supera el umbral	Si la memoria heap de procesos de JobHistory2x disponible es insuficiente, se produce un desbordamiento de memoria y el servicio se interrumpe.	95%
	Estadísticas de uso de memoria directa de IndexServer2x	43021	El uso de memoria directa del proceso IndexServer2x supera el umbral	Si la memoria directa del proceso IndexServer2x disponible es insuficiente, se produce un desbordamiento de memoria y el servicio se interrumpe.	95%
	Estadísticas de uso de memoria heap de IndexServer2x	43019	El uso de memoria heap del proceso IndexServer2x supera el umbral	Si la memoria heap de procesos de IndexServer2x disponible es insuficiente, se produce un desbordamiento de memoria y el servicio se interrumpe.	95%

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
	Estadísticas de uso de memoria no heap de IndexServer2x	43020	El uso de memoria no heap del proceso IndexServer2x supera el umbral	Si la memoria no heap del proceso IndexServer2x disponible es insuficiente, se produce un desbordamiento de memoria y el servicio se interrumpe.	95%
	Número de Full GC de JDBCServer2x	43017	El número de Full GC del proceso JDBCServer2x supera el umbral	El rendimiento del proceso JDBCServer2x se ve afectado, o incluso el proceso JDBCServer2x no está disponible.	12
	Número de Full GC de JobHistory2x	43018	El número de Full GC de proceso JobHistory2x supera el umbral	El rendimiento del proceso de JobHistory2x se ve afectado, o incluso el proceso de JobHistory2x no está disponible.	12
	Número de Full GC de IndexServer2x	43023	El número de Full GC del proceso IndexServer2x supera el umbral	Si el número de GC excede el umbral, IndexServer2x puede ejecutarse con bajo rendimiento o incluso no está disponible.	12
	Duración total de GC (en milisegundos) de JDBCServer2x	43013	La duración de GC del proceso JDBCServer2x supera el umbral	Si la duración de GC excede el umbral, JDBCServer2x puede ejecutarse con bajo rendimiento.	12000ms
	Duración total de GC (en milisegundos) de JobHistory2x	43009	La duración de GC del proceso JobHistory2x supera el umbral	Si la duración de GC excede el umbral, JobHistory2x puede ejecutarse con bajo rendimiento.	12000ms

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
	Duración total de GC (en milisegundos) de IndexServer2x	43022	La duración de GC del proceso IndexServer2x supera el umbral	Si la duración de GC excede el umbral, IndexServer2x puede ejecutarse con bajo rendimiento o incluso no estar disponible.	12000ms
Storm	Número de Supervisor disponible	26052	El número de supervisor disponible del servicio de Storm es menor que el umbral	No se pueden realizar tareas existentes en el clúster. El clúster puede recibir nuevas tareas de Storm, pero no puede realizar estas tareas.	1
	Uso de Slot	26053	El uso de Storm Slot supera el umbral	No se pueden realizar nuevas tareas de Storm.	80.0%
	Uso de memoria heap de Nimbus	26054	El uso de memoria heap de Nimbus supera el umbral	Cuando el uso de memoria heap de Storm Nimbus es demasiado alto, se producen GC frecuentes. Además, puede producirse un desbordamiento de memoria para que el servicio Yarn no esté disponible.	80%
Yarn	Estadísticas de uso de memoria directa de NodeManager	18014	El uso de memoria directa de NodeManager supera el umbral	Si la memoria directa disponible de NodeManager es insuficiente, se produce un desbordamiento de memoria y el servicio se interrumpe.	90%

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
	Estadísticas de uso de memoria heap de NodeManager	18018	El uso de memoria heap de NodeManager supera el umbral	Cuando el uso de memoria heap de Yarn NodeManager es demasiado alto, el rendimiento del envío y la operación de la tarea de Yarn se ve afectado. Además, puede producirse un desbordamiento de memoria para que el servicio Yarn no esté disponible.	95%
	Estadísticas de uso de memoria no heap de NodeManager	18017	El uso de memoria no heap de NodeManager supera el umbral	Cuando el uso de memoria heap de Yarn NodeManager es demasiado alto, el rendimiento del envío y la operación de la tarea de Yarn se ve afectado. Además, puede producirse un desbordamiento de memoria para que el servicio Yarn no esté disponible.	90%
	Estadísticas de uso de memoria directa de ResourceManager	18013	El uso de memoria directa de ResourceManager supera el umbral	Si la memoria directa disponible de ResourceManager es insuficiente, se produce un desbordamiento de memoria y el servicio se interrumpe.	90%

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
	Estadísticas de uso de memoria heap de ResourceManager	18008	El uso de memoria heap de ResourceManager supera el umbral	Cuando el uso de memoria heap de Yarn ResourceManager es demasiado alto, el rendimiento del envío y la operación de la tarea de Yarn se ve afectado. Además, puede producirse un desbordamiento de memoria para que el servicio Yarn no esté disponible.	95%
	Estadísticas de uso de memoria no heap de ResourceManager	18016	El uso de memoria no heap de ResourceManager supera el umbral	Cuando el uso de memoria no heap del Yarn ResourceManager es demasiado alto, el rendimiento del envío y operación de la tarea de Yarn se ve afectado. Además, puede producirse un desbordamiento de memoria para que el servicio Yarn no esté disponible.	90%
	Estadísticas de duración de GC de NodeManager	18011	La duración de GC de NodeManager supera el umbral	Una larga duración GC del proceso NodeManager puede interrumpir los servicios.	12000ms
	Estadísticas de duración de GC de ResourceManager	18010	La duración de GC de ResourceManager supera el umbral	Una larga duración de GC del proceso de ResourceManager puede interrumpir los servicios.	12000ms

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
	Número de tareas fallidas en la cola raíz	18026	El número de tareas de Yarn fallidas supera el umbral	No se puede ejecutar un gran número de tareas de aplicación. Las tareas fallidas deben enviarse de nuevo.	50
	Aplicaciones terminadas de la cola raíz	18025	El número de tareas de Yarn terminadas supera el umbral	Un gran número de tareas de aplicación se detienen forzosamente.	50
	Memoria pendiente	18024	El uso de memoria de Yarn pendiente supera el umbral	Se necesita mucho tiempo para finalizar una solicitud. Una nueva aplicación no se puede ejecutar después del envío.	83886080MB
	Tareas pendientes	18023	El número de tareas pendientes de Yarn supera el umbral	Se necesita mucho tiempo para finalizar una solicitud. Una nueva aplicación no puede ejecutarse durante mucho tiempo después del envío.	60
ZooKeeper	Uso de conexiones de ZooKeeper	13001	Las conexiones de ZooKeeper disponibles son insuficientes	Las conexiones de ZooKeeper disponibles son insuficientes. Cuando el uso de la conexión alcanza el 100%, las conexiones externas no se pueden manejar.	80%

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
	Uso de memoria heap de ZooKeeper	13004	El uso de memoria heap de ZooKeeper supera el umbral	Si la memoria ZooKeeper disponible es insuficiente, se produce un desbordamiento de memoria y el servicio se interrumpe.	95%
	Uso de memoria directa de ZooKeeper	13002	El uso de memoria directa de ZooKeeper supera el umbral	Si la memoria ZooKeeper disponible es insuficiente, se produce un desbordamiento de memoria y el servicio se interrumpe.	80%
	Duración de GC de ZooKeeper por minuto	13003	La duración GC del proceso ZooKeeper supera el umbral	Una larga duración GC del proceso ZooKeeper puede interrumpir los servicios.	12000ms
Ranger	Duración de GC de UserSync	45284	La duración de GC de UserSync supera el umbral	UserSync responde lentamente.	12000ms
	Duración de GC de PolicySync	45292	La duración de GC de PolicySync supera el umbral	PolicySync responde lentamente.	12000ms
	Duración de GC de RangerAdmin	45280	La duración de GC de RangerAdmin supera el umbral	RangerAdmin responde lentamente.	12000ms
	Duración de GC de TagSync	45288	La duración de GC de TagSync supera el umbral	TagSync responde lentamente.	12000ms

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
	Uso de memoria no heap de UserSync	45283	El uso de memoria no heap de UserSync supera el umbral	El desbordamiento de la memoria no heap puede provocar una falla en el servicio.	80.0%
	Uso de memoria directa de UserSync	45282	El uso de memoria directa de UserSync supera el umbral	El desbordamiento de la memoria directa puede provocar una falla en el servicio.	80.0%
	Uso de memoria heap de UserSync	45281	El uso de memoria heap de UserSync supera el umbral	El desbordamiento de la memoria heap puede causar una falla en el servicio.	95.0%
	Uso de memoria directa de PolicySync	45290	El uso de memoria directa de PolicySync supera el umbral	El desbordamiento de la memoria directa puede provocar una falla en el servicio.	80.0%
	Uso de memoria heap de PolicySync	45289	El uso de memoria heap de PolicySync supera el umbral	El desbordamiento de la memoria heap puede causar una falla en el servicio.	95.0%
	Uso de memoria no heap de PolicySync	45291	El uso de memoria no heap de PolicySync supera el umbral	El desbordamiento de la memoria no heap puede provocar una falla en el servicio.	80.0%
	Uso de memoria no heap de RangerAdmin	45279	El uso de memoria no heap de RangerAdmin supera el umbral	El desbordamiento de la memoria no heap puede provocar una falla en el servicio.	80.0%

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
	Uso de memoria heap de RangerAdmin	45277	El uso de memoria heap de RangerAdmin supera el umbral	El desbordamiento de la memoria heap puede causar una falla en el servicio.	95.0%
	Uso de memoria directa de RangerAdmin	45278	El uso de memoria directa de RangerAdmin supera el umbral	El desbordamiento de la memoria directa puede provocar una falla en el servicio.	80.0%
	Uso de memoria directa de TagSync	45286	El uso de memoria directa de TagSync supera el umbral	El desbordamiento de la memoria directa puede provocar una falla en el servicio.	80.0%
	Uso de memoria no heap de TagSync	45287	El uso de memoria no heap de TagSync supera el umbral	El desbordamiento de la memoria no heap puede provocar una falla en el servicio.	80.0%
	Uso de memoria heap de TagSync	45285	El uso de memoria heap de TagSync supera el umbral	El desbordamiento de la memoria heap puede causar una falla en el servicio.	95.0%
ClickHouse	Uso de la cuota de cantidad de servicio de Clickhouse en ZooKeeper	45426	El uso de la cuota de cantidad de servicio ClickHouse en ZooKeeper supera el umbral	Una vez que la cuota de cantidad de ZooKeeper del servicio ClickHouse supera el umbral, no puede realizar operaciones de clúster en el servicio ClickHouse en FusionInsight Manager. Como resultado, no se puede utilizar el servicio ClickHouse.	90%

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
	Uso de cuota de capacidad de servicio ClickHouse en ZooKeeper	45427	El uso de la cuota de capacidad de servicio ClickHouse en ZooKeeper supera el umbral	Una vez que la cuota de capacidad de ZooKeeper del servicio ClickHouse supera el umbral, no puede realizar operaciones de clúster en el servicio ClickHouse en FusionInsight Manager. Como resultado, no se puede utilizar el servicio ClickHouse.	90%
IoTDB	Latencia máxima de fusión (fusión intraespacio)	45594	La duración de la fusión intraespacio de IoTDBServer supera el umbral	La escritura de datos se bloquea y el rendimiento de la operación de escritura se ve afectado.	300000 ms
	Latencia máxima de fusión (Flush)	45593	La duración de ejecución de IoTDBServer Flush supera el umbral	La escritura de datos se bloquea y el rendimiento de la operación de escritura se ve afectado.	300000 ms
	Latencia máxima de fusión (fusión de espacio cruzado)	45595	La duración de la fusión entre espacios de IoTDBServer supera el umbral	La escritura de datos se bloquea y el rendimiento de la operación de escritura se ve afectado.	300000 ms
	Latencia máxima de RPC (executeStatement)	45592	La duración de ejecución de IoTDBServer RPC supera el umbral	El rendimiento de ejecución del proceso IoTDBServer se ve afectado.	10000s
	Duración total del GC de IoTDBServer	45587	La duración de GC de IoTDBServer supera el umbral	Una larga duración de GC del proceso IoTDBServer puede interrumpir los servicios.	12000ms

Servicio	Métrica	ID	Nombre de la alarma	Impacto en el sistema	Umbral predeterminado
	Duración total del GC de ConfigNode	45590	La duración de GC de ConfigNode supera el umbral	Una larga duración de GC del proceso ConfigNode puede interrumpir los servicios.	12000ms
	Uso de la memoria heap de IoTDBServer	45586	El uso de la memoria heap de IoTDBServer supera el umbral	Si la memoria heap de procesos IoTDBServer disponible es insuficiente, se produce un desbordamiento de memoria y el servicio se interrumpe.	90%
	Uso de la memoria directa de IoTDBServer	45588	El uso de la memoria directa de IoTDBServer supera el umbral	El desbordamiento de la memoria directa puede provocar una falla en el servicio.	90%
	Uso de memoria heap de ConfigNode	45589	El uso de memoria heap de ConfigNode supera el umbral	Si la memoria heap de proceso ConfigNode disponible es insuficiente, se produce un desbordamiento de memoria y el servicio se interrumpe.	90%
	Uso de memoria directa de ConfigNode	45591	El uso de memoria directa de ConfigNode supera el umbral	El desbordamiento de la memoria directa puede hacer que la instancia de IoTDB no esté disponible.	90%

12 Desarrollo de aplicaciones de componentes MRS

12.1 Desarrollo de aplicaciones de HBase

HBase es un sistema de almacenamiento distribuido basado en columnas que ofrece alta confiabilidad, rendimiento y escalabilidad. Está diseñado para eliminar las limitaciones de las bases de datos relacionales en el procesamiento de cantidades masivas de datos.

Los escenarios de aplicación de HBase tienen las siguientes características:

- Procesamiento masivo de datos (más alto que el nivel TB o PB)
- Alto rendimiento
- Lectura aleatoria altamente eficiente de datos masivos
- Excelente escalabilidad
- Procesamiento simultáneo de datos estructurados y no estructurados

MRS proporciona ejemplos de proyectos de desarrollo de aplicaciones basados en HBase. Esta práctica proporciona orientación para que obtenga e importe un proyecto de muestra después de crear un clúster MRS y, a continuación, realice la construcción y puesta en marcha localmente. En este proyecto de ejemplo, puede crear tablas HBase, insertar datos, crear índices y eliminar tablas en el clúster MRS.

Creación de un clúster MRS HBase

1. Cree y compre un clúster MRS que contenga HBase. Para obtener más información, consulte **Compra de un clúster personalizado**.

   NOTA

   En esta práctica, se utiliza como ejemplo un clúster MRS 3.1.0, con Hadoop y HBase instalados y con la autenticación Kerberos habilitada.

2. Haga clic en **Buy Now** y espere hasta que se cree el clúster MRS.

Figura 12-1 Clúster adquirido

Preparación del archivo de configuración de desarrollo de aplicaciones

Paso 1 Una vez creado el clúster, inicie sesión en FusionInsight Manager y cree un usuario del clúster para la autenticación de seguridad del proyecto de ejemplo.

1. Elija **System** > **Permission** > **User**. En el panel derecho, haga clic en **Create**. En la página mostrada, cree un usuario hombre-máquina, por ejemplo, **developuser**.

   Añada el grupo de usuarios **hadoop** a **User Group**.

   Una vez creado el usuario, inicie sesión en FusionInsight Manager como **developuser** y cambie la contraseña inicial según se le solicite.
2. Inicie sesión en la interfaz de usuario web de Ranger como administrador de Ranger **rangeradmin**.

   La contraseña predeterminada del **rangeradmin** de usuario es **Rangeradmin@123**. Para obtener más información, consulte **Lista de cuenta de usuario**.
3. En la página de inicio del Ranger, haga clic en el nombre del complemento del componente en el área **HBASE**, por ejemplo, **HBase**.
4. Haga clic en en la columna **Action** de la fila que contiene la política **all - table, column-family, column**.
5. En el área **Allow Conditions**, agregue una condición de permiso. Seleccione el usuario creado para **Select User** y seleccione **Select/Deselect All** para **Permissions**.
6. Haga clic en **Save**.

Paso 2 Inicie sesión en FusionInsight Manager como usuario **admin** y elija **System** > **Permission** > **User**. En la columna **Operation** de **developuser**, elija **More** > **Download Authentication Credential**. Guarde el archivo y descomprímalo para obtener los archivos **user.keytab** y **krb5.conf** del usuario.

Paso 3 Elija **Cluster**. En la pestaña **Dashboard**, haga clic en **More** y seleccione **Download Client**. En el cuadro de diálogo que se muestra, establezca **Select Client Type** en **Configuration Files Only** y haga clic en **OK**. Después de generar el paquete cliente, descargue el paquete como se le indique y descomprima.

Por ejemplo, si el paquete del archivo de configuración del cliente es **FusionInsight_Cluster_1_Services_Client.tar**, descomprima para obtener **FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles.tar**. A continuación, continúe para descomprimir este archivo.

1. Vaya al directorio **FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles\HBase\config** y obtenga los archivos de configuración que aparecen en la lista de **Tabla 12-1**.

Tabla 12-1 Archivos de configuración

Archivo de configuración	Descripción
core-site.xml	Configura los parámetros de Hadoop Core.
hbase-site.xml	Configura parámetros de HBase.
hdfs-site.xml	Configura parámetros de HDFS.

2. Copie todo el contenido del archivo **hosts** en el directorio de descompresión al archivo **hosts** local. Asegúrese de que el PC local pueda comunicarse con los hosts que figuran en el archivo **hosts** del directorio de descompresión.

   NOTA

   - En esta práctica, asegúrese de que el entorno local puede comunicarse con el plano de red donde reside el clúster MRS. En general, puede acceder al clúster MRS a través de una EIP.
   - Si el entorno local no puede comunicarse con los nodos del clúster MRS, puede crear primero el proyecto de ejemplo y cargar el paquete JAR en el clúster para ejecutarlo. .
   - **C:\WINDOWS\system32\drivers\etc\hosts** es un directorio de ejemplo en un entorno Windows para almacenar el archivo **hosts** local.

----Fin

Obtención del proyecto de muestra

Paso 1 Obtenga el proyecto de muestra de Huawei Mirrors.

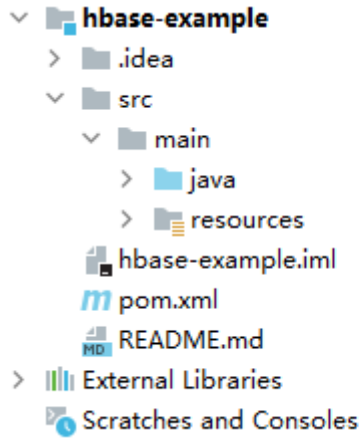
Descargue el código fuente del proyecto Maven y los archivos de configuración del proyecto de ejemplo, y configure las herramientas de desarrollo relacionadas en su PC local. Para obtener más información, consulte **Obtención de proyectos de muestra desde Huawei Mirros**.

Seleccione una rama basada en la versión del clúster y descargue el proyecto de muestra de MRS requerido.

Por ejemplo, el proyecto de muestra adecuado para esta práctica es **hbase-example**, que se puede obtener en **https://github.com/huaweicloud/huaweicloud-mrs-example/tree/mrs-3.1.0/src/hbase-examples/hbase-example**.

Paso 2 Utilice IDEA para importar el proyecto de ejemplo y espere a que el proyecto Maven descargue los paquetes de dependencias. Para obtener más información, consulte **Configuración e importación de proyectos de muestra**.

Figura 12-2 Proyecto de muestra HBase



Después de configurar los parámetros Maven y SDK en el PC local, el proyecto de ejemplo carga automáticamente paquetes de dependencias relacionados.

Paso 3 Coloque los archivos de configuración del clúster y las credenciales de autenticación de usuario obtenidas en **Preparación del archivo de configuración de desarrollo de aplicaciones** al directorio **../src/main/resources/conf** del proyecto de ejemplo.

