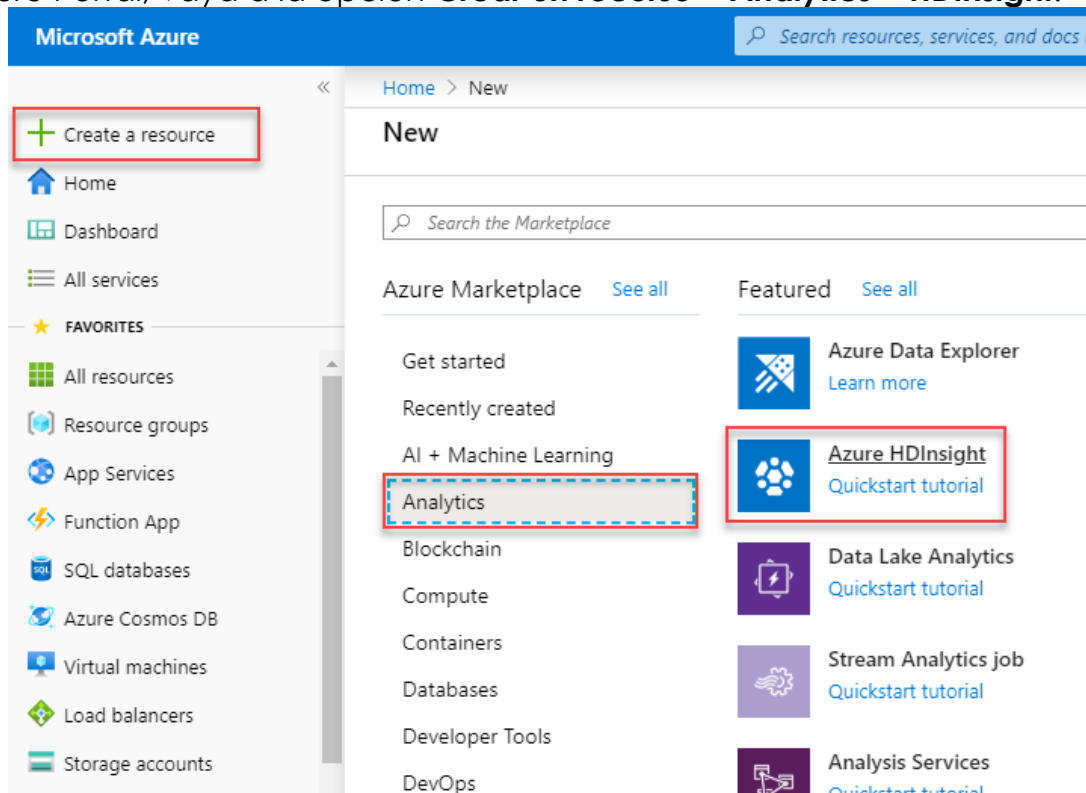


## LAB 02 – Creación de un clúster de Apache Hadoop en Azure HDInsight

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
<b>Subscription</b>	Seleccione su suscripción a Azure.
<b>Resource group</b>	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.
<b>Nombre del clúster</b>	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.
<b>Location</b>	Seleccione una ubicación de Azure en la que quiera crear el clúster. Elija una ubicación más cercana a usted para mejorar el rendimiento.
<b>Tipo de clúster</b>	Seleccione el tipo de cluster Hadoop.
<b>Nombre de usuario y contraseña de inicio de sesión del clúster</b>	El nombre de inicio de sesión predeterminado es <b>admin</b> . 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 <b>no proporcionar</b> contraseñas comunes, como "Pass@word1".
<b>Nombre de usuario de Secure Shell (SSH)</b>	El nombre de usuario predeterminado es <b>sshuser</b> . Puede proporcionar otro nombre para el nombre de usuario de SSH.
<b>Uso de la contraseña de inicio de sesión del clúster para SSH</b>	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.



Subscription \*

Resource group \*   
[Create new](#)

**Cluster details**  
Name your cluster, pick a region, and choose a cluster type and version. [Learn more](#)

Cluster name \*

Region \*

Cluster type \* **Hadoop**  
[Change](#)

Version \*

**Cluster credentials**  
Enter new credentials that will be used to administer or access the cluster.

Cluster login username \* ⓘ

Cluster login password \*  ✓

Confirm cluster login password \*  ✓

Secure Shell (SSH) username \* ⓘ

Use cluster login password for SSH ☒

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

Propiedad	DESCRIPCIÓN
<b>Tipo de clúster</b>	Seleccione <b>Hadoop</b>
<b>Versión</b>	Seleccione <b>Hadoop 3.1.0 (HDI 4.0)</b> .

