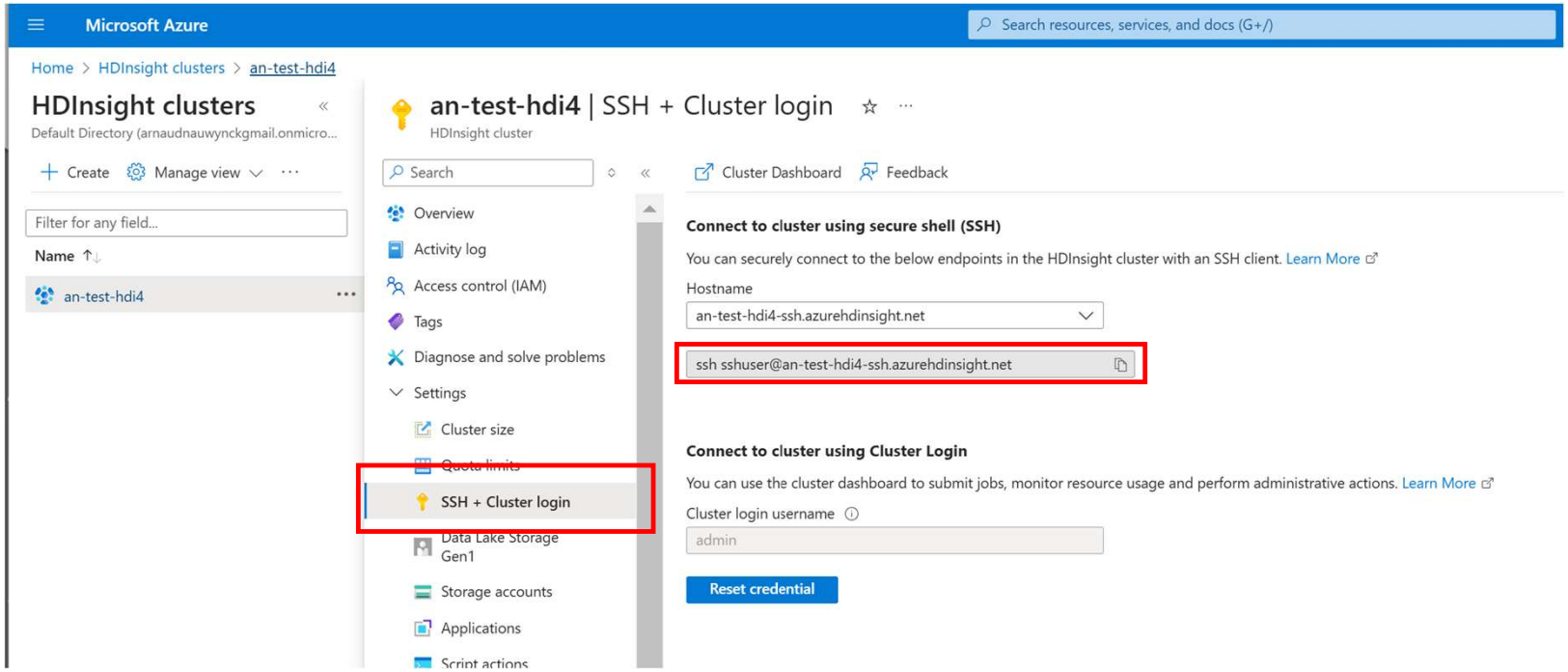
HD-Insight

arnaud.nauwynck@gmail.com

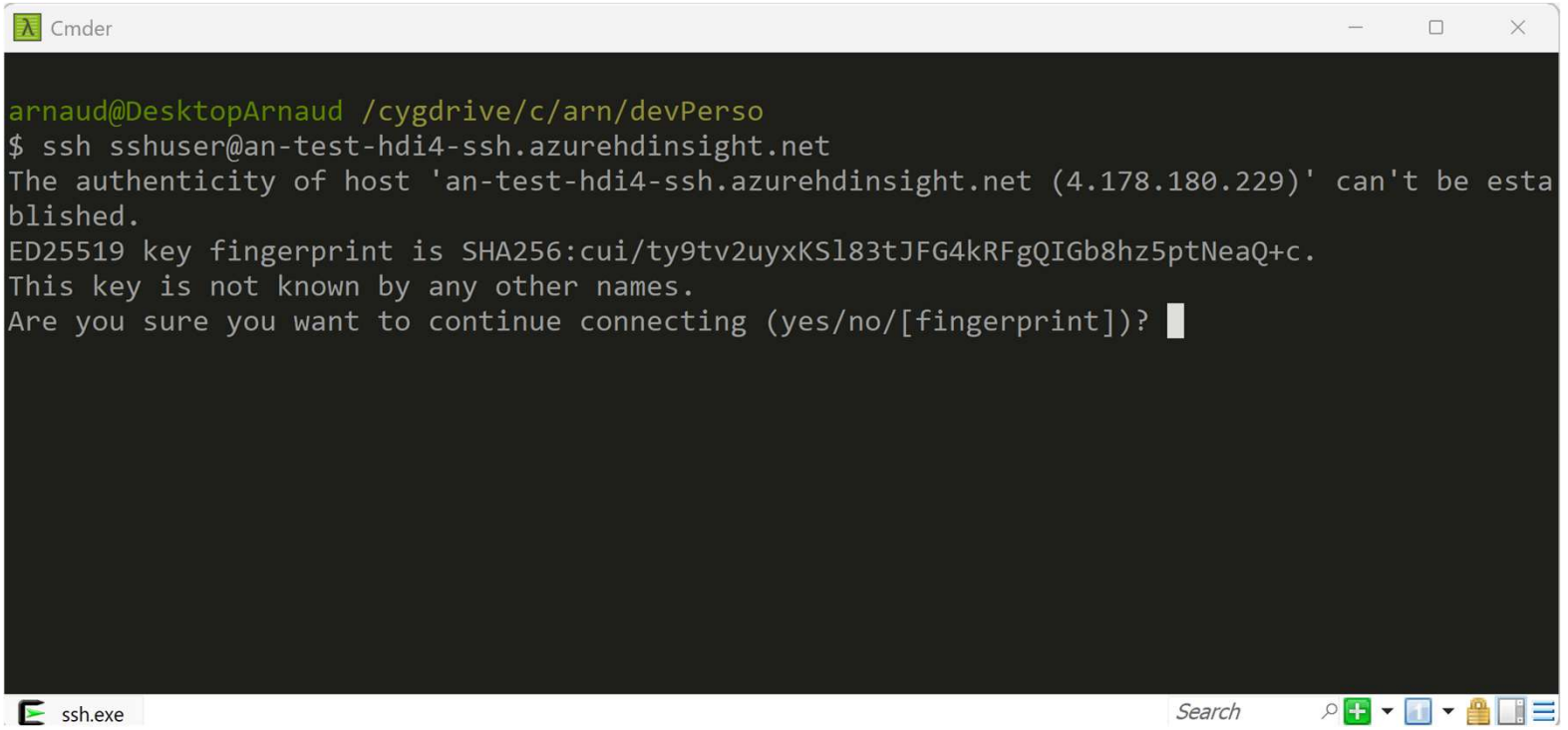
Course Esilv 2024

Exploring Linux HeadNode Server via SSH

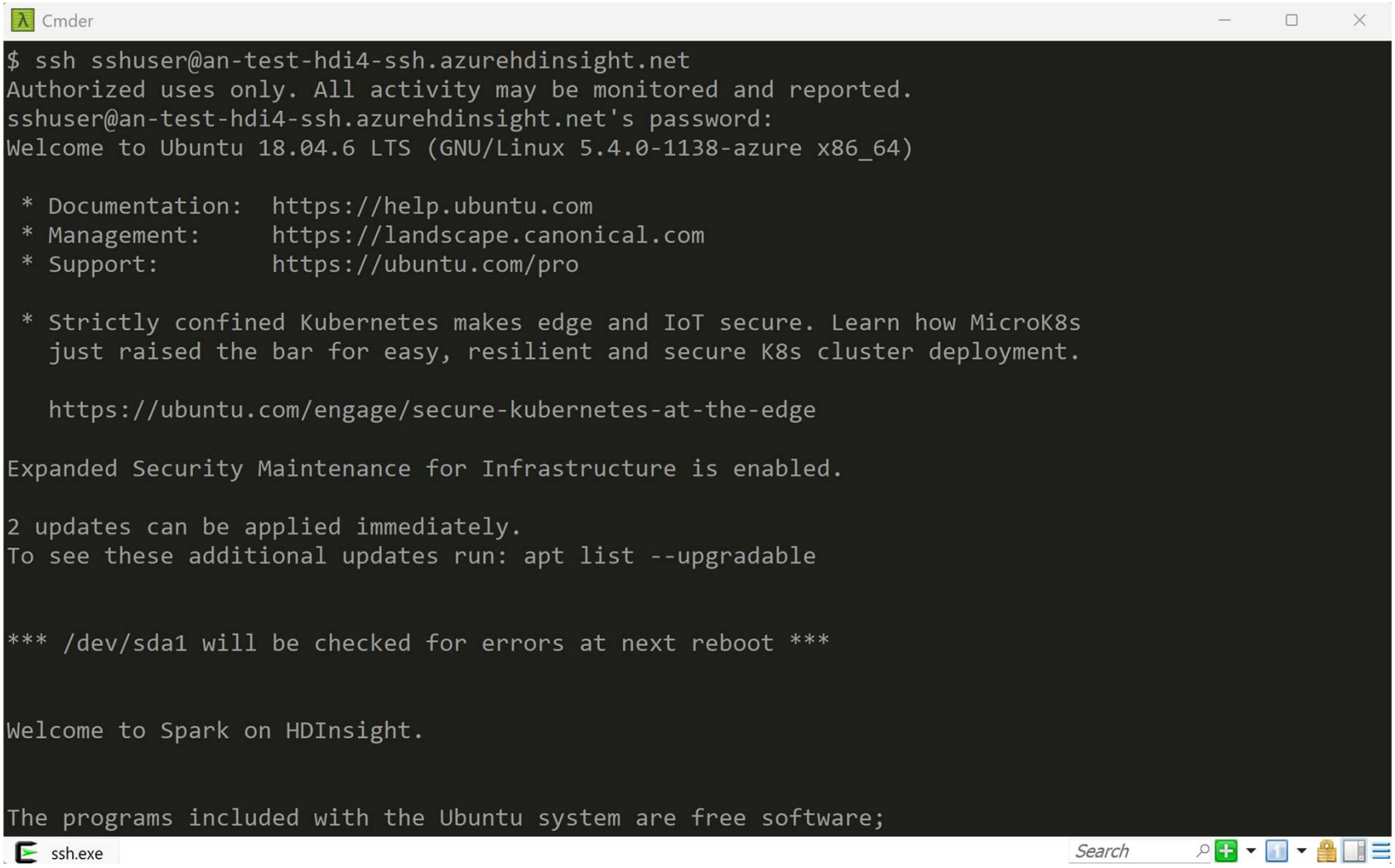
Connecting to SSH on HD-Insight



SSH


Cmder
arnaud@DesktopArnaud /cygdrive/c/arn/devPerso
$ ssh sshuser@an-test-hdi4-ssh.azurehdinsight.net
The authenticity of host 'an-test-hdi4-ssh.azurehdinsight.net (4.178.180.229)' can't be esta
blished.
ED25519 key fingerprint is SHA256:cui/ty9tv2uyxKSl83tJFG4kRFgQIGb8hz5ptNeaQ+c.
This key is not known by any other names.
Are you sure you want to continue connecting (yes/no/[fingerprint])?
ssh.exe
Search

SSH


Cmder
$ ssh sshuser@an-test-hdi4-ssh.azurehdinsight.net
Authorized uses only. All activity may be monitored and reported.
sshuser@an-test-hdi4-ssh.azurehdinsight.net's password:
Welcome to Ubuntu 18.04.6 LTS (GNU/Linux 5.4.0-1138-azure x86_64)
* Documentation: https://help.ubuntu.com
* Management: https://landscape.canonical.com
* Support: https://ubuntu.com/pro
* Strictly confined Kubernetes makes edge and IoT secure. Learn how MicroK8s
just raised the bar for easy, resilient and secure K8s cluster deployment.
https://ubuntu.com/engage/secure-kubernetes-at-the-edge
Expanded Security Maintenance for Infrastructure is enabled.
2 updates can be applied immediately.
To see these additional updates run: apt list --upgradable
*** /dev/sda1 will be checked for errors at next reboot ***
Welcome to Spark on HDInsight.
The programs included with the Ubuntu system are free software;
ssh.exe
Search

Who am I ... sudo su

```
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

sshuser@hn0-an-tes:~$
sshuser@hn0-an-tes:~$ pwd
/home/sshuser
sshuser@hn0-an-tes:~$ who am i
sshuser  pts/0        2024-11-11 17:18 (91.167.220.101)
sshuser@hn0-an-tes:~$
sshuser@hn0-an-tes:~$ sudo su
root@hn0-an-tes:/home/sshuser#
root@hn0-an-tes:/home/sshuser# who am i
sshuser  pts/0        2024-11-11 17:18 (91.167.220.101)
root@hn0-an-tes:/home/sshuser#
```

/etc/hosts
... Hostname are using "hn|wn"
then cluster name truncated to 6 letters

```
root@hn0-an-tes:/home/sshuser# cat /etc/hosts | grep an-tes
        10.0.0.16       hn0-an-tes.fibfli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net        hea
                headnodehost.   hn0-an-tes      hn0-an-tes.fibfli32hrrenczv02qeic3dbb.parx.internal
                # SlaveNodeManager
10.0.0.4 wn1-an-tes.fibfli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net wn1-an-tes wn1-an-tes.fib
fli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net.
10.0.0.8 zk0-an-tes.fibfli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net zk0-an-tes zk0-an-tes.fib
fli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net.
10.0.0.15 hn1-an-tes.fibfli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net hn1-an-tes hn1-an-tes.fi
bfli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net.
10.0.0.9 zk2-an-tes.fibfli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net zk2-an-tes zk2-an-tes.fib
fli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net.
10.0.0.6 zk3-an-tes.fibfli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net zk3-an-tes zk3-an-tes.fib
fli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net.
10.0.0.13 gw0-an-tes.fibfli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net gw0-an-tes gw0-an-tes.fi
bfli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net.
10.0.0.14 gw2-an-tes.fibfli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net gw2-an-tes gw2-an-tes.fi
bfli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net.
root@hn0-an-tes:/home/sshuser#
```

ssh -> ssh on hn0-

```
root@hn0-an-tes:/home/sshuser# ssh sshuser@hn0-an-tes.fibfli32hrrenczv02qeic3dbb.parx.internal.clo
udapp.net
