빅데이터 플랫폼

HDFS 실습 과제

소프트웨어전공
20152791 강길웅

Download MovieLens Data

실습에 필요한 데이터 셋 다운로드. `wget <url>` 명령어 통해서 다운로드.

```
[hadoop@ip-172-31-62-109 ~]$ wget https://s3.ap-northeast-2.amazonaws.com/kmubigdata-movielensdata/ml-20m.zip
--2020-10-27 10:36:16--  https://s3.ap-northeast-2.amazonaws.com/kmubigdata-movielensdata/ml-20m.zip
Resolving s3.ap-northeast-2.amazonaws.com (s3.ap-northeast-2.amazonaws.com)... 52.219.56.53
Connecting to s3.ap-northeast-2.amazonaws.com (s3.ap-northeast-2.amazonaws.com)|52.219.56.53|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 142787556 (136M) [application/zip]
Saving to: 'ml-20m.zip'

100%[======================================>] 142,787,556 19.6MB/s   in 9.0s

2020-10-27 10:36:26 (15.1 MB/s) - 'ml-20m.zip' saved [142787556/142787556]

[hadoop@ip-172-31-62-109 ~]$ unzip ml-20m.zip
Archive:  ml-20m.zip
  inflating: ml-20m/README.txt
  inflating: ml-20m/movies.csv
  inflating: ml-20m/ratings.csv
  inflating: ml-20m/tags.csv
  inflating: ml-20m/tags_1.csv
  inflating: ml-20m/tags_2.csv
  inflating: ml-20m/tags_3.csv
[hadoop@ip-172-31-62-109 ~]$ rm -f ml-20.zip
[hadoop@ip-172-31-62-109 ~]$ head ml-20m/movies.csv
movieId,title,genres
1,Toy Story (1995),Adventure|Animation|Children|Comedy|Fantasy
2,Jumanji (1995),Adventure|Children|Fantasy
3,Grumpier Old Men (1995),Comedy|Romance
4,Waiting to Exhale (1995),Comedy|Drama|Romance
5,Father of the Bride Part II (1995),Comedy
6,Heat (1995),Action|Crime|Thriller
7,Sabrina (1995),Comedy|Romance
8,Tom and Huck (1995),Adventure|Children
9,Sudden Death (1995),Action
[hadoop@ip-172-31-62-109 ~]$
```

hdfs dfs

HDFS의 기본 명령어 포맷은 `hdfs [SHELL_OPTION] COMMAND` 이다. 이중 대부분의 명령은 `hdfs dfs [COMMAND[COMMAND_OPTION]]` 형식이다. 단순히 `hdfs dfs` 명령을 입력하면 HDFS의 파일 시스템과 연관된 옵션들을 보여준다.

```
[hadoop@ip-172-31-62-109 ~]$ hdfs dfs
Usage: hadoop fs [generic options]
        [-appendToFile <localsrc> ... <dst>]
        [-cat [-ignoreCrc] <src> ...]
        [-checksum <src> ...]
        [-chgrp [-R] GROUP PATH...]
        [-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]
        [-chown [-R] [OWNER][:[GROUP]] PATH...]
        [-copyFromLocal [-f] [-p] [-l] [-d] <localsrc> ... <dst>]
        [-copyToLocal [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
        [-count [-q] [-h] [-v] [-t [<storage type>]] [-u] [-x] <path> ...]
        [-cp [-f] [-p | -p[topax]] [-d] <src> ... <dst>]
        [-createSnapshot <snapshotDir> [<snapshotName>]]
        [-deleteSnapshot <snapshotDir> <snapshotName>]
        [-df [-h] [<path> ...]]
        [-du [-s] [-h] [-x] <path> ...]
        [-expunge]
        [-find <path> ... <expression> ...]
        [-get [-f] [-p] [-ignoreCrc] [-crc] <src> ... <localdst>]
        [-getfacl [-R] <path>]
        [-getfattr [-R] {-n name | -d} [-e en] <path>]
        [-getmerge [-nl] [-skip-empty-file] <src> <localdst>]
        [-help [cmd ...]]
        [-ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [<path> ...]]
        [-mkdir [-p] <path> ...]
        [-moveFromLocal <localsrc> ... <dst>]
        [-moveToLocal <src> <localdst>]
        [-mv <src> ... <dst>]
        [-put [-f] [-p] [-l] [-d] <localsrc> ... <dst>]
        [-renameSnapshot <snapshotDir> <oldName> <newName>]
        [-rm [-f] [-r|-R] [-skipTrash] [-safely] <src> ...]
        [-rmdir [--ignore-fail-on-non-empty] <dir> ...]
        [-setfacl [-R] [{-b|-k} {-m|-x <acl_spec>} <path>]|[--set <acl_spec> <path>]]
        [-setfattr {-n name [-v value] | -x name} <path>]
        [-setrep [-R] [-w] <rep> <path> ...]
        [-stat [format] <path> ...]
        [-tail [-f] <file>]
        [-test -[defsz] <path>]
        [-text [-ignoreCrc] <src> ...]
        [-touchz <path> ...]
        [-truncate [-w] <length> <path> ...]
        [-usage [cmd ...]]

Generic options supported are:
-conf <configuration file>        specify an application configuration file
-D <property=value>               define a value for a given property
-fs <file:///|hdfs://namenode:port> specify default filesystem URL to use, overrides 'fs.defaultFS' property from configurations.
-jt <local|resourcemanager:port>  specify a ResourceManager
-files <file1,...>                specify a comma-separated list of files to be copied to the map reduce cluster
-libjars <jar1,...>               specify a comma-separated list of jar files to be included in the classpath
-archives <archive1,...>          specify a comma-separated list of archives to be unarchived on the compute machines

The general command line syntax is:
command [genericOptions] [commandOptions]

[hadoop@ip-172-31-62-109 ~]$
```