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
<b>Tipo de almacenamiento principal</b>	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.
<b>Método de selección</b>	Para este artículo, seleccione <b>Mis suscripciones</b> para usar una cuenta de almacenamiento de la suscripción de Azure. Para usar una cuenta de almacenamiento de otras suscripciones, seleccione <b>Clave de acceso</b> y, a continuación, proporcione la clave de acceso para esa cuenta.
<b>Selección de una cuenta de almacenamiento</b>	Elija <b>Seleccione una cuenta de Storage</b> para seleccionar una cuenta de almacenamiento existente o bien elija <b>Crear nuevo</b> . 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.

Basics **Storage** Security + networking Configuration + pricing Review + create

Select or create storage accounts that will be used for the cluster's logs, job input, and job output. Configure the cluster's access to these accounts, if needed.

**Primary storage**

Select or create a storage account that will be the default location for cluster logs and other output.

Primary storage type \*

Selection method \* ☒ Select from list ☐ Use access key

Primary storage account \*  [Create new](#)

Container \*

6. Seleccione siguiente para ir a la sección de **Security + Networking** deje los valores por defecto y luego de click en siguiente para ir a la sección **Configuración +**



**pricing** y valide la configuración del cluster a nivel de numero de nodos y de click en siguiente:

**i** This configuration will use 40 of 250 available cores in the East US region.  
[View cores usage](#)

**Add application**


Node type	Node size	Number of ...	Estimated cost/hour
Head node	D12 v2 (4 Cores, 28 GB RAM), 0.37 USD/... <a href="#">▼</a>	2	0.75 USD
Worker node	D4 v2 (8 Cores, 28 GB RAM), 0.59 USD/h... <a href="#">▼</a>	4	2.36 USD

☐ Enable autoscale  
[Learn more](#)




**Total estimated cost/hour 3.11 USD**

- En la pestaña **Review + Create**, compruebe los valores seleccionados en los pasos anteriores.
- Seleccione **Crear**. Se tarda aproximadamente 20 minutos en crear un clúster.

### ... Your deployment is underway

 Deployment name: HDInsight\_\_2019-11-21T05.09.09.894Z      Start time: 11/21/2019, 12:09:12 AM  
Subscription: [Pase para Azure: patrocinio \(969be351-6ad7-443b-9...](#)      Correlation ID: db8f66ad-4319-4200-9eee-f0591bcb!  
Resource group: [qualitycode](#)

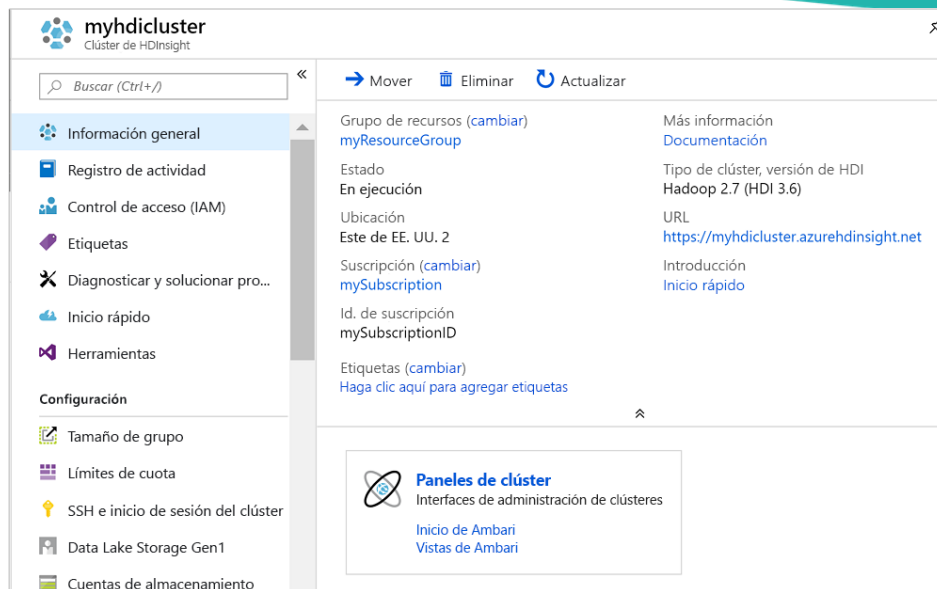
#### ^ Deployment details [\(Download\)](#)

Resource	Type	Status	Operation details
 qualitycluster11	Microsoft.HDInsight/clu...	OK	<a href="#">Operation details</a>
 qualityclustehdistorage	Microsoft.Storage/stora...	OK	<a href="#">Operation details</a>
 qualityclustehdistorage	Microsoft.Storage/stora...	OK	<a href="#">Operation details</a>

#### ^ Next steps

- 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 línea de comando.