The authenticity of host 'hn0-an-tes.fibfli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net (10.0.0.
16)' can't be established.
ECDSA key fingerprint is SHA256:P77UPPhmWS8M8fIPuuZ0vK2vwOdPBdd+dHD9XlOeFQE.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'hn0-an-tes.fibfli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net,10.0.0
.16' (ECDSA) to the list of known hosts.
Authorized uses only. All activity may be monitored and reported.
sshuser@hn0-an-tes.fibfli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net's password:
Welcome to Ubuntu 18.04.6 LTS (GNU/Linux 5.4.0-1138-azure x86_64)
```

Remark: direct SSH to hn0, hn1 … FAIL need ssh tunnel

```
$ ssh sshuser@hn1-an-tes.fibfli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net
ssh: Could not resolve hostname hn1-an-tes.fibfli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net: Name or service not known
```

Discovering processes running on hn0
sudo ps aux

```
yarn      2354  0.0  0.0  76760  7824 ?        Ss   14:17   0:00 /lib/systemd/systemd --user
yarn      2356  0.0  0.0 128756  3036 ?        S    14:17   0:00 (sd-pam)
ams       2406  0.0  0.0  13452  3436 ?        S    14:17   0:00 bash /usr/lib/ams-hbase/bin/hbase-daemon.sh --config
ams       2420  4.2  4.3 6527408 1245768 ?     Sl   14:17   8:00 /usr/lib/jvm/temurin-8-jdk-amd64/bin/java -Dproc_mas
yarn      2506  3.9  2.4 2991096 699008 ?      Sl   14:17   7:31 /usr/lib/jvm/temurin-8-jdk-amd64/bin/java -Dproc_tim
ams       2669  4.1  2.5 6084232 730796 ?      Sl   14:17   7:49 /usr/lib/jvm/temurin-8-jdk-amd64/bin/java -Xms2048m
ams       3068  0.0  0.0 778168 24712 ?        Sl   14:17   0:07 /usr/lib/ambari-metrics-grafana/bin/grafana-server -
livy      3111  0.0  0.0  76768  7768 ?        Ss   14:17   0:00 /lib/systemd/systemd --user
livy      3114  0.0  0.0 128756  3040 ?        S    14:17   0:00 (sd-pam)
livy      3201  3.6  1.4 5755976 426640 ?      Sl   14:17   6:59 /usr/lib/jvm/temurin-8-jdk-amd64/bin/java -Xmx2g -cp
root      3280  0.0  0.0      0     0 ?        S    14:06   0:00 [audit_prune_tre]
yarn      4132  7.4  2.3 2984684 680576 ?      Sl   14:17  13:57 /usr/lib/jvm/temurin-8-jdk-amd64/bin/java -Dproc_res
root      4144  0.0  0.0  27724  2980 ?        Ss   14:07   0:00 /usr/lib/ipsec/starter --daemon charon --nofork
root      4204  0.0  0.0 1306692 8364 ?        Ssl  14:07   0:00 /usr/lib/ipsec/charon
root      4861  0.0  0.0   9928  2844 ?        S    14:17   0:00 /bin/bash /var/lib/ambari-agent/ambari-sudo.sh su sp
root      4863  0.0  0.0  78772  3884 ?        S    14:17   0:00 su spark -l -s /bin/bash -c export  PATH=/usr/local/
spark     4867  0.0  0.0  76768  7712 ?        Ss   14:17   0:00 /lib/systemd/systemd --user
spark     4872  0.0  0.0 128756  3040 ?        S    14:17   0:00 (sd-pam)
spark     4889  0.0  0.0  13452  3356 ?        Ss   14:18   0:00 -su -c export  PATH=/usr/local/sbin:/usr/local/bin:/
spark     4903  0.0  0.0  26076  3112 ?        S    14:18   0:00 awk -v fname=/var/log/jupyter/2024-11-11-14-17-59- -
spark     4907  0.0  0.4 620656 125300 ?       Sl   14:18   0:02 /usr/bin/miniforge/envs/py38jupyter/bin/python3.8 /u
root      4930  0.0  0.0  60560 14464 ?        Ss   14:18   0:00 /usr/bin/python /var/lib/.jupyter/jupyter_logger.py
root      5175  0.1  0.7 9973704 205096 ?      Ssl  14:07   0:13 /usr/bin/java -Dlog4j.configuration=file:/etc/hdinsi
oozie     5201  0.0  0.0  76768  7824 ?        Ss   14:18   0:00 /lib/systemd/systemd --user
root      5217  0.1  0.1 397844 48816 ?        Ssl  14:07   0:13 /usr/bin/python /usr/local/lib/python2.7/dist-packag
oozie     5226  0.0  0.0 128756  3040 ?        S    14:18   0:00 (sd-pam)
root      5562  0.0  0.1  83292 33568 ?        Ss   14:08   0:01 /usr/bin/python -m hdinsight_agent.HdinsightAgent
```

