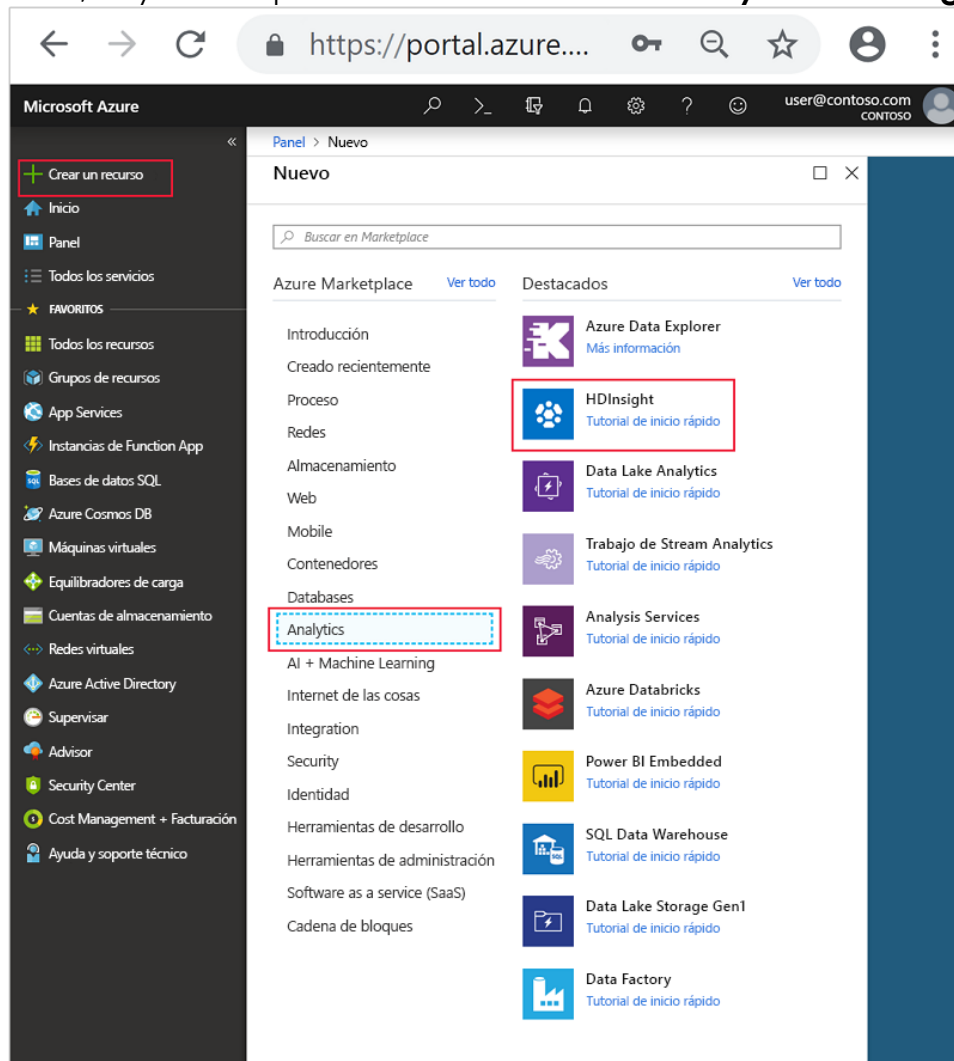


LAB 02 – Creación de un clúster de Apache Hadoop en Azure HDInsight

Aprenderá a crear clústeres de Apache Hadoop en HDInsight con Azure Portal y a ejecutar trabajos de Apache Hive en HDInsight. La mayoría de los trabajos de Hadoop son por lotes. Se crea un clúster, se ejecutan algunos trabajos y luego se elimina el clúster.

Paso 1: Creación de un clúster de Apache Hadoop

1. Inicie sesión en [Azure Portal](#).
2. En Azure Portal, vaya a la opción **Crear un recurso** > **Analytics** > **HDInsight**.