hdfs dfs -ls

디렉토리 상태 표기.

```
[hadoop@ip-172-31-62-109 ~]$ hdfs dfs -ls /
Found 3 items
drwxrwxrwt   - hdfs hadoop          0 2020-10-27 10:35 /tmp
drwxr-xr-x   - hdfs hadoop          0 2020-10-27 10:35 /user
drwxr-xr-x   - hdfs hadoop          0 2020-10-27 10:35 /var
[hadoop@ip-172-31-62-109 ~]$
```

hdfs dfs -mkdir [-p] <paths>

입력으로 들어온 경로에 지정 이름의 디렉토리를 생성한다.

```
[hadoop@ip-172-31-62-109 ~]$ hdfs dfs -mkdir -p /dataset/movielens/
[hadoop@ip-172-31-62-109 ~]$ hdfs dfs -ls /dataset/
Found 1 items
drwxr-xr-x   - hadoop hadoop          0 2020-10-27 10:41 /dataset/movielens
[hadoop@ip-172-31-62-109 ~]$
```

hdfs dfs -put <localsrc> <dst>

local의 파일을 hdfs로 복사해서 가져온다.

```
[hadoop@ip-172-31-62-109 ~]$ hdfs dfs -put ml-20m/ratings.csv /dataset/movielens/
[hadoop@ip-172-31-62-109 ~]$ hdfs dfs -ls /dataset/movielens/
Found 1 items
-rw-r--r--   1 hadoop hadoop  533444411 2020-10-27 10:42 /dataset/movielens/ratings.csv
[hadoop@ip-172-31-62-109 ~]$
```

hdfs fsck <path>

파일 시스템의 상태를 확인한다.

```
[hadoop@ip-172-31-62-109 ~]$ hdfs fsck /dataset/movielens/ratings.csv
Connecting to namenode via http://ip-172-31-62-109.ec2.internal:50070/fsck?ugi=hadoop&path=%2Fdataset%2Fmovielens%2Fratings.csv
FSCK started by hadoop (auth:SIMPLE) from /172.31.62.109 for path /dataset/movielens/ratings.csv at Tue Oct 27 10:43:19 UTC 2020
.Status: HEALTHY
 Total size:    533444411 B
 Total dirs:    0
 Total files:   1
 Total symlinks:                0
 Total blocks (validated):      4 (avg. block size 133361102 B)
 Minimally replicated blocks:   4 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       0 (0.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    1
 Average block replication:     1.0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)
 Number of data-nodes:          2
 Number of racks:               1
FSCK ended at Tue Oct 27 10:43:19 UTC 2020 in 9 milliseconds


The filesystem under path '/dataset/movielens/ratings.csv' is HEALTHY
[hadoop@ip-172-31-62-109 ~]$
```

- ratings.csv 파일의 전체 파일 크기는 얼마인가? 평균 블락 크기는 무엇인가? 블락의 갯수는 몇개이며, HDFS 에서의 기본 블락 크기는 얼마로 설정되어 있을까요?

> 전체 파일 크기 : 533444411 B 평균 블락 크기 : 133361102 B 블락의 갯수 : 4 HDFS 기본 블락 크기 : 128MB
>
> - 별도의 블락 크기를 설정하지 않았다면, HDFS의 기본 블락 크기는 128MB이다.