Paso 4 En la clase **TestMain** del paquete **com.huawei.bigdata.hbase.examples**, cambie **userName** por el nombre de usuario real, por ejemplo, **developuser**.

```
private static void login() throws IOException {
        if (User.isHBaseSecurityEnabled(conf)) {
            userName = "developuser";
            //In Windows environment
            String userdir =
TestMain.class.getClassLoader().getResource("conf").getPath() + File.separator;
            //In Linux environment
            //String userdir = System.getProperty("user.dir") + File.separator +
"conf" + File.separator;

            LoginUtil.setJaasConf(ZOOKEEPER_DEFAULT_LOGIN_CONTEXT_NAME, userName,
userKeytabFile);
            LoginUtil.login(userName, userKeytabFile, krb5File, conf);
        }
}
```

Supongamos que está desarrollando una aplicación para gestionar información sobre los usuarios del servicio A en una empresa. El proceso de operación es como sigue.

No.	Paso
1	Crear una tabla basada en la información existente.
2	Importar datos de usuario.
3	Agregar la familia de columnas **Education Information** y agregar los fondos educativos y los títulos de los usuarios a la tabla de información del usuario.
4	Consultar nombres de usuario y direcciones por ID de usuario.
5	Ejecutar consultas por nombre de usuario.

No.	Paso
6	Para mejorar el rendimiento de las consultas, cree o elimine índices secundarios.
7	Anular el registro de usuarios y eliminar los datos de usuario de la tabla de información de usuario.
8	Eliminar la tabla de información del usuario después de que finalice el servicio A.

Por ejemplo, el siguiente fragmento de código ejecuta el método **testCreateTable** en la clase **HBaseSample** del paquete **com.huawei.bigdata.hbase.examples** para crear una tabla de información de usuario.

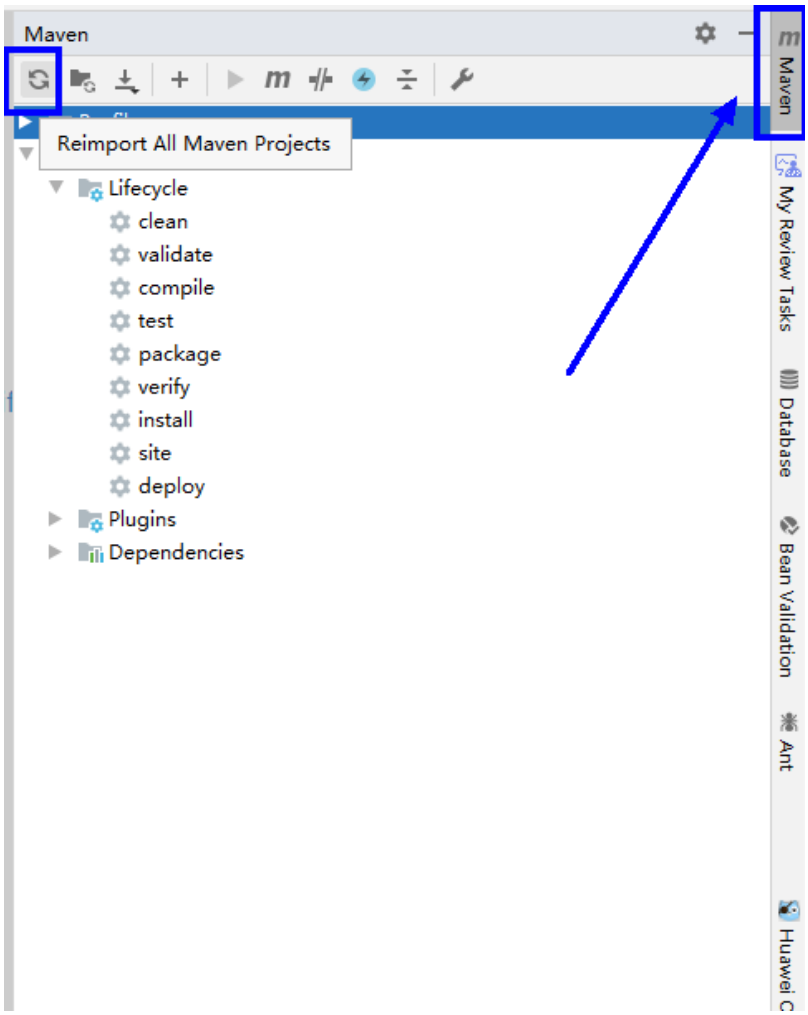
```
public void testCreateTable() {
             LOG.info("Entering testCreateTable.");
             TableDescriptorBuilder htd =
TableDescriptorBuilder.newBuilder(tableName); //Create a table descriptor.
             ColumnFamilyDescriptorBuilder hcd =
ColumnFamilyDescriptorBuilder.newBuilder(Bytes.toBytes("info")); //Create a
column family descriptor.
             hcd.setDataBlockEncoding(DataBlockEncoding.FAST_DIFF); //Set the
encoding algorithm. HBase provides DIFF, FAST_DIFF, and PREFIX encoding
algorithms.
             hcd.setCompressionType(Compression.Algorithm.SNAPPY);
             htd.setColumnFamily(hcd.build());  //Add the column family
descriptor to the table descriptor.
             Admin admin = null;
             try {
               admin = conn.getAdmin(); //Obtain the Admin object, which allows
you to create a table, create a column family, check whether the table exists,
change the table structure and column family structure, and delete the table.
               if (!admin.tableExists(tableName)) {
                 LOG.info("Creating table...");
                 admin.createTable(htd.build());//Call the createTable method of
Admin.
                 LOG.info(admin.getClusterMetrics().toString());
                 LOG.info(admin.listNamespaceDescriptors().toString());
                 LOG.info("Table created successfully.");
               } else {
                 LOG.warn("table already exists");
               }
             } catch (IOException e) {
                 LOG.error("Create table failed " ,e);
             } finally {
               if (admin != null) {
                 try {
                   admin.close();
                 } catch (IOException e) {
                   LOG.error("Failed to close admin " ,e);
                 }
               }
             }
             LOG.info("Exiting testCreateTable.");
  }
```

----Fin

Creación y ejecución de la aplicación

Paso 1 Haga clic en **Reimport All Maven Projects** en la ventana Maven a la derecha de IDEA para cargar las dependencias del proyecto Maven.

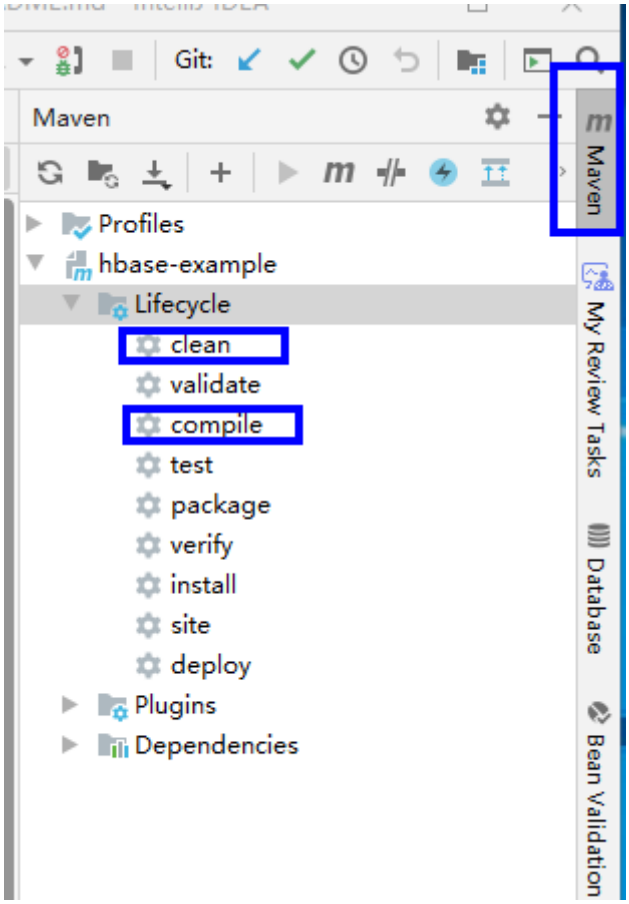
Figura 12-3 Cargar un proyecto de muestra



Paso 2 Construir la aplicación.

1. Elija **Maven**, busque el nombre del proyecto de destino y haga doble clic en **clean** en **Lifecycle** para ejecutar el comando **clean** de Maven.
2. Elija **Maven**, busque el nombre del proyecto de destino y haga doble clic en **compile** en **Lifecycle** para ejecutar el comando **compile** de Maven.

Figura 12-4 clean y compile de Maven



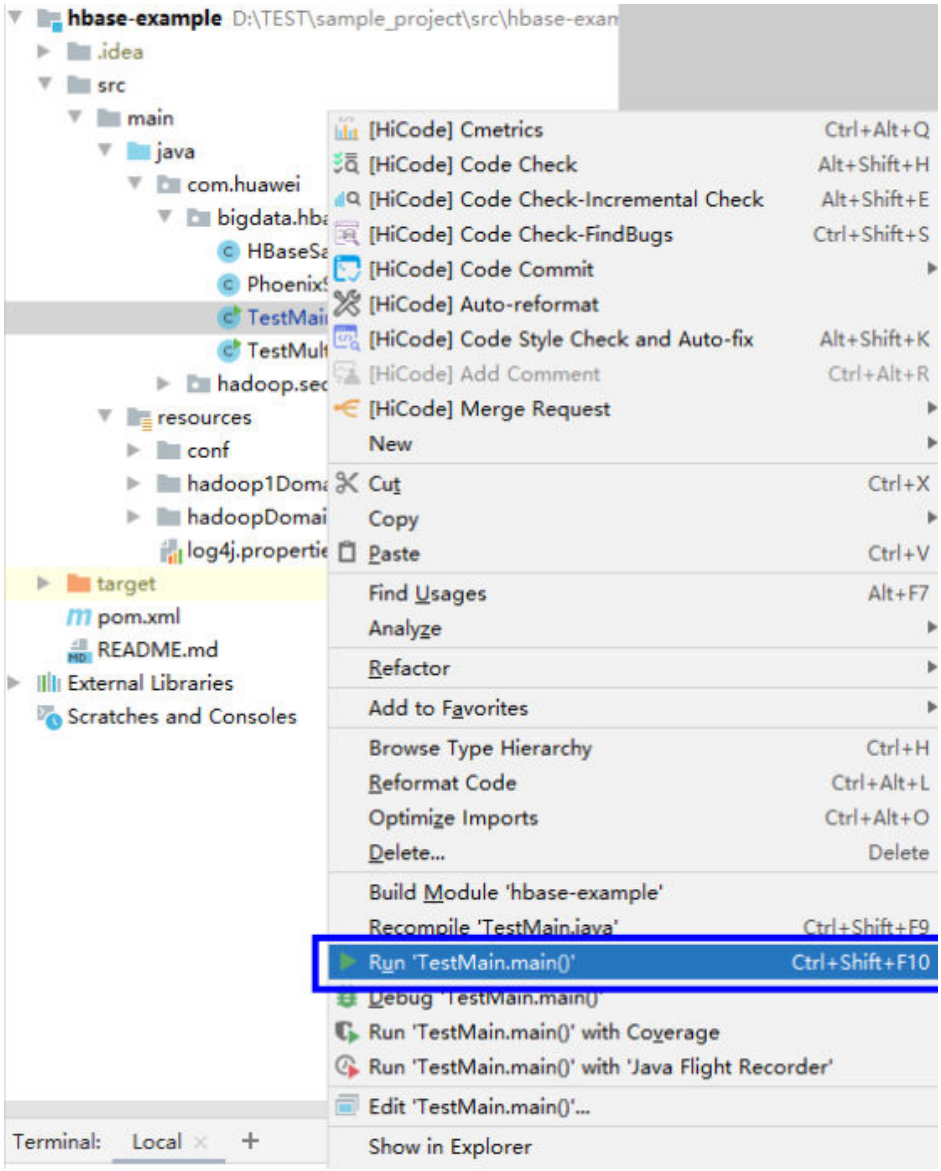
Una vez completada la construcción, se muestra el mensaje "Build Success" y se genera el directorio **target**.

```
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time:  21.276 s
[INFO] Finished at: 2023-05-05T14:36:39+08:00
[INFO] ------------------------------------------------------------------------
```

Paso 3 Ejecutar la aplicación.

Haga clic con el botón derecho del ratón en el archivo **TestMain.java** y elija **Run 'TestMain.main()**.

Figura 12-5 Ejecución de la aplicación



Paso 4 Compruebe la información de salida después de ejecutar la muestra **hbase-example**. La siguiente información indica que las operaciones de tabla relacionadas se ejecutan correctamente:

```
...
2023-05-05 15:05:27,050 INFO  [main] examples.HBaseSample: Table created
successfully.
2023-05-05 15:05:27,050 INFO  [main] examples.HBaseSample: Exiting
testCreateTable.
2023-05-05 15:05:27,050 INFO  [main] examples.HBaseSample: Entering
testMultiSplit.
2023-05-05 15:05:31,171 INFO  [main] client.HBaseAdmin: Operation:
MULTI_SPLIT_REGION, Table Name: default:hbase_sample_table, procId: 21 completed
2023-05-05 15:05:31,171 INFO  [main] examples.HBaseSample: MultiSplit
successfully.
2023-05-05 15:05:31,172 INFO  [main] examples.HBaseSample: Exiting testMultiSplit.
2023-05-05 15:05:31,172 INFO  [main] examples.HBaseSample: Entering testPut.
2023-05-05 15:05:32,862 INFO  [main] examples.HBaseSample: Put successfully.
2023-05-05 15:05:32,862 INFO  [main] examples.HBaseSample: Exiting testPut.
2023-05-05 15:05:32,862 INFO  [main] examples.HBaseSample: Entering createIndex.
2023-05-05 15:05:36,627 INFO  [main] examples.HBaseSample: Create index
```



```
successfully.
2023-05-05 15:05:36,627 INFO  [main] examples.HBaseSample: Exiting createIndex.
2023-05-05 15:05:36,627 INFO  [main] examples.HBaseSample: Entering createIndex.
2023-05-05 15:05:37,912 INFO  [main] examples.HBaseSample: Successfully enable
indices [index_name]  of the table hbase_sample_table
2023-05-05 15:05:37,912 INFO  [main] examples.HBaseSample: Entering
testScanDataByIndex.
2023-05-05 15:05:37,915 INFO  [main] examples.HBaseSample: Scan indexed data.
2023-05-05 15:05:39,939 INFO  [main] examples.HBaseSample: Scan data by index
successfully.
2023-05-05 15:05:39,939 INFO  [main] examples.HBaseSample: Exiting
testScanDataByIndex.
2023-05-05 15:05:39,941 INFO  [main] examples.HBaseSample: Entering
testModifyTable.
2023-05-05 15:05:40,191 INFO  [main] client.HBaseAdmin: Started disable of
hbase_sample_table
2023-05-05 15:05:41,322 INFO  [main] client.HBaseAdmin: Operation: DISABLE, Table
Name: default:hbase_sample_table, procId: 53 completed
2023-05-05 15:05:42,230 INFO  [main] client.HBaseAdmin: Started enable of
hbase_sample_table
2023-05-05 15:05:43,187 INFO  [main] client.HBaseAdmin: Operation: ENABLE, Table
Name: default:hbase_sample_table, procId: 65 completed
2023-05-05 15:05:43,187 INFO  [main] examples.HBaseSample: Modify table
successfully.
2023-05-05 15:05:43,187 INFO  [main] examples.HBaseSample: Exiting
testModifyTable.
2023-05-05 15:05:43,187 INFO  [main] examples.HBaseSample: Entering testGet.
2023-05-05 15:05:43,278 INFO  [main] examples.HBaseSample:
012005000201:info,address,Shenzhen, Guangdong
2023-05-05 15:05:43,279 INFO  [main] examples.HBaseSample:
012005000201:info,name,Zhang San
2023-05-05 15:05:43,279 INFO  [main] examples.HBaseSample: Get data successfully.
2023-05-05 15:05:43,279 INFO  [main] examples.HBaseSample: Exiting testGet.
2023-05-05 15:05:43,279 INFO  [main] examples.HBaseSample: Entering testScanData.
2023-05-05 15:05:43,576 INFO  [main] examples.HBaseSample:
012005000201:info,name,Zhang San
2023-05-05 15:05:43,576 INFO  [main] examples.HBaseSample:
012005000202:info,name,Li Wanting
2023-05-05 15:05:43,577 INFO  [main] examples.HBaseSample:
012005000203:info,name,Wang Ming
2023-05-05 15:05:43,577 INFO  [main] examples.HBaseSample:
012005000204:info,name,Li Gang
2023-05-05 15:05:43,578 INFO  [main] examples.HBaseSample:
012005000205:info,name,Zhao Enru
2023-05-05 15:05:43,578 INFO  [main] examples.HBaseSample:
012005000206:info,name,Chen Long
2023-05-05 15:05:43,578 INFO  [main] examples.HBaseSample:
012005000207:info,name,Zhou Wei
2023-05-05 15:05:43,578 INFO  [main] examples.HBaseSample:
012005000208:info,name,Yang Yiwen
2023-05-05 15:05:43,578 INFO  [main] examples.HBaseSample:
012005000209:info,name,Xu Bing
2023-05-05 15:05:43,578 INFO  [main] examples.HBaseSample:
012005000210:info,name,Xiao Kai
2023-05-05 15:05:43,578 INFO  [main] examples.HBaseSample: Scan data successfully.
2023-05-05 15:05:43,578 INFO  [main] examples.HBaseSample: Exiting testScanData.
2023-05-05 15:05:43,578 INFO  [main] examples.HBaseSample: Entering
testSingleColumnValueFilter.
2023-05-05 15:05:43,883 INFO  [main] examples.HBaseSample: Single column value
filter successfully.
2023-05-05 15:05:43,883 INFO  [main] examples.HBaseSample: Exiting
testSingleColumnValueFilter.
2023-05-05 15:05:43,884 INFO  [main] examples.HBaseSample: Entering
testFilterList.
2023-05-05 15:05:44,388 INFO  [main] examples.HBaseSample:
012005000201:info,name,Zhang San
2023-05-05 15:05:44,388 INFO  [main] examples.HBaseSample:
012005000202:info,name,Li Wanting
2023-05-05 15:05:44,388 INFO  [main] examples.HBaseSample:
```



```
012005000203:info,name,Wang Ming
2023-05-05 15:05:44,388 INFO  [main] examples.HBaseSample:
012005000204:info,name,Li Gang
2023-05-05 15:05:44,389 INFO  [main] examples.HBaseSample:
012005000205:info,name,Zhao Enru
2023-05-05 15:05:44,389 INFO  [main] examples.HBaseSample:
012005000206:info,name,Chen Long
2023-05-05 15:05:44,389 INFO  [main] examples.HBaseSample:
012005000207:info,name,Zhou Wei
2023-05-05 15:05:44,389 INFO  [main] examples.HBaseSample:
012005000208:info,name,Yang Yiwen
2023-05-05 15:05:44,389 INFO  [main] examples.HBaseSample:
012005000209:info,name,Xu Bing
2023-05-05 15:05:44,389 INFO  [main] examples.HBaseSample:
012005000210:info,name,Xiao Kai
2023-05-05 15:05:44,389 INFO  [main] examples.HBaseSample: Filter list
successfully.
2023-05-05 15:05:44,389 INFO  [main] examples.HBaseSample: Exiting testFilterList.
2023-05-05 15:05:44,389 INFO  [main] examples.HBaseSample: Entering testDelete.
2023-05-05 15:05:44,586 INFO  [main] examples.HBaseSample: Delete table
successfully.
2023-05-05 15:05:44,586 INFO  [main] examples.HBaseSample: Exiting testDelete.
2023-05-05 15:05:44,586 INFO  [main] examples.HBaseSample: Entering disableIndex.
2023-05-05 15:05:45,819 INFO  [main] examples.HBaseSample: Successfully disable
indices [index_name]  of the table hbase_sample_table
2023-05-05 15:05:45,819 INFO  [main] examples.HBaseSample: Entering dropIndex.
2023-05-05 15:05:48,084 INFO  [main] examples.HBaseSample: Drop index
successfully.
2023-05-05 15:05:48,084 INFO  [main] examples.HBaseSample: Exiting dropIndex.
2023-05-05 15:05:48,084 INFO  [main] examples.HBaseSample: Entering dropTable.
2023-05-05 15:05:48,237 INFO  [main] client.HBaseAdmin: Started disable of
hbase_sample_table
2023-05-05 15:05:49,543 INFO  [main] client.HBaseAdmin: Operation: DISABLE, Table
Name: default:hbase_sample_table, procId: 95 completed
2023-05-05 15:05:50,645 INFO  [main] client.HBaseAdmin: Operation: DELETE, Table
Name: default:hbase_sample_table, procId: 106 completed
2023-05-05 15:05:50,645 INFO  [main] examples.HBaseSample: Drop table
successfully.
2023-05-05 15:05:50,645 INFO  [main] examples.HBaseSample: Exiting dropTable.
2023-05-05 15:05:50,646 INFO  [main] client.ConnectionImplementation: Closing
master protocol: MasterService
2023-05-05 15:05:50,652 INFO  [main] client.ConnectionImplementation: Connection
has been closed by main.
2023-05-05 15:05:50,654 INFO  [main] hbase.ChoreService: Chore service for:
AsyncConn Chore Service had [[ScheduledChore: Name: RefreshCredentials Period:
30000 Unit: MILLISECONDS]] on shutdown
2023-05-05 15:05:50,655 INFO  [main] examples.TestMain: -----------finish HBase
-------------------
...
```

----Fin

12.2 Desarrollo de aplicaciones de HDFS

Hadoop Distribute File System (HDFS) es un sistema de archivos distribuido que se ejecuta en hardware universal. Cuenta con una alta tolerancia a fallos y admite acceso a datos de alto rendimiento. Es adecuado para procesar conjuntos de datos a gran escala.

