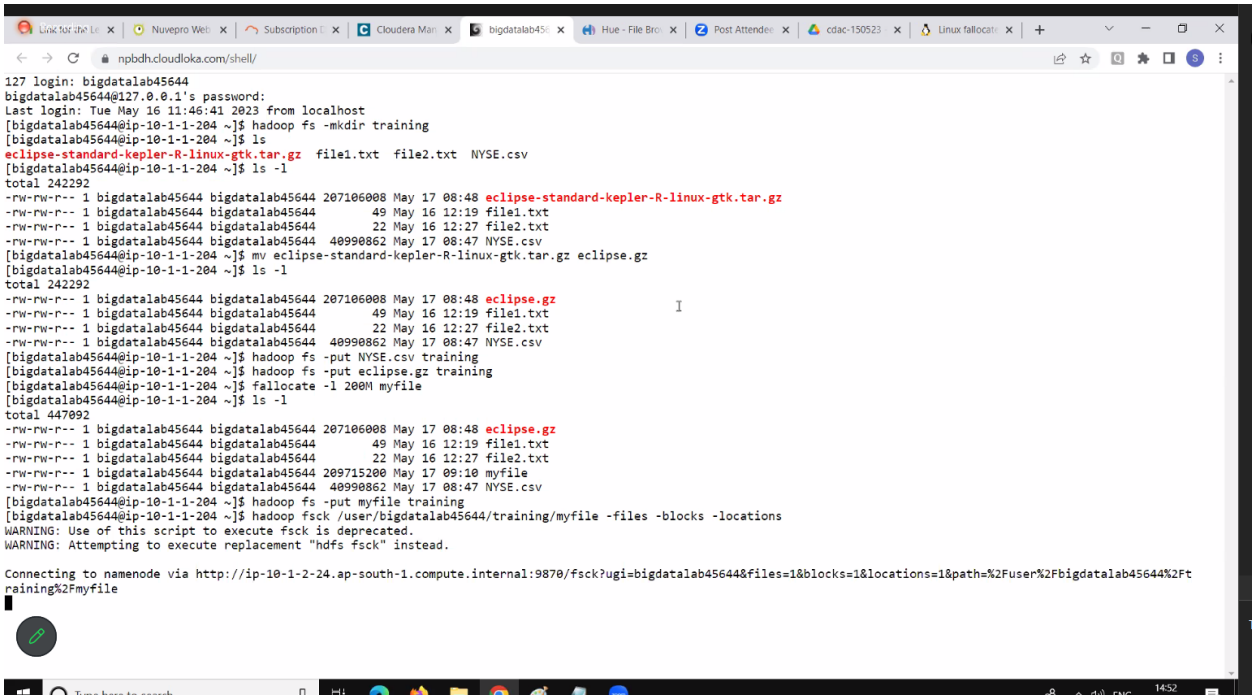
→

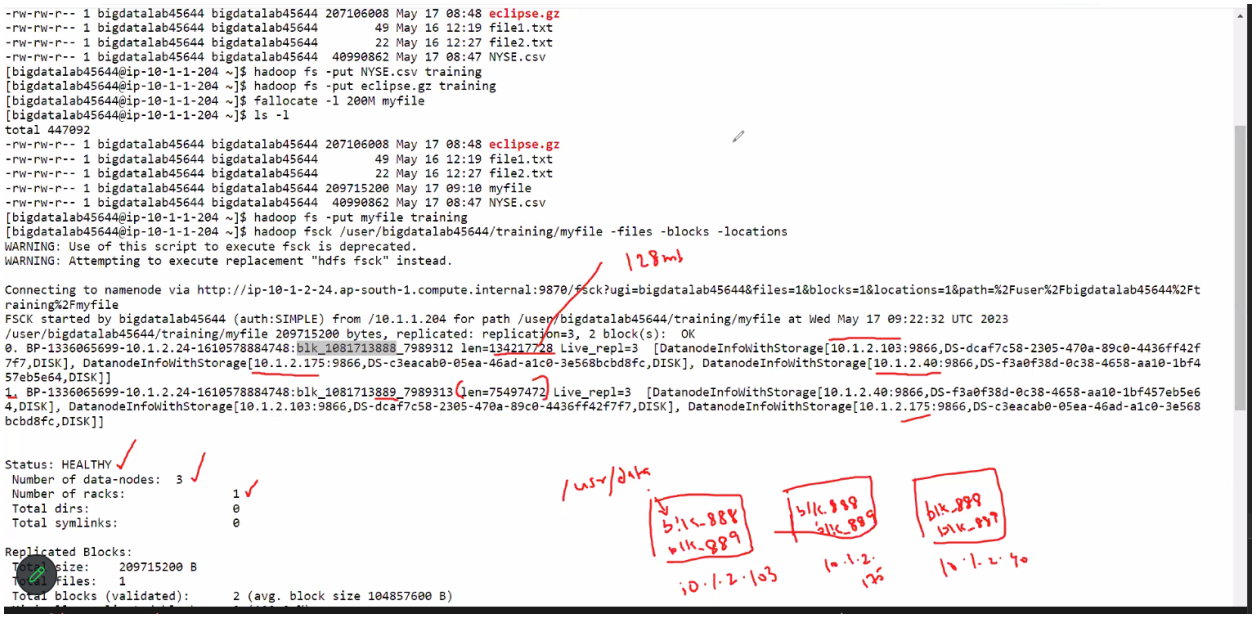



```
# create dir named training
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fs -mkdir training

[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fs -ls

Found 3 items
drwx------   - bigdatalab456422 bigdatalab456422          0
2023-05-16 14:00 .Trash
-rw-r--r--   3 bigdatalab456422 bigdatalab456422         50
2023-05-16 12:43 newfile.txt
drwxr-xr-x   - bigdatalab456422 bigdatalab456422          0
2023-05-17 08:37 training

# upload file using FTP

# rename eclipse file to eclipse.gz




# create 200M file using fallocate
[bigdatalab456422@ip-10-1-1-204 ~]$ fallocate -l 200M myfile
```

```
[bigdatalab456422@ip-10-1-1-204 ~]$ ll
total 244840
-rw-rw-r-- 1 bigdatalab456422 bigdatalab456422        50 May 16 12:19
file1.txt
-rw-rw-r-- 1 bigdatalab456422 bigdatalab456422        20 May 16 12:30
file2.txt
-rw-rw-r-- 1 bigdatalab456422 bigdatalab456422 209715200 May 17 09:16
myfile
-rw-rw-r-- 1 bigdatalab456422 bigdatalab456422  40990862 May 17 09:21
NYSE.csv

# using put command to put the NYSE.csv to hadoop fs
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fs -put NYSE.csv training
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fs -put myfile training
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fs -put eclipse.gz
training
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fs -ls training
Found 6 items
-rw-r--r--   3 bigdatalab456422 bigdatalab456422   40990862
2023-05-17 09:27 training/NYSE.csv
-rw-r--r--   3 bigdatalab456422 bigdatalab456422  207106008
2023-05-19 05:26 training/eclipse.gz
-rw-r--r--   4 bigdatalab456422 bigdatalab456422  209715200
2023-05-17 09:27 training/myfile
-rw-r--r--   3 bigdatalab456422 bigdatalab456422  524288000
2023-05-17 09:36 training/myfile2


# use hadoop fsck to check details of myfile file
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fsck training/myfile
-files -blocks -locations
WARNING: Use of this script to execute fsck is deprecated.
WARNING: Attempting to execute replacement "hdfs fsck" instead.

Connecting to namenode via
http://ip-10-1-2-24.ap-south-1.compute.internal:9870/fsck?ugi=bigdata
lab456422&files=1&blocks=1&locations=1&path=%2Fuser%2Fbigdatalab45642
2%2
Ftraining%2Fmyfile
FSCK started by bigdatalab456422 (auth:SIMPLE) from /10.1.1.204 for
path /user/bigdatalab456422/training/myfile at Wed May 17 09:28:35
UTC 2023
/user/bigdatalab456422/training/myfile 209715200 bytes, replicated:
replication=3, 2 block(s):  OK
```

```
0. BP-1336065699-10.1.2.24-1610578884748:blk_1081713995_7989428
len=134217728 Live_repl=3
[DatanodeInfoWithStorage[10.1.2.175:9866,DS-c3eacab0-05ea-46ad-a1c0-3
e568bcbd
8fc,DISK],
DatanodeInfoWithStorage[10.1.2.40:9866,DS-f3a0f38d-0c38-4658-aa10-1bf
457eb5e64,DISK],
DatanodeInfoWithStorage[10.1.2.103:9866,DS-dcaf7c58-2305-470a-89c0-44
36
ff42f7f7,DISK]]
1. BP-1336065699-10.1.2.24-1610578884748:blk_1081713996_7989429
len=75497472 Live_repl=3
