Prácticas de HDFS

Descripción

El alumno deberá manipular ficheros en HDFS utilizando la línea de comandos.

Solo se requieren los conceptos mostrados en la sesión teórica para completar el ejercicio.

Sigue las instrucciones del profesor para inicializar el entorno.

Ficheros usados en este ejercicio

https://github.com/ciceonline/cice/archive/master.zip

- cice/hdfs/data/quijote.txt

- cice/hdfs/data/access_log.txt

Explorar el HDFS

Algunos de los diferentes comandos presentados en las diapositivas. Todos los comandos son de la forma:

hdfs dfs <comando> <argumento 1> <argumento 2> .... <argumento n>

1. Abre un terminal en la máquina virtual.

2. Para obtener un listado de los diferentes comandos disponibles ejecuta:

```
$ hdfs dfs
```

3. Vamos a listar los contenidos del directorio raíz de nuestro HDFS.

```
$ hdfs dfs ls /
```

Se mostrará el contenido actual de dicho directorio.

4. Cada usuario con permisos de acceso al sistema tiene asignada una carpeta dentro de /users lo que podemos ver ejecutando:

```
$ hdfs dfs ls /users
```

5. Prueba a listar el contenido de algunos de los directorios que aparecen en el listado anterior.

6. ¿Qué ocurre si intentamos acceder a una ruta inexistente? ¿Y a una en la que no tenemos permisos?

Carga de ficheros

A continuación vamos a crear un directorio para nuestros datos y vamos a subir un par de ficheros a él:

1. Vamos a crear un directorio para albergar nuestros datos. El directorio se llamará data y estará en nuestra "home" en HDFS (/user/cloudera).

$ hdfs dfs mkdir data

Dado que estamos trabajando con el usuario cloudera, por defecto, todas las rutas relativas lo son a su carpeta de trabajo.

2. Si no los tienes ya copiados, copia a la máquina virtual los ficheros quijote.txt y access_log en la carpeta /home/cloudera/ejercicios/data del filesystem local (os los proporcionará el profesor).

3. Sube a la carpeta /user/cloudera/data dichos ficheros.

```
$ hdfs dfs put ~/ejercicios/data/quijote.txt data/.
```

```
$ hdfs dfs put ~/ejercicios/data/access_log data/.
```

Ver y manipular ficheros

Vamos a ver los ficheros que acabamos de cargar en nuestro HDFS.

1. Primero, comprobemos que realmente hemos cargado esos ficheros.

```
$ hdfs dfs ls data
```

2. Vamos a realizar una copia del fichero quijote.txt

```
$ hdfs dfs cp data/quijote.txt data/quijote_copia.txt
```

3. Volvemos a comprobar la carpeta

```
$ hdfs dfs ls data
```

4. Para borrar dicho fichero de copia utilizamos el comando:

```
$ hdfs dfs rm data/quijote_copia.txt
```

5. Veamos las primeras líneas del fichero access_log

```
$ hdfs dfs cat data/access_log | head 20
```

6. Para descargar el fichero access_log a la máquina local podemos usar el comando get:

```
$ hdfs dfs get data/access_log /tmp
```

Esto nos copia el fichero a la carpeta /tmp de nuestro filesystem local.

Otros comandos

Prueba a ejecutar alguno de los otros comandos que aparecen al ejecutar, o de los vistos en clase:

```
$ hdfs dfs
```