HDFS es adecuado para los siguientes escenarios de aplicación:

- Procesamiento de cantidades masivas de datos (TB o PB y mayores)
- Escenarios que requieren un alto rendimiento
- Escenarios que requieren alta confiabilidad
- Escenarios que requieren una excelente escalabilidad

MRS proporciona ejemplos de proyectos de desarrollo de aplicaciones basados en HBase. Esta práctica proporciona orientación para obtener e importar un proyecto de ejemplo después de crear un clúster MRS y, a continuación, compilar y depurar el código localmente. En este proyecto de ejemplo, puede crear directorios HDFS y escribir, leer y eliminar archivos.

Creación de un clúster MRS Hadoop

1. Cree y compre un clúster MRS que contenga Hadoop. Para obtener más información, consulte **Compra de un clúster personalizado**.

   NOTA

   En esta práctica, se utiliza como ejemplo un clúster MRS 3.2.0-LTS.1, con Hadoop instalado y con la autenticación Kerberos habilitada.

2. Haga clic en **Buy Now** y espere hasta que se cree el clúster MRS.

Preparación del archivo de configuración de desarrollo de aplicaciones

Paso 1 Inicie sesión en FusionInsight Manager y cree un usuario del clúster para la autenticación de seguridad del proyecto de ejemplo.

Elija **System** > **Permission** > **User**. En la página mostrada, haga clic en **Create**. En la página mostrada, cree un usuario máquina-máquina, por ejemplo, **developuser**.

Añada el grupo de usuarios **hadoop** a **User Group**.

Paso 2 Elija **System** > **Permission** > **User**. En la columna **Operation** de **developuser**, elija **More** > **Download Authentication Credential**. Guarde el archivo y descomprímalo para obtener los archivos **user.keytab** y **krb5.conf** del usuario.

Paso 3 Elija **Cluster**. En la pestaña **Dashboard**, haga clic en **More** y seleccione **Download Client**. En el cuadro de diálogo que se muestra, establezca **Select Client Type** en **Configuration Files Only** y haga clic en **OK**. Después de generar el paquete cliente, descargue el paquete como se le indique y descomprima.

Por ejemplo, si el paquete del archivo de configuración del cliente es **FusionInsight_Cluster_1_Services_Client.tar**, descomprima para obtener **FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles.tar**. A continuación, continúe para descomprimir este archivo.

1. Vaya al directorio **FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles\HDFS\config** y obtenga los archivos de configuración que aparecen en la lista de **Tabla 12-2**.

   Tabla 12-2 Archivo

Archivo	Descripción
core-site.xml	Parámetros de Hadoop Core
hdfs-site.xml	Parámetros de HDFS

2. Copie todo el contenido del archivo **hosts** en el directorio de descompresión al archivo **hosts** local. Asegúrese de que el PC local pueda comunicarse con los hosts que figuran en el archivo **hosts** del directorio de descompresión.

NOTA

- En esta práctica, asegúrese de que el entorno local puede comunicarse con el plano de red donde se despliega el clúster MRS. En general, puede acceder al clúster MRS a través de una EIP.
- Si el entorno local no puede comunicarse con los nodos del clúster MRS, puede crear primero el proyecto de ejemplo y cargar el paquete JAR en el clúster para ejecutarlo.
- **C:\WINDOWS\system32\drivers\etc\hosts** es un directorio de ejemplo en un entorno Windows para almacenar el archivo **hosts** local.

----Fin

Obtención del proyecto de muestra

Paso 1 Obtenga el proyecto de muestra de Huawei Mirrors.

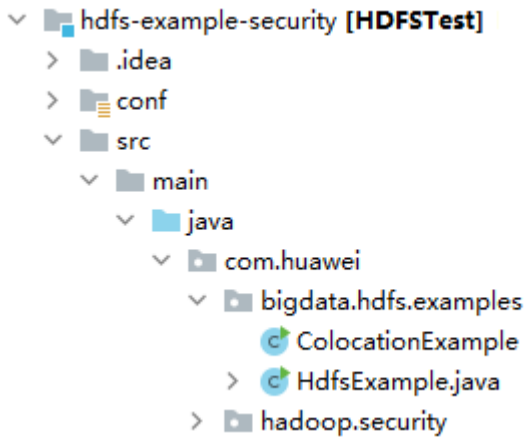
Descargue el código fuente y los archivos de configuración del proyecto de ejemplo, y configure las herramientas de desarrollo relacionadas en su PC local. Para obtener más información, consulte **Obtención de proyectos de muestra desde Huawei Mirros**.

Seleccione una rama basada en la versión del clúster y descargue el proyecto de muestra de MRS requerido.

Por ejemplo, el proyecto de muestra adecuado para esta práctica es **hdfs-example-security**, que se puede obtener en **https://github.com/huaweicloud/huaweicloud-mrs-example/tree/mrs-3.2.0.1/src/hdfs-example-security**.

Paso 2 Utilice IDEA para importar el proyecto de ejemplo y espere a que el proyecto Maven descargue los paquetes de dependencias. Para obtener más información, consulte **Configuración e importación de proyectos de muestra**.

Figura 12-6 Proyecto de muestra de HDFS



Después de configurar los parámetros Maven y SDK en el PC local, el proyecto de ejemplo carga automáticamente paquetes de dependencias relacionados.

Paso 3 Coloque los archivos de configuración del clúster y las credenciales de autenticación de usuario obtenidas en **Preparación del archivo de configuración de desarrollo de aplicaciones** al directorio **conf** del proyecto de ejemplo.

Paso 4 Utilice el código de autenticación requerido para el proyecto de ejemplo de HDFS. Generalmente, hay autenticación de seguridad y autenticación de ZooKeeper.

En este ejemplo, no es necesario acceder a ZooKeeper o HBase. Solo se requiere el código de autenticación de seguridad básico.

En la clase **HdfsExample** del paquete **com.huawei.bigdata.hdfs.examples**, cambie **PRNCIPAL_NAME** por el nombre de usuario que está utilizando, por ejemplo, **developuser**.

```
private static final String PATH_TO_HDFS_SITE_XML =
System.getProperty("user.dir") + File.separator + "conf"
        + File.separator + "hdfs-site.xml";
private static final String PATH_TO_CORE_SITE_XML =
System.getProperty("user.dir") + File.separator + "conf"
        + File.separator + "core-site.xml";
private static final String PRNCIPAL_NAME = "developuser";
private static final String PATH_TO_KEYTAB = System.getProperty("user.dir") +
File.separator + "conf"
        + File.separator + "user.keytab";
private static final String PATH_TO_KRB5_CONF = System.getProperty("user.dir") +
File.separator + "conf"
        + File.separator + "krb5.conf";
...
```

En este proyecto de ejemplo, la hoja de ruta de desarrollo basada en los requisitos de servicio es la siguiente.

En el ejemplo siguiente se describe cómo leer, escribir y eliminar el archivo **/user/hdfs-examples/test.txt** en HDFS.

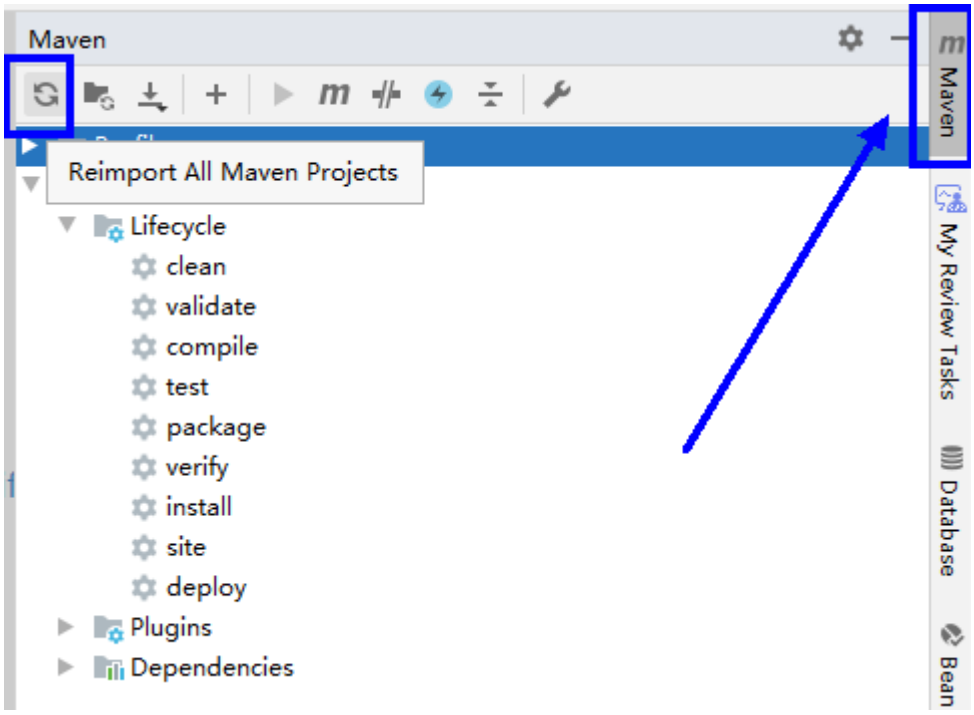
1. Pase la autenticación de seguridad del clúster.
2. Cree un objeto FileSystem: **fSystem**
3. Invoque a la API **mkdir** de **fSystem** para crear un directorio.
4. Invoque a **create** en **fSystem** para crear un objeto FSDataOutputStream **out**. Escriba datos en **out** invocando a **write**.
5. Invoque a **append** en **fSystem** para crear un objeto FSDataOutputStream **out**. Añada datos a **out** invocando a **write**.
6. Invoque a **open** en **fSystem** para crear un objeto FSDataInputStream **in**. Lea los archivos de **in** invocando a **read**.
7. Invoque **delete** en **fSystem** para eliminar un archivo.
8. Invoque **delete** en **fSystem** para eliminar una carpeta.

----Fin

Creación y ejecución de la aplicación

Paso 1 Haga clic en **Reimport All Maven Projects** en la ventana Maven a la derecha de IDEA para cargar las dependencias del proyecto Maven.

Figura 12-7 Cargar un proyecto de muestra



Paso 2 Compilar y ejecutar la aplicación.

1. Elija **Maven**, busque el nombre del proyecto de destino y haga doble clic en **clean** en **Lifecycle** para ejecutar el comando **clean** de Maven.
2. Elija **Maven**, busque el nombre del proyecto de destino y haga doble clic en **compile** en **Lifecycle** para ejecutar el comando **compile** de Maven.

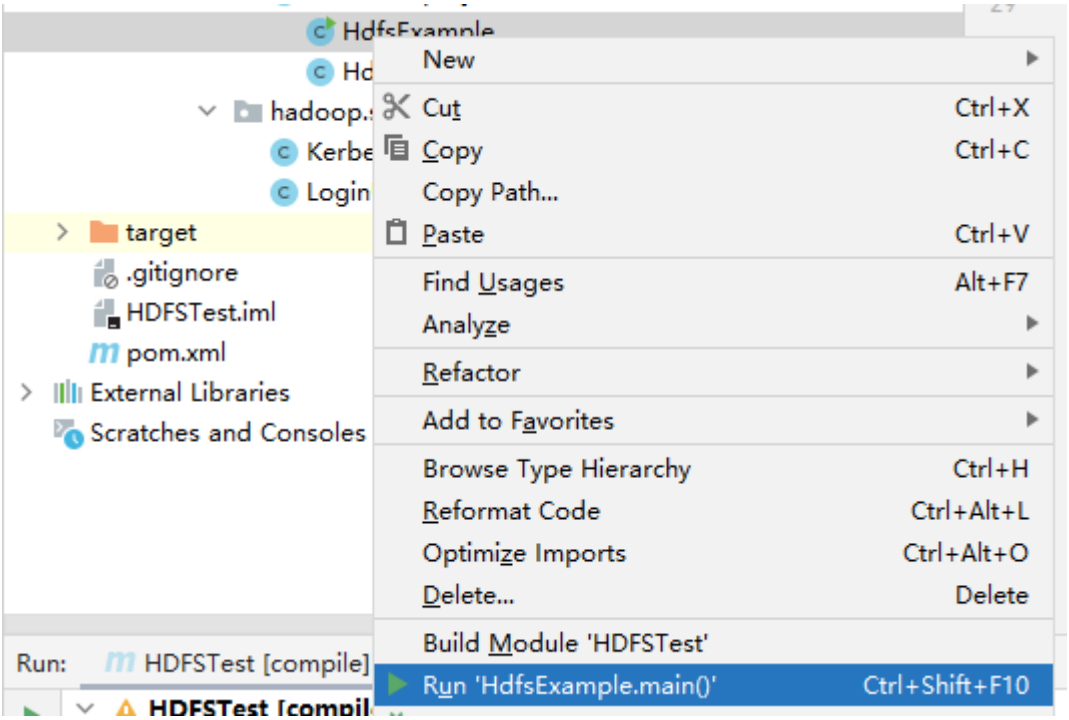
Una vez completada la construcción, se muestra el mensaje "Build Success" y se genera el directorio **target**.

```
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time:  21.276 s
[INFO] Finished at: 2023-05-05T14:36:39+08:00
[INFO] ------------------------------------------------------------------------
```

Paso 3 Ejecutar la aplicación.

Haga clic con el botón derecho en el archivo **HdfsExample.java** y elija **Run 'HdfsExample.main()'** en el menú contextual.

Figura 12-8 Ejecución de la aplicación



Paso 4 Compruebe la información de salida después de ejecutar la muestra. La siguiente información indica que las operaciones de archivo relacionadas se ejecutan correctamente:

```
...
2217 [main] INFO  org.apache.hadoop.security.UserGroupInformation  - Login
successful for user developuser using keytab file user.keytab. Keytab auto
renewal enabled : false
2217 [main] INFO  com.huawei.hadoop.security.LoginUtil  - Login
success!!!!!!!!!!!!!!
3529 [main] WARN  org.apache.hadoop.hdfs.shortcircuit.DomainSocketFactory  - The
short-circuit local reads feature cannot be used because UNIX Domain sockets are
not available on Windows.
4632 [main] INFO  com.huawei.bigdata.hdfs.examples.HdfsExample  - success to
create path /user/hdfs-examples
5392 [main] INFO  com.huawei.bigdata.hdfs.examples.HdfsExample  - success to
write.
8200 [main] INFO  com.huawei.bigdata.hdfs.examples.HdfsExample  - success to
append.
9384 [main] INFO  com.huawei.bigdata.hdfs.examples.HdfsExample  - result is : hi,
I am bigdata. It is successful if you can see me.I append this content.
9384 [main] INFO  com.huawei.bigdata.hdfs.examples.HdfsExample  - success to read.
9636 [main] INFO  com.huawei.bigdata.hdfs.examples.HdfsExample  - success to
delete the file /user/hdfs-examples\test.txt
9860 [main] INFO  com.huawei.bigdata.hdfs.examples.HdfsExample  - success to
delete path /user/hdfs-examples
10010 [hdfs_example_0] INFO  com.huawei.bigdata.hdfs.examples.HdfsExample  -
success to create path /user/hdfs-examples/hdfs_example_0
10069 [hdfs_example_1] INFO  com.huawei.bigdata.hdfs.examples.HdfsExample  -
success to create path /user/hdfs-examples/hdfs_example_1
10553 [hdfs_example_0] INFO  com.huawei.bigdata.hdfs.examples.HdfsExample  -
success to write.
10607 [hdfs_example_1] INFO  com.huawei.bigdata.hdfs.examples.HdfsExample  -
success to write.
13356 [hdfs_example_0] INFO  com.huawei.bigdata.hdfs.examples.HdfsExample  -
success to append.
13469 [hdfs_example_1] INFO  com.huawei.bigdata.hdfs.examples.HdfsExample  -
success to append.
13784 [hdfs_example_0] INFO  com.huawei.bigdata.hdfs.examples.HdfsExample  -
result is : hi, I am bigdata. It is successful if you can see me.I append this
```



```
content.
13784 [hdfs_example_0] INFO  com.huawei.bigdata.hdfs.examples.HdfsExample  -
success to read.
13834 [hdfs_example_1] INFO  com.huawei.bigdata.hdfs.examples.HdfsExample  -
result is : hi, I am bigdata. It is successful if you can see me.I append this
content.
13834 [hdfs_example_1] INFO  com.huawei.bigdata.hdfs.examples.HdfsExample  -
success to read.
13837 [hdfs_example_0] INFO  com.huawei.bigdata.hdfs.examples.HdfsExample  -
success to delete the file /user/hdfs-examples/hdfs_example_0\test.txt
13889 [hdfs_example_1] INFO  com.huawei.bigdata.hdfs.examples.HdfsExample  -
success to delete the file /user/hdfs-examples/hdfs_example_1\test.txt
14003 [hdfs_example_0] INFO  com.huawei.bigdata.hdfs.examples.HdfsExample  -
success to delete path /user/hdfs-examples/hdfs_example_0
14118 [hdfs_example_1] INFO  com.huawei.bigdata.hdfs.examples.HdfsExample  -
success to delete path /user/hdfs-examples/hdfs_example_1
...
```

----Fin

12.3 Desarrollo de aplicaciones de Hive JDBC

Hive es un marco de almacenamiento de datos de código abierto construido en Hadoop. Puede usarlo para almacenar datos estructurados y analizar datos con las sentencias del lenguaje de consulta Hive (HiveQL). Hive convierte las sentencias HiveQL en trabajos MapReduce o Spark para consultar y analizar cantidades masivas de datos almacenados en clústeres de Hadoop.

Puedes usar Hive para:

- Extraer, transformar y cargar datos (ETL) con HiveQL.
- Analizar cantidades masivas de datos estructurados con HiveQL.
- Procesar datos en una amplia gama de formatos, como JSON, CSV, TEXTFILE, RCFILE, ORCFILE y SEQUENCEFILE, y personalice las extensiones.
- Conectar el cliente de forma flexible e invocar a las API de JDBC.

Hive es bueno para el análisis masivo de datos fuera de línea (como el análisis de estado de registros y clústeres), la minería de datos a gran escala (como el análisis del comportamiento del usuario, el análisis de la región de interés y la visualización de la región) y otros escenarios.

MRS proporciona ejemplos de proyectos de desarrollo de aplicaciones basados en Hive. Esta práctica proporciona orientación para obtener e importar un proyecto de ejemplo después de crear un clúster MRS y, a continuación, compilar y depurar el código localmente. En este proyecto de ejemplo, puede crear tablas Hive, insertar datos y leer datos.

Creación de un clúster MRS Hive

1. Cree y compre un clúster MRS que contenga Hive. Para obtener más información, consulte **Compra de un clúster personalizado**.

   NOTA

   En esta práctica, se utiliza como ejemplo un clúster MRS 3.1.5, con Hadoop y Hive instalados y con la autenticación Kerberos habilitada.