[DatanodeInfoWithStorage[10.1.2.175:9866,DS-c3eacab0-05ea-46ad-a1c0-3
e568bcbd8
fc,DISK],
DatanodeInfoWithStorage[10.1.2.40:9866,DS-f3a0f38d-0c38-4658-aa10-1bf
457eb5e64,DISK],
DatanodeInfoWithStorage[10.1.2.103:9866,DS-dcaf7c58-2305-470a-89c0-44
36f
f42f7f7,DISK]]


Status: HEALTHY
Number of data-nodes:  3
Number of racks:               1
Total dirs:                    0
Total symlinks:                0

Replicated Blocks:
Total size:    209715200 B
Total files:   1
Total blocks (validated):      2 (avg. block size 104857600 B)
Minimally replicated blocks:   2 (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
Missing blocks:                0
Corrupt blocks:                0
Missing replicas:              0 (0.0 %)
Blocks queued for replication: 0

Erasure Coded Block Groups:
Total size:    0 B
```

```
Total files:   0
Total block groups (validated):        0
Minimally erasure-coded block groups:  0
Over-erasure-coded block groups:       0
Under-erasure-coded block groups:      0
Unsatisfactory placement block groups: 0
Average block group size:      0.0
Missing block groups:          0
Corrupt block groups:          0
Missing internal blocks:       0
Blocks queued for replication: 0
FSCK ended at Wed May 17 09:28:35 UTC 2023 in 1 milliseconds


The filesystem under path '/user/bigdatalab456422/training/myfile' is
HEALTHY


# create another file myfile2 of 500Mb using fallocate
[bigdatalab456422@ip-10-1-1-204 ~]$ fallocate -l 500M myfile2
[bigdatalab456422@ip-10-1-1-204 ~]$ ll
total 756844
-rw-rw-r-- 1 bigdatalab456422 bigdatalab456422        50 May 16 12:19
file1.txt
-rw-rw-r-- 1 bigdatalab456422 bigdatalab456422        20 May 16 12:30
file2.txt
-rw-rw-r-- 1 bigdatalab456422 bigdatalab456422 209715200 May 17 09:16
myfile
-rw-rw-r-- 1 bigdatalab456422 bigdatalab456422 524288000 May 17 09:35
myfile2
-rw-rw-r-- 1 bigdatalab456422 bigdatalab456422  40990862 May 17 09:21
NYSE.csv


# now put it into hadoop fs training folder
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fs -put NYSE.csv training
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fs -put myfile training




# check fsck for myfile file
```




```
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fsck training/myfile
-files -blocks -locations
WARNING: Use of this script to execute fsck is deprecated.
WARNING: Attempting to execute replacement "hdfs fsck" instead.

Connecting to namenode via
http://ip-10-1-2-24.ap-south-1.compute.internal:9870/fsck?ugi=bigdata
lab456422&files=1&blocks=1&locations=1&path=%2Fuser%2Fbigdatalab45642
2%2
Ftraining%2Fmyfile
FSCK started by bigdatalab456422 (auth:SIMPLE) from /10.1.1.204 for
path /user/bigdatalab456422/training/myfile at Wed May 17 09:28:35
UTC 2023
/user/bigdatalab456422/training/myfile 209715200 bytes, replicated:
replication=3, 2 block(s):  OK
0. BP-1336065699-10.1.2.24-1610578884748:blk_1081713995_7989428
len=134217728 Live_repl=3
[DatanodeInfoWithStorage[10.1.2.175:9866,DS-c3eacab0-05ea-46ad-a1c0-3
e568bcbd
8fc,DISK],
DatanodeInfoWithStorage[10.1.2.40:9866,DS-f3a0f38d-0c38-4658-aa10-1bf