3. En **HDInsight > Creación rápida > Aspectos básicos**, escriba o seleccione los valores siguientes:

Propiedad	DESCRIPCIÓN
Nombre del clúster	Escriba el nombre del clúster de Hadoop. Dado que todos los clústeres de HDInsight comparten el mismo espacio de nombres de DNS, este nombre debe ser único. El nombre puede tener un máximo de 59 caracteres, letras, números y guiones incluidos. El primer y el último carácter del nombre no pueden ser guiones.
Subscription	Seleccione su suscripción a Azure.
Tipo de clúster	Omítalo por ahora. Proporcione esta entrada en el paso siguiente de este procedimiento.
Nombre de usuario y contraseña de inicio de sesión del clúster	El nombre de inicio de sesión predeterminado es admin . La contraseña debe tener un mínimo de 10 caracteres y contener al menos un dígito, una letra mayúscula y una letra minúscula, y un carácter no alfanumérico (excepto los caracteres ' " y `). Asegúrese de no proporcionar contraseñas comunes, como "Pass@word1".
Nombre de usuario de Secure Shell (SSH)	El nombre de usuario predeterminado es sshuser . Puede proporcionar otro nombre para el nombre de usuario de SSH.
Uso de la contraseña de inicio de sesión del clúster para SSH	Seleccione esta casilla para que el usuario de SSH tenga la misma contraseña que la proporcionada para el usuario de inicio de sesión del clúster.
Resource group	Cree un grupo de recursos o seleccione uno existente. Un grupo de recursos es un contenedor de componentes de Azure. En este caso, el grupo de recursos contiene el clúster de HDInsight y la cuenta de Azure Storage dependiente.
Location	Seleccione una ubicación de Azure en la que quiera crear el clúster. Elija una ubicación más cercana a usted para mejorar el rendimiento.



HDInsight

by Microsoft

Quick create

Custom (size, settings, apps)

1

Basics

Configure basic settings

>

2

Storage

Set storage settings

>

3

Summary

Confirm configurations

>

This cluster may take up to 20 minutes to create.

Basics

* Cluster name

myhadoopcluster

.azurehdinsight.net

* Subscription

mySubscription

* Cluster type

Configure required settings

* Cluster login username

admin

* Cluster login password

.....

Secure Shell (SSH) username

sshuser

☒

Use cluster login password for SSH

* Resource group

Select existing...

Create new

* Location

East US

Click here to view cores usage.

Next

4. Seleccione **Tipo de clúster** para abrir la página **Configuración de clúster** e indique luego los valores siguientes:

Propiedad	DESCRIPCIÓN
Tipo de clúster	Seleccione Hadoop
Versión	Seleccione Hadoop 2.7.3 (HDI 3.6) .



Cluster configuration

Learn about HDInsight and cluster versions. →

Cluster configuration

* Cluster type ⓘ

Hadoop

* Operating system

Linux

* Version

Hadoop 2.7.3 (HDI 3...

Hadoop : Petabyte-scale processing with Hadoop components like MapReduce, Hive (SQL on Hadoop), Pig, Sqoop and Oozie.

Features

* denotes preview feature

+ Apache Ranger

+ Enterprise Security Package

+ HDInsight IO Cache

+ Secure shell (SSH) access

+ HDInsight applications

+ Custom virtual network

+ Custom Hive metastore

+ Custom Oozie metastore

+ Data Lake Storage Gen1 access

+ Data Lake Storage Gen2 access

+ Data Lake Storage Gen1 as primary data storage

Select

Elija **Seleccionar** y, a continuación, seleccione **Siguiente** para avanzar a la configuración de almacenamiento.