2. Haga clic en **Buy Now** y espere hasta que se cree el clúster MRS.

Preparación del archivo de configuración de desarrollo de aplicaciones

Paso 1 Inicie sesión en FusionInsight Manager y cree un usuario del clúster para la autenticación de seguridad del proyecto de ejemplo.

Elija **System** > **Permission** > **User**. En la página mostrada, haga clic en **Create**. En la página mostrada, cree un usuario máquina-máquina, por ejemplo, **developuser**.

Añada **hive** y **supergroup** a **User Group**.

Paso 2 Elija **System** > **Permission** > **User**. En la columna **Operation** de **developuser**, elija **More** > **Download Authentication Credential**. Guarde el archivo y descomprímalo para obtener los archivos **user.keytab** y **krb5.conf** del usuario.

Paso 3 Elija **Cluster**. En la pestaña **Dashboard**, haga clic en **More** y seleccione **Download Client**. En el cuadro de diálogo que se muestra, establezca **Select Client Type** en **Configuration Files Only** y haga clic en **OK**. Después de generar el paquete cliente, descargue el paquete como se le indique y descomprima.

Por ejemplo, si el paquete del archivo de configuración del cliente es **FusionInsight_Cluster_1_Services_Client.tar**, descomprima para obtener **FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles.tar**. A continuación, continúe para descomprimir este archivo.

1. Vaya al directorio **FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles\Hive\config** y obtenga los archivos de configuración.
2. Copie todo el contenido del archivo **hosts** en el directorio de descompresión al archivo **hosts** local. Asegúrese de que el PC local pueda comunicarse con los hosts que figuran en el archivo **hosts** del directorio de descompresión.

   NOTA

   - En esta práctica, asegúrese de que el entorno local puede comunicarse con el plano de red donde se despliega el clúster MRS. En general, puede acceder al clúster MRS a través de una EIP.
   - **C:\WINDOWS\system32\drivers\etc\hosts** es un directorio de ejemplo en un entorno Windows para almacenar el archivo **hosts** local.

----Fin

Obtención del proyecto de muestra

Paso 1 Obtenga el proyecto de muestra de Huawei Mirrors.

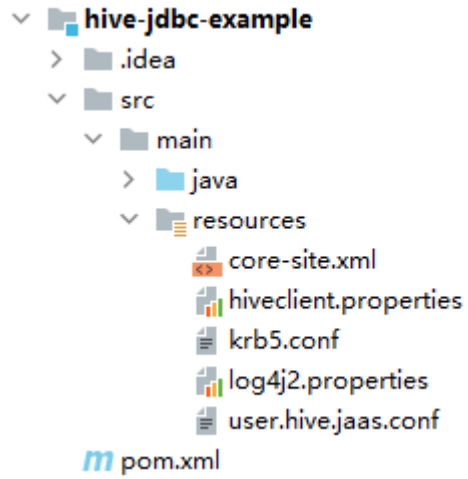
Descargue el código fuente y los archivos de configuración del proyecto de ejemplo, y configure las herramientas de desarrollo relacionadas en su PC local. Para obtener más información, consulte **Obtención de proyectos de muestra desde Huawei Mirros**.

Seleccione una rama basada en la versión del clúster y descargue el proyecto de muestra de MRS requerido.

Por ejemplo, el proyecto de muestra adecuado para esta práctica es **hive-jdbc-example**, que se puede obtener en **https://github.com/huaweicloud/huaweicloud-mrs-example/tree/mrs-3.1.5/src/hive-examples/hive-jdbc-example**.

Paso 2 Utilice IDEA para importar el proyecto de ejemplo y espere a que el proyecto Maven descargue los paquetes de dependencias. Para obtener más información, consulte **Configuración de proyecto de muestra de JDBC**.

Figura 12-9 Proyecto de muestra Hive



Después de configurar los parámetros Maven y SDK en el PC local, el proyecto de ejemplo carga automáticamente paquetes de dependencias relacionados.

Paso 3 Coloque los archivos de configuración del clúster y las credenciales de autenticación de usuario obtenidas en **Preparación del archivo de configuración de desarrollo de aplicaciones** al directorio **resources** del proyecto de ejemplo.

Paso 4 Para conectarse a un clúster MRS con la autenticación Kerberos habilitada, especifique la información de autenticación relacionada en el código de ejemplo.

En la clase **JDBCExample** del paquete **com.huawei.bigdata.hive.examples**, cambie **USER_NAME** por el nombre de usuario que está utilizando, por ejemplo, **developuser**.

```
KRB5_FILE = userdir + "krb5.conf";
System.setProperty("java.security.krb5.conf", KRB5_FILE);
USER_NAME = "developuser";
if ("KERBEROS".equalsIgnoreCase(auth)) {
    USER_KEYTAB_FILE = "src/main/resources/user.keytab";
    ZOOKEEPER_DEFAULT_SERVER_PRINCIPAL = "zookeeper/" + getUserRealm();
    System.setProperty(ZOOKEEPER_SERVER_PRINCIPAL_KEY,
ZOOKEEPER_DEFAULT_SERVER_PRINCIPAL);
}
...
```

En este proyecto de ejemplo, la hoja de ruta de desarrollo basada en los requisitos de servicio es la siguiente.

1. Preparar datos.
   a. Cree una tabla de información de empleados **employees_info**.
   b. Cargue la información de los empleados a **employees_info**.
2. Analizar datos.
   Recopile el número de registros de la tabla **employees_info**.
3. Eliminar la tabla.

----Fin

Creación y ejecución de la aplicación

Paso 1 Compilar el programa de muestra JDBC.

Haga clic en **Terminal** en la esquina inferior izquierda de la página IDEA para acceder al terminal. Ejecute el comando **mvn clean package** para realizar la compilación.

Si se muestra "BUILD SUCCESS", la compilación se realiza correctamente. Un archivo JAR que contiene el campo **-with-dependencies** se genera en el directorio **target** del proyecto de ejemplo.

```
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time:  03:30 min
[INFO] Finished at: 2023-05-17T20:22:44+08:00
[INFO] ------------------------------------------------------------------------
```

Paso 2 Cree un directorio como directorio de tiempo de ejecución, por ejemplo **D:\jdbc_example** en su entorno local, guarde los paquetes JAR generados cuyos nombres contienen el campo **-with-dependencies** en el directorio y cree el subdirectorio **src/main/resources** en el directorio. Copie todos los archivos del directorio **resources** del proyecto de ejemplo en este subdirectorio local.

Paso 3 Ejecute los siguientes comandos en el entorno de Windows CMD:

cd /d d:\jdbc_example

java -jar hive-jdbc-example-*XXX*-with-dependencies.jar

Paso 4 Compruebe la información de salida después de ejecutar la muestra. La siguiente información indica que las operaciones de tabla Hive se ejecutan correctamente:

```
...
2023-05-17 20:25:09,421 INFO  HiveConnection - Login timeout is 0
2023-05-17 20:25:09,656 INFO  HiveConnection - user login success.
2023-05-17 20:25:09,685 INFO  HiveConnection - Will try to open client transport
with JDBC Uri: jdbc:hive2://192.168.64.216:21066/;principal=hive/
hadoop.hadoop.com@HADOOP.COM;sasl.qop=auth-
conf;serviceDiscoveryMode=zooKeeper;auth=KERBEROS;zooKeeperNamespace=hiveserver2;u
ser.principal=developuser;user.keytab=src/main/resources/user.keytab
2023-05-17 20:25:30,294 INFO  JDBCExample - Create table success!
2023-05-17 20:26:34,032 INFO  JDBCExample - _c0
2023-05-17 20:26:34,266 INFO  JDBCExample - 0
2023-05-17 20:26:35,199 INFO  JDBCExample - Delete table success!
...
```

----Fin

12.4 Desarrollo de aplicaciones de Hive HCatalog

Hive es un marco de almacenamiento de datos de código abierto construido en Hadoop. Puede usarlo para almacenar datos estructurados y analizar datos con las sentencias del lenguaje de consulta Hive (HiveQL). Hive convierte las sentencias HiveQL en trabajos MapReduce o Spark para consultar y analizar cantidades masivas de datos almacenados en clústeres de Hadoop.

Puedes usar Hive para:

- Extraer, transformar y cargar datos (ETL) con HiveQL.
- Analizar cantidades masivas de datos estructurados con HiveQL.
- Procesar datos en una amplia gama de formatos, como JSON, CSV, TEXTFILE, RCFILE, ORCFILE y SEQUENCEFILE, y personalice las extensiones.
- Conectar el cliente de forma flexible e invocar a las API de JDBC.

Hive es bueno para el análisis masivo de datos fuera de línea (como el análisis de estado de registros y clústeres), la minería de datos a gran escala (como el análisis del comportamiento del usuario, el análisis de la región de interés y la visualización de la región) y otros escenarios.

MRS proporciona ejemplos de proyectos de desarrollo de aplicaciones basados en Hive. Esta práctica proporciona orientación para obtener e importar un proyecto de ejemplo después de crear un clúster MRS y, a continuación, compilar y depurar el código localmente. En este proyecto de ejemplo, puede crear tablas Hive, insertar datos y leer datos.

Creación de un clúster MRS Hive

1. Cree y compre un clúster MRS que contenga Hive. Para obtener más información, consulte **Compra de un clúster personalizado**.

   NOTA

   En esta práctica, se utiliza como ejemplo un clúster MRS 3.1.5, con Hadoop y Hive instalados y con la autenticación Kerberos habilitada.

2. Haga clic en **Buy Now** y espere hasta que se cree el clúster MRS.

Preparación del archivo de configuración de desarrollo de aplicaciones

Paso 1 Inicie sesión en FusionInsight Manager para crear un usuario del clúster para crear tablas de datos de Hive y enviar el programa HCatalog.

Elija **System** > **Permission** > **User**. En la página mostrada, haga clic en **Create**. En la página mostrada, cree un usuario máquina-máquina, por ejemplo, **hiveuser**.

Añada **hive** y **supergroup** a **User Group**.

Paso 2 Descargue e instale el cliente de clúster para ejecutar el programa HCatalog. Por ejemplo, el directorio de instalación es **/opt/client**.

----Fin

Obtención del proyecto de muestra

Paso 1 Obtenga el proyecto de muestra de Huawei Mirrors.

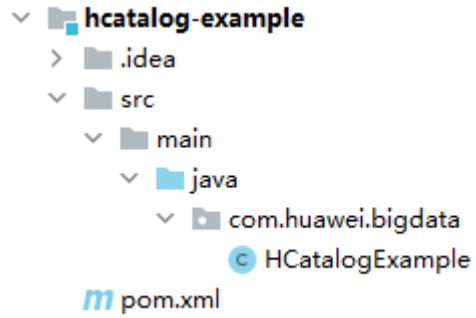
Descargue el código fuente y los archivos de configuración del proyecto de ejemplo, y configure las herramientas de desarrollo relacionadas en su PC local. Para obtener más información, consulte **Obtención de proyectos de muestra desde Huawei Mirros**.

Seleccione una rama basada en la versión del clúster y descargue el proyecto de muestra de MRS requerido.

Por ejemplo, el proyecto de muestra adecuado para esta práctica es **hcatalog-example**, que se puede obtener en **https://github.com/huaweicloud/huaweicloud-mrs-example/tree/mrs-3.1.5/src/hive-examples/hcatalog-example**.

Paso 2 Utilice IDEA para importar el proyecto de ejemplo y espere a que el proyecto Maven descargue los paquetes de dependencias. Para obtener más información, consulte **Configuración de proyecto de muestra de JDBC**.

Figura 12-10 Proyecto de muestra Hive HCatalog



Después de configurar los parámetros Maven y SDK en el PC local, el proyecto de ejemplo carga automáticamente paquetes de dependencias relacionados.

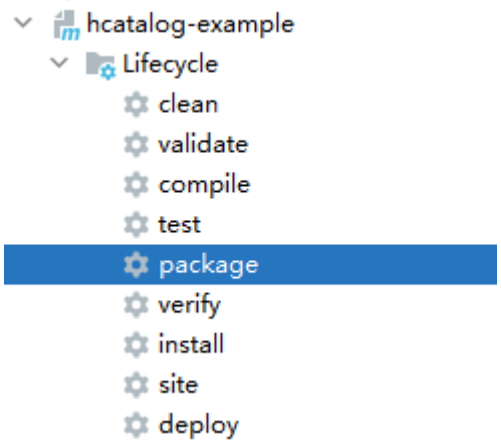
----Fin

Creación y ejecución de la aplicación

Paso 1 Compilar el programa de muestra HCatalog.

1. En la ventana de la herramienta Maven, seleccione **clean** en **Lifecycle** para ejecutar el proceso de construcción de Maven.
2. Seleccionar **package** de **Lifecycle** y ejecutar el proceso de compilación de Maven

Figura 12-11 Empaquetado del programa de muestra



Si se muestra "BUILD SUCCESS", la compilación se realiza correctamente.

El paquete **hcatalog-example-*XXX*.jar** se genera en el directorio **target** del proyecto de ejemplo.

```
[INFO]
------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO]
------------------------------------------------------------------------
[INFO] Total time:  03:30 min
[INFO] Finished at: 2023-05-17T20:22:44+08:00
[INFO]
------------------------------------------------------------------------
```


Paso 2 Inicie sesión en la Hive Beeline CLI y cree tablas de origen y tablas de datos para el análisis de HCatalog.

source /opt/client/bigdata_env

kinit hiveuser

beeline

create table t1(col1 int);

create table t2(col1 int,col2 int);

Inserte los datos de prueba en la tabla de datos de origen **t1**.

insert into table t1 select 1 union all select 1 union all select 2 union all select 2 union all select 3;

select * from t1;

```
+----------+
| t1.col1  |
+----------+
| 1        |
| 1        |
| 2        |
| 2        |
| 3        |
+----------+
```

Paso 3 Cargue el paquete JAR exportado a la ruta de acceso especificada del nodo Linux donde se despliega el cliente de clúster, por ejemplo, **/opt/hive_demo**.

Paso 4 Para facilitar las operaciones posteriores, configure el directorio de programa de ejemplo y el directorio de componentes del cliente como variables públicas.

Salga de la Beeline CLI y ejecute los siguientes comandos:

export HCAT_CLIENT=/opt/hive_demo

export HADOOP_HOME=/opt/client/HDFS/hadoop

export HIVE_HOME=/opt/client/Hive/Beeline

export HCAT_HOME=$HIVE_HOME/../HCatalog

export LIB_JARS=$HCAT_HOME/lib/hive-hcatalog-core-*XXX*.jar,$HCAT_HOME/lib/hive-metastore-*XXX*.jar,$HCAT_HOME/lib/hive-standalone-metastore-*XXX*.jar,$HIVE_HOME/lib/hive-exec-*XXX*.jar,$HCAT_HOME/lib/libfb303-*XXX*.jar,$HCAT_HOME/lib/slf4j-api-*XXX*.jar,$HCAT_HOME/lib/jdo-api-*XXX*.jar,$HCAT_HOME/lib/antlr-runtime-*XXX*.jar,$HCAT_HOME/lib/datanucleus-api-jdo-*XXX*.jar,$HCAT_HOME/lib/datanucleus-core-*XXX*.jar,$HCAT_HOME/lib/datanucleus-rdbms-fi-*XXX*.jar,$HCAT_HOME/lib/log4j-api-*XXX*.jar,$HCAT_HOME/lib/log4j-core-*XXX*.jar,$HIVE_HOME/lib/commons-lang-*XXX*.jar,$HIVE_HOME/lib/hive-exec-*XXX*.jar

export HADOOP_CLASSPATH=$HCAT_HOME/lib/hive-hcatalog-core-*XXX*.jar:$HCAT_HOME/lib/hive-metastore-*XXX*.jar:$HCAT_HOME/lib/hive-standalone-metastore-*XXX*.jar:$HIVE_HOME/lib/hive-exec-*XXX*.jar:$HCAT_HOME/lib/libfb303-*XXX*.jar:$HADOOP_HOME/etc/hadoop:$HCAT_HOME/conf:$HCAT_HOME/lib/slf4j-api-*XXX*.jar:$HCAT_HOME/lib/jdo-api-*XXX*.jar:$HCAT_HOME/lib/antlr-runtime-*XXX*.jar:$HCAT_HOME/lib/datanucleus-api-jdo-*XXX*.jar:$HCAT_HOME/lib/

datanucleus-core-*XXX*.jar:$HCAT_HOME/lib/datanucleus-rdbms-fi-*XXX*.jar:$HCAT_HOME/lib/log4j-api-*XXX*.jar:$HCAT_HOME/lib/log4j-core-*XXX*.jar:$HIVE_HOME/lib/commons-lang-*XXX*.jar:$HIVE_HOME/lib/hive-exec-*XXX*.jar

NOTA

El número de versión *XXX* del paquete JAR especificado en **LIB_JARS** y **HADOOP_CLASSPATH** debe cambiarse a la versión que está utilizando.

Paso 5 Utilice el cliente Yarn para enviar una tarea.

yarn --config $HADOOP_HOME/etc/hadoop jar $HCAT_CLIENT/hcatalog-example-*XXX*.jar com.huawei.bigdata.HCatalogExample -libjars $LIB_JARS t1 t2

```
...
2023-05-18 20:05:56,691 INFO mapreduce.Job: The url to track the job: https://
host-192-168-64-122:26001/proxy/application_1683438782910_0008/
2023-05-18 20:05:56,692 INFO mapreduce.Job: Running job: job_1683438782910_0008
2023-05-18 20:06:07,250 INFO mapreduce.Job: Job job_1683438782910_0008 running in
uber mode : false
2023-05-18 20:06:07,253 INFO mapreduce.Job:  map 0% reduce 0%
2023-05-18 20:06:15,362 INFO mapreduce.Job:  map 25% reduce 0%
2023-05-18 20:06:16,386 INFO mapreduce.Job:  map 50% reduce 0%
2023-05-18 20:06:35,999 INFO mapreduce.Job:  map 100% reduce 0%
2023-05-18 20:06:42,048 INFO mapreduce.Job:  map 100% reduce 100%
2023-05-18 20:06:43,136 INFO mapreduce.Job: Job job_1683438782910_0008 completed
successfully
2023-05-18 20:06:44,118 INFO mapreduce.Job: Counters: 54
...
```

Paso 6 Después de completar el trabajo, vaya a la Hive Beeline CLI, consulte los datos en la tabla **t2** y vea el resultado del análisis de datos.

select * from t2;

```
+----------+----------+
| t2.col1  | t2.col2  |
+----------+----------+
| 1        | 2        |
| 2        | 2        |
| 3        | 1        |
+----------+----------+
```

----Fin

12.5 Desarrollo de aplicaciones de Kafka

Kafka es un sistema de publicación y suscripción de mensajes distribuidos. Con características similares a JMS, Kafka procesa datos de streaming activos.

Kafka se aplica a muchos escenarios, como la cola de mensajes, el seguimiento de comportamiento, la supervisión de datos de O&M, la recopilación de registros, el procesamiento de secuencias, el seguimiento de eventos y la persistencia de registros.

Kafka tiene las siguientes características:

- Alto rendimiento
- Persistencia de mensajes a los discos
- Sistema distribuido escalable
- Alta tolerancia a fallos

- Soporte para escenarios en línea y fuera de línea

MRS proporciona ejemplos de proyectos de desarrollo de aplicaciones basados en Kafka. Esta práctica proporciona orientación para que obtenga e importe un proyecto de muestra después de crear un clúster MRS y, a continuación, realice la construcción y puesta en marcha localmente. En este proyecto de ejemplo, puede implementar el procesamiento de datos de streaming.

Las directrices para el proyecto de muestra en esta práctica son las siguientes:

1. Cree dos topics en el cliente Kafka para que sirvan como topic de entrada y salida.
2. Desarrolle Kafka Streams para contar palabras en cada mensaje leyendo mensajes en el topic de entrada y para generar el resultado en pares clave-valor consumiendo datos en el topic de salida.

Creación de un clúster de MRS

Paso 1 Cree y compre un clúster MRS que contenga Kafka. Para obtener más información, consulte **Compra de un clúster personalizado**.

NOTA

En esta práctica, se utiliza como ejemplo un clúster MRS 3.1.0, con Hadoop y Kafka instalados y con la autenticación Kerberos deshabilitada.