457eb5e64,DISK],
DatanodeInfoWithStorage[10.1.2.103:9866,DS-dcaf7c58-2305-470a-89c0-44
36
ff42f7f7,DISK]]
1. BP-1336065699-10.1.2.24-1610578884748:blk_1081713996_7989429
len=75497472 Live_repl=3
```

```
[DatanodeInfoWithStorage[10.1.2.175:9866,DS-c3eacab0-05ea-46ad-a1c0-3
e568bcbd8
fc,DISK],
DatanodeInfoWithStorage[10.1.2.40:9866,DS-f3a0f38d-0c38-4658-aa10-1bf
457eb5e64,DISK],
DatanodeInfoWithStorage[10.1.2.103:9866,DS-dcaf7c58-2305-470a-89c0-44
36f
f42f7f7,DISK]]


Status: HEALTHY
Number of data-nodes:  3
Number of racks:               1
Total dirs:                    0
Total symlinks:                0

Replicated Blocks:
Total size:    209715200 B
Total files:   1
Total blocks (validated):      2 (avg. block size 104857600 B)
Minimally replicated blocks:   2 (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
Missing blocks:                0
Corrupt blocks:                0
Missing replicas:              0 (0.0 %)
Blocks queued for replication: 0

Erasure Coded Block Groups:
Total size:    0 B
Total files:   0
Total block groups (validated):        0
Minimally erasure-coded block groups:  0
Over-erasure-coded block groups:       0
Under-erasure-coded block groups:      0
Unsatisfactory placement block groups: 0
Average block group size:      0.0
Missing block groups:          0
Corrupt block groups:          0
Missing internal blocks:       0
Blocks queued for replication: 0
FSCK ended at Wed May 17 09:28:35 UTC 2023 in 1 milliseconds
```

```
The filesystem under path '/user/bigdatalab456422/training/myfile' is
HEALTHY


# create another file myfile2 of 500Mb using fallocate
[bigdatalab456422@ip-10-1-1-204 ~]$ fallocate -l 500M myfile2
[bigdatalab456422@ip-10-1-1-204 ~]$ ll
total 756844
-rw-rw-r-- 1 bigdatalab456422 bigdatalab456422        50 May 16 12:19
file1.txt
-rw-rw-r-- 1 bigdatalab456422 bigdatalab456422        20 May 16 12:30
file2.txt
-rw-rw-r-- 1 bigdatalab456422 bigdatalab456422 209715200 May 17 09:16
myfile
-rw-rw-r-- 1 bigdatalab456422 bigdatalab456422 524288000 May 17 09:35
myfile2
-rw-rw-r-- 1 bigdatalab456422 bigdatalab456422  40990862 May 17 09:21
NYSE.csv


# now put it into hadoop fs training folder
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fs -put myfile2 training

# to check fsck for myfile2
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fsck training/myfile2
-files -blocks -locations
WARNING: Use of this script to execute fsck is deprecated.
WARNING: Attempting to execute replacement "hdfs fsck" instead.

Connecting to namenode via
http://ip-10-1-2-24.ap-south-1.compute.internal:9870/fsck?ugi=bigdata
lab456422&files=1&blocks=1&locations=1&path=%2Fuser%2Fbigdatalab45642
2%2
Ftraining%2Fmyfile2
FSCK started by bigdatalab456422 (auth:SIMPLE) from /10.1.1.204 for
path /user/bigdatalab456422/training/myfile2 at Wed May 17 09:38:11
UTC 2023
/user/bigdatalab456422/training/myfile2 524288000 bytes, replicated:
replication=3, 4 block(s):  OK
0. BP-1336065699-10.1.2.24-1610578884748:blk_1081714135_7989571