5. En la pestaña **Almacenamiento**, proporcione los valores siguientes:

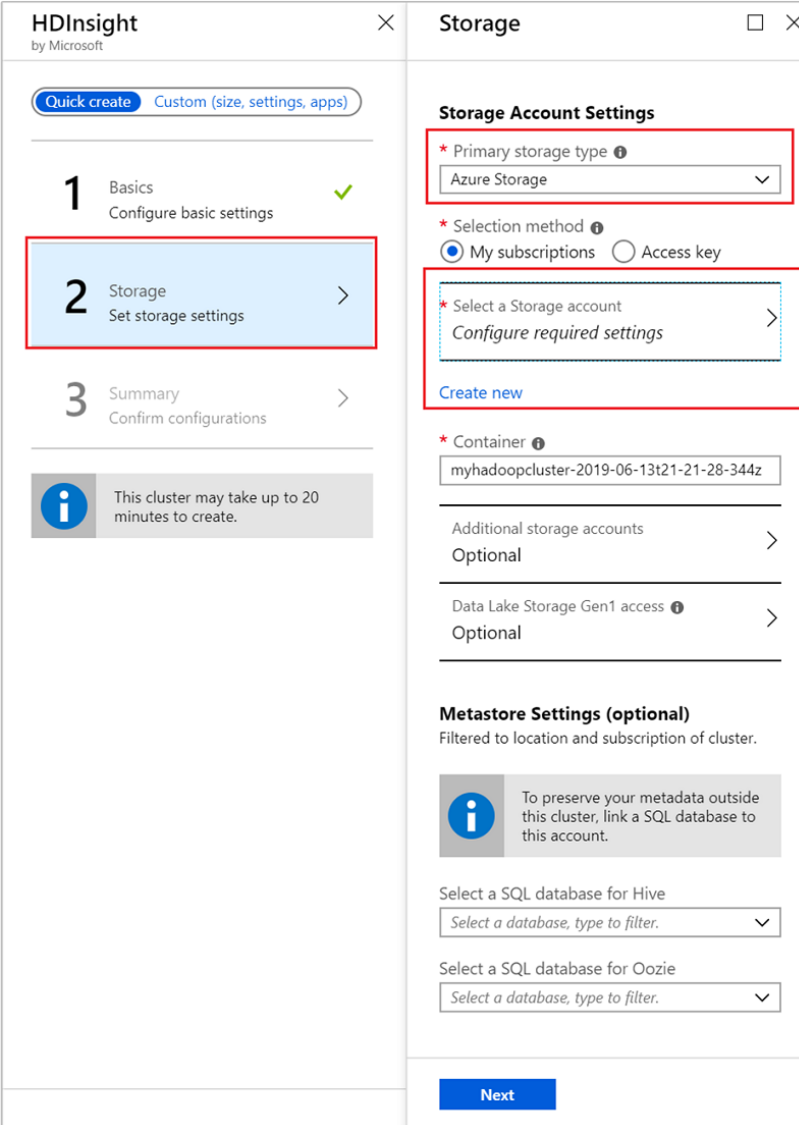
Propiedad	DESCRIPCIÓN
Tipo de almacenamiento principal	Para este artículo, seleccione Azure Storage para usar Azure Storage Blob como cuenta de almacenamiento predeterminada. También puede usar Azure Data Lake Store como almacenamiento predeterminado.
Método de selección	Para este artículo, seleccione Mis suscripciones para usar una cuenta de almacenamiento de la suscripción de Azure. Para usar una cuenta de almacenamiento de otras suscripciones, seleccione Clave de acceso y, a continuación, proporcione la clave de acceso para esa cuenta.



Selección de una cuenta de almacenamiento

Elija **Seleccione una cuenta de Storage** para seleccionar una cuenta de almacenamiento existente o bien elija **Crear nuevo**. Si crea una cuenta nueva, el nombre debe tener una longitud de entre 3 y 24 caracteres y solo puede contener números y letras minúsculas.

Acepte todos los demás valores predeterminados y, a continuación, seleccione **Siguiente** para avanzar a la página de resumen.



HDInsight
by Microsoft

Quick create Custom (size, settings, apps)

- 1 Basics
Configure basic settings ✓
- 2 **Storage**
Set storage settings >
- 3 Summary
Confirm configurations >

Storage

Storage Account Settings

- * Primary storage type ⓘ
Azure Storage
- * Selection method ⓘ
☒ My subscriptions ☐ Access key
- * Select a Storage account
Configure required settings >
- Create new
- * Container ⓘ
myhadoopcluster-2019-06-13t21-21-28-344z
- Additional storage accounts
Optional >
- Data Lake Storage Gen1 access ⓘ
Optional >

Metastore Settings (optional)
Filtered to location and subscription of cluster.

i To preserve your metadata outside this cluster, link a SQL database to this account.