HDFS 블락 크기 설정

`hdfs dfs -Ddfs.block.size=[size] [OPTION]` 명령어를 통해 블럭사이즈를 변경 가능. 아래 결과는 블락 사이즈를 1MB로 변경한 결과 값이다. **기존 4개의 블락 갯수가 509개로 늘어난 것을 확인 할 수 있다.**

```
[hadoop@ip-172-31-62-109 ~]$ hdfs dfs -Ddfs.block.size=1048576 -put ml-20m/ratings.csv /
[hadoop@ip-172-31-62-109 ~]$ hdfs fsck /ratings.csv
Connecting to namenode via http://ip-172-31-62-109.ec2.internal:50070/fsck?ugi=hadoop&path=%2Fratings.csv
FSCK started by hadoop (auth:SIMPLE) from /172.31.62.109 for path /ratings.csv at Tue Oct 27 10:50:47 UTC 2020
.Status: HEALTHY
 Total size:    533444411 B
 Total dirs:    0
 Total files:   1
 Total symlinks:                0
 Total blocks (validated):      509 (avg. block size 1048024 B)
 Minimally replicated blocks:   509 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       0 (0.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    1
 Average block replication:     1.0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)
 Number of data-nodes:          2
 Number of racks:               1
FSCK ended at Tue Oct 27 10:50:47 UTC 2020 in 19 milliseconds


The filesystem under path '/ratings.csv' is HEALTHY
```

hdfs dfs -get <localsrc> <dst>

hdfs 에 존재하는 파일을 로컬 디스크로 복사한다.

```
[hadoop@ip-172-31-62-109 ~]$ hdfs dfs -get /dataset/movielens/ratings.csv /tmp/
[hadoop@ip-172-31-62-109 ~]$ ls /tmp/ratings.csv
/tmp/ratings.csv
[hadoop@ip-172-31-62-109 ~]$
```

hdfs dfs -cat URI

stdout에 파일 내용을 출력한다.

```
userId,movieId,rating,timestamp
1,2,3.5,1112486027
1,29,3.5,1112484676
1,32,3.5,1112484819
1,47,3.5,1112484727
1,50,3.5,1112484580
1,112,3.5,1094785740
1,151,4.0,1094785734
1,223,4.0,1112485573
1,253,4.0,1112484940
1,260,4.0,1112484826
1,293,4.0,1112484703
1,296,4.0,1112484767
1,318,4.0,1112484798
1,337,3.5,1094785709
1,367,3.5,1112485980
1,541,4.0,1112484603
1,589,3.5,1112485557
1,593,3.5,1112484661
1,653,3.0,1094785691
1,919,3.5,1094785621
1,924,3.5,1094785598
1,1009,3.5,1112486013
1,1036,4.0,1112485480
1,1079,4.0,1094785665
1,1080,3.5,1112485375
1,1089,3.5,1112484669
1,1090,4.0,1112485453
1,1097,4.0,1112485701
1,1136,3.5,1112484609
1,1193,3.5,1112484690
1,1196,4.5,1112484742
1,1198,4.5,1112484624
1,1200,4.0,1112484560
1,1201,3.0,1112484642
1,1208,3.5,1112484815
1,1214,4.0,1094785977
1,1215,4.0,1094786082
1,1217,3.5,1112484810
1,1219,4.0,1094785994
1,1222,3.5,1112484637
1,1240,4.0,1112485401
1,1243,3.0,1112485567
1,1246,3.5,1094785759
1,1249,4.0,1112485382
1,1258,4.0,1094785994
1,1259,4.0,1112485440
1,1261,3.5,1094786113
1,1262,3.5,1112484735
1,1266,4.0,1112485371
1,1278,4.0,1094785986
1,1291,3.5,1112485525
1,1304,3.0,1094785720
1,1321,4.0,1094786062
1,1333,4.0,1112484990
1,1348,3.5,1094786056
1,1350,3.5,1094786158
1,1358,4.0,1112485419
1,1370,3.0,1094785764
1,1374,4.0,1094785746
1,1387,4.0,1112484913
1,1525,3.0,1112486150
1,1584,3.5,1094785656
1,1750,3.5,1112486201
1,1848,3.5,1112486032
1,1920,3.5,1112486098
1,1967,4.0,1112485739
1,1994,3.5,1094786087
1,1997,3.5,1094786034
1,2021,4.0,1112485929
1,2100,4.0,1112485955
1,2118,4.0,1094786092
:
```