len=134217728 Live_repl=3
[DatanodeInfoWithStorage[10.1.2.103:9866,DS-dcaf7c58-2305-470a-89c0-4
436ff42f
```

```
7f7,DISK],
DatanodeInfoWithStorage[10.1.2.175:9866,DS-c3eacab0-05ea-46ad-a1c0-3e
568bcbd8fc,DISK],
DatanodeInfoWithStorage[10.1.2.40:9866,DS-f3a0f38d-0c38-4658-aa10-1bf
4
57eb5e64,DISK]]
1. BP-1336065699-10.1.2.24-1610578884748:blk_1081714136_7989572
len=134217728 Live_repl=3
[DatanodeInfoWithStorage[10.1.2.175:9866,DS-c3eacab0-05ea-46ad-a1c0-3
e568bcbd
8fc,DISK],
DatanodeInfoWithStorage[10.1.2.103:9866,DS-dcaf7c58-2305-470a-89c0-44
36ff42f7f7,DISK],
DatanodeInfoWithStorage[10.1.2.40:9866,DS-f3a0f38d-0c38-4658-aa10-1bf
4
57eb5e64,DISK]]
2. BP-1336065699-10.1.2.24-1610578884748:blk_1081714137_7989573
len=134217728 Live_repl=3
[DatanodeInfoWithStorage[10.1.2.103:9866,DS-dcaf7c58-2305-470a-89c0-4
436ff42f
7f7,DISK],
DatanodeInfoWithStorage[10.1.2.40:9866,DS-f3a0f38d-0c38-4658-aa10-1bf
457eb5e64,DISK],
DatanodeInfoWithStorage[10.1.2.175:9866,DS-c3eacab0-05ea-46ad-a1c0-3e
56
8bcbd8fc,DISK]]
3. BP-1336065699-10.1.2.24-1610578884748:blk_1081714138_7989574
len=121634816 Live_repl=3
[DatanodeInfoWithStorage[10.1.2.103:9866,DS-dcaf7c58-2305-470a-89c0-4
436ff42f
7f7,DISK],
DatanodeInfoWithStorage[10.1.2.175:9866,DS-c3eacab0-05ea-46ad-a1c0-3e
568bcbd8fc,DISK],
DatanodeInfoWithStorage[10.1.2.40:9866,DS-f3a0f38d-0c38-4658-aa10-1bf
4
57eb5e64,DISK]]


Status: HEALTHY
Number of data-nodes:  3
Number of racks:               1
Total dirs:                    0
Total symlinks:                0

Replicated Blocks:
```

```
Total size:    524288000 B
Total files:   1
Total blocks (validated):      4 (avg. block size 131072000 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:    3
Average block replication:     3.0
Missing blocks:                0
Corrupt blocks:                0
Missing replicas:              0 (0.0 %)
Blocks queued for replication: 0

Erasure Coded Block Groups:
Total size:    0 B
Total files:   0
Total block groups (validated):        0
Minimally erasure-coded block groups:  0
Over-erasure-coded block groups:       0
Under-erasure-coded block groups:      0
Unsatisfactory placement block groups: 0
Average block group size:      0.0
Missing block groups:          0
Corrupt block groups:          0
Missing internal blocks:       0
Blocks queued for replication: 0
FSCK ended at Wed May 17 09:38:11 UTC 2023 in 0 milliseconds


The filesystem under path '/user/bigdatalab456422/training/myfile2'
is HEALTHY


# to change replication factor
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fs -setrep 4
training/myfile
Replication 4 set: training/myfile
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fs -ls training
Found 3 items
-rw-r--r--   3 bigdatalab456422 bigdatalab456422   40990862
2023-05-17 09:27 training/NYSE.csv
-rw-r--r--   4 bigdatalab456422 bigdatalab456422  209715200
2023-05-17 09:27 training/myfile
```

```
-rw-r--r--   3 bigdatalab456422 bigdatalab456422  524288000
2023-05-17 09:36 training/myfile2


