# Tecnoloxías de computación para datos masivos [P4181103] [2021/2022]



## Tema 3 - Práctica: Descripción y entrega

**Pendente:** Domingo, 31 de Outubro de 2021, 23:59

---

### Gestión de HDFS

En esta práctica veremos algunos comandos para gestionar el HDFS.

**Primera parte**: Probar el comandos **hdfs dfsadmin**

1. En el NameNode, como usuario hdadmin, crea un directorio en HDFS y ponle una cuota de solo 4 ficheros. Comprueba cuántos ficheros puedes copiar a ese directorio. Explica a qué se debe este comportamiento.

**Segunda parte:** Probar el comando **hdfs fsck**

1. En el NameNode (como usuario hdadmin) haz un chequeo de todo el HDFS, y comprueba si te da errores:
   - Intenta determinar las causas de los posibles errores
2. Detén datanodes de forma brusca, parando los dockers, (sin detener los demonios, simplemente haciendo p.e. `docker container stop datanode1` datanode2 datanode4) hasta que te queden **solo 2 datanodes vivos en dos racks diferentes**. Espera unos 10 minutos[1] y comprueba que el comando `hdfs dfsadmin -report` muestra que, efectivamente, solo quedan 2 datanodes activos.
3. Realiza de nuevo el chequeo de disco en el NameNode y comprueba la salida. ¿Cuántos bloques aparecen *under-replicated*?
4. Prueba a hacer un `get` del fichero `random_words.txt.bz2` para ver si se hace correctamente.
5. Sigue los pasos que vistes en la práctica 1 para añadir un datanode nuevo (datanode6). Comprueba, haciendo de nuevo el chequeo, que los datos se replican en el nuevo nodo hasta alcanzar el factor de replicación por defecto (tarda un rato en alcanzar el nivel de replicación 3, si no avanza, comprueba en el nuevo datanode que el demonio esté funcionando).

**Tercera parte:** Probar el uso de **códigos de borrado (erasure codes o EC)**
Para poder utilizar EC en vez de replicación es necesario tener activos como mínimo 5 datanodes. EC se aplica a ficheros nuevos que se guarden en carpetas en las que se haya especificado una política de EC.

Existen diferentes políticas de EC, que garantizan una mayor o menor protección frente a fallos, con menor o mayor ahorro de espacio. Se pueden ver las políticas disponibles con el comando hdfs ec -listPolicies

La política que se usa por defecto se define a través de la propiedad 'dfs.namenode.ec.system.default.policy' , y es "RS-6-3-1024k" por defecto.

Lo primero que haremos será crear una carpeta para la que especificaremos una política de EC. Sigue los siguientes pasos:

1. Inicia los dockers necesarios para tener 5 datanodes
2. En el NameNode, como usuario hdadmin, comprueba las políticas disponibles con el comando antes indicado
3. Habilita la politica "RS-3-2-1024k" ejecutando:
   - `hdfs ec -enablePolicy -policy RS-3-2-1024k`
4. Crea una carpeta /user/grandes en HDFS, para la que vamos a indicar que se aplique una política EC
5. Aplica la política EC con:
   - `hdfs ec -setPolicy -path /user/grandes -policy RS-3-2-1024k`

Con esos pasos, aplicamos la política EC indicada en el directorio /user/grandes. Vamos a ver que funciona. Ejecuta los siguientes pasos en el NameNode como hdadmin:

1. Comprueba con `hdfs dfsadmin -report` el espacio ocupado en DFS, y apunta el valor (captura de pantalla)
2. Ejecuta `hdfs dfs -get libros/random_words.txt.bz2` para obtener ese fichero grande desde HDFS
3. Borra ese fichero de HDFS y vacía la papelera (`hdfs dfs -expunge`)
4. Pon el fichero en la carpeta /user/grandes con `hdfs dfs -put random_words.txt.bz2 /user/grandes` (da un warning, pero la copia se realiza igual)
5. Comprueba de nuevo el espacio ocupado en DFS y comparalo con el valor que había antes.

### Entrega

1. Un documento que muestre que se han ejecutado los diferentes apartados. Mostrar capturas de pantalla, incluyendo una explicación y justificación de lo que aparece en las mismas.

[1] El tiempo depende de los parámetros dfs.namenode.stale.datanode.interval y dfs.namenode.heartbeat.recheck-interval. El primero indica el tiempo sin detectar actividad del DataNode para que el NameNode lo considere en estado stale (por defecto, 30 segundos). Un nodo en estado stale tiene menor prioridad en lecturas y escrituras. El segundo parámetro indica el intervalo de chequeo en busca de DataNodes expirados (valor por defecto, 5 minutos). Un DataNode se pasa al estado Dead cuando el tiempo sin detectar actividad es superior a dfs.namenode.stale.datanode.interval + 2 * dfs.namenode.heartbeat.recheck-interval.

## Estado da entrega

| | |
|---|---|
| **Estado da entrega** | Sen intentos |
| **Estado das cualificacións** | Sen cualificar |
| **Tempo restante** | 10 días 8 horas |
| **Última modificación** | - |
| **Comentarios a entrega** | Comentarios (0) |

Engadir entrega

Vostede aínda non fixo ningunha entrega

◄ Grabación clase 04/10/21

Ir a...

Grabación clase 14/10/21 ►

Тоҷикӣ (tg)
Українська (uk)
עברית (he)
日本語 (ja_old)
正體中文 (zh_tw)
简体中文 (zh_cn)