Assignment:

Hdfs task-1

1.What is the Namenode's URI and which file is it configured in?

Namenodes URI is 9000, it is configured in coresite.xml

2. Where on a local file system will Namenode store its image and which file is it configured in?

Hadoop-env.sh

3. Where on a local file system will Datanode store its blocks and which file is it configured in?

Hadoop-env.sh

4. What is the block replication and which file is it configured in?

Block replication stores large amount of data in sequence of blocks around cluster.

Perform:

Start HDFS and verify that it's running

Check the entire filesystem for inconsistencies/problems

```
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.6$ bin/hdfs fsck /exercise1
Connecting to namenode via http://localhost:9870/fsck?ugi=hduser&path=%2Fexercise1
FSCK started by hduser (auth:SIMPLE) from /127.0.0.1 for path /exercise1 at Wed Jun 02 17:26:01 IST 2021


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

Replicated Blocks:
 Total size:     743 B
 Total files:    1
 Total blocks (validated):      1 (avg. block size 743 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
 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 Jun 02 17:26:01 IST 2021 in 27 milliseconds


The filesystem under path '/exercise1' is HEALTHY
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.6$
```

Copy deckofcards.txt back to local file system and name it deckofcards.copy.txt

```
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.6$ bin/hadoop fs -get /exercise1/deckofcards.txt /home/hduser/sunill
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.6$
```

Copy deckofcards.txt to deckofcardsCopy.txt

```
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.0$ bin/hadoop distcp /exercise1/deckofcards.txt /new_fol/copydeckofcards.txt
2021-06-02 17:02:34,431 INFO tools.DistCp: Input Options: DistCpOptions{atomicCommit=false, syncFolder=false, deleteMissing=false, ignoreFailures=false, overwrite=false
, append=false, useDiff=false, useRdiff=false, fromSnapshot=null, toSnapshot=null, skipCRC=false, blocking=true, numListstatusThreads=0, maxMaps=20, mapBandwidth=0.0, c
opyStrategy='uniformsize', preserveStatus=[BLOCKSIZE], atomicWorkPath=null, logPath=null, sourceFileListing=null, sourcePaths=[/exercise1/deckofcards.txt], targetPath=/
new_fol/copydeckofcards.txt, filtersFile='null', blocksPerChunk=0, copyBufferSize=8192, verboseLog=false, directWrite=false}, sourcePaths=[/exercise1/deckofcards.txt],
targetPathExists=false, preserveRawXattrsfalse
2021-06-02 17:02:34,958 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-06-02 17:02:40,236 INFO tools.SimpleCopyListing: Paths (files+dirs) cnt = 1; dirCnt = 0
2021-06-02 17:02:40,237 INFO tools.SimpleCopyListing: Build file listing completed.
2021-06-02 17:02:40,240 INFO Configuration.deprecation: io.sort.mb is deprecated. Instead, use mapreduce.task.io.sort.mb
2021-06-02 17:02:40,241 INFO Configuration.deprecation: io.sort.factor is deprecated. Instead, use mapreduce.task.io.sort.factor
2021-06-02 17:02:40,981 INFO tools.DistCp: Number of paths in the copy list: 1
2021-06-02 17:02:41,151 INFO tools.DistCp: Number of paths in the copy list: 1
2021-06-02 17:02:41,183 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2021-06-02 17:02:43,397 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hduser/.staging/job_1622630291232_0001
2021-06-02 17:02:44,122 INFO mapreduce.JobSubmitter: number of splits:1
2021-06-02 17:02:46,471 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1622630291232_0001
2021-06-02 17:02:46,474 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-06-02 17:02:47,657 INFO conf.Configuration: resource-types.xml not found
2021-06-02 17:02:47,658 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-06-02 17:02:51,037 INFO impl.YarnClientImpl: Submitted application application_1622630291232_0001
2021-06-02 17:02:51,483 INFO mapreduce.Job: The url to track the job: http://LAPTOP-C8CQB2TI.localdomain:8088/proxy/application_1622630291232_0001/
2021-06-02 17:02:51,484 INFO tools.DistCp: DistCp job-id: job_1622630291232_0001
2021-06-02 17:02:51,504 INFO mapreduce.Job: Running job: job_1622630291232_0001
2021-06-02 17:04:32,235 INFO mapreduce.Job: Job job_1622630291232_0001 running in uber mode : false
2021-06-02 17:04:32,283 INFO mapreduce.Job:  map 0% reduce 0%
2021-06-02 17:05:28,217 INFO mapreduce.Job:  map 100% reduce 0%
2021-06-02 17:05:36,410 INFO mapreduce.Job: Job job_1622630291232_0001 completed successfully
2021-06-02 17:05:36,769 INFO mapreduce.Job: Counters: 37
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=267605
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1130
                HDFS: Number of bytes written=743
                HDFS: Number of read operations=16
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=5
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
```

```
                HDFS: Number of bytes written=743
                HDFS: Number of read operations=16
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=5
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=49089
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=49089
                Total vcore-milliseconds taken by all map tasks=49089
                Total megabyte-milliseconds taken by all map tasks=50267136
        Map-Reduce Framework
                Map input records=1
                Map output records=0
                Input split bytes=136
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=94
                CPU time spent (ms)=2080
                Physical memory (bytes) snapshot=165036032
                Virtual memory (bytes) snapshot=1674669531136
                Total committed heap usage (bytes)=69206016
                Peak Map Physical memory (bytes)=165036032
                Peak Map Virtual memory (bytes)=1674669531136
        File Input Format Counters
                Bytes Read=251
        File Output Format Counters
                Bytes Written=0
        DistCp Counters
                Bandwidth in Bytes=371
                Bytes Copied=743
                Bytes Expected=743
                Files Copied=1
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.0$ bin/hadoop fs -ls /exercise1
Found 1 items
-rw-r--r--   1 hduser supergroup        743 2021-06-02 16:38 /exercise1/deckofcards.txt
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.0$ bin/hadoop fs -ls /new_fol
Found 2 items
-rw-r--r--   1 hduser supergroup      28629 2021-05-25 19:13 /new_fol/copy_titanic.csv
-rw-r--r--   1 hduser supergroup        743 2021-06-02 17:05 /new_fol/copydeckofcards.txt
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.0$
```