# now check fsck for myfile again
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fsck training/myfile
-files -blocks -locations
WARNING: Use of this script to execute fsck is deprecated.
WARNING: Attempting to execute replacement "hdfs fsck" instead.

Connecting to namenode via
http://ip-10-1-2-24.ap-south-1.compute.internal:9870/fsck?ugi=bigdata
lab456422&files=1&blocks=1&locations=1&path=%2Fuser%2Fbigdatalab45642
2%2
Ftraining%2Fmyfile
FSCK started by bigdatalab456422 (auth:SIMPLE) from /10.1.1.204 for
path /user/bigdatalab456422/training/myfile at Wed May 17 09:44:30
UTC 2023
/user/bigdatalab456422/training/myfile 209715200 bytes, replicated:
replication=4, 2 block(s):  Under replicated
BP-1336065699-10.1.2.24-1610578884748:blk_1081713995_79
89428. Target Replicas is 4 but found 3 live replica(s), 0
decommissioned replica(s), 0 decommissioning replica(s).
Under replicated
BP-1336065699-10.1.2.24-1610578884748:blk_1081713996_7989429. Target
Replicas is 4 but found 3 live replica(s), 0 decommissioned
replica(s), 0 decommissioning replica(s).
0. BP-1336065699-10.1.2.24-1610578884748:blk_1081713995_7989428
len=134217728 Live_repl=3
[DatanodeInfoWithStorage[10.1.2.175:9866,DS-c3eacab0-05ea-46ad-a1c0-3
e568bcbd
8fc,DISK],
DatanodeInfoWithStorage[10.1.2.40:9866,DS-f3a0f38d-0c38-4658-aa10-1bf
457eb5e64,DISK],
DatanodeInfoWithStorage[10.1.2.103:9866,DS-dcaf7c58-2305-470a-89c0-44
36
ff42f7f7,DISK]]
1. BP-1336065699-10.1.2.24-1610578884748:blk_1081713996_7989429
len=75497472 Live_repl=3
[DatanodeInfoWithStorage[10.1.2.175:9866,DS-c3eacab0-05ea-46ad-a1c0-3
e568bcbd8
fc,DISK],
DatanodeInfoWithStorage[10.1.2.40:9866,DS-f3a0f38d-0c38-4658-aa10-1bf
457eb5e64,DISK],
```

```
DatanodeInfoWithStorage[10.1.2.103:9866,DS-dcaf7c58-2305-470a-89c0-44
36f
f42f7f7,DISK]]


Status: HEALTHY
Number of data-nodes:  3
Number of racks:               1
Total dirs:                    0
Total symlinks:                0

Replicated Blocks:
Total size:    209715200 B
Total files:   1
Total blocks (validated):      2 (avg. block size 104857600 B)
Minimally replicated blocks:   2 (100.0 %)
Over-replicated blocks:        0 (0.0 %)
Under-replicated blocks:       2 (100.0 %)
Mis-replicated blocks:         0 (0.0 %)
Default replication factor:    3
Average block replication:     3.0
Missing blocks:                0
Corrupt blocks:                0
Missing replicas:              2 (25.0 %)
Blocks queued for replication: 0

Erasure Coded Block Groups:
Total size:    0 B
Total files:   0
Total block groups (validated):        0
Minimally erasure-coded block groups:  0
Over-erasure-coded block groups:       0
Under-erasure-coded block groups:      0
Unsatisfactory placement block groups: 0
Average block group size:      0.0
Missing block groups:          0
Corrupt block groups:          0
Missing internal blocks:       0
Blocks queued for replication: 0
FSCK ended at Wed May 17 09:44:30 UTC 2023 in 0 milliseconds


The filesystem under path '/user/bigdatalab456422/training/myfile' is
HEALTHY
```