Paso 2 Después de comprar el clúster, instale el cliente en cualquier nodo del clúster. Para obtener más información, consulte **Instalación y uso de cliente de clúster**.

Por ejemplo, instale el cliente en el directorio **/opt/client** en el nodo de gestión activo.

Paso 3 Después de instalar el cliente, cree el directorio **lib** en el cliente para almacenar los paquetes JAR relacionados.

Copie a **lib** los paquetes JAR de Kafka en el directorio descomprimido durante la instalación del cliente.

Por ejemplo, si la ruta de descarga del paquete de software cliente es **/tmp/FusionInsight-Client** en el nodo de gestión activa, ejecute los siguientes comandos:

mkdir /opt/client/lib

cd /tmp/FusionInsight-Client/FusionInsight_Cluster_1_Services_ClientConfig

scp Kafka/install_files/kafka/libs/* /opt/client/lib

----Fin

Desarrollo de la aplicación

Paso 1 Obtenga el proyecto de muestra de Huawei Mirrors.

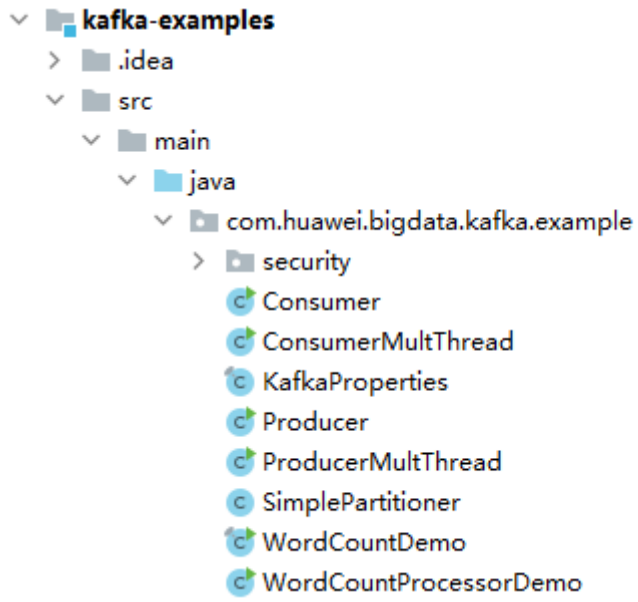
Descargue el código fuente del proyecto Maven y los archivos de configuración del proyecto de ejemplo, y configure las herramientas de desarrollo relacionadas en su PC local. Para obtener más información, consulte **Obtención de proyectos de muestra desde Huawei Mirros**.

Seleccione una rama basada en la versión del clúster y descargue el proyecto de muestra de MRS requerido.

Por ejemplo, el proyecto de muestra adecuado para esta práctica es **WordCountDemo**, que se puede obtener en **https://github.com/huaweicloud/huaweicloud-mrs-example/tree/mrs-3.1.0/src/kafka-examples**.

Paso 2 Utilice IDEA para importar el proyecto de ejemplo y espere a que el proyecto Maven descargue los paquetes de dependencias.

Después de configurar los parámetros Maven y SDK en el PC local, el proyecto de ejemplo carga automáticamente paquetes de dependencias relacionados. Para obtener más información, consulte **Configuración e importación de proyectos de muestra**.



El proyecto de ejemplo **WordCountDemo** invoca a las API de Kafka para obtener y ordenar registros de palabras y luego obtener los registros de cada palabra. El fragmento de código es el siguiente:

```
...
    static Properties getStreamsConfig() {
        final Properties props = new Properties();
        KafkaProperties kafkaProc = KafkaProperties.getInstance();
        //Set broker addresses based on site requirements.
        props.put(BOOTSTRAP_SERVERS, kafkaProc.getValues(BOOTSTRAP_SERVERS, "node-
group-1kLFk.mrs-rbmq.com:9092"));
        props.put(SASL_KERBEROS_SERVICE_NAME, "kafka");
        props.put(KERBEROS_DOMAIN_NAME, kafkaProc.getValues(KERBEROS_DOMAIN_NAME,
"hadoop.hadoop.com"));
        props.put(APPLICATION_ID, kafkaProc.getValues(APPLICATION_ID, "streams-
wordcount"));
        //Set the protocol type, which can be SASL_PLAINTEXT or PLAINTEXT.
        props.put(SECURITY_PROTOCOL, kafkaProc.getValues(SECURITY_PROTOCOL,
"PLAINTEXT"));
        props.put(CACHE_MAX_BYTES_BUFFERING, 0);
        props.put(DEFAULT_KEY_SERDE, Serdes.String().getClass().getName());
        props.put(DEFAULT_VALUE_SERDE, Serdes.String().getClass().getName());
        props.put(ConsumerConfig.AUTO_OFFSET_RESET_CONFIG, "earliest");
        return props;
    }
    static void createWordCountStream(final StreamsBuilder builder) {
        //Receives input records from the input topic.
        final KStream<String, String> source = builder.stream(INPUT_TOPIC_NAME);
        //Aggregates the calculation results of the key-value pairs.
        final KTable<String, Long> counts = source
                .flatMapValues(value ->
Arrays.asList(value.toLowerCase(Locale.getDefault()).split(REGEX_STRING)))
```



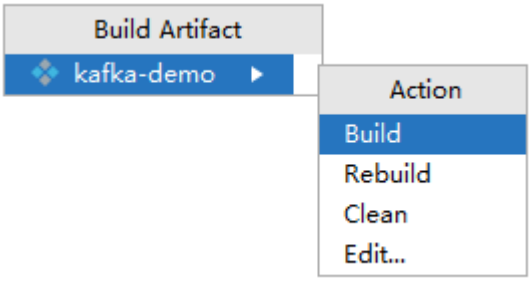
```
                .groupBy((key, value) -> value)
                .count();
        //Outputs the key-value pairs from the output topic.
        counts.toStream().to(OUTPUT_TOPIC_NAME, Produced.with(Serdes.String(),
Serdes.Long()));
    }
...
```

NOTA

- Establezca **BOOTSTRAP_SERVERS** en el nombre del host y el número de puerto del nodo del broker Kafka según los requisitos del sitio. Puede elegir **Cluster** > **Services** > **Kafka** > **Instance** en FusionInsight Manager para ver la información de la instancia del broker.
- Establezca **SECURITY_PROTOCOL** en el protocolo para conectarse a Kafka. En este ejemplo, establezca este parámetro en **PLAINTEXT**.

Paso 3 Después de confirmar que los parámetros de **WordCountDemo.java** son correctos, construya el proyecto y empaquetelo en un archivo JAR.

Para obtener más información sobre cómo crear un archivo JAR, consulte **Puesta en marcha de una aplicación de Linux**.



Por ejemplo, el archivo JAR es **kafka-demo.jar**.

----Fin

Cargar el paquete JAR y los datos de origen

Paso 1 Cargue el paquete JAR a un directorio, por ejemplo, **/opt/client/lib** en el nodo cliente.

NOTA

Si no puede acceder directamente al nodo cliente para cargar archivos a través de la red local, cargue el paquete JAR o los datos de origen a OBS, impórtelos a HDFS en la pestaña **Files** de la consola MRS. Y ejecute el comando **hdfs dfs -get** en el cliente HDFS para descargarlo al nodo cliente.

----Fin

Ejecución de un trabajo y visualización del resultado

Paso 1 Inicie sesión en el nodo donde está instalado el cliente de clúster como usuario **root**.

cd /opt/client

source bigdata_env

Paso 2 Cree un topic de entrada y un topic de salida. Asegúrese de que los nombres de los topics son los mismos que los especificados en el código de ejemplo. Establezca la política de limpieza del topic de salida en **compact**.

kafka-topics.sh --create --zookeeper *IP address of the quorumpeer instance:ZooKeeper client connection port*/**kafka --replication-factor 1 --partitions 1 --topic** *Topic name*

Para consultar la dirección IP de la instancia de quorumpeer, inicie sesión en el FusionInsight Manager del clúster, elija **Cluster** > **Services** > **ZooKeeper** y haga clic en la pestaña **Instance**. Utilice comas (,) para separar varias direcciones IP. Puede obtener el puerto de conexión del cliente ZooKeeper consultando el parámetro de configuración ZooKeeper **clientPort**. El valor predeterminado es **2181**.

Por ejemplo, ejecute los siguientes comandos:

kafka-topics.sh --create --zookeeper 192.168.0.17:2181/kafka --replication-factor 1 --partitions 1 --topic streams-wordcount-input

kafka-topics.sh --create --zookeeper 192.168.0.17:2181/kafka --replication-factor 1 --partitions 1 --topic streams-wordcount-output --config cleanup.policy=compact

Paso 3 Después de crear los topics, ejecute el siguiente comando para ejecutar la aplicación:

java -cp .:/opt/client/lib/* com.huawei.bigdata.kafka.example.WordCountDemo

Paso 4 Abra una nueva ventana de cliente y ejecute los siguientes comandos para usar **kafka-console-producer.sh** para escribir mensajes en el topic de entrada:

cd /opt/client

source bigdata_env

kafka-console-producer.sh --broker-list *IP address of the broker instance:Kafka connection port(for example, 192.168.0.13:9092)* **--topic streams-wordcount-input --producer.config /opt/client/Kafka/kafka/config/producer.properties**

Paso 5 Abra una nueva ventana de cliente y ejecute los siguientes comandos para usar **kafka-console-consumer.sh** para consumir datos del tema de salida y ver el resultado:

cd /opt/client

source bigdata_env

kafka-console-consumer.sh --topic streams-wordcount-output --bootstrap-server *IP address of the broker instance:Kafka connection port* **--consumer.config /opt/client/Kafka/kafka/config/consumer.properties --from-beginning --property print.key=true --property print.value=true --property key.deserializer=org.apache.kafka.common.serialization.StringDeserializer --property value.deserializer=org.apache.kafka.common.serialization.LongDeserializer --formatter kafka.tools.DefaultMessageFormatter**

Escriba un mensaje al topic de entrada.

```
>This is Kafka Streams test
>test starting
>now Kafka Streams is running
>test end
```

La información que se muestra es la siguiente:

```
this    1
is      1
kafka   1
streams 1
test    1
test    2
```



```
starting 1
now      1
kafka    2
streams  2
is       2
running  1
test     3
end      1
```

----Fin

12.6 Desarrollo de aplicaciones de Flink

Flink es un marco de computación unificado que soporta tanto el procesamiento por lotes como el procesamiento de flujo. Proporciona un motor de procesamiento de datos de flujo que admite la distribución de datos y la computación en paralelo. Flink cuenta con procesamiento de flujo y es un motor de procesamiento de flujo de código abierto superior en la industria.

Flink proporciona procesamiento de datos de canalización de alta concurrencia, latencia de nivel de milisegundos y alta confiabilidad, lo que lo hace adecuado para el procesamiento de datos de baja latencia.

El sistema Flink consta de las siguientes partes:

- Client

  Flink client se utiliza para enviar trabajos de streaming a Flink.
- TaskManager

  TaskManager es un nodo de ejecución de servicio de Flink, que ejecuta tareas específicas. Puede haber muchos TaskManagers y son equivalentes entre sí.
- JobManager

  JobManager es un nodo de gestión de Flink. Gestiona todas las TaskManagers y programa las tareas enviadas por los usuarios a TaskManagers específicos. En el modo de alta disponibilidad (HA), se despliega múltiples JobManagers. Entre estos JobManagers se selecciona uno como el JobManager activo, y los otros están en espera.

MRS proporciona ejemplos de proyectos de desarrollo de aplicaciones basados en múltiples componentes de Flink. Esta práctica proporciona orientación para que obtenga e importe un proyecto de muestra después de crear un clúster MRS y, a continuación, realice la construcción y puesta en marcha localmente. En este proyecto de ejemplo, puede implementar Flink DataStream para procesar datos.

Creación de un clúster MRS Flink

1. Cree y compre un clúster MRS que contenga Hive. Para obtener más información, consulte **Compra de un clúster personalizado**.

   NOTA

   En esta práctica, se utiliza como ejemplo un clúster MRS 3.2.0-LTS.1, con Hadoop y Flink instalados y con la autenticación Kerberos habilitada.
2. Haga clic en **Buy Now** y espere hasta que se cree el clúster MRS.

Preparación del archivo de configuración de clúster

Paso 1 Una vez creado el clúster, inicie sesión en FusionInsight Manager y cree un usuario del clúster para enviar trabajos de Flink.

Elija **System** > **Permission** > **User**. En la página mostrada, haga clic en **Create**. En la página mostrada, cree un usuario máquina-máquina, por ejemplo, **flinkuser**.

Agregue el grupo de usuarios **supergroup** y asocie el rol **System_administrator**.

Paso 2 Elija **System** > **Permission** > **User**. En la columna **Operation** de **flinkuser**, elija **More** > **Download Authentication Credential**. Guarde el archivo y descomprímalo para obtener los archivos **user.keytab** y **krb5.conf** del usuario.

Paso 3 Elija **Cluster**. En la pestaña **Dashboard**, haga clic en **More** y seleccione **Download Client**. En el cuadro de diálogo que se muestra, establezca **Select Client Type** en **Configuration Files Only** y haga clic en **OK**. Después de generar el paquete cliente, descargue el paquete como se le indique y descomprima.

Por ejemplo, si el paquete del archivo de configuración del cliente es **FusionInsight_Cluster_1_Services_Client.tar**, descomprima para obtener **FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles.tar**. A continuación, continúe para descomprimir este archivo.

Vaya al directorio **FusionInsight_Cluster_1_Services_ClientConfig_ConfigFiles\Flink \config** y obtenga los archivos de configuración.

----Fin

Obtención del proyecto de muestra

Paso 1 Obtenga el proyecto de muestra de Huawei Mirrors.

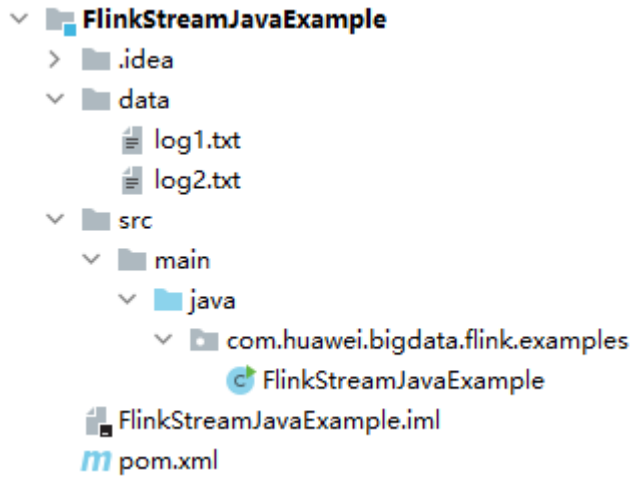
Descargue el código fuente y los archivos de configuración del proyecto de ejemplo, y configure las herramientas de desarrollo relacionadas en su PC local. Para obtener más información, consulte **Obtención de proyectos de muestra desde Huawei Mirros**.

Seleccione una rama basada en la versión del clúster y descargue el proyecto de muestra de MRS requerido.

Por ejemplo, el proyecto de muestra adecuado para esta práctica es **FlinkStreamJavaExample**, que se puede obtener en **https://github.com/huaweicloud/huaweicloud-mrs-example/tree/mrs-3.2.0.1/src/flink-examples/flink-examples-security/FlinkStreamJavaExample**.

Paso 2 Utilice IDEA para importar el proyecto de ejemplo y espere a que el proyecto Maven descargue los paquetes de dependencias. Para obtener más información, consulte **Configuración e importación de proyectos de muestra**.

Figura 12-12 Proyecto de muestra de Flink



Después de configurar los parámetros Maven y SDK en el PC local, el proyecto de ejemplo carga automáticamente paquetes de dependencias relacionados.

Paso 3 Utilice Flink client para enviar el programa DataStream desarrollado para que no se requiera autenticación de seguridad en el código.

Supongamos que hay un archivo de registro de tiempo en el sitio durante los fines de semana de un sitio web de compras en línea. Escriba el programa DataStream para recopilar estadísticas en tiempo real sobre información detallada sobre las usuarias femeninas que pasan más de 2 horas en compras en línea.

La primera columna del archivo de registro registra los nombres, la segunda columna registra el sexo y la tercera columna registra el tiempo en el sitio (en minutos). Tres columnas están separadas por comas (,).

- **log1.txt**: registros recogidos el sábado

```
LiuYang,female,20
YuanJing,male,10
GuoYijun,male,5
CaiXuyu,female,50
Liyuan,male,20
FangBo,female,50
LiuYang,female,20
YuanJing,male,10
GuoYijun,male,50
CaiXuyu,female,50
FangBo,female,60
```

log2.txt: registros recogidos el domingo

```
LiuYang,female,20
YuanJing,male,10
CaiXuyu,female,50
FangBo,female,50
GuoYijun,male,5
CaiXuyu,female,50
Liyuan,male,20
CaiXuyu,female,50
FangBo,female,50
LiuYang,female,20
YuanJing,male,10
FangBo,female,50
GuoYijun,male,50
CaiXuyu,female,50
FangBo,female,60
```


El procedimiento de desarrollo es el siguiente:

1. Lea los datos de texto, genere DataStreams y analice los datos para generar UserRecord.
2. Busque los datos de destino (tiempo en el sitio de usuarios femeninos).
3. Realice la operación keyby basada en nombres y géneros, y calcule el tiempo total que cada usuario femenino pasa en línea dentro de una ventana de tiempo.
4. Busque usuarios cuya duración consecutiva en línea exceda el umbral.

```
public class FlinkStreamJavaExample {
    public static void main(String[] args) throws Exception {
        // Print the command reference for flink run.
        System.out.println("use command as: ");
        System.out.println("./bin/flink run --class
com.huawei.bigdata.flink.examples.FlinkStreamJavaExample /opt/test.jar --
filePath /opt/log1.txt,/opt/log2.txt --windowTime 2");

System.out.println("*****************************************************************
**************************");
        System.out.println("<filePath> is for text file to read data, use comma
to separate");
        System.out.println("<windowTime> is the width of the window, time as
minutes");

System.out.println("*****************************************************************
**************************");

        // Read text pathes and separate them with commas (,). If the source file
is in the HDFS, set this parameter to a specific HDFS path, for example, hdfs://
hacluster/tmp/log1.txt,hdfs://hacluster/tmp/log2.txt.
        final String[] filePaths = ParameterTool.fromArgs(args).get("filePath",
"/opt/log1.txt,/opt/log2.txt").split(",");
        assert filePaths.length > 0;

       // Set the time window. The default value is 2 minutes per time window.
One time window is sufficient to read all data in the text.
        final int windowTime = ParameterTool.fromArgs(args).getInt("windowTime",
2);

       // Construct an execution environment and use eventTime to process the
data obtained in a time window.
        final StreamExecutionEnvironment env =
StreamExecutionEnvironment.getExecutionEnvironment();
        env.setStreamTimeCharacteristic(TimeCharacteristic.EventTime);
        env.setParallelism(1);

       // Read the text data stream.
        DataStream<String> unionStream = env.readTextFile(filePaths[0]);
        if (filePaths.length > 1) {
            for (int i = 1; i < filePaths.length; i++) {
                unionStream = unionStream.union(env.readTextFile(filePaths[i]));
            }
        }

       // Convert the data, construct data processing logic, and calculate and
print the results.
        unionStream.map(new MapFunction<String, UserRecord>() {
            @Override
            public UserRecord map(String value) throws Exception {
                return getRecord(value);
            }
        }).assignTimestampsAndWatermarks(
                new Record2TimestampExtractor()
        ).filter(new FilterFunction<UserRecord>() {
            @Override
            public boolean filter(UserRecord value) throws Exception {
                return value.sexy.equals("female");
            }
```



```
        }).keyBy(
            new UserRecordSelector()
        ).window(
            TumblingEventTimeWindows.of(Time.minutes(windowTime))
        ).reduce(new ReduceFunction<UserRecord>() {
            @Override
            public UserRecord reduce(UserRecord value1, UserRecord value2)
                    throws Exception {
                value1.shoppingTime += value2.shoppingTime;
                return value1;
            }
        }).filter(new FilterFunction<UserRecord>() {
            @Override
            public boolean filter(UserRecord value) throws Exception {
                return value.shoppingTime > 120;
            }
        }).print();

        // Call execute to trigger the execution.
        env.execute("FemaleInfoCollectionPrint java");
    }

    // Construct a keyBy keyword for grouping.
    private static class UserRecordSelector implements KeySelector<UserRecord,
Tuple2<String, String>> {
        @Override
        public Tuple2<String, String> getKey(UserRecord value) throws Exception {
            return Tuple2.of(value.name, value.sexy);
        }
    }

    // Resolve text line data and construct the UserRecord data structure.
    private static UserRecord getRecord(String line) {
        String[] elems = line.split(",");
        assert elems.length == 3;
        return new UserRecord(elems[0], elems[1], Integer.parseInt(elems[2]));
    }

    // Define the UserRecord data structure and override the toString printing
method.
    public static class UserRecord {
        private String name;
        private String sexy;
        private int shoppingTime;

        public UserRecord(String n, String s, int t) {
            name = n;
            sexy = s;
            shoppingTime = t;
        }

        public String toString() {
            return "name: " + name + "  sexy: " + sexy + "  shoppingTime: " +
shoppingTime;
        }
    }

    // Construct a class inherited from AssignerWithPunctuatedWatermarks to set
eventTime and waterMark.
    private static class Record2TimestampExtractor implements
AssignerWithPunctuatedWatermarks<UserRecord> {

        // add tag in the data of datastream elements
        @Override
        public long extractTimestamp(UserRecord element, long previousTimestamp) {
            return System.currentTimeMillis();
        }

        // give the watermark to trigger the window to execute, and use the value
```



```
to check if the window elements is ready
        @Override
        public Watermark checkAndGetNextWatermark(UserRecord element, long
extractedTimestamp) {
            return new Watermark(extractedTimestamp - 1);
        }
    }
}
```

