PRACTICAL-1: - Hadoop installation and basics of HDFS

Task 1: Identify the local file system in your Linux OS system. (Command followed by the output screenshot is expected).

Solution: **df -Th**

```
hirwuser150430@ip-172-31-45-217:~$ df -Th
Filesystem     Type      Size  Used Avail Use% Mounted on
udev           devtmpfs  2.0G  8.0K  2.0G   1% /dev
tmpfs          tmpfs     396M  476K  395M   1% /run
/dev/xvda1     ext4       99G   28G   67G  30% /
none           tmpfs     4.0K     0  4.0K   0% /sys/fs/cgroup
none           tmpfs     5.0M     0  5.0M   0% /run/lock
none           tmpfs     2.0G     0  2.0G   0% /run/shm
none           tmpfs     100M     0  100M   0% /run/user
hirwuser150430@ip-172-31-45-217:~$
```

Task 2: Study of Hadoop installation. Study of Hadoop configurations files. Enlist the contents of the files core-site.xml and hdfs-site.xml

Solution:

1) Content of core-site.xml

```
<property>
        <name>hadoop.tmp.dir</name>
        <value>/data/hdfs/tmp</value>
        <description>Where Hadoop will place all of its working files</description>
</property>
<property>
        <name>fs.defaultFS</name>
        <value>hdfs://master:9000</value>
        <description>Where HDFS NameNode can be found on the network</description>
</property>
<property>
        <name>hadoop.proxyuser.hduser.groups</name>
        <value>*</value>
        <description>
                What user groups are allow to connect to the HDFS proxy.
```

```
            * for all.</description>
    </property>
    <property>
        <name>hadoop.proxyuser.hduser.hosts</name>
        <value>*</value>
        <description>
            What user hosts are allow to connect to the HDFS proxy.
            * for all.
        </description>
    </property>
```

2) Content of hdfs-site.xml

```
    <property>
        <name>dfs.replication</name>
        <value>2</value>
        <description>The default replication factor of files on HDFS</description>
    </property>
    <property>
        <name>dfs.block.size</name>
        <value>16777216</value>
        <description>The default block size in bytes of data saved to HDFS</description>
    </property>
    <property>
        <name>dfs.namenode.rpc-bind-host</name>
        <value>0.0.0.0</value>
        <description>
            controls what IP address the NameNode binds to.
            0.0.0.0 means all available.
        </description>
    </property>
    <property>
        <name>dfs.namenode.servicerpc-bind-host</name>
```

```
        <value>0.0.0.0</value>
        <description>
            controls what IP address the NameNode binds to.
            0.0.0.0 means all available.
        </description>
    </property>
    <property>
        <name>dfs.namenode.http-bind-host</name>
        <value>0.0.0.0</value>
        <description>
            controls what IP address the NameNode binds to.
            0.0.0.0 means all available.
        </description>
    </property>
    <property>
        <name>dfs.namenode.https-bind-host</name>
        <value>0.0.0.0</value>
        <description>
            controls what IP address the NameNode binds to.
            0.0.0.0 means all available.
        </description>
    </property>
    <property>
        <name>nfs.dump.dir</name>
        <value>/tmp/.hdfs-nfs</value>
        <description>A temporary working directory for files coming into the HDFS proxy.</description>
    </property>
    <property>
        <name>nfs.metrics.percentiles.intervals</name>
        <value>100</value>
```

```
        <description>
                Enable the latency histograms for read, write and commit requests.
                The time unit is 100 seconds in this example.
        </description>
</property>
<property>
        <name>nfs.exports.allowed.hosts</name>
        <value>* rw</value>
        <description>Host permissions for connecting to the proxy.</description>
</property>
<property>
        <name>dfs.permissions</name>
        <value>true</value>
        <description>Enforce permissions</description>
</property>
<property>
        <name>dfs.permissions.supergroup</name>
        <value>hadoop</value>
        <description>The name of the group of Hadoop super-users.</description>
</property>
```

Task 3: Study and run following commands on the hadoop cluster and show the output.