Conf files(for default values):

1. core-default..xml
2. hdfs–default.xml for using default values
   https://hadoop.apache.org/docs/r2.4.1/hadoop-project-dist/hadoop-hdfs/hdfs-default.xml

3. mapred-default.xml
4. yarn-default.xml

COnf files (For changes/modified values):

a. core-site.xml
   i. <name>hadoop.tmp.dir</name>
      <value>/app/hadoop/tmp</value>
      <description>A base for other temporary directories.</ description>
   ii. <name>fs.default.name</name>
      <value>hdfs://localhost:54310</value>
      <description>The name of the default file system. A URI whose scheme and authority determine the FileSystem implementation. The uri's scheme determines the config property (fs.SCHEME.impl) naming the FileSystem implementation class. The uri's authority is used to determine the host, port, etc. for a filesystem.</description>
b. mapred-site.xml
   i. <name>mapred.job.tracker</name>
      <value>localhost:54311</value>
      <description>The host and port that the MapReduce job tracker runs at. If "local", then jobs are run in-process as a single map and reduce task.</description>
c. hdfs-site.xml
   i. <name>dfs.replication</name>
      <value>1</value>
      <description>Default block replication. The actual number of replications can be specified when the file is created. The default is used if replication is not specified in create time.</description>
   ii. <name>dfs.blocksize</name>
      <value>27M</value>
   iii. <name>dfs.namenode.name.dir</name>
      <value>file:/usr/local/hadoop_store/hdfs/namenode</value>
   iv. <name>dfs.datanode.data.dir</name>
      <value>file:/usr/local/hadoop_store/hdfs/datanode</value>

To start hadoop system

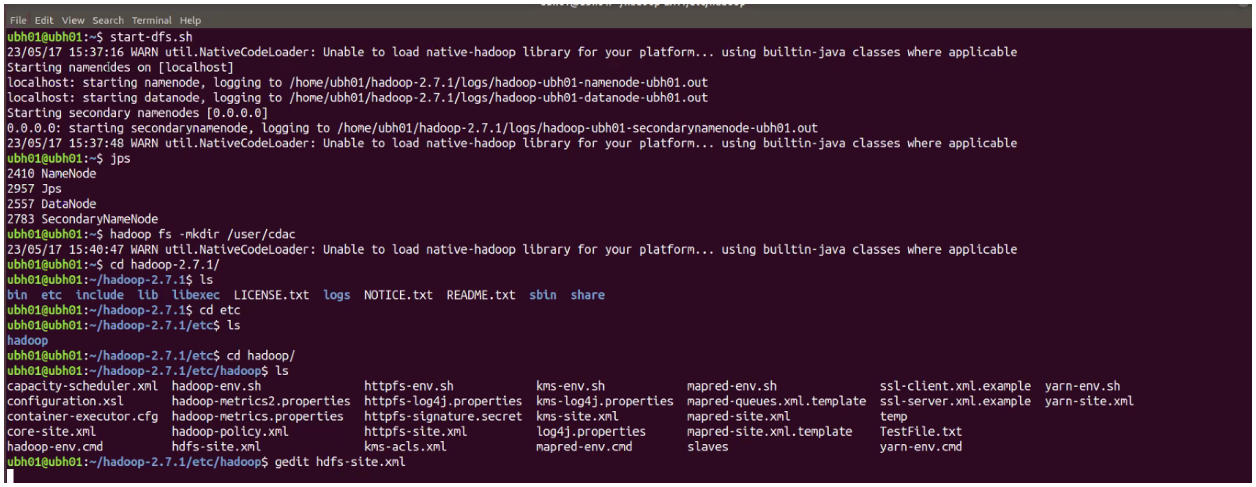
```
$ start-dfs.sh
```

Jps stands for java processes

```
$ jps
```

To open webview for hadoop system, use this

```
localhost:50070
```



→ Map task : filters data
→ Reduce task : process filtered data

→ HDFS: NameNode + DataNode [Read+Write]

→ YARN: ResourceManager (RM) + NodeManager(NM) [Process]

→ Resource manager (RM)

a. works with nameNode to handle query,
b. schedules jobs
c. Starts separate instances of temporary service (Application master) for each of the query/job received by RM from client

