

Apuntes databricks navegar en ficheros

Sistema de archivos de Databricks (DBFS)

- Artículo
- 08/04/2022
- Tiempo de lectura: 13 minutos
- 3 colaboradores

El sistema de archivos de Databricks (DBFS) es un sistema de archivos distribuido montado en un área de trabajo de Azure Databricks y está disponible en los clústeres de Azure Databricks. DBFS es una abstracción sobre el almacenamiento de objetos escalable y ofrece las ventajas siguientes:

- Permite montar objetos de almacenamiento, para que pueda acceder sin problemas a los datos, sin la necesidad de usar credenciales.
- Le permite interactuar con el almacenamiento de objetos mediante la semántica de archivos y directorios en lugar de las direcciones URL de almacenamiento.
- Conserva los archivos en el almacenamiento de objetos, por lo que no perderá los datos después de finalizar un clúster.

Información importante sobre los permisos de DBFS

Todos los usuarios tienen acceso de lectura y escritura a los objetos del almacenamiento de objetos montados en DBFS, a excepción de la raíz de DBFS.

Raíz de DBFS

La ubicación de almacenamiento predeterminada de DBFS se conoce como la raíz de DBFS. Varios tipos de datos se almacenan en las siguientes ubicaciones raíz de DBFS:

- `/FileStore`: archivos de datos importados, trazados generados y bibliotecas cargadas. Consulte Ubicaciones raíz de DBFS especiales.
- `/databricks-datasets`: conjuntos de datos públicos de ejemplo. Consulte Ubicaciones raíz de DBFS especiales.
- `/databricks-results`: archivos generados mediante la descarga de los resultados completos de una consulta.
- `/databricks/init`: scripts de inicialización globales y con nombre de clúster (en desuso).
- `/user/hive/warehouse`: datos y metadatos para tablas de Hive no externas.

En una nueva área de trabajo, la raíz de DBFS tiene las siguientes carpetas predeterminadas:

1	%fs ls /FileStore/tables			
	path	name	size	modificationTime
1	dbfs:/FileStore/tables/el_quijote-1.txt	el_quijote-1.txt	1060259	1651254890000
2	dbfs:/FileStore/tables/el_quijote.txt	el_quijote.txt	1060259	1651254652000
3	dbfs:/FileStore/tables/mnm_dataset.csv	mnm_dataset.csv	1284872	1651569144000

Showing all 3 rows.

Command took 2.67 seconds -- by josemanuel.cerrato@bosonit.com at 4/5/2022, 9:19:13 on My Cluster

La raíz de DBFS también contiene datos —incluidos los metadatos y credenciales de punto de montaje y determinados tipos de registros— que no son visibles y a los que no se puede acceder directamente.

Acceso a archivos en DBFS

La ruta de acceso al almacenamiento de blogs predeterminado (raíz) es `dbfs:/`.

La ubicación predeterminada de `%fs` y `dbutils.fs` es raíz. Por lo tanto, para leer o escribir en la raíz o en un cubo externo:

Bash

```
%fs <command> /<path>
```

Python

```
dbutils.fs.<command> ("/<path>/")
```

`%sh` lee el sistema de archivos local de manera predeterminada. Para acceder a las rutas de acceso raíz o montadas, en la raíz, con `%sh`, coloque `/dbfs/` antes de la ruta de acceso. Un caso de uso típico es si está trabajando con bibliotecas de nodo único, como TensorFlow o scikit-learn, y quiere leer y escribir datos en el almacenamiento en la nube.

Bash

```
%sh <command> /dbfs/<path>/
```

También puede usar las API del sistema de archivos de nodo único:

Python

```
import os
```

```
os.<command>('/dbfs/tmp')
```

Ejemplos

Bash

```
# Default location for %fs is root
%fs ls /tmp/
%fs mkdirs /tmp/my_cloud_dir
%fs cp /tmp/test_dbfs.txt /tmp/file_b.txt
```

Python

```
# Default location for dbutils.fs is root
dbutils.fs.ls ("/tmp/")
dbutils.fs.put("/tmp/my_new_file", "This is a file in cloud storage.")
```

Bash

```
# Default location for %sh is the local filesystem
%sh ls /dbfs/tmp/
```

Python

```
# Default location for os commands is the local filesystem
import os
os.listdir('/dbfs/tmp')
```

Acceso a archivos en el sistema de archivos local

%fs y dbutils.fs leen de manera predeterminada desde la raíz (dbfs:/). Para leer desde el sistema de archivos local, debe usar file:/.

Bash

```
%fs <command> file:./<path>
dbutils.fs.<command> ("file:./<path>/")
%sh lee el sistema de archivos local de manera predeterminada, así que no use file:/.
```

Bash

```
%sh <command> /<path>
```

Ejemplos

Bash

```
# With %fs and dbutils.fs, you must use file:/ to read from local filesystem
%fs ls file:/tmp
%fs mkdirs file:/tmp/my_local_dir
dbutils.fs.ls ("file:/tmp/")
dbutils.fs.put("file:/tmp/my_new_file", "This is a file on the local driver node.")
```

Bash

%sh reads from the local filesystem by default

%sh `ls` /tmp

Acceso a archivos en el almacenamiento de objetos montados

El montaje del almacenamiento de objetos en DBFS permite acceder a objetos en el almacenamiento de objetos, como si estuvieran en el sistema de archivos local.

Ejemplos

Python

```
dbutils.fs.ls("/mnt/mymount")
```

```
df = spark.read.text("dbfs:/mymount/my_file.txt")
```