JPS: to check out all the **Hadoop** daemons like DataNode, NodeManager, NameNode, and ResourceManager that are currently running on the machine.

```
hirwuser150430@ip-172-31-45-217:~$ jps
27606 Jps
```

Fsck: to check health of the HDFS.

```
hirwuser150430@ip-172-31-45-217:~$ hdfs fsck /user/hirwuser150430/MYNEWDIRECTORY -files -blocks -locations
Connecting to namenode via http://ec2-54-92-244-237.compute-1.amazonaws.com:50070
FSCK started by hirwuser150430 (auth:SIMPLE) from /172.31.45.217 for path /user/hirwuser150430/MYNEWDIRECTORY at
Sat Jun 12 07:23:13 UTC 2021
/user/hirwuser150430/MYNEWDIRECTORY <dir>
/user/hirwuser150430/MYNEWDIRECTORY/dwp-payments-april10.csv 3326129 bytes, 1 block(s):  OK
0. BP-2125152513-172.31.45.216-1410037307133:blk_1075416868_1676131 len=3326129 Live_repl=2 [DatanodeInfoWithStor
age[172.31.46.124:50010,DS-fe24aecb-f56f-4c9c-8cf9-a3b1259bc0d0,DISK], DatanodeInfoWithStorage[172.31.45.216:5001
0,DS-d0d9eb5c-f35f-4a12-bfdf-544085d693a3,DISK]]

Status: HEALTHY
 Total size:    3326129 B
 Total dirs:    1
 Total files:   1
 Total symlinks:                0
 Total blocks (validated):      1 (avg. block size 3326129 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:     2.0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)
 Number of data-nodes:          3
 Number of racks:               1
FSCK ended at Sat Jun 12 07:23:13 UTC 2021 in 1 milliseconds


The filesystem under path '/user/hirwuser150430/MYNEWDIRECTORY' is HEALTHY
```

Touchz: to create an empty file.

```
hirwuser150430@ip-172-31-45-217:~$ hdfs dfs -touchz emptyexample.txt
hirwuser150430@ip-172-31-45-217:~$
```

copyFromlocal: to copy a file from local file system to HDFS.

```
hirwuser150430@ip-172-31-45-217:~$ hadoop fs -copyFromLocal /hirw-starterkit/hdfs/commands/dwp-payments-april10.csv ILOVEBIGDATA
hirwuser150430@ip-172-31-45-217:~$ hadoop fs -ls ILOVEBIGDATA
Found 1 items
-rw-r--r--   3 hirwuser150430 hirwuser150430    3326129 2021-06-12 06:45 ILOVEBIGDATA/dwp-payments-april10.csv
```

copyToLocal: to copy a file from HDFS to local file system.

```
hirwuser150430@ip-172-31-45-217:~$ hadoop fs -copyToLocal ILOVEBIGDATA/dwp-payments-april10.csv .
copyToLocal: /home/hirwuser150430/dwp-payments-april10.csv._COPYING_ (Permission denied)
```

cat: to print the content of a specific file.

```
hirwuser150430@ip-172-31-45-217:~$ hadoop fs -cat ILOVEBIGDATA/dwp-payments-april10.csv
```

```
Department for Work and Pensions ,Jobcentre Plus,30/04/2010,PRINTING STATIONERY IT & CONSUMABLES,JOBCENTRE PLUS,XEROX UK LIMITED,2015
125301,91.38,,
Department for Work and Pensions ,Jobcentre Plus,30/04/2010,PRINTING STATIONERY IT & CONSUMABLES,JOBCENTRE PLUS,XEROX UK LIMITED,2015
125301,14.17,,
Department for Work and Pensions ,Jobcentre Plus,30/04/2010,PRINTING STATIONERY IT & CONSUMABLES,JOBCENTRE PLUS,XEROX UK LIMITED,2015
125301,91.66,,
Department for Work and Pensions ,Jobcentre Plus,30/04/2010,PRINTING STATIONERY IT & CONSUMABLES,JOBCENTRE PLUS,XEROX UK LIMITED,2015
125301,15.07,,
Department for Work and Pensions ,Jobcentre Plus,30/04/2010,PRINTING STATIONERY IT & CONSUMABLES,JOBCENTRE PLUS,XEROX UK LIMITED,2015
125301,10.18,,
```

moveFromLocal: to move a file from local file system to HDFS.

Hadoop fs -moveFromLocal <local source> <destination>

Put:

Similar to **copyFromLocal.**

Get:

Similar to **copyToLocal.**

Rmr: to remove a file or directory recursively.

```
hirwuser150430@ip-172-31-45-217:~$ hdfs dfs -rm -r test-antony-derectory
Deleted test-antony-derectory
```

Setrep: to change the replication factor of a file/directory in HDFS. By default the value is 3.

```
hirwuser150430@ip-172-31-45-217:~$ hadoop fs -setrep 2 MYNEWDIRECTORY/dwp-payments-april10.csv
Replication 2 set: MYNEWDIRECTORY/dwp-payments-april10.csv
```