

Урок 1. Базовые операции в HDFS. Консольные утилиты.

Выполнил Колеганов Н.Д.

Задание

1. [Исследовательское] Сколько узлов одновременно можно потерять без потери данных в кластере из 10 узлов? Из 100 узлов?
2. Опробовать консольные утилиты для работы с кластером
3. Создать/скопировать/удалить папку
4. Положить в HDFS любой файл
5. Скопировать/удалить этот файл
6. Просмотреть размер любой папки
7. Посмотреть как файл хранится на файловой системе (см. команду fsck)
8. Установить нестандартный фактор репликации (см. команду setrep)
9. Опробовать rest-доступ для работы с кластером
10. Используя утилиту CURL
11. Используя python3 и библиотеку requests
12. [Для любителей администрирования] Опробовать NFS доступ. Предварительно связаться со мной чтобы я открыл нужные порты.
13. [Для любителей программирования] Достучаться до файловой системы используя python и библиотеку libhdfs3

Решение

Сколько можно одновременно потерять узлов без потери данных при 10? при 100?

Многие распределенные системы хранения используют репликацию для сохранности данных. Если один диск в ноде отказывает, то данные этого диска просто теряются. Чтобы предотвратить безвозвратную потерю данных, СУБД хранит копию (реплику) данных где-то на дисках в другой ноде. Самым распространённым фактором репликации является 3 — это значит, что база данных хранит три копии каждого фрагмента данных на разных дисках, подключенных к трём разным компьютерам. Объяснение этому примерно такое: диски выходят из строя редко. Если диск вышел из строя, то есть время заменить его, и в это время у вас ещё две копии, с которых можно восстановить данные на новый диск. Риск выхода из строя второго диска, пока вы восстанавливаете первый, достаточно низок, а вероятность смерти всех трёх дисков одновременно настолько мала, что более вероятно погибнуть от попадания астероида.

Опробовать консольные утилиты для работы с кластером

Создать/скопировать/удалить папку

```
[student4_2@manager ~]$ hdfs dfs -mkdir /123
[student4_2@manager ~]$ ls
[student4_2@manager ~]$ hdfs dfs -ls /
Found 56 items
drwxr-xr-x   - student4_3  supergroup          0 2020-05-23 09:32 /0000
drwxr-xr-x   - student4_2  supergroup          0 2020-05-23 16:59 /123
drwxrwxrwx   - student3_2  supergroup          0 2020-05-14 18:35 /ALI_folder
drwxr-xr-x   - student3_2  supergroup          0 2020-03-28 13:10 /BD
drwxrwxrwx   - dekan       supergroup          0 2020-05-14 18:48 /NetCat
drwxr-xr-x+  - centos      supergroup          0 2019-11-25 15:33 /acldir
drwxrwxrwx   - student3_16 supergroup          0 2020-04-08 06:13 /border
```

```
drwxr-xr-x   - student4_2  supergroup          0 2020-05-23 17:01 /1234
[student4_2@manager ~]$ hdfs dfs -cp /123 /1234
[student4_2@manager ~]$ hdfs dfs -ls /
Found 57 items
drwxr-xr-x   - student4_3  supergroup          0 2020-05-23 09:32 /0000
drwxr-xr-x   - student4_2  supergroup          0 2020-05-23 16:59 /123
drwxr-xr-x   - student4_2  supergroup          0 2020-05-23 17:01 /1234
```

```
[student4_2@manager ~]$ hdfs dfs -rm -r /123
20/05/23 17:02:39 INFO fs.TrashPolicyDefault: Moved: 'hdfs://manager.novalocal:8020/123' to trash at: hdfs://manager.novalocal:8020/user/student4_2/.Trash/Current/123
[student4_2@manager ~]$ hdfs dfs -ls /
Found 56 items
drwxr-xr-x   - student4_3  supergroup          0 2020-05-23 09:32 /0000
drwxr-xr-x   - student4_2  supergroup          0 2020-05-23 17:01 /1234
drwxrwxrwx   - student3_2  supergroup          0 2020-05-14 18:35 /ALI_folder
drwxr-xr-x   - student3_2  supergroup          0 2020-03-28 13:10 /BD
```