Lot of Java processes ..
running as user "yarn", "oozie", livy", ...

```
yarn      2354  0.0  0.0  76760  7824 ?        Ss   14:17   0:00 /lib/systemd/systemd --user
yarn      2356  0.0  0.0 128756  3036 ?        S    14:17   0:00 (sd-pam)
ams       2406  0.0  0.0  13452  3436 ?        S    14:17   0:00 bash /usr/lib/ams-hbase/bin/hbase-daemon.sh --config
ams       2420  4.2  4.3 6527408 1245768 ?     Sl   14:17   8:00 /usr/lib/jvm/temurin-8-jdk-amd64/bin/java -Dproc_mas
yarn      2506  3.9  2.4 2991096 699008 ?      Sl   14:17   7:31 /usr/lib/jvm/temurin-8-jdk-amd64/bin/java -Dproc_tim
ams       2669  4.1  2.5 6084232 730796 ?      Sl   14:17   7:49 /usr/lib/jvm/temurin-8-jdk-amd64/bin/java -Xms2048m
ams       3068  0.0  0.0 778168 24712 ?        Sl   14:17   0:07 /usr/lib/ambari-metrics-grafana/bin/grafana-server -
livy      3111  0.0  0.0  76768  7768 ?        Ss   14:17   0:00 /lib/systemd/systemd --user
livy      3114  0.0  0.0 128756  3040 ?        S    14:17   0:00 (sd-pam)
livy      3201  3.6  1.4 5755976 426640 ?      Sl   14:17   6:59 /usr/lib/jvm/temurin-8-jdk-amd64/bin/java -Xmx2g -cp
root      3280  0.0  0.0      0     0 ?        S    14:06   0:00 [audit_prune_tre]
yarn      4132  7.4  2.3 2984684 680576 ?      Sl   14:17  13:57 /usr/lib/jvm/temurin-8-jdk-amd64/bin/java -Dproc_res
root      4144  0.0  0.0  27724  2980 ?        Ss   14:07   0:00 /usr/lib/ipsec/starter --daemon charon --nofork
root      4204  0.0  0.0 1306692 8364 ?        Ssl  14:07   0:00 /usr/lib/ipsec/charon
root      4861  0.0  0.0   9928  2844 ?        S    14:17   0:00 /bin/bash /var/lib/ambari-agent/ambari-sudo.sh su sp
root      4863  0.0  0.0  78772  3884 ?        S    14:17   0:00 su spark -l -s /bin/bash -c export  PATH=/usr/local/
spark     4867  0.0  0.0  76768  7712 ?        Ss   14:17   0:00 /lib/systemd/systemd --user
spark     4872  0.0  0.0 128756  3040 ?        S    14:17   0:00 (sd-pam)
spark     4889  0.0  0.0  13452  3356 ?        Ss   14:18   0:00 -su -c export  PATH=/usr/local/sbin:/usr/local/bin:/
spark     4903  0.0  0.0  26076  3112 ?        S    14:18   0:00 awk -v fname=/var/log/jupyter/2024-11-11-14-17-59- -
spark     4907  0.0  0.4 620656 125300 ?       Sl   14:18   0:02 /usr/bin/miniforge/envs/py38jupyter/bin/python3.8 /u
root      4930  0.0  0.0  60560 14464 ?        Ss   14:18   0:00 /usr/bin/python /var/lib/.jupyter/jupyter_logger.py
root      5175  0.1  0.7 9973704 205096 ?      Ssl  14:07   0:13 /usr/bin/java -Dlog4j.configuration=file:/etc/hdinsi
oozie     5201  0.0  0.0  76768  7824 ?        Ss   14:18   0:00 /lib/systemd/systemd --user
root      5217  0.1  0.1 397844 48816 ?        Ssl  14:07   0:13 /usr/bin/python /usr/local/lib/python2.7/dist-packag
oozie     5226  0.0  0.0 128756  3040 ?        S    14:18   0:00 (sd-pam)
root      5562  0.0  0.1  83292 33568 ?        Ss   14:08   0:01 /usr/bin/python -m hdinsight_agent.HdinsightAgent
```

Focus on Yarn processes
sudo ps aux | grep ^yarn

We see HDInsightApplicationHistoryServer & ResourceManager

```
root@hn0-an-tes:/hadoop# ps aux | grep ^yarn
yarn      2354  0.0  0.0  76760  7824 ?        Ss   14:17   0:00 /lib/systemd/systemd --user
yarn      2356  0.0  0.0 128756  3036 ?        S    14:17   0:00 (sd-pam)
yarn      2506  3.9  2.4 2991096 699008 ?      Sl   14:17   7:35 /usr/lib/jvm/temurin-8-jdk-amd64/bin/java -Dproc_timelinese
ion=5.1.6.7 -Dwhitelist.filename=core-whitelist.res,coremanual-whitelist.res -Dcomponent=apptimelineserver -Dyarn.log.dir=/v
arn/yarn -Dyarn.log.file=hadoop-yarn-timelineserver-hn0-an-tes.log -Dyarn.home.dir=/usr/hdp/5.1.6.7/hadoop-yarn -Dyarn.root.
sole -Djava.library.path=:/usr/hdp/5.1.6.7/hadoop/lib/native/Linux-amd64-64:/var/lib/ambari-agent/tmp/hadoop_java_io_tmpdir:
7/hadoop/lib/native -Xmx1024m -Dhadoop.log.dir=/var/log/hadoop-yarn/yarn -Dhadoop.log.file=hadoop-yarn-timelineserver-hn0-an
op.home.dir=/usr/hdp/5.1.6.7/hadoop -Dhadoop.id.str=yarn -Dhadoop.root.logger=INFO,RFA -Dhadoop.policy.file=hadoop-policy.xm
rity.logger=INFO,NullAppender org.apache.hadoop.yarn.server.applicationhistoryservice.HDInsightApplicationHistoryServer
yarn      4132  7.3  2.3 2984684 680576 ?      Sl   14:17  14:05 /usr/lib/jvm/temurin-8-jdk-amd64/bin/java -Dproc_resourcema
sion=5.1.6.7 -Dservice.libdir=/usr/hdp/5.1.6.7/hadoop-yarn/./,/usr/hdp/5.1.6.7/hadoop-yarn/lib,/usr/hdp/5.1.6.7/hadoop-hdfs/
.6.7/hadoop-hdfs/lib,/usr/hdp/5.1.6.7/hadoop/./,/usr/hdp/5.1.6.7/hadoop/lib -Dyarn.server.resourcemanager.appsummary.logger=
-Dyarn.server.resourcemanager.appsummary.logger=INFO,RMSUMMARY -Drm.audit.logger=INFO,RMAUDIT -Dwhitelist.filename=core-whit
anual-whitelist.res -Dcomponent=resourcemanager -Dyarn.log.dir=/var/log/hadoop-yarn/yarn -Dyarn.log.file=hadoop-yarn-resourc
-tes.log -Dyarn.home.dir=/usr/hdp/5.1.6.7/hadoop-yarn -Dyarn.root.logger=INFO,console -Djava.library.path=:/usr/hdp/5.1.6.7/
ve/Linux-amd64-64:/var/lib/ambari-agent/tmp/hadoop_java_io_tmpdir:/usr/hdp/5.1.6.7/hadoop/lib/native -Xmx1024m -Dhadoop.log.
doop-yarn/yarn -Dhadoop.log.file=hadoop-yarn-resourcemanager-hn0-an-tes.log -Dhadoop.home.dir=/usr/hdp/5.1.6.7/hadoop -Dhadc
-Dhadoop.root.logger=INFO,RFA -Dhadoop.policy.file=hadoop-policy.xml -Dhadoop.security.logger=INFO,NullAppender org.apache.h
er.resourcemanager.ResourceManager
root@hn0-an-tes:/hadoop#
```