hdfs dfs -tail URI

파일의 끝부분을 출력한다.

```
[hadoop@ip-172-31-62-109 ~]$ hdfs dfs -put ml-20m/movies.csv /dataset/movielens/
[hadoop@ip-172-31-62-109 ~]$ hdfs dfs -tail /dataset/movielens/movies.csv
n puñado de besos (2014),Drama|Romance
131164,Vietnam in HD (2011),War
131166,WWII IN HD (2009),(no genres listed)
131168,Phoenix (2014),Drama
131170,Parallels (2015),Sci-Fi
131172,Closed Curtain (2013),(no genres listed)
131174,Gentlemen (2014),Drama|Romance|Thriller
131176,A Second Chance (2014),Drama
131180,Dead Rising: Watchtower (2015),Action|Horror|Thriller
131231,Standby (2014),Comedy|Romance
131237,What Men Talk About (2010),Comedy
131239,Three Quarter Moon (2011),Comedy|Drama
131241,Ants in the Pants (2000),Comedy|Romance
131243,Werner - Gekotzt wird später (2003),Animation|Comedy
131248,Brother Bear 2 (2006),Adventure|Animation|Children|Comedy|Fantasy
131250,No More School (2000),Comedy
131252,Forklift Driver Klaus: The First Day on the Job (2001),Comedy|Horror
131254,Kein Bund für's Leben (2007),Comedy
131256,Feuer Eis & Dosenbier (2002),Comedy
131258,The Pirates (2014),Adventure
131260,Rentun Ruusu (2001),(no genres listed)
131262,Innocence (2014),Adventure|Fantasy|Horror
[hadoop@ip-172-31-62-109 ~]$
```

hdfs dfs -df [-h] URI

hdfs에서 디스크 가용 용량을 보여준다.

```
[hadoop@ip-172-31-62-109 ~]$ hdfs dfs -df -h /
Filesystem                                Size   Used  Available  Use%
hdfs://ip-172-31-62-109.ec2.internal:8020  53.0 G  1.0 G     51.9 G    2%
[hadoop@ip-172-31-62-109 ~]$
```

hdfs dfs -cp <source> <dest>

hdfs내부 source경로에서 dest경로로 파일을 복제한다.

```
[hadoop@ip-172-31-62-109 ~]$ hdfs dfs -cp /dataset/movielens/*.csv /
hdcp: `/ratings.csv': File exists
fs[hadoop@ip-172-31-62-109 ~]$ hdfs dfs -ls /
Found 6 items
drwxr-xr-x   - hadoop hadoop          0 2020-10-27 10:41 /dataset
-rw-r--r--   1 hadoop hadoop    1377677 2020-10-27 10:57 /movies.csv
-rw-r--r--   1 hadoop hadoop  533444411 2020-10-27 10:50 /ratings.csv
drwxrwxrwt   - hdfs   hadoop          0 2020-10-27 10:35 /tmp
drwxr-xr-x   - hdfs   hadoop          0 2020-10-27 10:35 /user
drwxr-xr-x   - hdfs   hadoop          0 2020-10-27 10:35 /var
[hadoop@ip-172-31-62-109 ~]$
```

hdfs dfs -getfacl <path>

파일과 디렉토리의 Access Control 을 출력한다.

```
[hadoop@ip-172-31-62-109 ~]$ hdfs dfs -getfacl /movies.csv
# file: /movies.csv
# owner: hadoop
# group: hadoop
user::rw-
group::r--
other::r--

[hadoop@ip-172-31-62-109 ~]$
```