→ NodeManager (NM)

a. works on slave node,
b. communicates with Resource Manager

→ Application Master (AM)

a. Manages containers for each node
b. Checks details of blocks from name node
c. Shares task (Mapper code) to node manager to process

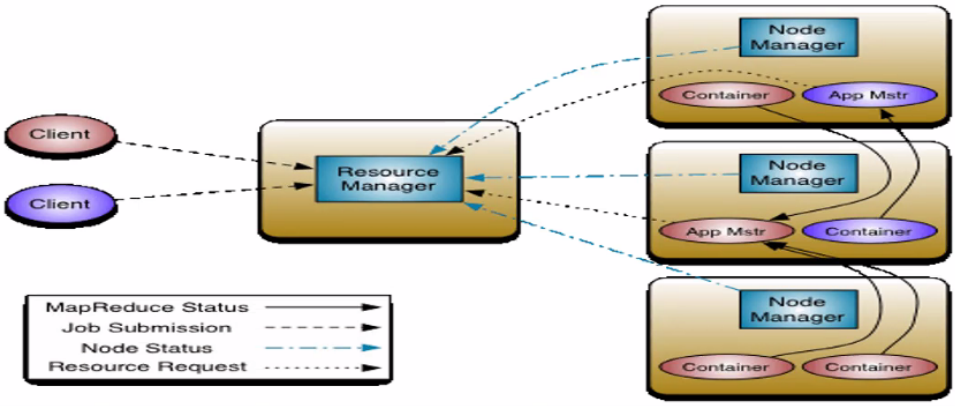
YARN High Level Architecture

Resource Manager: Runs on master node, Global resource scheduler, Independent system resource in application computing

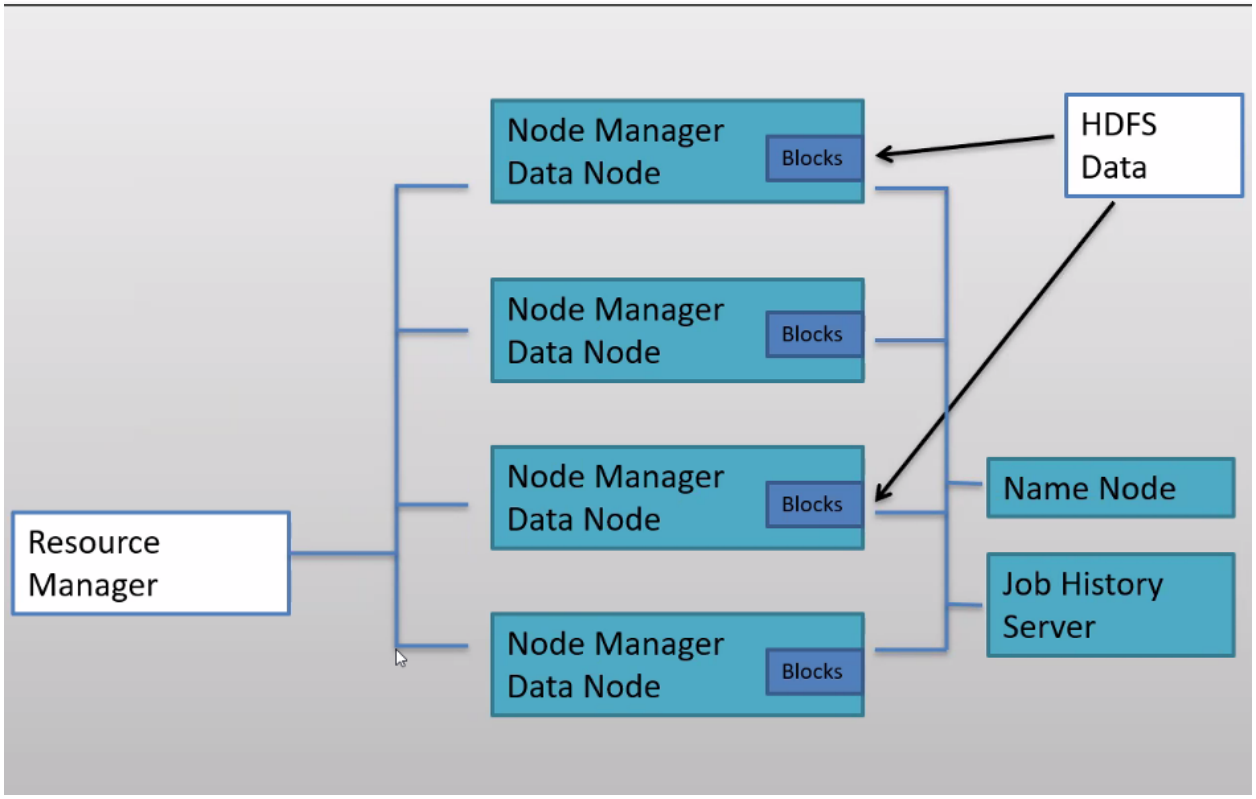
Node Manager: Runs on slave nodes and communicates with resource manager