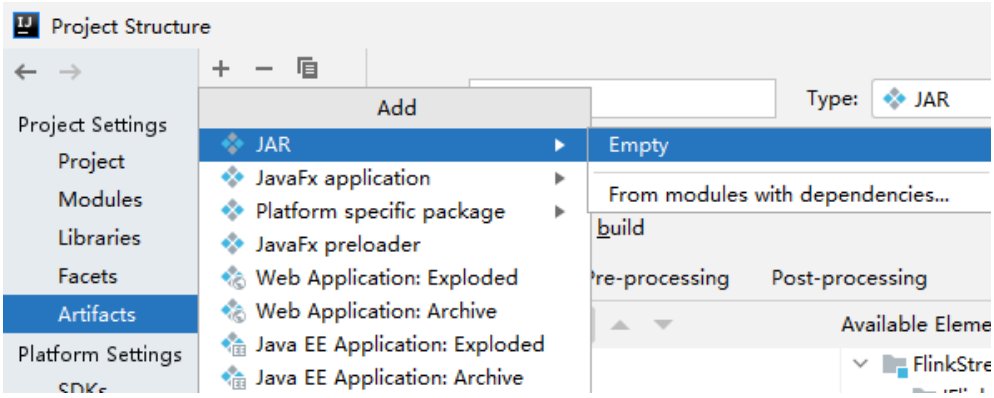
----Fin

Creación y ejecución de la aplicación

Paso 1 En IntelliJ IDEA, configure la información de Artifacts del proyecto.

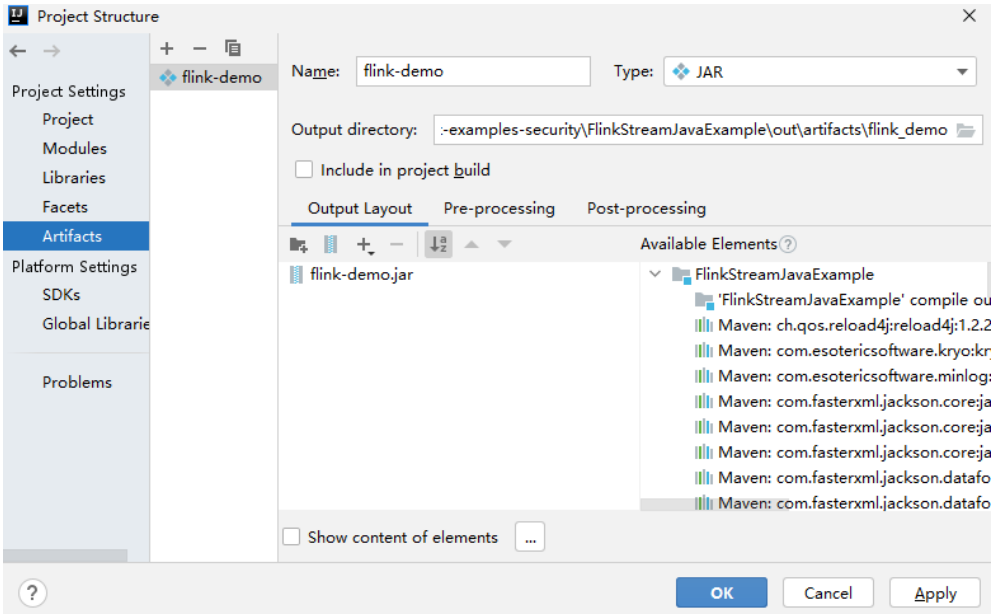
1. En la página de inicio de IDEA, seleccione **File** > **Project Structures...**.
2. En la página **Project Structure**, seleccione **Artifacts**, haga clic en + y elija **JAR** > **Empty**.

Figura 12-13 Adición de Artifacts



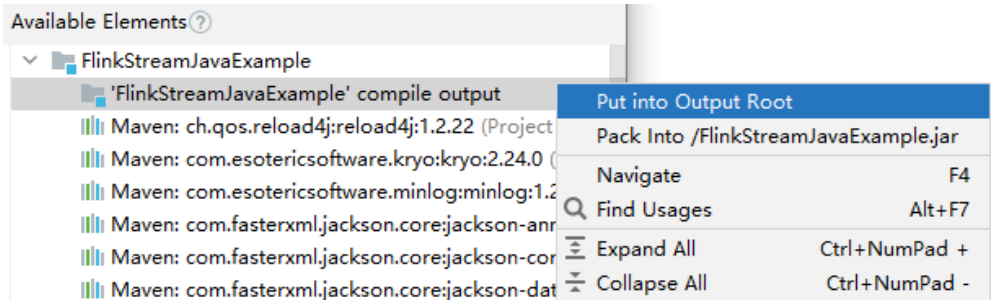
3. Establezca el nombre, el tipo y la ruta de salida del paquete JAR, por ejemplo, **flink-demo**.

Figura 12-14 Configuración de información básica



4. Haga clic con el botón derecho en **'FlinkStreamJavaExample' compile output** y elija **Put into Output Root** en el menú contextual. A continuación, haga clic en **Apply**.

Figura 12-15 Put into Output Root

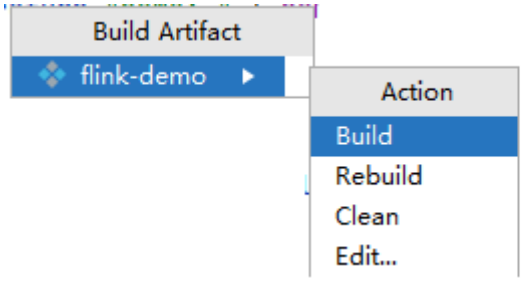


5. Haga clic en **OK**.

Paso 2 Generar un archivo JAR.

1. En la página de inicio de IDEA, seleccione **Build** > **Build Artifacts...**.
2. En el menú que se muestra, elija **FlinkStreamJavaExample** > **Build** para generar el archivo JAR.

Figura 12-16 Build



3. Obtenga el archivo **flink-demo.jar** de la ruta de acceso configurada en **Paso 1.3**.

Paso 3 Instalar y configurar el Flink client.

1. Instale el cliente de clúster MRS, por ejemplo, en **/opt/hadoopclient**.
2. Descomprima el paquete de credenciales de autenticación descargado de **Preparación del archivo de configuración de clúster** y copie el archivo obtenido en un directorio en el nodo cliente, por ejemplo, **/opt/hadoopclient/Flink/flink/conf**.
3. Ejecute el siguiente comando para establecer los parámetros de configuración del cliente de Flink y guardar la configuración:

   vi /opt/hadoopclient/Flink/flink/conf/flink-conf.yaml

   Agregue la dirección IP del servicio del nodo cliente y la dirección IP flotante del FusionInsight Manager al elemento de configuración **jobmanager.web.allow-access-address** y agregue la ruta **keytab** y el nombre de usuario a los elementos de configuración correspondientes.

```
...
jobmanager.web.allow-access-address:
192.168.64.122,192.168.64.216,192.168.64.234
...
security.kerberos.login.keytab: /opt/client/Flink/flink/conf/user.keytab
security.kerberos.login.principal: flinkuser
...
```


4. Configurar la autenticación de seguridad.

   a. Ejecute los siguientes comandos para generar un archivo de autenticación de seguridad de Flink client:

      cd /opt/hadoopclient/Flink/flink/bin

      sh generate_keystore.sh

      Introduzca una contraseña definida por el usuario para la autenticación.

   b. Configure las rutas para que el cliente acceda a los archivos **flink.keystore** y **flink.truststore**.

      cd /opt/hadoopclient/Flink/flink/conf/

      mkdir ssl

      mv flink.keystore ssl/

      mv flink.truststore ssl/

      vi /opt/hadoopclient/Flink/flink/conf/flink-conf.yaml

      Cambie las rutas de los dos parámetros siguientes a rutas relativas:

      ```
      security.ssl.keystore: ssl/flink.keystore
      security.ssl.truststore: ssl/flink.truststore
      ```

Paso 4 Cargue el paquete JAR generado en **Paso 2** al directorio relacionado en el nodo de Flink client, por ejemplo, **/opt/hadoopclient**.

Cree el directorio **conf** en el directorio donde se encuentra el paquete JAR y cargue los archivos de configuración en **Flink/config** del paquete de archivos de configuración del cliente de clúster obtenido en **Preparación del archivo de configuración de clúster** al directorio **conf**.

Paso 5 Cargue los archivos de datos de origen de aplicación al nodo donde se despliega la instancia NodeManager.

En este ejemplo, los archivos de datos de origen **log1.txt** y **log2.txt** se almacenan en el host local. Debe cargar los archivos en el directorio **/opt** en todos los nodos de Yarn NodeManager y establecer el permiso de archivo en **755**.

Paso 6 En el Flink client, ejecute el comando **yarn session** para iniciar el clúster de Flink.

cd /opt/hadoopclient/Flink/flink

bin/yarn-session.sh -jm 1024 -tm 1024 -t conf/ssl/

```
...
Cluster started: Yarn cluster with application id application_1683438782910_0009
JobManager Web Interface: http://192.168.64.10:32261
```

Paso 7 Abra una nueva ventana de conexión de cliente, vaya al directorio de Flink client y ejecute el programa.

source /opt/hadoopclient/bigdata_env

cd /opt/hadoopclient/Flink/flink

bin/flink run --class com.huawei.bigdata.flink.examples.FlinkStreamJavaExample /opt/hadoopclient/flink-demo.jar --filePath /opt/log1.txt,/opt/log2.txt --windowTime 2

```
...
2023-05-26 19:56:52,068 | INFO  | [main] | Found Web Interface
host-192-168-64-10:32261 of application 'application_1683438782910_0009'. |
org.apache.flink.yarn.YarnClusterDescriptor.setClusterEntrypointInfoToConfig(YarnC
```

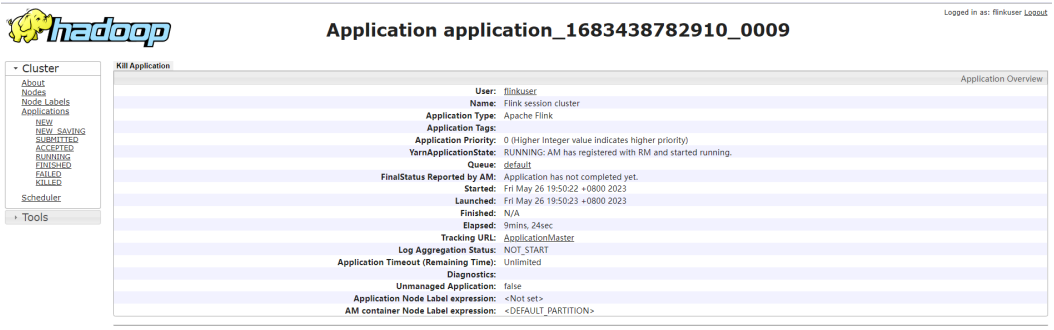


```
lusterDescriptor.java:1854)
Job has been submitted with JobID 7647255752b09456d5a580e33a8529f5
Program execution finished
Job with JobID 7647255752b09456d5a580e33a8529f5 has finished.
Job Runtime: 36652 ms
```

Paso 8 Comprobar los resultados de la ejecución.

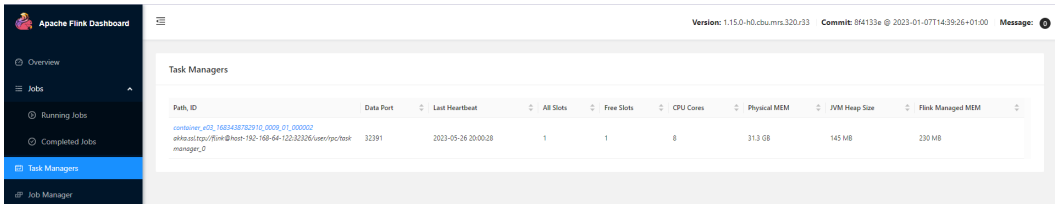
Inicie sesión en FusionInsight Manager como usuario **flinkuser** y elija **Cluster** > **Service** > **Yarn**. En la página **Applications**, haga clic en un nombre de trabajo para ir a la página de detalles del trabajo.

Figura 12-17 Consulta de detalles de trabajo de Yarn



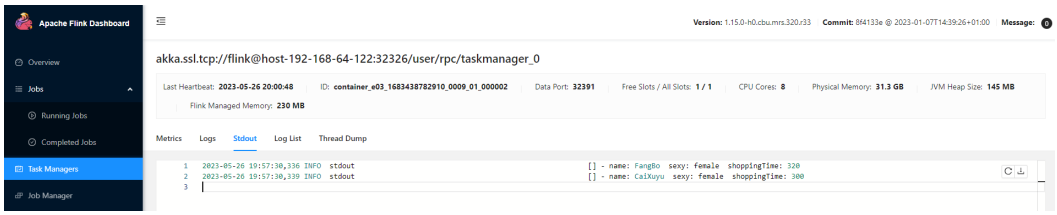
Para el trabajo enviado en una session, puede hacer clic en **Tracking URL** para iniciar sesión en la página nativa del servicio de Flink para ver la información del trabajo.

Figura 12-18 Ver detalles del trabajo de Flink



En este proyecto de ejemplo, haga clic en **Task Managers** y vea el resultado en ejecución en la pestaña **Stdout** del trabajo.

Figura 12-19 Ver los resultados en ejecución de la aplicación



----Fin

12.7 Desarrollo de aplicaciones de ClickHouse

ClickHouse es una base de datos orientada a columnas para el procesamiento analítico en línea. Soporta consultas SQL y proporciona un buen rendimiento de consultas. El análisis de

agregación y el rendimiento de las consultas basadas en tablas grandes y amplias es excelente, que es un orden de magnitud más rápido que otras bases de datos analíticas.

Características de ClickHouse:

- Alta relación de compresión de datos
- Computación paralela multinúcleo
- Motor de computación vectorizado
- Soporte para estructura de datos anidados
- Soporte para índices dispersos
- Soporte para INSERT y UPDATE

Escenarios de aplicaciones de ClickHouse:

- Almacenamiento de datos en tiempo real

  El motor de computación de streaming (como Flink) se usa para escribir datos en tiempo real en ClickHouse. Con el excelente rendimiento de la consulta de ClickHouse, las consultas y las solicitudes de análisis en tiempo real multidimensionales y multimodo pueden responderse en subsegundos.

- Consulta sin conexión

  Los datos de servicio a gran escala se importan a ClickHouse para construir una gran tabla amplia con cientos de millones a decenas de miles de millones de registros y cientos de dimensiones. Es compatible con la recopilación de estadísticas personalizadas y consultas y análisis exploratorios continuos en cualquier momento para ayudar a la toma de decisiones de negocios y proporcionar una excelente experiencia de consulta.

MRS proporciona ejemplos de proyectos de desarrollo de aplicaciones basados en ClickHouse JDBC. Esta práctica proporciona orientación para que obtenga e importe un proyecto de muestra después de crear un clúster MRS y, a continuación, realice la construcción y puesta en marcha localmente. En este proyecto de ejemplo, puede crear y eliminar tablas ClickHouse e insertar y consultar datos en el clúster MRS.

Creación de un clúster de MRS ClickHouse

1. Cree y compre un clúster MRS que contenga ClickHouse. Para obtener más información, consulte **Compra de un clúster personalizado**.

   NOTA

   En esta práctica, se utiliza como ejemplo un clúster MRS 3.2.0-LTS.1, con ClickHouse instalado y con autenticación Kerberos habilitada.

2. Haga clic en **Buy Now** y espere hasta que se cree el clúster MRS.

Preparación de un usuario de autenticación de aplicación

Para un clúster MRS con autenticación Kerberos habilitada, prepare un usuario que tenga el permiso de operación en componentes relacionados para la autenticación de aplicaciones.

El siguiente ejemplo de configuración de permisos de ClickHouse es solo para referencia. Puede modificar la configuración según lo necesite.

Paso 1 Una vez creado el clúster, inicie sesión en FusionInsight Manager.

Paso 2 Elija **System** > **Permission** > **Role** y haga clic en **Create Role** en el panel derecho.

1. Escriba un nombre de rol, por ejemplo, **developrole** y haga clic en **OK**.
2. En el cuadro de diálogo **Configure Resource Permission**, seleccione el clúster deseado, elija **ClickHouse** y **SUPER_USER_GROUP**.

Paso 3 Elija **System** > **Permission** > **User**, haga clic en **Create** en el panel derecho para crear un usuario humano-máquina, por ejemplo, **developuser** y agregue el rol **developrole**.

Una vez creado el usuario, inicie sesión en FusionInsight Manager como **developuser** y cambie la contraseña inicial según se le solicite.

----Fin

Obtención del proyecto de muestra

Paso 1 Obtenga el proyecto de muestra de Huawei Mirrors.

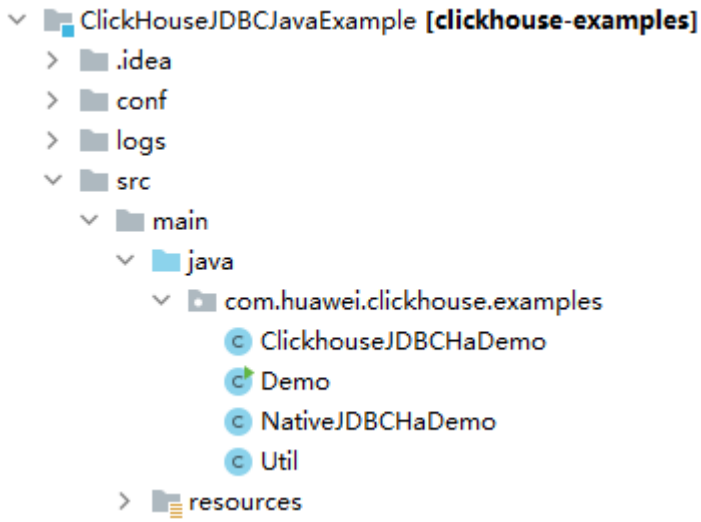
Descargue el código fuente del proyecto Maven y los archivos de configuración del proyecto de ejemplo, y configure las herramientas de desarrollo relacionadas en su PC local. Para obtener más información, consulte **Obtención de proyectos de muestra desde Huawei Mirros**.

Seleccione una rama basada en la versión del clúster y descargue el proyecto de muestra de MRS requerido.

Por ejemplo, el proyecto de muestra adecuado para esta práctica es **clickhouse-examples**, que se puede obtener en **https://github.com/huaweicloud/huaweicloud-mrs-example/tree/mrs-3.2.0.1/src/clickhouse-examples**.