hdfs dfs -setrep <numReplicas> <path>

hdfs 파일 replica갯수를 바꿀 수 있다.

```
[hadoop@ip-172-31-62-109 ~]$ hdfs dfs -put ml-20m/tags.csv /dataset/movielens/
[hadoop@ip-172-31-62-109 ~]$ hdfs fsck /dataset/movielens/tags.csv
Connecting to namenode via http://ip-172-31-62-109.ec2.internal:50070/fsck?ugi=hadoop&path=%2Fdataset%2Fmovielens%2Ftags.csv
FSCK started by hadoop (auth:SIMPLE) from /172.31.62.109 for path /dataset/movielens/tags.csv at Tue Oct 27 10:59:01 UTC 2020
.Status: HEALTHY
 Total size:    16603996 B
 Total dirs:    0
 Total files:   1
 Total symlinks:                0
 Total blocks (validated):      1 (avg. block size 16603996 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:    1
 Average block replication:     1.0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)
 Number of data-nodes:          2
 Number of racks:               1
FSCK ended at Tue Oct 27 10:59:01 UTC 2020 in 1 milliseconds


The filesystem under path '/dataset/movielens/tags.csv' is HEALTHY
[hadoop@ip-172-31-62-109 ~]$ hdfs dfs -setrep 3 /dataset/movielens/tags.csv
Replication 3 set: /dataset/movielens/tags.csv
[hadoop@ip-172-31-62-109 ~]$ hdfs fsck /dataset/movielens/tags.csv
Connecting to namenode via http://ip-172-31-62-109.ec2.internal:50070/fsck?ugi=hadoop&path=%2Fdataset%2Fmovielens%2Ftags.csv
FSCK started by hadoop (auth:SIMPLE) from /172.31.62.109 for path /dataset/movielens/tags.csv at Tue Oct 27 11:02:26 UTC 2020
.
/dataset/movielens/tags.csv:  Under replicated BP-1349649011-172.31.62.109-1603794911083:blk_1073742340_1516. Target Replicas is 3 but f
ound 2 live replica(s), 0 decommissioned replica(s), 0 decommissioning replica(s).
Status: HEALTHY
 Total size:    16603996 B
 Total dirs:    0
 Total files:   1
 Total symlinks:                0
 Total blocks (validated):      1 (avg. block size 16603996 B)
 Minimally replicated blocks:   1 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       1 (100.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    1
 Average block replication:     2.0
 Corrupt blocks:                0
 Missing replicas:              1 (33.333332 %)
 Number of data-nodes:          2
 Number of racks:               1
FSCK ended at Tue Oct 27 11:02:26 UTC 2020 in 1 milliseconds


The filesystem under path '/dataset/movielens/tags.csv' is HEALTHY
[hadoop@ip-172-31-62-109 ~]$
```

파일의 replica 갯수를 바꾸는 명령어를 실행. 만일 replica 갯수를 3으로 설정한다면 어떻게 될까요? 그 이유를 서술해 주세요.

한개의 레플리카가 부족하다고 출력된다. 파일의 레플리카 갯수를 3개로 설정해 3개의 레플리카가 필요한데 현재는 2개의 레플리카만 가지고 있으므로 1개의 레플리카가 부족하다는 출력이 확인된다. (Misssing replicas) 이경우 복제본을 더 생성해 해당 문제를 해결 할 수 있다.

hdfs dfs -rm [-f] [-r|-R] URI

주어진 파일을 지움.

```
[hadoop@ip-172-31-62-109 ~]$ hdfs dfs -rm /ratings.csv /movies.csv
Deleted /ratings.csv
Deleted /movies.csv
[hadoop@ip-172-31-62-109 ~]$
```

hdfs dfsadmin -report

hdfs 시스템 상태를 알려주는 명령어.

```
[hadoop@ip-172-31-62-109 ~]$ hdfs dfsadmin -report
Configured Capacity: 56877875200 (52.97 GB)
Present Capacity: 56776208192 (52.88 GB)
DFS Remaining: 56201560064 (52.34 GB)
DFS Used: 574648128 (548.03 MB)
DFS Used%: 1.01%
Under replicated blocks: 1
Blocks with corrupt replicas: 0
Missing blocks: 0
Missing blocks (with replication factor 1): 0
Pending deletion blocks: 0

-------------------------------------------------
Live datanodes (2):

Name: 172.31.54.159:50010 (ip-172-31-54-159.ec2.internal)
Hostname: ip-172-31-54-159.ec2.internal
Decommission Status : Normal
Configured Capacity: 28438937600 (26.49 GB)
DFS Used: 421584884 (402.05 MB)
Non DFS Used: 50929676 (48.57 MB)
DFS Remaining: 27966423040 (26.05 GB)
DFS Used%: 1.48%
DFS Remaining%: 98.34%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Tue Oct 27 11:04:45 UTC 2020
Last Block Report: Tue Oct 27 10:35:21 UTC 2020


Name: 172.31.59.225:50010 (ip-172-31-59-225.ec2.internal)
Hostname: ip-172-31-59-225.ec2.internal
Decommission Status : Normal
Configured Capacity: 28438937600 (26.49 GB)
DFS Used: 153063244 (145.97 MB)
Non DFS Used: 50737332 (48.39 MB)
DFS Remaining: 28235137024 (26.30 GB)
DFS Used%: 0.54%
DFS Remaining%: 99.28%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 1
Last contact: Tue Oct 27 11:04:45 UTC 2020
Last Block Report: Tue Oct 27 10:35:21 UTC 2020


[hadoop@ip-172-31-62-109 ~]$
```