Yarn HN Servers logs files

for ResourceManager => see java argument ...
-Dyarn.log.dir=/var/log/hadoop-yarn/yarn

root@hn0-an-tes:/var/log/hadoop-yarn/yarn# **cd /var/log/hadoop-yarn/yarn**

root@hn0-an-tes:/var/log/hadoop-yarn/yarn# ls -ltr
total 2076
-rw-r--r-- 1 yarn hadoop   0 Nov 11 14:17 hadoop-mapreduce.jobsummary.log
-rw-r--r-- 1 yarn hadoop   2640 Nov 11 14:18 hadoop-yarn-resourcemanager-hn0-an-tes.out
-rw-r--r-- 1 yarn hadoop 117503 Nov 11 14:18 **hadoop-yarn-resourcemanager-hn0-an-tes.log**
-rw-r--r-- 1 yarn hadoop   257 Nov 11 14:18 rm-audit.log
-rw-r--r-- 1 yarn hadoop   49309 Nov 11 16:49 hadoop-yarn-timelineserver-hn0-an-tes.out
-rw-r--r-- 1 yarn hadoop 1940054 Nov 11 17:32 hadoop-yarn-timelineserver-hn0-an-tes.log

Sample ResourceManager Logs

 # less /var/log/hadoop-yarn/yarn/hadoop-yarn-resourcemanager-hn0-an-tes.log

2024-11-11 14:17:49,421 INFO resourcemanager.ResourceManager - STARTUP_MSG:
/**
STARTUP_MSG: Starting ResourceManager
STARTUP_MSG: host = hn0-an-tes/10.0.0.16
STARTUP_MSG: args = []
STARTUP_MSG: version = 3.3.4.5.1.6.7
STARTUP_MSG: classpath = /usr/hdp/5.1.6.7/hadoop/conf:

... (truncated)

2024-11-11 14:18:11,789 INFO ipc.Server - IPC Server Responder: starting
2024-11-11 14:18:11,797 INFO ipc.Server - IPC Server listener on 8141: starting
2024-11-11 14:18:11,804 INFO azurebfs.AbfsConfiguration - AbfsClientRetryPolicy: default
2024-11-11 14:18:11,812 INFO conf.Configuration - found resource yarn-site.xml at file:/etc/hadoop/5.1.6.7/0/yarn-site.xml
2024-11-11 14:18:11,898 INFO resourcemanager.ResourceManager - **Already in standby state**

Sample ResourceManager Logs on Container

 # less /var/log/hadoop-yarn/yarn/hadoop-yarn-resourcemanager-hn0-an-tes.log

.... (truncated)

2024-11-11 14:23:16,376 INFO capacity.CapacityScheduler - Allocation proposal accepted
2024-11-11 14:23:16,427 INFO rmcontainer.RMContainerImpl - container_1731334645766_0003_01_000002 Container Transitioned from ALLOCATED to **ACQUIRED**
2024-11-11 14:23:16,428 INFO rmcontainer.RMContainerImpl - container_1731334645766_0003_01_000003 Container Transitioned from ALLOCATED to ACQUIRED
2024-11-11 14:23:16,874 INFO rmcontainer.RMContainerImpl - container_1731334645766_0003_01_000002 Container Transitioned from ACQUIRED to **RUNNING**
2024-11-11 14:23:16,875 INFO rmcontainer.RMContainerImpl - container_1731334645766_0003_01_000003 Container Transitioned from ACQUIRED to RUNNING
2024-11-11 14:23:17,489 INFO scheduler.AppSchedulingInfo - checking for deactivate of application :application_173133464576
2024-11-11 14:23:35,003 INFO rmcontainer.RMContainerImpl - container_1731334645766_0003_01_000002 Container Transitioned from RUNNING to **COMPLETED**

Exploring Shell > Hdfs commands

$ hdfs

```
sshuser@hn1-an-tes:~$ hdfs
Usage: hdfs [OPTIONS] SUBCOMMAND [SUBCOMMAND OPTIONS]

  OPTIONS is none or any of:

--buildpaths                       attempt to add class files from build tree
--config dir                       Hadoop config directory
--daemon (start|status|stop)       operate on a daemon
--debug                            turn on shell script debug mode
--help                             usage information
--hostnames list[,of,host,names]   hosts to use in worker mode
--hosts filename                   list of hosts to use in worker mode
--loglevel level                   set the log4j level for this command
--workers                          turn on worker mode

  SUBCOMMAND is one of:


    Admin Commands:

cacheadmin           configure the HDFS cache
crypto               configure HDFS encryption zones
debug                run a Debug Admin to execute HDFS debug commands
dfsadmin             run a DFS admin client
dfsrouteradmin       manage Router-based federation
```

$ Hdfs dfs

```
    Client Commands:

classpath            prints the class path needed to get the hadoop jar and the required libraries
dfs                  run a filesystem command on the file system
envvars              display computed Hadoop environment variables
fetchdt              fetch a delegation token from the NameNode
getconf              get config values from configuration
groups               get the groups which users belong to
lsSnapshottableDir   list all snapshottable dirs owned by the current user
snapshotDiff         diff two snapshots of a directory or diff the current directory contents with a snapshot
version              print the version
```

$ hdfs dfs help <<subcommand>>

$ hdfs dfs -help ls

-ls [-C] [-d] [-h] [-q] [-R] [-t] [-S] [-r] [-u] [-e] [<path> ...] :

List the contents that match the specified file pattern. If path is not specified, the contents of /user/<currentUser> will be listed. For a directory a list of its direct children is returned (unless -d option is specified).

Directory entries are of the form: permissions - userId groupId sizeOfDirectory(in bytes) modDate directoryName and file entries are of the form: permissions numberOfReplicas userId groupId sizeOfFile(in bytes) modDate fileName

-C Display the paths of files and directories only.
-d Directories are listed as plain files.
-h Formats the sizes of files in a human-readable fashion rather than a number of bytes.
-q Print ? instead of non-printable characters.
-R Recursively list the contents of directories.
-t Sort files by modification time (most recent first).
-S Sort files by size.
-r Reverse the order of the sort.
-u Use time of last access instead of modification for display and sorting.
-e Display the erasure coding policy of files and directories.

Focus on main "hdfs dfs " commands

$ **hdfs dfs**
Usage: hadoop fs [generic options]

-**put** [-f] <localsrc> ... <dst>
-**get** [-f] <src> ... <localdst>

-**ls** [-C] [-d] [-h] [-R] [<path> ...
-**count** <path> ...
-**find** <path> ... <expression> ...

-**cp** [-f] [-d] <src> ... <dst>
-**mv** <src> ... <dst>
-**rm** [-f] [-r] <src> ...
-**rmdir** <dir> ...
-**mkdir** <path>

sudo hdfs dfs -ls /

sshuser@hn1-an-tes:~$ sudo su hdfs
hdfs@hn1-an-tes:/home/sshuser$ **hdfs dfs -ls /**
Found 19 items

```
-rwxrwxrwx   1 hdfs hadoop          0 2024-11-11 14:02 /HDInsight_TestAccessiblityBlobName
drwxrwxrwx   - hdfs hadoop          0 2024-11-11 14:17 /HdiNotebooks
drwxrwxrwx   - hdfs hadoop          0 2024-11-11 14:22 /HdiSamples
drwxrwxrwx   - hdfs hadoop          0 2024-11-11 14:03 /ams
drwxrwxrwx   - hdfs hadoop          0 2024-11-11 14:03 /amshbase
drwxrwxrwx   - hdfs hadoop          0 2024-11-11 14:03 /app-logs
drwxrwxrwx   - hdfs hadoop          0 2024-11-11 14:03 /apps
drwxrwxrwx   - hdfs hadoop          0 2024-11-11 14:03 /atshistory
drwxrwxrwx   - hdfs hadoop          0 2024-11-11 14:22 /custom-scriptaction-logs
drwxrwxrwx   - hdfs hadoop          0 2024-11-11 14:22 /example
drwxrwxrwx   - hdfs hadoop          0 2024-11-11 14:03 /hbase
drwxrwxrwx   - hdfs hadoop          0 2024-11-11 14:03 /hdp
drwxrwxrwx   - hdfs hadoop          0 2024-11-11 14:03 /hive
drwxrwxrwx   - hdfs hadoop          0 2024-11-11 14:03 /mapred
drwxrwxrwx   - hdfs hadoop          0 2024-11-11 14:03 /mr-history
drwxrwxrwx   - hdfs hadoop          0 2024-11-11 14:03 /tmp
drwxrwxrwx   - hdfs hadoop          0 2024-11-11 14:03 /user
drwxrwxrwx   - hdfs hadoop          0 2024-11-11 14:03 /warehouse
drwxrwxrwx   - hdfs hadoop          0 2024-11-11 14:17 /yarn
```

$ hdfs dfs -mkdir
$ hdfs dfs -put | -get | -cat | ..

h$ hdfs dfs **-mkdir** /user/test1

$ hdfs dfs **-ls** /user/test1

$ echo "Hello Hdfs, here is a file" > test-file.txt

$ hdfs dfs **-put** test-file.txt /user/test1/test-file.txt

$ hdfs dfs **-ls** /user/test1
Found 1 items
-rwxrwxrwx   1 hdfs hadoop   27 2024-11-11 18:05 /user/test1/test-file.txt