Upload GitHub repo sample_data/deckofcards.txt to HDFS under /exercise1 directory

```
drwxr-xr-x   - hduser supergroup          0 2021-05-25 19:13 /sunil
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.0$ bin/hdfs dfs -copyFromLocal /home/hduser/deckofcards.txt /exercise1
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.0$ bin/hdfs dfs -ls /exercise1
Found 1 items
-rw-r--r--   1 hduser supergroup          743 2021-06-02 16:38 /exercise1/deckofcards.txt
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.0$
```

Create a new directory /exercise1 on HDFS

```
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.0$ bin/hdfs dfs -ls /sunil
Found 1 items
-rw-r--r--   1 hduser supergroup        28629 2021-05-25 19:13 /sunil/titanic_test.csv
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.0$ bin/hdfs dfs -mkdir /exercise1
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.0$ bin/hdfs dfs -ls /
Found 3 items
drwxr-xr-x   - hduser supergroup          0 2021-06-02 16:18 /exercise1
drwxr-xr-x   - hduser supergroup          0 2021-05-27 15:57 /new_fol
drwxr-xr-x   - hduser supergroup          0 2021-05-25 19:13 /sunil
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.0$
```

Delete the /exercise1 directory from HDFS

```
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.0$ bin/hdfs dfs -ls /
Found 4 items
drwxr-xr-x   - hduser supergroup          0 2021-06-02 17:27 /exercise1
drwxr-xr-x   - hduser supergroup          0 2021-06-02 17:05 /new_fol
drwxr-xr-x   - hduser supergroup          0 2021-06-02 16:53 /sunil
drwx------   - hduser supergroup          0 2021-06-02 17:02 /tmp
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.0$ bin/hdfs dfs -rm -r /exercise1
Deleted /exercise1
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.0$ bin/hdfs dfs -ls /
Found 3 items
drwxr-xr-x   - hduser supergroup          0 2021-06-02 17:05 /new_fol
drwxr-xr-x   - hduser supergroup          0 2021-06-02 16:53 /sunil
drwx------   - hduser supergroup          0 2021-06-02 17:02 /tmp
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.0$
```

Delete deckofcards.txt from HDFS

```
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.6$ bin/hdfs dfs -ls /exercise1
Found 1 items
-rw-r--r--   1 hduser supergroup        743 2021-06-02 16:38 /exercise1/deckofcards.txt
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.6$ bin/hdfs dfs -rm /exercise1/deckofcards.txt
Deleted /exercise1/deckofcards.txt
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.6$ bin/hdfs dfs -ls /exercise1
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.6$
```

Print the first 25 lines to the screen from deckofcards.txt on HDFS

```
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.6$ bin/hdfs dfs -cat /exercise1/deckofcards.txt | head -n 25
BLACK|SPADE|2
BLACK|SPADE|3
BLACK|SPADE|4
BLACK|SPADE|5
BLACK|SPADE|6
BLACK|SPADE|7
BLACK|SPADE|8
BLACK|SPADE|9
BLACK|SPADE|10
BLACK|SPADE|J
BLACK|SPADE|Q
BLACK|SPADE|K
BLACK|SPADE|A
BLACK|CLUB|2
BLACK|CLUB|3
BLACK|CLUB|4
BLACK|CLUB|5
BLACK|CLUB|6
BLACK|CLUB|7
BLACK|CLUB|8
BLACK|CLUB|9
BLACK|CLUB|10
BLACK|CLUB|J
BLACK|CLUB|Q
BLACK|CLUB|K
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.6$
```

View the content of the /exercise1 directory

```
hduser@LAPTOP-C8CQB2TI:~/hadoop/hadoop-3.3.6$ bin/hdfs dfs -cat /exercise1/deckofcards.txt
BLACK|SPADE|2
BLACK|SPADE|3
BLACK|SPADE|4
BLACK|SPADE|5
BLACK|SPADE|6
BLACK|SPADE|7
BLACK|SPADE|8
BLACK|SPADE|9
BLACK|SPADE|10
BLACK|SPADE|J
BLACK|SPADE|Q
BLACK|SPADE|K
BLACK|SPADE|A
BLACK|CLUB|2
BLACK|CLUB|3
BLACK|CLUB|4
BLACK|CLUB|5
BLACK|CLUB|6
BLACK|CLUB|7
BLACK|CLUB|8
BLACK|CLUB|9
BLACK|CLUB|10
BLACK|CLUB|J
BLACK|CLUB|Q
BLACK|CLUB|K
BLACK|CLUB|A
RED|DIAMOND|2
RED|DIAMOND|3
RED|DIAMOND|4
```