1. Desde Power Shell o desde el Shell de Azure, ejecute el siguiente comando.

**ssh sshuser@CLUSTERNAME-ssh.azurehdinsight.net**

2. Cuando pregunte que si desea continuar, presione **YES**.
3. Ingrese la contraseña que configuró durante la creación del cluster.

Deberá visualizar una pantalla como la siguiente:

```
sshuser@hn0-cluste:~$
```

4. Para acceder a la consola de consulta de HIVE, deberá ejecutar el siguiente comando a través del utilitario **BEELINE**.

**beeline -u 'jdbc:hive2://localhost:10001/;transportMode=http'**

5. Podrá visualizar una ventana como la siguiente:



```
sshuser@hnn0-cluste: $ beeline -u 'jdbc:hive2://localhost:10001;transportMode=http'
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/hdp/3.1.2.2-1/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/3.1.2.2-1/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Connecting to jdbc:hive2://localhost:10001;transportMode=http
Connected to: Apache Hive (version 3.1.0.3.1.2.2-1)
Driver: Hive JDBC (version 3.1.0.3.1.2.2-1)
Transaction isolation: TRANSACTION_REPEATABLE_READ
beeline version 3.1.0.3.1.2.2-1 by Apache Hive
0: jdbc:hive2://localhost:10001/>
```

- Para visualizar las tablas que actualmente existen dentro de HIVE ejecute la siguiente instrucción:

**SHOW TABLES;**

- El resultado será el siguiente:

```
0: jdbc:hive2://localhost:10001/> SHOW TABLES;
INFO : Compiling command(queryId=hive_20191125050245_729652f6-4b95-4f90-9095-ee06717c6856): SHOW TABLES
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, type:string, comment:from deserializer)], properties:null)
INFO : Completed compiling command(queryId=hive_20191125050245_729652f6-4b95-4f90-9095-ee06717c6856); Time taken: 0.128 seconds
INFO : Executing command(queryId=hive_20191125050245_729652f6-4b95-4f90-9095-ee06717c6856): SHOW TABLES
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20191125050245_729652f6-4b95-4f90-9095-ee06717c6856); Time taken: 0.068 seconds
INFO : OK
+-----+
| tab_name |
+-----+
| hivesampletable |
+-----+
1 row selected (0.425 seconds)
0: jdbc:hive2://localhost:10001/>
```

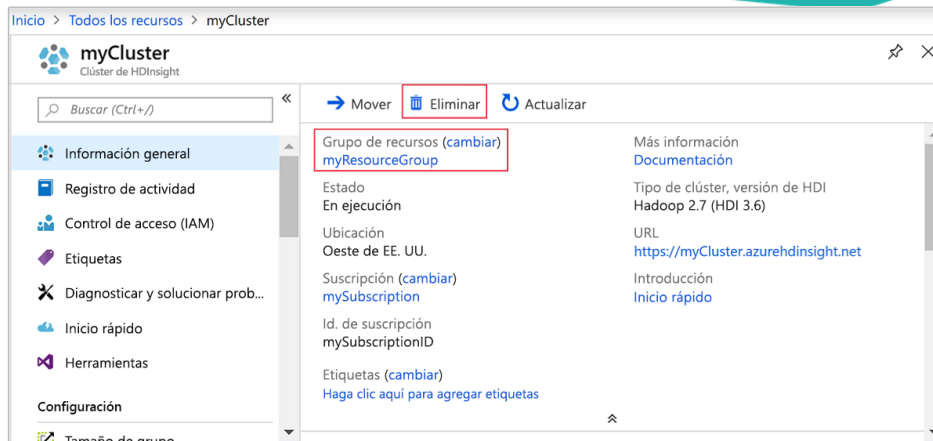
- Obtenga todos los registros de la tabla de ejemplo con la siguiente instrucción:

**SELECT \* FROM hivesampletable;**

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

- 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.