$ hdfs dfs **-cat** /user/test1/test-file.txt
Hello Hdfs, here is a file

$ hdfs dfs **-get** /user/test1/test-file.txt test-file2.txt

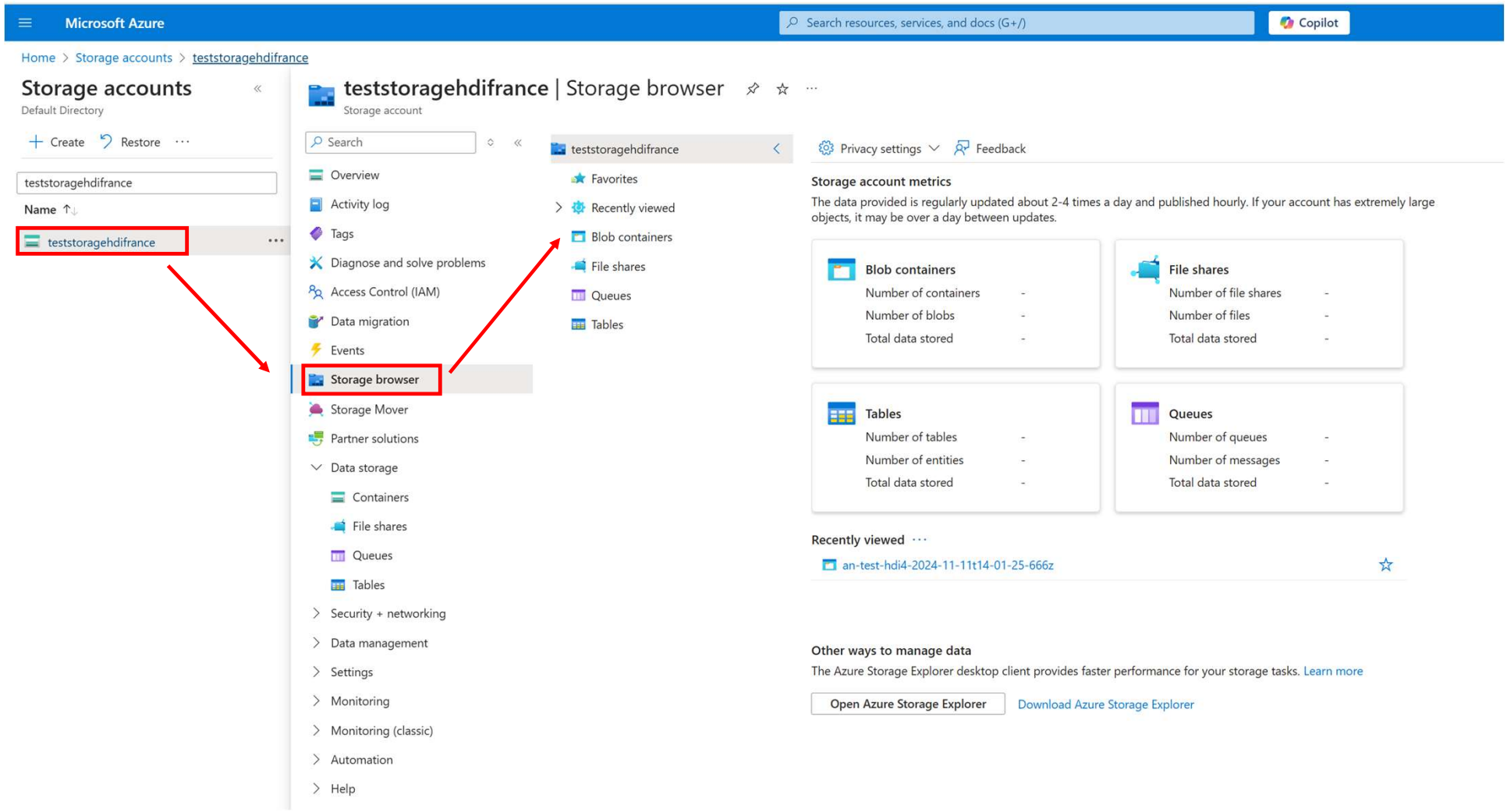
Hdfs command : using implicitly "fs.defaultFS"

```
hdfs@hn1-an-tes:~$ cat /etc/hadoop/conf/core-site.xml | grep -C3 fs.defaultF
  </property>

  <property>
    <name>fs.defaultFS</name>
    <value>abfss://an-test-hdi4-2024-11-11t14-01-25-666z@teststoragehdifrance.dfs.core.windows.net</value>
    <final>true</final>
  </property>
```

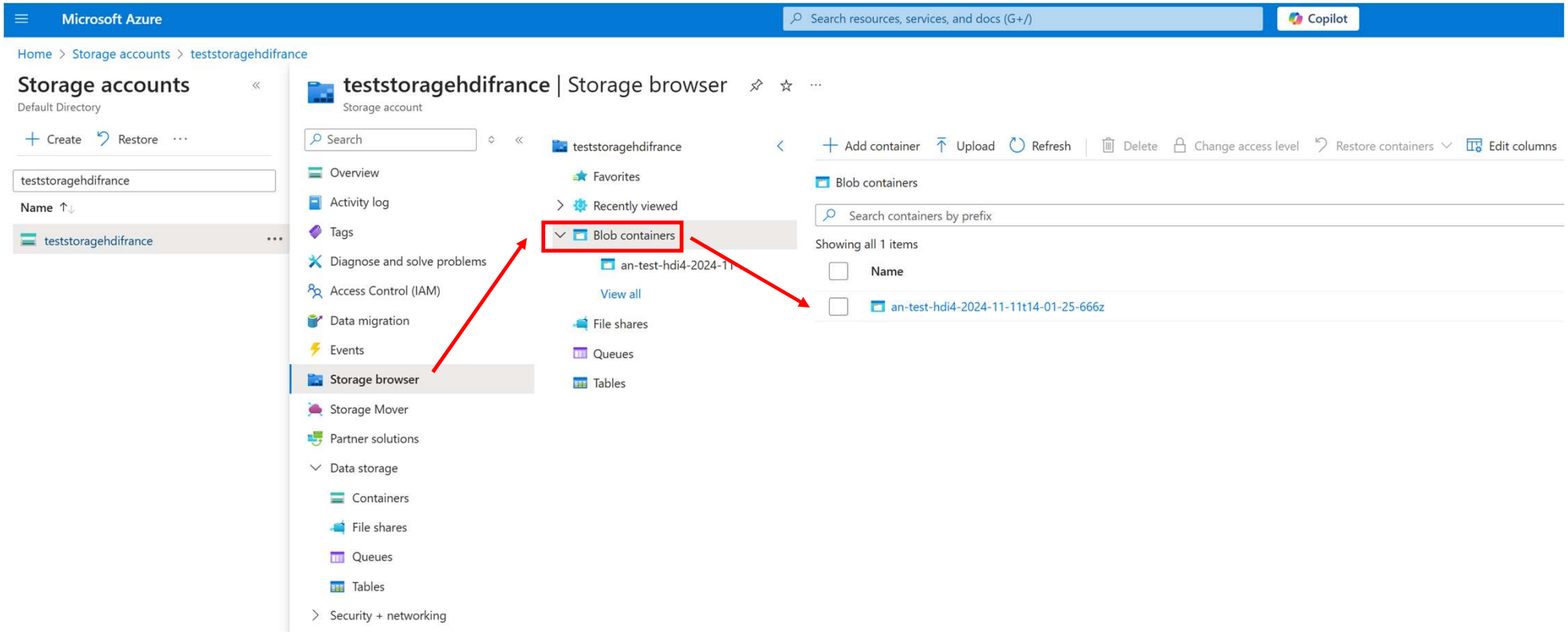
Explicit command :
$ hdfs dfs -ls abfss://

$ hdfs dfs -ls abfss://an-test-hdi4-2024-11-11t14-01-25-666z@teststoragehdifrance.dfs.core.windows.net/user/test1
Found 1 items
-rwxrwxrwx   1 hdfs hadoop   27 2024-11-11 18:05 abfss://an-test-hdi4-2024-11-11t14-01-25-666z@teststoragehdifrance.dfs.core.windows.net/user/test1/test-file.txt