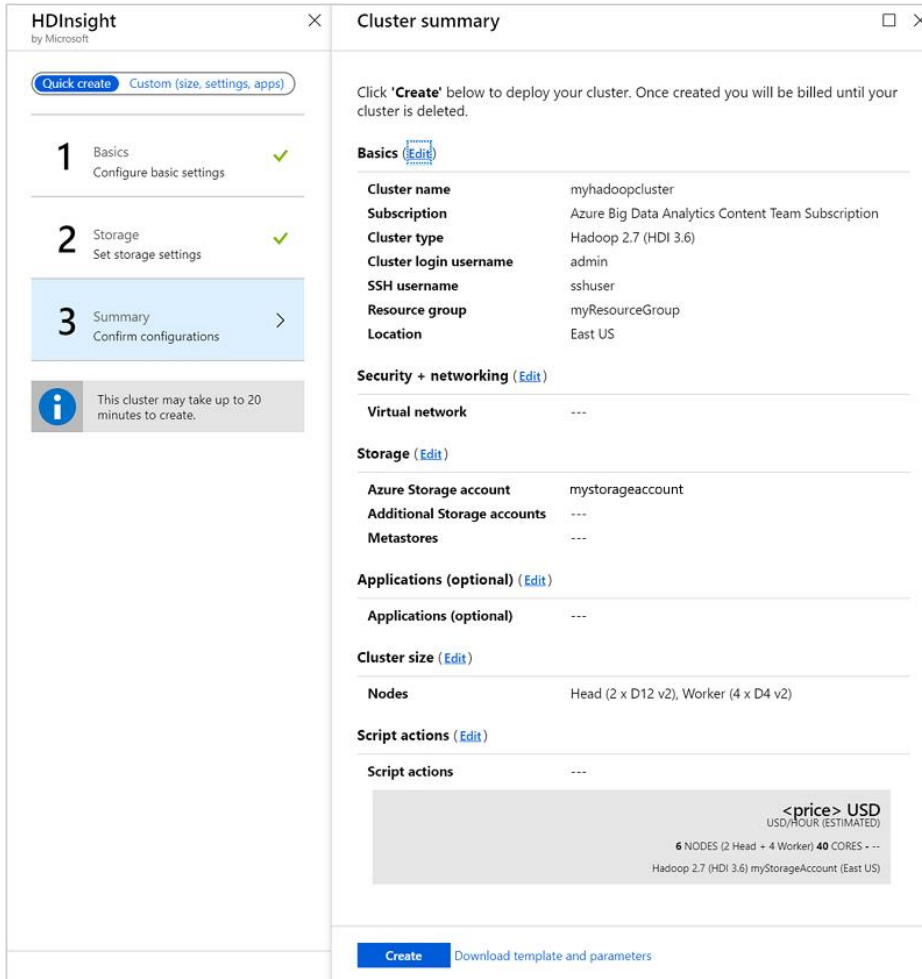
Select a SQL database for Hive
Select a database, type to filter. >

Select a SQL database for Oozie
Select a database, type to filter. >

Next



6. En la pestaña **Resumen**, compruebe los valores seleccionados en los pasos anteriores.



HDInsight by Microsoft

Quick create Custom (size, settings, apps)

- 1 Basics Configure basic settings ✓
- 2 Storage Set storage settings ✓
- 3 Summary Confirm configurations >

Cluster summary

Click 'Create' below to deploy your cluster. Once created you will be billed until your cluster is deleted.

Basics (Edit)

Cluster name	myhadoopcluster
Subscription	Azure Big Data Analytics Content Team Subscription
Cluster type	Hadoop 2.7 (HDI 3.6)
Cluster login username	admin
SSH username	sshuser
Resource group	myResourceGroup
Location	East US

Security + networking (Edit)

Virtual network ---

Storage (Edit)

Azure Storage account	mystorageaccount
Additional Storage accounts	---
Metastores	---

Applications (optional) (Edit)

Applications (optional) ---

Cluster size (Edit)

Nodes	Head (2 x D12 v2), Worker (4 x D4 v2)
-------	---------------------------------------

Script actions (Edit)

Script actions ---

Price

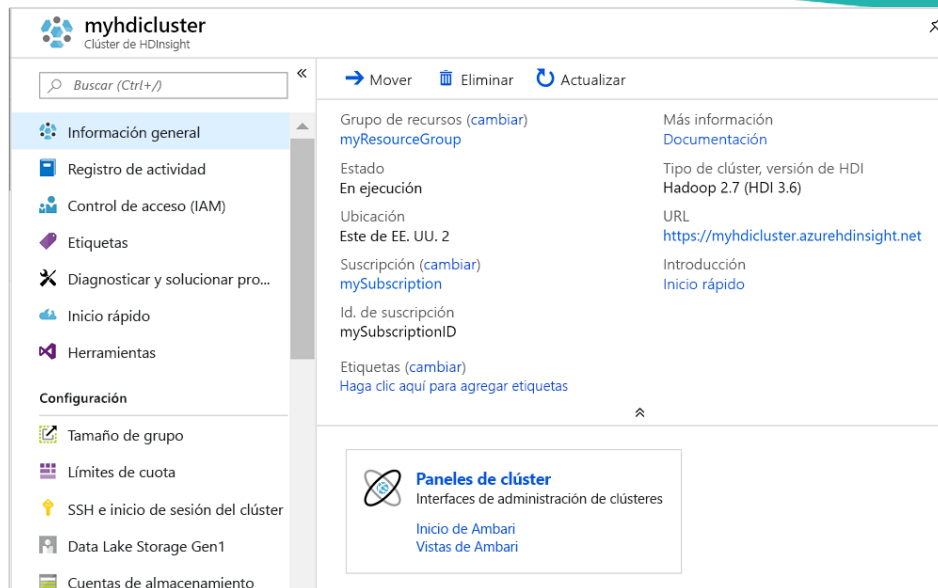
<price> USD
USD/HOUR (ESTIMATED)