Положить в HDFS любой файл

Скопировать/удалить этот файл

Просмотреть размер любой папки

```
[student4_2@manager ~]$ ls
test.txt
[student4_2@manager ~]$ cat test.txt
hello word

[student4_2@manager ~]$ hdfs dfs -copyFromLocal test.txt /1234
[student4_2@manager ~]$ hdfs dfs -ls /1234
Found 2 items
drwxr-xr-x  - student4_2 supergroup          0 2020-05-23 17:01 /1234/123
-rw-r--r--  3 student4_2 supergroup        22 2020-05-23 17:08 /1234/test.txt
[student4_2@manager ~]$ hdfs dfs -rm -r /1234/test.txt
20/05/23 17:09:35 INFO fs.TrashPolicyDefault: Moved: 'hdfs://manager.novalocal:8020/1234/test.txt' to trash at: hdfs://manager.novalocal:8020/user/student4_2/.Trash/Current/1234/test.txt
[student4_2@manager ~]$ hdfs dfs -ls /1234
Found 1 items
drwxr-xr-x  - student4_2 supergroup          0 2020-05-23 17:01 /1234/123
[student4_2@manager ~]$ hdfs dfs -du -h -s /1234
0 0 /1234
```

Посмотреть как файл хранится на файловой системе (см. команду fsck)

```
[student4_2@manager ~]$ hdfs fsck /1234/test.txt
Connecting to namenode via http://manager.novalocal:50070/fsck?ugi=student4_2&path=%2F1234%2Ftest.txt
FSCK started by student4_2 (auth:SIMPLE) from /89.208.221.132 for path /1234/test.txt at Sat May 23 17:11:25 UTC 2020
Status: HEALTHY
Total size:      22 B
Total dirs:      0
Total files:      1
Total symlinks:    0
Total blocks (validated): 1 (avg. block size 22 B)
Minimally replicated blocks: 1 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 3
Average block replication: 3.0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Number of data-nodes: 3
Number of racks: 1
FSCK ended at Sat May 23 17:11:25 UTC 2020 in 1 milliseconds

The filesystem under path '/1234/test.txt' is HEALTHY
```

Установить нестандартный фактор репликации (см. команду setrep)

сделал репликацию равную 4

```
[student4_2@manager ~]$ hdfs dfs -setrep 4 /1234/test.txt  
Replication 4 set: /1234/test.txt
```

Опробовать rest-доступ для работы с кластером

Используя утилиту CURL

Используя python3 и библиотеку requests

```
[student4_2@manager ~]$ curl -X GET 'http://node2.novalocal:14000/webhdfs/v1/acldir?user.name=hdfs&op=LISTSTATUS'  
{  
  "FileStatuses": {  
    "FileStatus": [ {  
      "pathSuffix": "etc",  
      "type": "DIRECTORY",  
      "length": 0,  
      "owner": "centos",  
      "group": "supergroup",  
      "permission": "755",  
      "accessTime": 0,  
      "modificationTime": 1574696487181,  
      "blockSize": 0,  
      "replication": 0  
    } ]  
  }  
}  
[student4_2@manager ~]$ curl -X GET 'http://node2.novalocal:14000/webhdfs/v1/acldir?user.name=student4_2&op=LISTSTATUS'  
{  
  "FileStatuses": {  
    "FileStatus": [ {  
      "pathSuffix": "etc",  
      "type": "DIRECTORY",  
      "length": 0,  
      "owner": "centos",  
      "group": "supergroup",  
      "permission": "755",  
      "accessTime": 0,  
      "modificationTime": 1574696487181,  
      "blockSize": 0,  
      "replication": 0  
    } ]  
  }  
}
```

```
[student4_2@manager ~]$ curl -X GET 'http://node2.novalocal:14000/webhdfs/v1/acldir?user.name=hdfs&op=GETCONTENTSUMMARY'  
{"ContentSummary": {"directoryCount": 123, "fileCount": 801, "length": 10210950, "quota": -1, "spaceConsumed": 30632850, "spaceQuota": -1}}  
[student4_2@manager ~]$
```