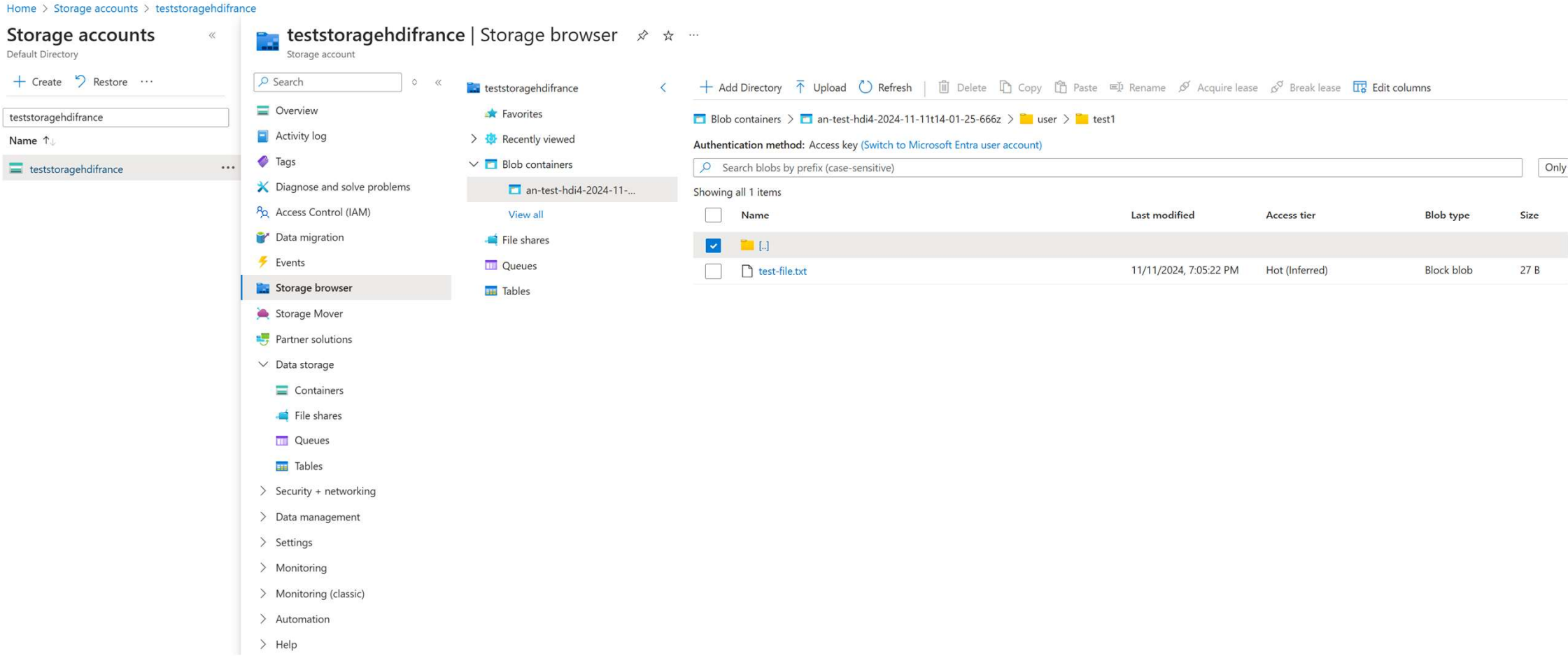
Storage Browser



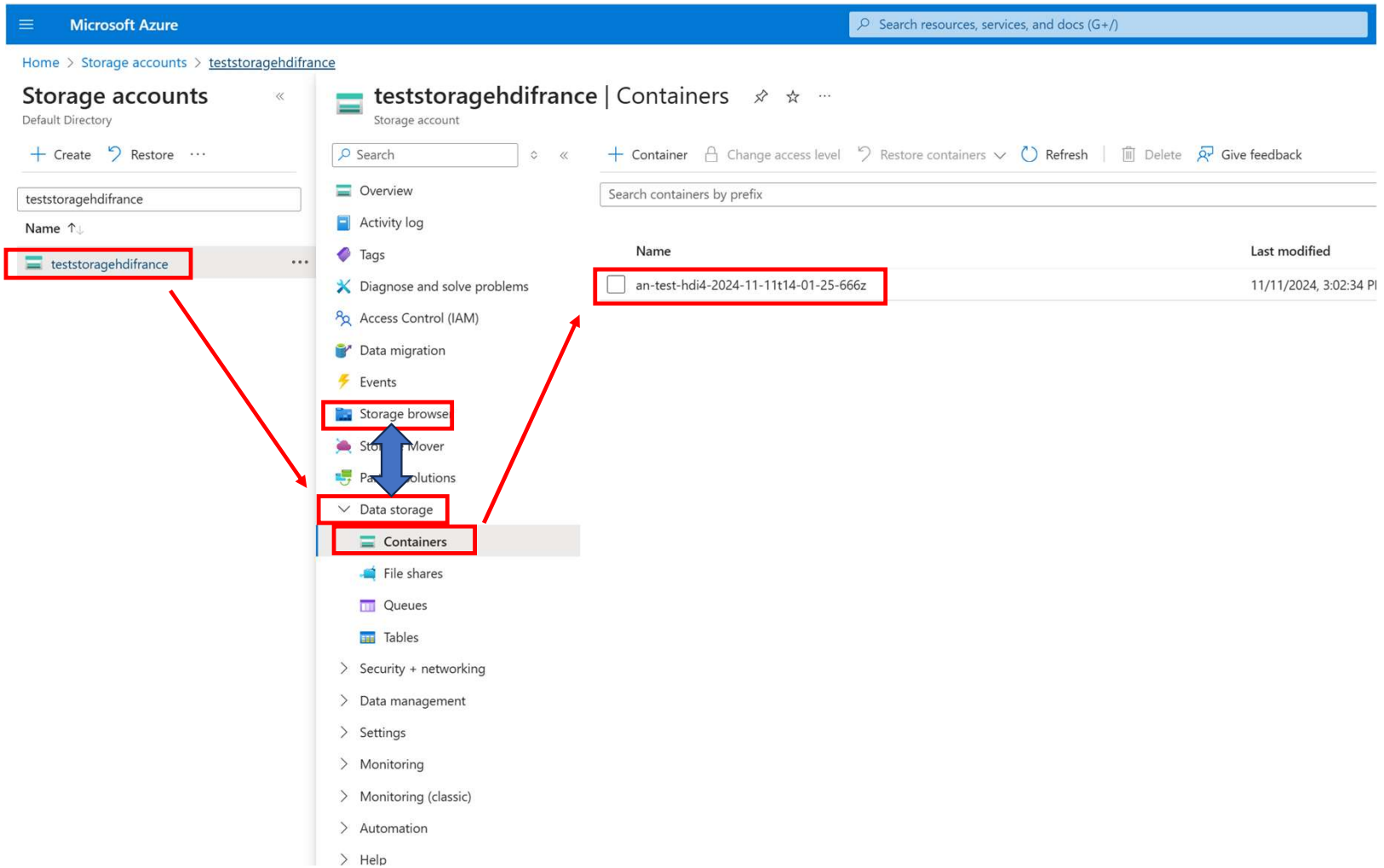
Storage > Containers > Directories > Files