Paso 2 Utilice IDEA para importar el proyecto de ejemplo y espere a que el proyecto Maven descargue los paquetes de dependencias. Para obtener más información, consulte **Configuración e importación de proyectos de muestra**.

Figura 12-20 Proyecto de muestra de ClickHouse



Después de configurar los parámetros Maven y SDK en el PC local, el proyecto de ejemplo carga automáticamente paquetes de dependencias relacionados.

Paso 3 En este proyecto de ejemplo, la aplicación se conecta al servidor ClickHouse a través de la dirección IP y la información del usuario en el archivo de configuración. Por lo tanto, después

de importar el proyecto, debe modificar el archivo **clickhouse-example.properties** en el directorio **conf** del proyecto de ejemplo basado en la información real del entorno.

```
loadBalancerIPList=192.168.64.10,192.168.64.122
sslUsed=true
loadBalancerHttpPort=21425
loadBalancerHttpsPort=21426
CLICKHOUSE_SECURITY_ENABLED=true
user=developuser
password=Bigdata_!@#
isMachineUser=false
isSupportMachineUser=false
clusterName=default_cluster
databaseName=testdb
tableName=testtb
batchRows=10000
batchNum=10
clickhouse_dataSource_ip_list=192.168.64.10:21426,192.168.64.122:21426
native_dataSource_ip_list=192.168.64.10:21424,192.168.64.122:21424
```

Tabla 12-3 Descripción de configuración

Elemento de configuración	Descripción
loadBalancerIPList	Direcciones de las instancias ClickHouseBalancer. Para ver las direcciones IP de instancia, inicie sesión en FusionInsight Manager, elija **Cluster** > **Services** > **ClickHouse**, y haga clic en **Instance**. En este ejemplo, establezca este parámetro en **192.168.64.10,192.168.64.122**.
sslUsed	Si se debe habilitar la encriptación SSL. Establezca este parámetro en **true** para clústeres en modo de seguridad.
loadBalancerHttpPort	Números de puertos HTTP y HTTPS del balanceador de carga. Inicie sesión en FusionInsight Manager y elija **Cluster** > **Services** > **ClickHouse**. Haga clic en **Logical Cluster**, busque la fila que contiene el clúster lógico deseado y vea **Port** y **Ssl Port** en la columna **HTTP Balancer Port**.
loadBalancerHttpsPort	
CLICKHOUSE_SECURITY_ENABLED	Si se debe habilitar el modo de seguridad ClickHouse. En este ejemplo, establezca este parámetro en **true**.
user	Información de autenticación del usuario en desarrollo. Para un usuario de máquina-máquina, deje el **password** vacío.
password	
isMachineUser	Si el usuario de autenticación es un usuario máquina-máquina.
isSupportMachineUser	Si se admite la autenticación de un usuario máquina-máquina. En este ejemplo, establezca este parámetro en **false**.
clusterName	Nombre del clúster lógico ClickHouse conectado a la aplicación. En este ejemplo, conserve el valor predeterminado **default_cluster**.

Elemento de configuración	Descripción
databaseName	Nombres de la base de datos y de la tabla que se van a crear en el proyecto de ejemplo. Puede cambiar los nombres según los requisitos del sitio.
tableName	
batchRows	Número de registros de datos que se escribirán en un lote. En este ejemplo, establezca este parámetro en **10**.
batchNum	Número total de lotes para escribir datos. Conservar el valor predeterminado en este ejemplo.
clickhouse_dataSource_ip_list	Direcciones y puertos HTTP de las instancias ClickHouseBalancer. Inicie sesión en FusionInsight Manager, seleccione **Cluster** > **Services** > **ClickHouse**, and click **Logical Cluster**. En este ejemplo se utiliza un clúster en modo de seguridad. Por lo tanto, localice la fila que contiene el clúster lógico deseado y vea **Ssl Port** en la columna **HTTP Balancer Port**. En este ejemplo, establezca este parámetro en **192.168.64.10:21426,192.168.64.122:21426**.
native_dataSource_ip_list	Direcciones y puertos TCP de las instancias ClickHouseBalancer. Inicie sesión en FusionInsight Manager y elija **Cluster** > **Services** > **ClickHouse**. Haga clic en **Logical Cluster**, busque la fila que contiene el clúster lógico deseado y vea **Port** en la columna **TCP Balancer Port**. En este ejemplo, establezca este parámetro en **192.168.64.10:21424,192.168.64.122:21424**.

Paso 4 Desarrolle la aplicación en este proyecto de ejemplo a través de la clickhouse-jdbc API. Para obtener detalles sobre los fragmentos de código de cada función, consulte **Código de muestra de ClickHouse**.

- Configuración de una conexión: Configure una conexión a la instancia de servicio ClickHouse.

  Durante la configuración de la conexión, la información de usuario configurada en **Tabla 12-3** se pasa como la credencial de autenticación para la autenticación de seguridad en el servidor.

  ```
  clickHouseProperties.setPassword(userPass);
  clickHouseProperties.setUser(userName);
  BalancedClickhouseDataSource balancedClickhouseDataSource = new
  BalancedClickhouseDataSource(JDBC_PREFIX + UriList, clickHouseProperties);
  ```

- Creación de una base de datos: Crear una base de datos ClickHouse.

  Ejecute la sentencia **on cluster** para crear una base de datos en el clúster.

  ```
  private void createDatabase(String databaseName, String clusterName) throws
  Exception  {
       String createDbSql = "create database if not exists " + databaseName + "
  on cluster " + clusterName;
       util.exeSql(createDbSql);
  }
  ```


- Creación de una tabla: Crear una tabla en la base de datos ClickHouse.

  Ejecute la sentencia **on cluster** para crear una tabla **ReplicatedMerge** y una tabla **Distributed** en el clúster.

```
private void createTable(String databaseName, String tableName, String
clusterName) throws Exception {
   String createSql = "create table " + databaseName + "." + tableName + " on
cluster " + clusterName  + " (name String, age UInt8, date
Date)engine=ReplicatedMergeTree('/clickhouse/tables/{shard}/" + databaseName
+ "." + tableName + "'," + "'{replica}') partition by toYYYYMM(date) order by
age";
   String createDisSql = "create table " + databaseName + "." + tableName +
"_all" + " on cluster " + clusterName + " as " + databaseName + "." +
tableName + " ENGINE = Distributed(default_cluster," + databaseName + "," +
tableName + ", rand());";    ArrayList<String> sqlList = new
ArrayList<String>();
   sqlList.add(createSql);
   sqlList.add(createDisSql);
   util.exeSql(sqlList);
}
```

- Inserción de datos: Inserte datos en la tabla ClickHouse.

  Insertar datos en la tabla creada. La tabla creada en este ejemplo tiene tres columnas: **String**, **UInt8** y **Date**.

```
String insertSql = "insert into " + databaseName + "." + tableName + " values
(?,?,?)";
PreparedStatement preparedStatement = connection.prepareStatement(insertSql);
long allBatchBegin = System.currentTimeMillis();
for (int j = 0; j < batchNum; j++) {
    for (int i = 0; i < batchRows; i++) {
       preparedStatement.setString(1, "huawei_" + (i + j * 10));
       preparedStatement.setInt(2, ((int) (Math.random() * 100)));
       preparedStatement.setDate(3, generateRandomDate("2018-01-01",
"2021-12-31"));
       preparedStatement.addBatch();
    }
   long begin = System.currentTimeMillis();
   preparedStatement.executeBatch();
   long end = System.currentTimeMillis();
   log.info("Inert batch time is {} ms", end - begin);
}
long allBatchEnd = System.currentTimeMillis();
log.info("Inert all batch time is {} ms", allBatchEnd - allBatchBegin);
```

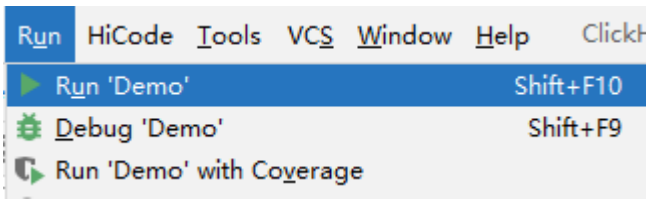
----Fin

Creación y ejecución de la aplicación

Si puede acceder al clúster MRS desde su PC local, puede poner en marcha y ejecutar la aplicación localmente.

Paso 1 En el proyecto **clickhouse-examples** de IntelliJ IDEA, haga clic en **Run 'Demo'** para ejecutar el proyecto de aplicación.

Figura 12-21 Ejecución de la aplicación de ClickHouse Demo



Paso 2 Vea la salida en la consola, como se muestra en la siguiente figura. Puede ver que la tabla ClickHouse se crea y que los datos se insertan correctamente.

```
...
2023-06-03 11:30:27,127 | INFO  | main | Execute query:create table testdb.testtb
on cluster default_cluster (name String, age UInt8, date
Date)engine=ReplicatedMergeTree('/clickhouse/tables/{shard}/
testdb.testtb','{replica}') partition by toYYYYMM(date) order by age |
com.huawei.clickhouse.examples.Util.exeSql(Util.java:68)
2023-06-03 11:30:27,412 | INFO  | main | Execute time is 284 ms |
com.huawei.clickhouse.examples.Util.exeSql(Util.java:72)
2023-06-03 11:30:27,412 | INFO  | main | Current load balancer is
192.168.64.10:21426 | com.huawei.clickhouse.examples.Util.exeSql(Util.java:63)
2023-06-03 11:30:28,426 | INFO  | main | Execute query:create table
testdb.testtb_all on cluster default_cluster as testdb.testtb ENGINE =
Distributed(default_cluster,testdb,testtb, rand()); |
com.huawei.clickhouse.examples.Util.exeSql(Util.java:68)
2023-06-03 11:30:28,686 | INFO  | main | Execute time is 259 ms |
com.huawei.clickhouse.examples.Util.exeSql(Util.java:72)
2023-06-03 11:30:28,686 | INFO  | main | Current load balancer is
192.168.64.10:21426 | com.huawei.clickhouse.examples.Util.insertData(Util.java:
137)
2023-06-03 11:30:29,784 | INFO  | main | Insert batch time is 227 ms |
com.huawei.clickhouse.examples.Util.insertData(Util.java:154)
2023-06-03 11:30:31,490 | INFO  | main | Insert batch time is 200 ms |
com.huawei.clickhouse.examples.Util.insertData(Util.java:154)
2023-06-03 11:30:33,337 | INFO  | main | Insert batch time is 335 ms |
com.huawei.clickhouse.examples.Util.insertData(Util.java:154)
2023-06-03 11:30:35,295 | INFO  | main | Insert batch time is 454 ms |
com.huawei.clickhouse.examples.Util.insertData(Util.java:154)
2023-06-03 11:30:37,077 | INFO  | main | Insert batch time is 275 ms |
com.huawei.clickhouse.examples.Util.insertData(Util.java:154)
2023-06-03 11:30:38,811 | INFO  | main | Insert batch time is 218 ms |
com.huawei.clickhouse.examples.Util.insertData(Util.java:154)
2023-06-03 11:30:40,468 | INFO  | main | Insert batch time is 144 ms |
com.huawei.clickhouse.examples.Util.insertData(Util.java:154)
2023-06-03 11:30:42,216 | INFO  | main | Insert batch time is 238 ms |
com.huawei.clickhouse.examples.Util.insertData(Util.java:154)
2023-06-03 11:30:43,977 | INFO  | main | Insert batch time is 257 ms |
com.huawei.clickhouse.examples.Util.insertData(Util.java:154)
2023-06-03 11:30:45,756 | INFO  | main | Insert batch time is 277 ms |
com.huawei.clickhouse.examples.Util.insertData(Util.java:154)
2023-06-03 11:30:47,270 | INFO  | main | Inert all batch time is 17720 ms |
com.huawei.clickhouse.examples.Util.insertData(Util.java:158)
2023-06-03 11:30:47,271 | INFO  | main | Current load balancer is
192.168.64.10:21426 | com.huawei.clickhouse.examples.Util.exeSql(Util.java:63)
2023-06-03 11:30:47,828 | INFO  | main | Execute query:select * from
testdb.testtb_all order by age limit 10 |
com.huawei.clickhouse.examples.Util.exeSql(Util.java:68)
2023-06-03 11:30:47,917 | INFO  | main | Execute time is 89 ms |
com.huawei.clickhouse.examples.Util.exeSql(Util.java:72)
2023-06-03 11:30:47,918 | INFO  | main | Current load balancer is
192.168.64.10:21426 | com.huawei.clickhouse.examples.Util.exeSql(Util.java:63)
2023-06-03 11:30:48,580 | INFO  | main | Execute query:select
toYYYYMM(date),count(1) from testdb.testtb_all group by toYYYYMM(date) order by
count(1) DESC limit 10 | com.huawei.clickhouse.examples.Util.exeSql(Util.java:68)
2023-06-03 11:30:48,680 | INFO  | main | Execute time is 99 ms |
com.huawei.clickhouse.examples.Util.exeSql(Util.java:72)
2023-06-03 11:30:48,682 | INFO  | main | name    age    date     |
com.huawei.clickhouse.examples.Demo.queryData(Demo.java:159)
2023-06-03 11:30:48,682 | INFO  | main | huawei_89    3    2021-02-21     |
com.huawei.clickhouse.examples.Demo.queryData(Demo.java:159)
2023-06-03 11:30:48,682 | INFO  | main | huawei_81    3    2020-05-27     |
com.huawei.clickhouse.examples.Demo.queryData(Demo.java:159)
2023-06-03 11:30:48,682 | INFO  | main | huawei_70    4    2021-10-28     |
com.huawei.clickhouse.examples.Demo.queryData(Demo.java:159)
2023-06-03 11:30:48,682 | INFO  | main | huawei_73    4    2020-03-23     |
com.huawei.clickhouse.examples.Demo.queryData(Demo.java:159)
2023-06-03 11:30:48,683 | INFO  | main | huawei_44    5    2020-12-10     |
com.huawei.clickhouse.examples.Demo.queryData(Demo.java:159)
```



```
2023-06-03 11:30:48,683 | INFO  | main | huawei_29    6    2021-10-12     |
com.huawei.clickhouse.examples.Demo.queryData(Demo.java:159)
2023-06-03 11:30:48,683 | INFO  | main | huawei_74    6    2021-03-03     |
com.huawei.clickhouse.examples.Demo.queryData(Demo.java:159)
2023-06-03 11:30:48,683 | INFO  | main | huawei_38    7    2020-05-30     |
com.huawei.clickhouse.examples.Demo.queryData(Demo.java:159)
2023-06-03 11:30:48,683 | INFO  | main | huawei_57    8    2020-09-27     |
com.huawei.clickhouse.examples.Demo.queryData(Demo.java:159)
2023-06-03 11:30:48,683 | INFO  | main | huawei_23    8    2020-08-08     |
com.huawei.clickhouse.examples.Demo.queryData(Demo.java:159)
2023-06-03 11:30:48,683 | INFO  | main | toYYYYMM(date)    count()     |
com.huawei.clickhouse.examples.Demo.queryData(Demo.java:159)
2023-06-03 11:30:48,684 | INFO  | main | 202005    8     |
com.huawei.clickhouse.examples.Demo.queryData(Demo.java:159)
2023-06-03 11:30:48,684 | INFO  | main | 202007    7     |
com.huawei.clickhouse.examples.Demo.queryData(Demo.java:159)
2023-06-03 11:30:48,684 | INFO  | main | 202004    6     |
com.huawei.clickhouse.examples.Demo.queryData(Demo.java:159)
2023-06-03 11:30:48,684 | INFO  | main | 202009    6     |
com.huawei.clickhouse.examples.Demo.queryData(Demo.java:159)
2023-06-03 11:30:48,684 | INFO  | main | 202103    6     |
com.huawei.clickhouse.examples.Demo.queryData(Demo.java:159)
2023-06-03 11:30:48,685 | INFO  | main | 202012    6     |
com.huawei.clickhouse.examples.Demo.queryData(Demo.java:159)
2023-06-03 11:30:48,685 | INFO  | main | 202010    5     |
com.huawei.clickhouse.examples.Demo.queryData(Demo.java:159)
2023-06-03 11:30:48,685 | INFO  | main | 202112    5     |
com.huawei.clickhouse.examples.Demo.queryData(Demo.java:159)
2023-06-03 11:30:48,685 | INFO  | main | 202003    5     |
com.huawei.clickhouse.examples.Demo.queryData(Demo.java:159)
2023-06-03 11:30:48,685 | INFO  | main | 202104    4     |
com.huawei.clickhouse.examples.Demo.queryData(Demo.java:159)
2023-06-03 11:30:48,689 | INFO  | main | Use HA module. |
ru.yandex.clickhouse.BalancedClickhouseDataSource.<init>(BalancedClickhouseDataSou
rce.java:122)
2023-06-03 11:30:51,651 | INFO  | main | Name is: huawei_89, age is: 3 |
com.huawei.clickhouse.examples.ClickhouseJDBCHaDemo.queryData(ClickhouseJDBCHaDemo
.java:73)
2023-06-03 11:30:51,651 | INFO  | main | Name is: huawei_81, age is: 3 |
com.huawei.clickhouse.examples.ClickhouseJDBCHaDemo.queryData(ClickhouseJDBCHaDemo
.java:73)
2023-06-03 11:30:51,651 | INFO  | main | Name is: huawei_70, age is: 4 |
com.huawei.clickhouse.examples.ClickhouseJDBCHaDemo.queryData(ClickhouseJDBCHaDemo
.java:73)
2023-06-03 11:30:51,651 | INFO  | main | Name is: huawei_73, age is: 4 |
com.huawei.clickhouse.examples.ClickhouseJDBCHaDemo.queryData(ClickhouseJDBCHaDemo
.java:73)
2023-06-03 11:30:51,652 | INFO  | main | Name is: huawei_44, age is: 5 |
com.huawei.clickhouse.examples.ClickhouseJDBCHaDemo.queryData(ClickhouseJDBCHaDemo
.java:73)
2023-06-03 11:30:51,652 | INFO  | main | Name is: huawei_29, age is: 6 |
com.huawei.clickhouse.examples.ClickhouseJDBCHaDemo.queryData(ClickhouseJDBCHaDemo
.java:73)
2023-06-03 11:30:51,652 | INFO  | main | Name is: huawei_74, age is: 6 |
com.huawei.clickhouse.examples.ClickhouseJDBCHaDemo.queryData(ClickhouseJDBCHaDemo
.java:73)
2023-06-03 11:30:51,652 | INFO  | main | Name is: huawei_38, age is: 7 |
com.huawei.clickhouse.examples.ClickhouseJDBCHaDemo.queryData(ClickhouseJDBCHaDemo
.java:73)
2023-06-03 11:30:51,654 | INFO  | main | Name is: huawei_57, age is: 8 |
com.huawei.clickhouse.examples.ClickhouseJDBCHaDemo.queryData(ClickhouseJDBCHaDemo
.java:73)
2023-06-03 11:30:51,654 | INFO  | main | Name is: huawei_23, age is: 8 |
com.huawei.clickhouse.examples.ClickhouseJDBCHaDemo.queryData(ClickhouseJDBCHaDemo
.java:73)
...
```

Paso 3 Instale el cliente de clúster MRS e inicie sesión en el cliente ClickHouse.

Por ejemplo, si el directorio de instalación del cliente es **/opt/client**, inicie sesión en el nodo donde está instalado el cliente como usuario de instalación del cliente.

cd /opt/client

source bigdata_env

kinit developuser

Paso 4 Ejecute el siguiente comando para conectarse al servidor ClickHouse:

clickhouse client --host *IP address of the ClickHouseServer instance* **--port** *Connection port* **--secure**

Para obtener la dirección IP de la instancia ClickHouse, inicie sesión en FusionInsight Manager, seleccione **Cluster** > **Services** > **ClickHouse** y haga clic en la pestaña **Instance**. Puede obtener el puerto de conexión buscando el parámetro **tcp_port_secure** en las configuraciones de servicio ClickHouse.