6 NODES (2 Head + 4 Worker) 40 CORES ---
Hadoop 2.7 (HDI 3.6) myStorageAccount (East US)

Create Download template and parameters

7. Seleccione **Crear**. Se tarda aproximadamente 20 minutos en crear un clúster.
8. Una vez creado el clúster, verá la página de información general del clúster en Azure Portal.



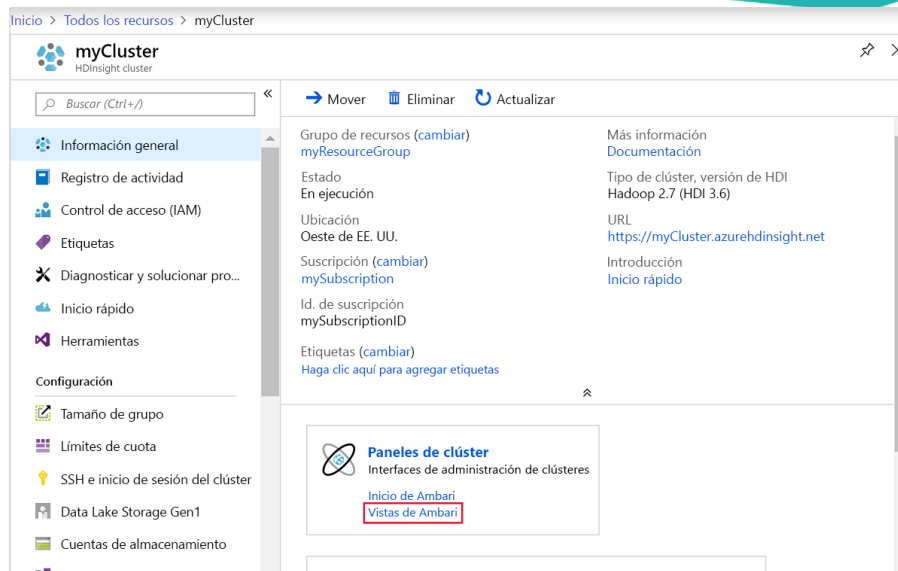


Paso 2: Ejecución de consultas de Apache Hive

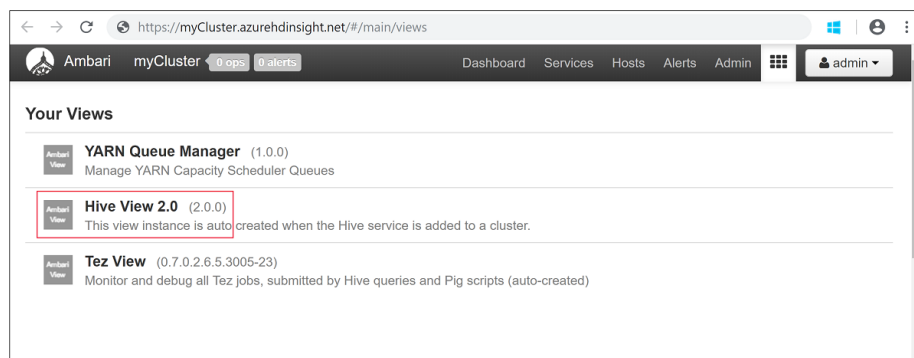
Apache Hive es el componente más popular de los que se usan en HDInsight. Hay muchas maneras de ejecutar trabajos de Hive en HDInsight. En este inicio rápido se usa la vista de Hive de Ambari desde el portal..