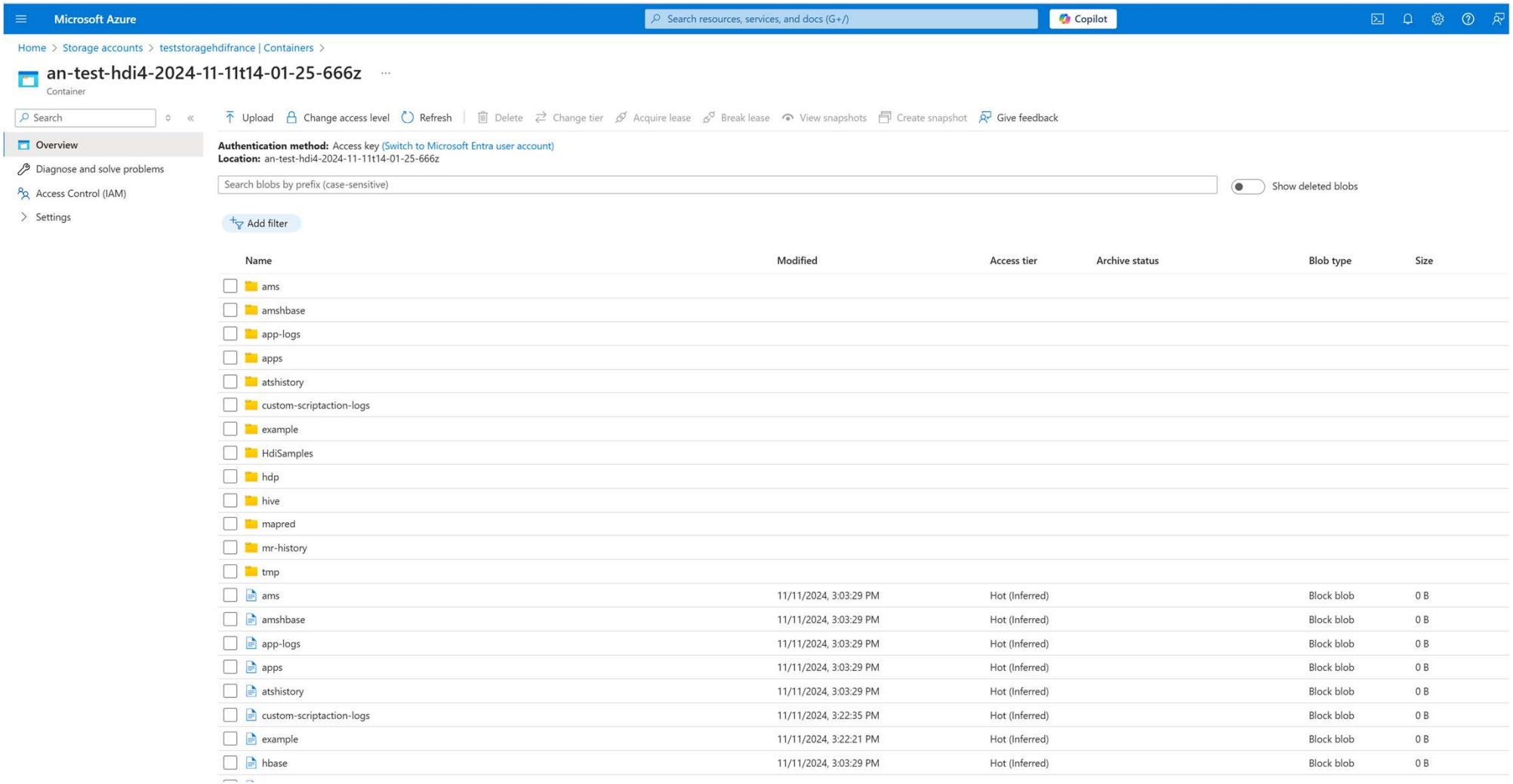
Storage Browser



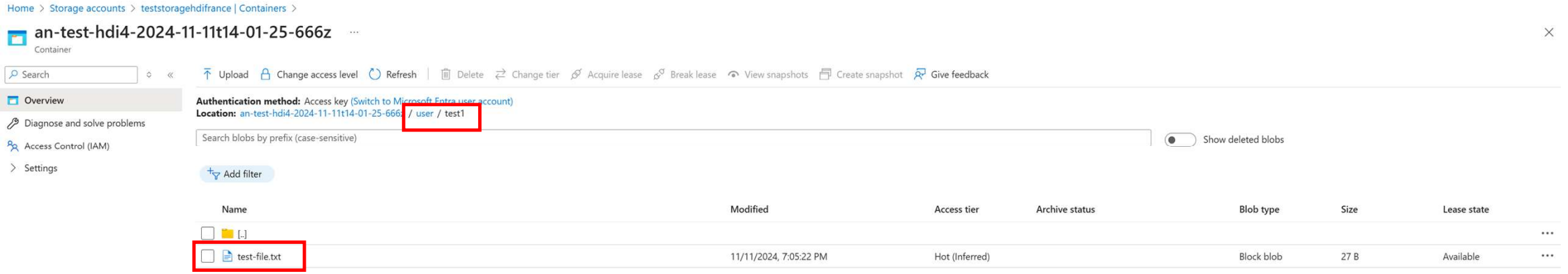
Other Interface : "Data Storage" (!= Browser)



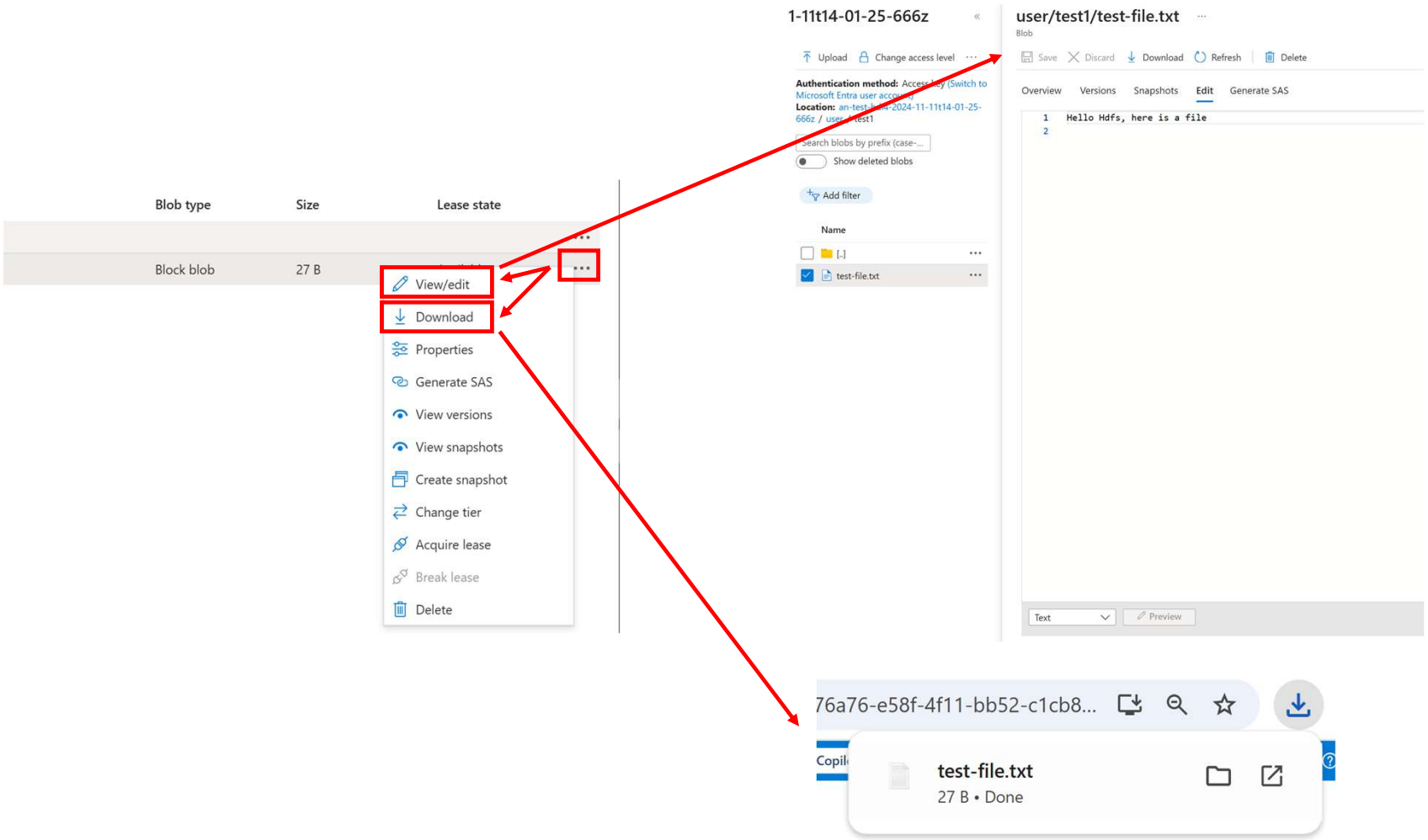
Data Explorer
(redundant, and less practical than Browser)



my created dir "/user/test1"
... and file "/user/test1/test-file.txt"



View / Download File from Portal



Using Shell > Yarn commands

(to list / kill Jobs, not to start one !)

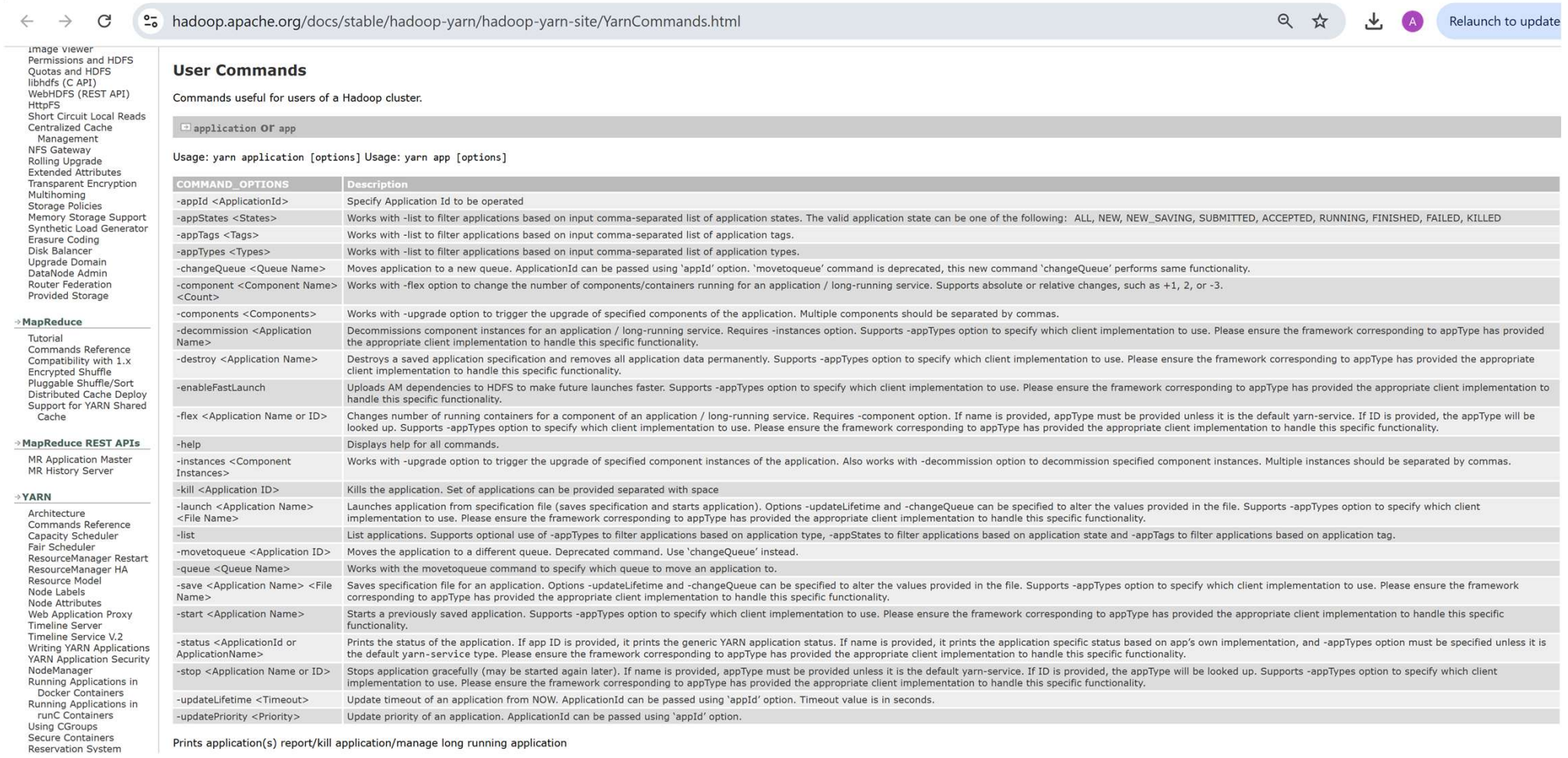
$ yarn <<sub-command>>

```
hdfs@hn1-an-tes:~$ yarn
WARNING: YARN_OPTS has been replaced by HADOOP_OPTS. Using value of YARN_OPTS.
Usage: yarn [OPTIONS] SUBCOMMAND [SUBCOMMAND OPTIONS]
 or    yarn [OPTIONS] CLASSNAME [CLASSNAME OPTIONS]
  where CLASSNAME is a user-provided Java class

  OPTIONS is none or any of:

--buildpaths                       attempt to add class files from build tree
--config dir                       Hadoop config directory
--daemon (start|status|stop)       operate on a daemon
--debug                            turn on shell script debug mode
--help                             usage information
--hostnames list[,of,host,names]   hosts to use in worker mode
--hosts filename                   list of hosts to use in worker mode
--loglevel level                   set the log4j level for this command
--workers                          turn on worker mode

  SUBCOMMAND is one of:


    Admin Commands:

daemonlog            get/set the log level for each daemon
```