Por ejemplo, ejecute el siguiente comando:

clickhouse client --host 192.168.64.10 --port 21427 --secure

Paso 5 Ejecute el siguiente comando para ver el contenido de la tabla creado por la aplicación:

select * from testdb.testtb;

```
┌─name─────┬─age─┬─────date─┐
│ huawei_70 │   4 │ 2021-10-28 │
│ huawei_29 │   6 │ 2021-10-12 │
│ huawei_16 │  28 │ 2021-10-04 │
│ huawei_15 │  29 │ 2021-10-03 │
└──────────┴─────┴────────────┘
...
```

----Fin

12.8 Desarrollo de aplicaciones de Spark

Spark es un marco de procesamiento por lotes distribuido. Proporciona capacidades de análisis y minería y computación de memoria iterativa y soporta el desarrollo de aplicaciones en múltiples lenguajes de programación. Se aplica a los siguientes escenarios:

- Procesamiento de datos: Spark puede procesar datos rápidamente y tiene tolerancia a fallos y escalabilidad.
- Cálculo iterativo: Spark admite el cálculo iterativo para mantenerse al día con la lógica de procesamiento de datos de varios pasos.
- Minería de datos: Basada en datos masivos, Spark puede manejar la minería y el análisis de datos complejos y admite múltiples algoritmos de minería de datos y aprendizaje automático.
- Procesamiento de streaming: Spark admite el procesamiento de streaming con solo una latencia de nivel de segundos y admite múltiples fuentes de datos externas.
- Análisis de consultas: Spark admite el análisis de consultas SQL estándar, proporciona el DSL (DataFrame) y admite múltiples entradas externas.

MRS proporciona ejemplos de proyectos de desarrollo de aplicaciones basados en Spark. Esta práctica proporciona orientación para que obtenga e importe un proyecto de muestra después de crear un clúster MRS y, a continuación, realice la construcción y puesta en marcha

localmente. En este proyecto de ejemplo, puede leer datos de tablas Hive y volver a escribir los datos en tablas HBase.

Las directrices para el proyecto de muestra en esta práctica son las siguientes:

1. Consultar datos en una tabla Hive especificada.
2. Consultar datos en una tabla HBase especificada basándose en la clave de los datos en la tabla Hive.
3. Agregar registros de datos relacionados y escribirlos en la tabla HBase.

Creación de un clúster de MRS

Paso 1 Cree y compre un clúster MRS que contenga Spark. Para obtener más información, consulte **Compra de un clúster personalizado**.

NOTA

En esta práctica, se utiliza como ejemplo un clúster MRS 3.1.5, con Spark2x, Hive y HBase instalados y con la autenticación Kerberos habilitada.

Paso 2 Después de comprar el clúster, instale el cliente en cualquier nodo del clúster. Para obtener más información, consulte **Instalación y uso de cliente de clúster**.

Por ejemplo, instale el cliente en el directorio **/opt/client** en el nodo de gestión activo.

----Fin

Preparación del archivo de configuración de clúster

Paso 1 Una vez creado el clúster, inicie sesión en FusionInsight Manager y cree un usuario del clúster para enviar trabajos de Flink.

Elija **System** > **Permission** > **User**. En el panel derecho, haga clic en **Create**. En la página mostrada, cree un usuario hombre-máquina, por ejemplo, **sparkuser**.

Agregue el grupo de usuarios **supergroup** y asocie el rol **System_administrator**.

Paso 2 Inicie sesión en FusionInsight Manager como nuevo usuario y cambie la contraseña inicial según se le solicite.

Paso 3 Elija **System** > **Permission** > **User**. En la columna **Operation** de **sparkuser**, elija **More** > **Download Authentication Credential**. Guarde el archivo y descomprímalo para obtener los archivos **user.keytab** y **krb5.conf** del usuario.

----Fin

Desarrollo de la aplicación

Paso 1 Obtenga el proyecto de muestra de Huawei Mirrors.

Descargue el código fuente del proyecto Maven y los archivos de configuración del proyecto de ejemplo, y configure las herramientas de desarrollo relacionadas en su PC local. Para obtener más información, consulte **Obtención de proyectos de muestra desde Huawei Mirros**.

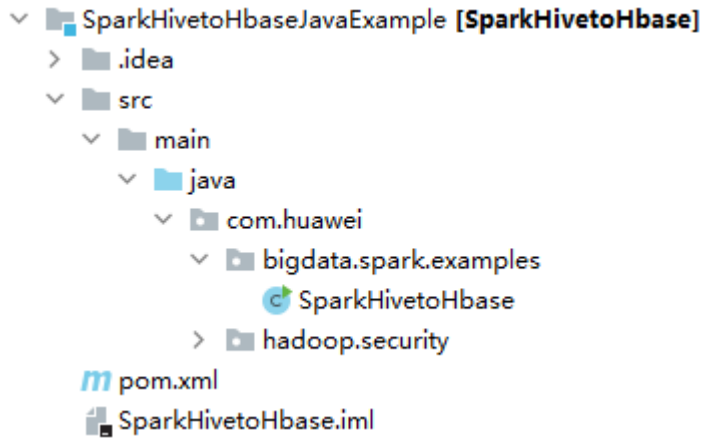
Seleccione una rama basada en la versión del clúster y descargue el proyecto de muestra de MRS requerido.

Por ejemplo, el proyecto de muestra adecuado para esta práctica es **SparkHivetoHbase**, que se puede obtener en **https://github.com/huaweicloud/huaweicloud-mrs-example/tree/mrs-3.1.5/src/spark-examples/sparksecurity-examples/SparkHivetoHbaseJavaExample**.

Paso 2 Utilice IDEA para importar el proyecto de ejemplo y espere a que el proyecto Maven descargue los paquetes de dependencias.

Después de configurar los parámetros Maven y SDK en el PC local, el proyecto de ejemplo carga automáticamente paquetes de dependencias relacionados. Para obtener más información, consulte **Configuración e importación de proyectos de muestra**.

Figura 12-22 Proyecto de muestra Spark Hive a HBase



La clase **SparkHivetoHbase** del proyecto de ejemplo utiliza Spark para invocar a las API de Hive para operar una tabla Hive, obtener el registro correspondiente de una tabla HBase basada en la clave, realizar operaciones en los dos registros de datos y actualizar los datos a la tabla HBase.

El fragmento de código es el siguiente:

```
...
public class SparkHivetoHbase {
    public static void main(String[] args) throws Exception {
        String userPrincipal = "sparkuser";     //Specifies the cluster user
information and keytab file address used for authentication.
        String userKeytabPath = "/opt/client/user.keytab";
        String krb5ConfPath = "/opt/client/krb5.conf";
        Configuration hadoopConf = new Configuration();
        LoginUtil.login(userPrincipal, userKeytabPath, krb5ConfPath, hadoopConf);
        //Calls the Spark API to obtain table data.
        SparkConf conf = new SparkConf().setAppName("SparkHivetoHbase");
        JavaSparkContext jsc = new JavaSparkContext(conf);
        HiveContext sqlContext = new org.apache.spark.sql.hive.HiveContext(jsc);
        Dataset<Row> dataFrame = sqlContext.sql("select name, account from
person");
        //Traverses partitions in the Hive table and updates the partitions to
the HBase table.
                dataFrame
                .toJavaRDD()
                .foreachPartition(
                        new VoidFunction<Iterator<Row>>() {
                            public void call(Iterator<Row> iterator) throws
Exception {
                                hBaseWriter(iterator);
                            }
                        });
        jsc.stop();
```



```
    }
    //Updates records in the HBase table on the executor.
    private static void hBaseWriter(Iterator<Row> iterator) throws IOException {
        //Reads the HBase table.
        String tableName = "table2";
        String columnFamily = "cf";
        Configuration conf = HBaseConfiguration.create();
        Connection connection = ConnectionFactory.createConnection(conf);
        Table table = connection.getTable(TableName.valueOf(tableName));
        try {
            connection = ConnectionFactory.createConnection(conf);
            table = connection.getTable(TableName.valueOf(tableName));
            List<Row> table1List = new ArrayList<Row>();
            List<Get> rowList = new ArrayList<Get>();
            while (iterator.hasNext()) {
                Row item = iterator.next();
                Get get = new Get(item.getString(0).getBytes());
                table1List.add(item);
                rowList.add(get);
            }
            //Obtains the records in the HBase table.
            Result[] resultDataBuffer = table.get(rowList);
            //Modifies records in the HBase table.
            List<Put> putList = new ArrayList<Put>();
            for (int i = 0; i < resultDataBuffer.length; i++) {
                Result resultData = resultDataBuffer[i];
                if (!resultData.isEmpty()) {
                    int hiveValue = table1List.get(i).getInt(1);
                    String hbaseValue =
Bytes.toString(resultData.getValue(columnFamily.getBytes(), "cid".getBytes()));
                    Put put = new Put(table1List.get(i).getString(0).getBytes());
                    //Calculates the result.
                    int resultValue = hiveValue + Integer.valueOf(hbaseValue);
                    put.addColumn(
                            Bytes.toBytes(columnFamily),
                            Bytes.toBytes("cid"),
                            Bytes.toBytes(String.valueOf(resultValue)));
                    putList.add(put);
                }
            }
            if (putList.size() > 0) {
                table.put(putList);
            }
        } catch (IOException e) {
            e.printStackTrace();
        } finally {
            if (table != null) {
                try {
                    table.close();
                } catch (IOException e) {
                    e.printStackTrace();
                }
            }
            if (connection != null) {
                try {
                    //Closes the HBase connection.
                    connection.close();
                } catch (IOException e) {
                    e.printStackTrace();
                }
            }
        }
    }
...
```


> NOTA
>
> Para un clúster MRS con autenticación Kerberos activada, la aplicación debe realizar la autenticación de usuario en el servidor. En este proyecto de ejemplo, configure la información de autenticación en código. Establezca **userPrincipal** en el nombre de usuario para la autenticación y cambie **userKeytabPath** y **krb5ConfPath** a las rutas de archivos reales en el servidor de cliente.

Paso 3 Después de confirmar que los parámetros en el proyecto son correctos, compilar el proyecto y empaquetarlo en un archivo JAR.

En la ventana Maven, seleccione **clean** en **Lifecycle** para ejecutar el proceso de construcción de Maven. Seleccione **package** y obtenga el paquete JAR del directorio **target**.

```
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time:  02:36 min
[INFO] Finished at: 2023-06-12T20:46:24+08:00
[INFO] ------------------------------------------------------------------------
```

Por ejemplo, el archivo JAR es **SparkHivetoHbase-1.0.jar**.

----Fin

Carga del paquete JAR y preparación de datos de origen

Paso 1 Cargue el paquete JAR a un directorio, por ejemplo, **/opt/client/sparkdemo** en el nodo cliente.

> NOTA
>
> Si no puede acceder directamente al nodo cliente para cargar archivos a través de la red local, cargue el paquete JAR o los datos de origen a OBS, impórtelos a HDFS en la pestaña **Files** de la consola MRS. Y ejecute el comando **hdfs dfs -get** en el cliente HDFS para descargarlo al nodo cliente.

Paso 2 Cargue el archivo keytab usado para la autenticación a la ubicación especificada en el código, por ejemplo, **/opt/client**.

Paso 3 Inicie sesión en el nodo donde está instalado el cliente de clúster como usuario **root**.

cd /opt/client

source bigdata_env

kinit sparkuser

Paso 4 Cree una tabla Hive e inserte datos en la tabla.

beeline

En Hive Beeline, ejecute los siguientes comandos para crear una tabla e insertar datos:

create table person (name STRING, account INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' ESCAPED BY '\\' STORED AS TEXTFILE;

insert into table person(name,account) values("1","100");

select * from person;

```
+--------------+-----------------+
| person.name  | person.account  |
+--------------+-----------------+
| 1            | 100             |
+--------------+-----------------+
```


Paso 5 Cree una tabla HBase e inserte datos en la tabla.

Salga de Hive Beeline, ejecute el comando **spark-beeline** y ejecute el siguiente comando para crear una tabla HBase:

create table table2 (key string, cid string) using org.apache.spark.sql.hbase.HBaseSource options(hbaseTableName "table2", keyCols "key", colsMapping "cid=cf.cid");

Salga de Spark Beeline, ejecute el comando **hbase shell** para ir al HBase Shell y ejecute los siguientes comandos para insertar datos:

put 'table2', '1', 'cf:cid', '1000'

scan 'table2'

```
ROW                                                    COLUMN
+CELL

 1                                                     column=cf:cid,
timestamp=2023-06-12T21:12:50.711,
value=1000

1 row(s)
```

----Fin

Ejecución de la aplicación y visualización del resultado

Paso 1 En el nodo donde está instalado el cliente de clúster, ejecute los siguientes comandos para ejecutar el paquete JAR exportado desde el proyecto de ejemplo:

cd /opt/client

source bigdata_env

cd Spark2x/spark

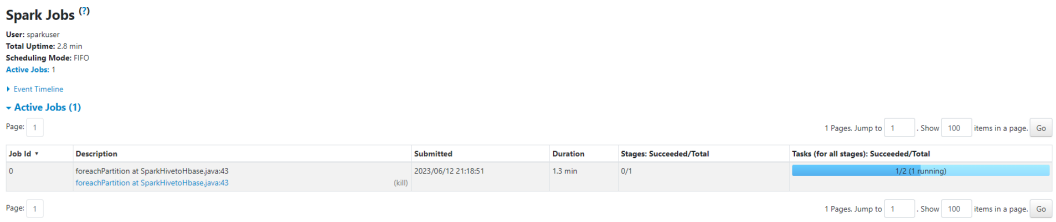
vi conf/spark-defaults.conf

Cambie el valor de **spark.yarn.security.credentials.hbase.enabled** a **true**.

bin/spark-submit --class com.huawei.bigdata.spark.examples.SparkHivetoHbase --master yarn --deploy-mode client /opt/client/sparkdemo/SparkHivetoHbase-1.0.jar

Paso 2 Una vez enviada la tarea, inicie sesión en FusionInsight Manager como usuario **sparkuser**, elija **Cluster** > **Services** > **Yarn** y vincule a la interfaz de usuario web ResourceManager. A continuación, localice la información del trabajo de la aplicación de Spark y haga clic en **ApplicationMaster** en la última columna de la información de la aplicación para ir a la interfaz de usuario de Spark y ver detalles.

Figura 12-23 Consulta de los detalles de la tarea Spark



Paso 3 Una vez completada la tarea, consulte el contenido de la tabla HBase en el HBase shell. Puede ver que los registros han sido actualizados.

cd /opt/client

source bigdata_env

hbase shell

scan 'table2'

```
ROW                                                    COLUMN
+CELL

 1                                                     column=cf:cid,
timestamp=2023-06-12T21:22:50.711,
value=1100

1 row(s)
```

----Fin

13 Prácticas

Después de desplegar un clúster MRS, puede probar algunas prácticas proporcionadas por MRS para satisfacer sus requisitos de servicio.

Tabla 13-1 Prácticas recomendadas

Práctica		Descripción
Análisis de datos	**Uso de Spark2x para analizar el comportamiento de conducción de los conductores de IoV**	Esta práctica describe cómo usar Spark para analizar el comportamiento de conducción. Puede familiarizarse con las funciones básicas de MRS utilizando el componente Spark2x para analizar y recopilar estadísticas sobre el comportamiento de conducción, obtener el resultado del análisis y recopilar estadísticas sobre el número de violaciones, tales como aceleración y desaceleración repentinas, inercia, exceso de velocidad, y la fatiga de conducir en un período determinado.
	Uso de Hive para cargar datos HDFS y analizar las puntuaciones del libro	Esta práctica describe cómo usar Hive para importar y analizar datos sin procesar y cómo crear análisis de big data fuera de línea elásticos y asequibles. En esta práctica, la lectura de comentarios del fondo de un sitio web de un libro se utiliza como datos brutos. Después de importar los datos a una tabla Hive, puede ejecutar comandos SQL para consultar los libros más vendidos más populares.
	Uso de Hive para cargar datos OBS y analizar información de empleados empresariales	Esta práctica describe cómo usar Hive para importar y analizar datos sin procesar de OBS y cómo crear análisis de big data elásticos y asequibles basados en recursos de procesamiento y almacenamiento desacoplados. Esta práctica describe cómo desarrollar una aplicación de análisis de datos de Hive y cómo ejecutar sentencias HQL para acceder a los datos de Hive almacenados en OBS después de conectarse a Hive a través del cliente. Por ejemplo, gestionar y consultar la información de los empleados de la empresa.

Práctica		Descripción
	Uso de trabajos de Flink para procesar datos de OBS	Esta práctica describe cómo utilizar el programa WordCount Flink integrado de un clúster MRS para analizar los datos de origen almacenados en el sistema de archivos OBS y calcular el número de ocurrencias de palabras especificadas en el origen de datos. MRS admite almacenamiento y cómputo desacoplados en escenarios en los que se requiere una gran capacidad de almacenamiento y los recursos de cómputo deben escalarse según la demanda. Esto le permite almacenar sus datos en OBS y usar un clúster MRS solo para computación de datos.
Migración de datos	**Solución de migración de datos**	Esta práctica describe cómo migrar datos HDFS, HBase y Hive a un clúster MRS en diferentes escenarios. Intentará prepararse para la migración de datos, exportar metadatos, copiar datos y restaurar datos.
	Migración de datos de Hadoop a MRS	En esta práctica, CDM se utiliza para migrar datos (decenas de terabytes o menos) de clústeres Hadoop a MRS.
	Migración de datos de HBase a MRS	En esta práctica, CDM se utiliza para migrar datos (decenas de terabytes o menos) de clústeres HBase a MRS. HBase almacena datos en HDFS, incluidos los archivos HFile y WAL. El elemento de configuración **hbase.rootdir** especifica la ruta de acceso de HDFS. De forma predeterminada, los datos se almacenan en la carpeta **/hbase** en MRS. Algunos mecanismos y comandos de herramientas de HBase también se pueden utilizar para migrar datos. Por ejemplo, puede migrar datos exportando instantáneas, exportando e importando datos y CopyTable.
	Migración de datos de Hive a MRS	En esta práctica, CDM se utiliza para migrar datos (decenas de terabytes o menos) de clústeres Hive a MRS. La migración de datos de Hive consta de dos partes: ● Metadatos de Hive, que se almacenan en las bases de datos como MySQL. De forma predeterminada, los metadatos del clúster de Hive de MRS se almacenan en MRS DBService (base de datos de GaussDB de Huawei). También puede utilizar RDS for MySQL como base de datos de metadatos externa. ● Datos de servicio Hive, que se almacenan en HDFS u OBS

Práctica		Descripción
	Migración de datos de MySQL a una tabla particionada de subárbol MRS	Esta práctica demuestra cómo usar CDM para importar datos MySQL a la tabla de particiones de Hive en un clúster MRS. Hive admite SQL para ayudarle a realizar operaciones de extracción, transformación y carga (ETL) en conjuntos de datos a gran escala. Las consultas en conjuntos de datos a gran escala tardan mucho tiempo. En muchos escenarios, puede crear particiones Hive para reducir la cantidad total de datos que se analizarán cada vez. Esto mejora significativamente el rendimiento de las consultas.
	Migración de datos de MRS HDFS a OBS	Esta práctica demuestra cómo migrar datos de archivos desde MRS HDFS a OBS usando CDM.
Interconexión del sistema	**Uso de DBeaver para acceder a Phoenix**	Esta práctica describe cómo usar DBeaver para acceder a Phoenix.
	Uso de DBeaver para acceder a HetuEngine	Esta práctica describe cómo usar DBeaver para acceder a HetuEngine.
	Interconexión de Hive con bases de datos relacionales autoconstruidas externas	Esta práctica describe cómo usar Hive para conectarse a bases de datos MySQL y Postgres de código abierto. Después de desplegar una base de datos de metadatos externa en un clúster que tiene datos de Hive, las tablas de metadatos originales no se sincronizarán automáticamente. Antes de instalar Hive, determine si desea almacenar metadatos en una base de datos externa o DBService. Para el primero, despliegue una base de datos externa al instalar Hive o cuando no hay datos de Hive. Después de la instalación de Hive, no se puede cambiar la ubicación de almacenamiento de metadatos. De lo contrario, se perderán los metadatos originales.
	Interconexión de Hive con CSS	Esta práctica describe cómo usar Hive para interconectarse con CSS Elasticsearch. En esta práctica, utilizará el complemento Elasticsearch-Hadoop para intercambiar datos entre Hive y Elasticsearch of Cloud Search Service (CSS) para que los datos del índice de Elasticsearch puedan asignarse a las tablas de Hive.