Containers: Created by the resource manager upon request, allocate resources on slave node, application run one or more containers

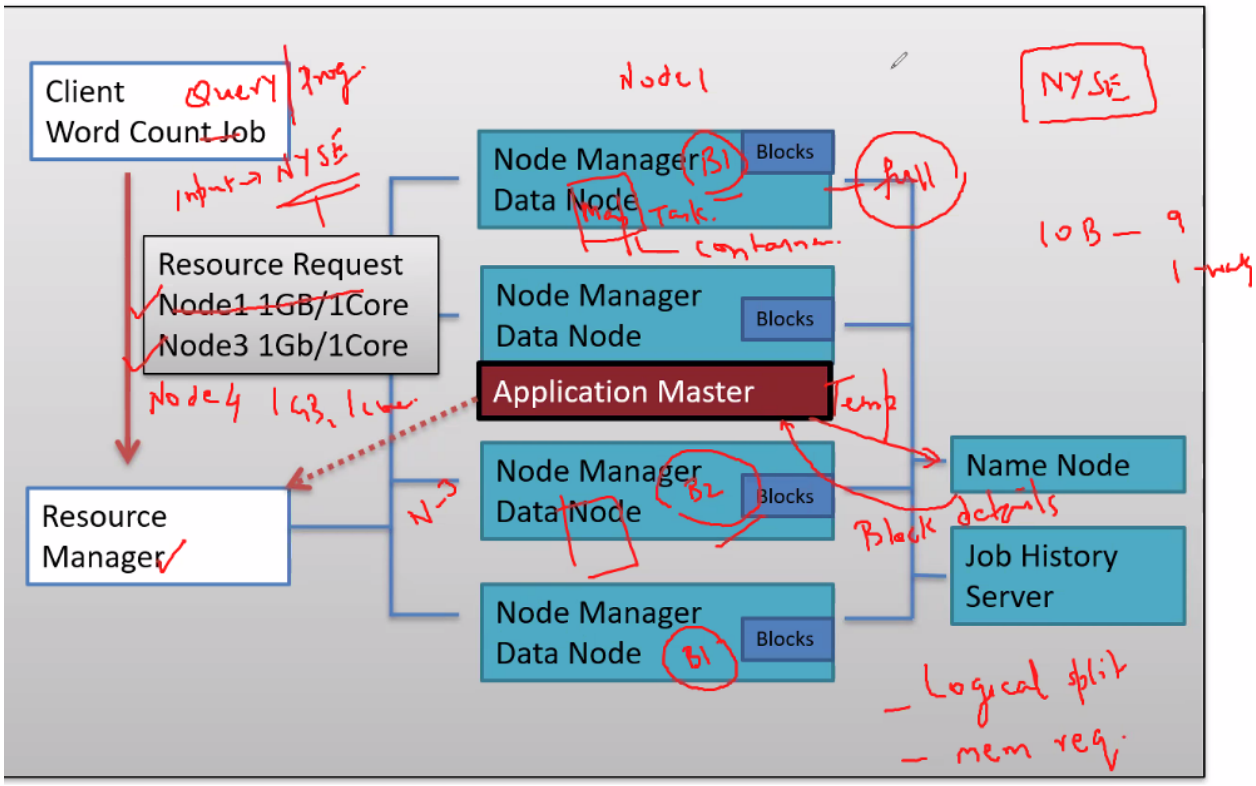
Application Master: One per application, application specific, run in container, request more containers to run application tasks



MR Job lifecycle on Yarn Cluster (cont'd)



MR Job lifecycle on Yarn Cluster (cont'd)



→ flow of job in YARN cluster

a. Job/query starts from client
b. YARN keeps Job History server keeps track of jobs being launched, but job history server is optional
c. Application master actually runs the job/query
d. YARN launches separate application managers for each job/query
e. Application Manager approaches NodeManager for block details
f. NameNode provides metadata/block details to Application manager
g. Now logical split is evaluated, and required resources(memory) is calculated by Application Master
h. Based on resources calculated by ApplicationManager, ApplicationManager requests resources from ResourceManager

i. YARN job/query reaches RM at NameNode to launch containers with a temporary service i.e. application manager which will run job/query

j. ApplicationManager sends the MapperCode to NodeManager on the DataNode where data is located, so that data is processed by MapperCode