1. Para abrir Ambari, desde la captura de pantalla anterior, seleccione **Panel de clúster**. También puede ir a <https://ClusterName.azurehdinsight.net>, donde ClusterName es el clúster que creó en la sección anterior.





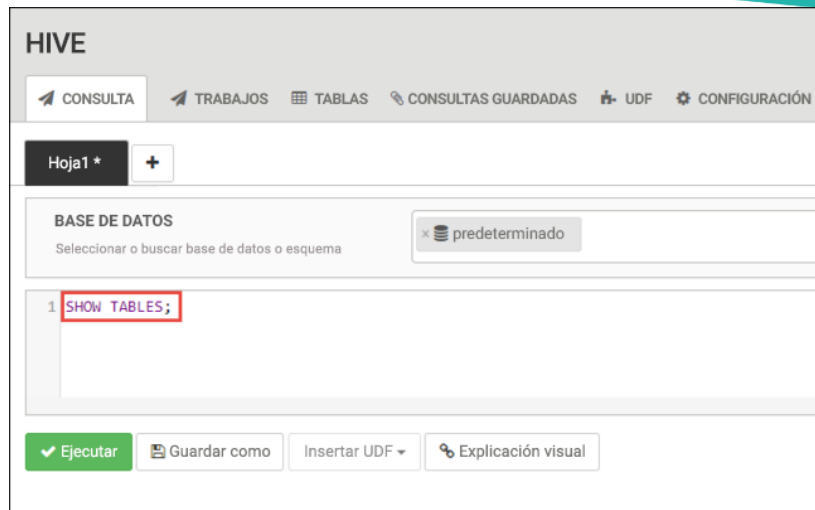
2. Escriba el nombre de usuario de Hadoop y la contraseña que especificó al crear el clúster. El nombre de usuario predeterminado es **admin..**
3. Abra la **vista de Hive** como se muestra en la siguiente captura de pantalla:



4. En la pestaña **CONSULTA**, pegue las instrucciones HiveQL siguientes en la hoja de cálculo:

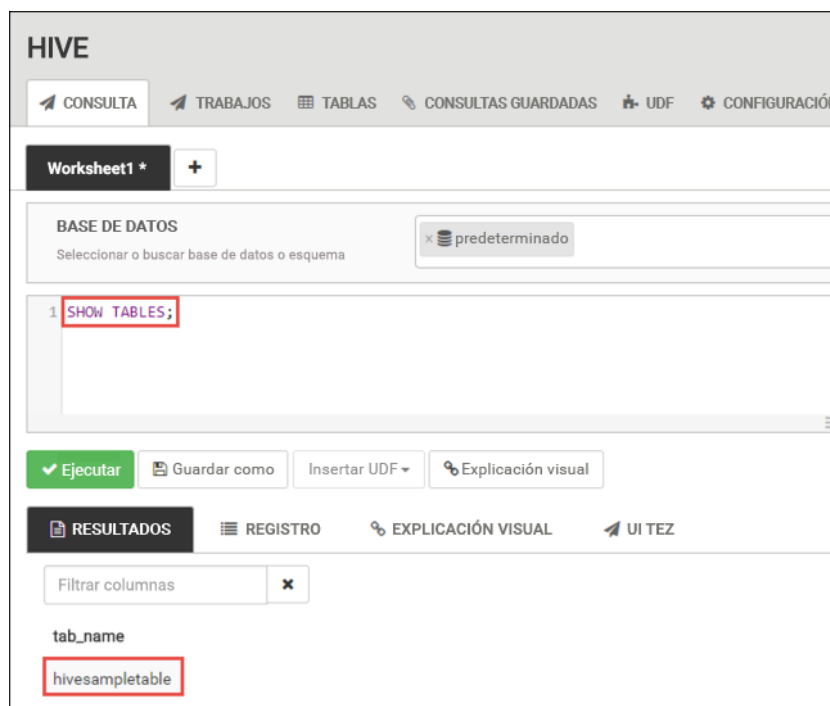
SHOW TABLES;





5. Seleccione **Execute** (Ejecutar). Aparecerá una pestaña **RESULTADOS** en la pestaña **CONSULTA** que mostrará información sobre el trabajo.

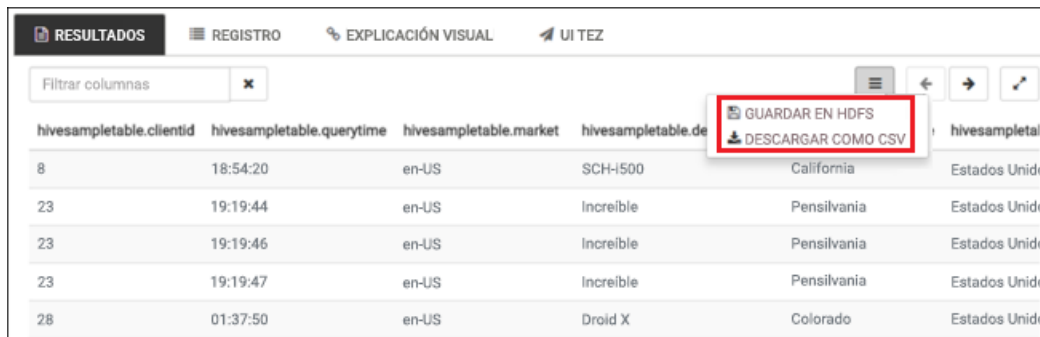
Cuando haya finalizado la consulta, la pestaña **CONSULTA** muestra los resultados de la operación. Verá una tabla denominada **hivesampletable**. Esta es una tabla de Hive de ejemplo que viene integrada en todos los clústeres de HDInsight.



6. Repita los pasos 4 y 5 para ejecutar la consulta siguiente:

```
SELECT * FROM hivesampletable;
```

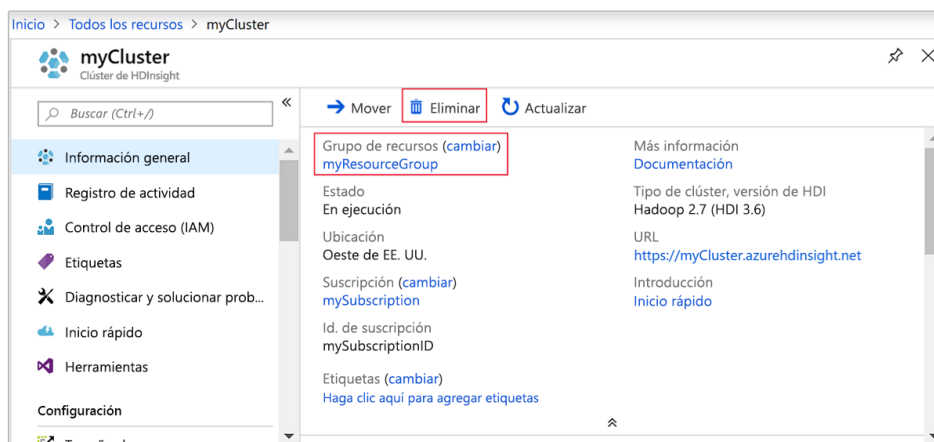
7. También puede guardar los resultados de la consulta. Seleccione el botón de menú de la derecha y especifique si quiere descargar los resultados como archivo CSV o almacenarlos en la cuenta de almacenamiento asociada al clúster.



hivesampletable.clientid	hivesampletable.querytime	hivesampletable.market	hivesampletable.de	hivesampletable.location	hivesampletable.country
8	18:54:20	en-US	SCH-I500	California	Estados Unidos
23	19:19:44	en-US	Incredible	Pensilvania	Estados Unidos
23	19:19:46	en-US	Incredible	Pensilvania	Estados Unidos
23	19:19:47	en-US	Incredible	Pensilvania	Estados Unidos
28	01:37:50	en-US	Droid X	Colorado	Estados Unidos

Paso 3: Para eliminar el clúster o la cuenta de almacenamiento predeterminada

1. Vuelva a la pestaña de explorador en la que tenga Azure Portal. Estará en la página de información general del clúster. Si solo quiere eliminar el clúster, pero desea seguir conservando la cuenta de almacenamiento predeterminada, seleccione **Eliminar**.



2. Si quiere eliminar el clúster y la cuenta de almacenamiento predeterminada, seleccione el nombre del grupo de recursos (resaltado en la captura de pantalla anterior) para abrir la página del grupo de recursos.
3. Seleccione **Eliminar grupo de recursos** para eliminar el grupo de recursos, que contiene el clúster y la cuenta de almacenamiento predeterminada. Tenga en cuenta que, al eliminar el grupo de recursos, se elimina también la cuenta de almacenamiento. Si desea mantener la cuenta de almacenamiento, elija eliminar solo el clúster.