$ yarn app <<sub-app-command>>

```
    Client Commands:

applicationattempt   prints applicationattempt(s) report
app|application      prints application(s) report/kill application/manage long running application
classpath            prints the class path needed to get the hadoop jar and the required libraries
cluster              prints cluster information
container            prints container(s) report
envvars              display computed Hadoop environment variables
fs2cs                converts Fair Scheduler configuration to Capacity Scheduler (EXPERIMENTAL)
jar <jar>            run a jar file
logs                 dump container logs
nodeattributes       node attributes cli client
queue                prints queue information
schedulerconf        Updates scheduler configuration
timelinereader       run the timeline reader server
top                  view cluster information
version              print the version
```

$ yarn app { -list | -status | -kill …}
see doc https://hadoop.apache.org/docs



$ yarn app -list

$ **yarn app -list**

Total number of applications (application-types: [], states: [SUBMITTED, ACCEPTED, RUNNING] and tags: []):2
Application-Id Application-Name Application-Type User Queue State Progress Tracking-URL

application_1731334645766_0002 Thrift JDBC/ODBC Server SPARK spark thriftsvr RUNNING UNDEFINED 10% http://hn0-an-tes.fibfli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net:5040

application_1731334645766_0001 Thrift JDBC/ODBC Server SPARK spark thriftsvr RUNNING 10% http://hn1-an-tes.fibfli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net:5040

$ yarn app -status <<app_id>>

$ **yarn app -status application_1731334645766_0002**
Application Report :
Application-Id : application_1731334645766_0002
Application-Name : Thrift JDBC/ODBC Server
Application-Type : SPARK
User : spark
Queue : thriftsvr
Application Priority : 0
Start-Time : 1731334779687
Finish-Time : 0
Progress : 10%
State : RUNNING
Final-State : UNDEFINED
Tracking-URL : http://hn0-an-tes.fibfli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net:5040
RPC Port : -1
AM Host : 10.0.0.4
Aggregate Resource Allocation : 24100938 MB-seconds, 15690 vcore-seconds
Aggregate Resource Preempted : 0 MB-seconds, 0 vcore-seconds
Log Aggregation Status : NOT_START
Diagnostics :
Unmanaged Application : false
Application Node Label Expression : <Not set>
AM container Node Label Expression : <DEFAULT_PARTITION>
TimeoutType : LIFETIME ExpiryTime : UNLIMITED RemainingTime : -1seconds

$ yarn logs -applicationId <<app_id>>

$ **yarn logs -applicationId** application_1731334645766_0002

Container: container_1731334645766_0002_01_000001 on wn1-an-tes.fibfli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net:30050
LogAggregationType: LOCAL
==========
LogType:prelaunch.out
LogLastModifiedTime:Mon Nov 11 14:19:40 +0000 2024
LogLength:100
LogContents:

... (truncated)

ExecutorLauncher --arg 'hn0-an-tes.fibfli32hrrenczv02qeic3dbb.parx.internal.cloudapp.net:46739' --properties-file $PWD/__spark_conf__/__spark_conf__.properties --dist-cache-conf $PWD/__spark_conf__/__spark_dist_cache__.properties 1> /mnt/resource/hadoop/yarn/log/application_1731334645766_0002/container_1731334645766_0002_01_000001/stdout 2> /mnt/resource/hadoop/yarn/log/application_1731334645766_0002/container_1731334645766_0002_01_000001/stderr"
End of LogType:launch_container.sh.This log file belongs to a running container (container_1731334645766_0002_01_000001) and so may not be complete.
****

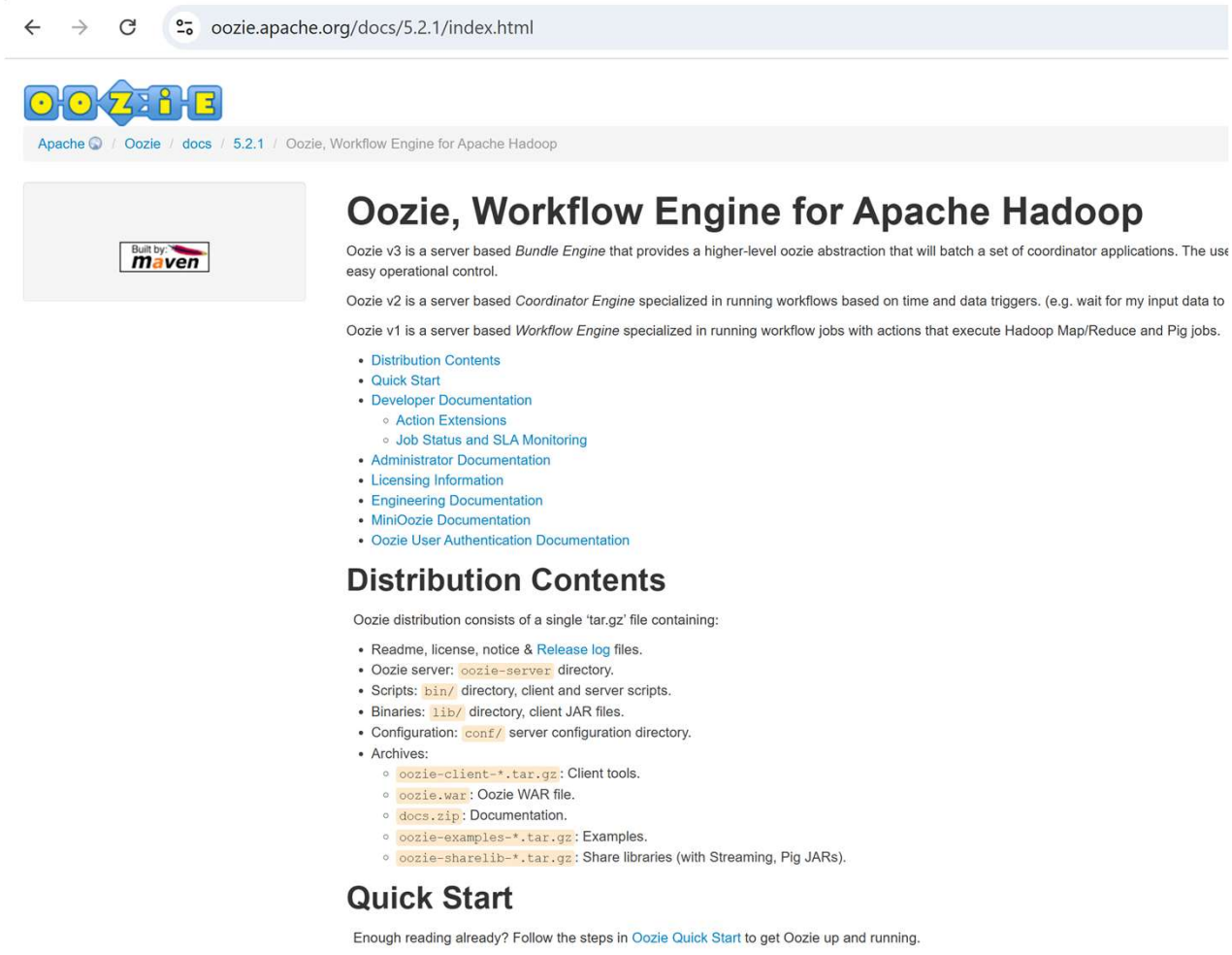
With Yarn ...
You can NOT start a new application,
ONLY restart (clone) an existing one !!

It has to be > 1000 lines of java code
OR using Oozie / Livy / spark-submit / ...

Submitting a Oozie Job

Oozie

https://oozie.apache.org/docs/5.2.1/index.html



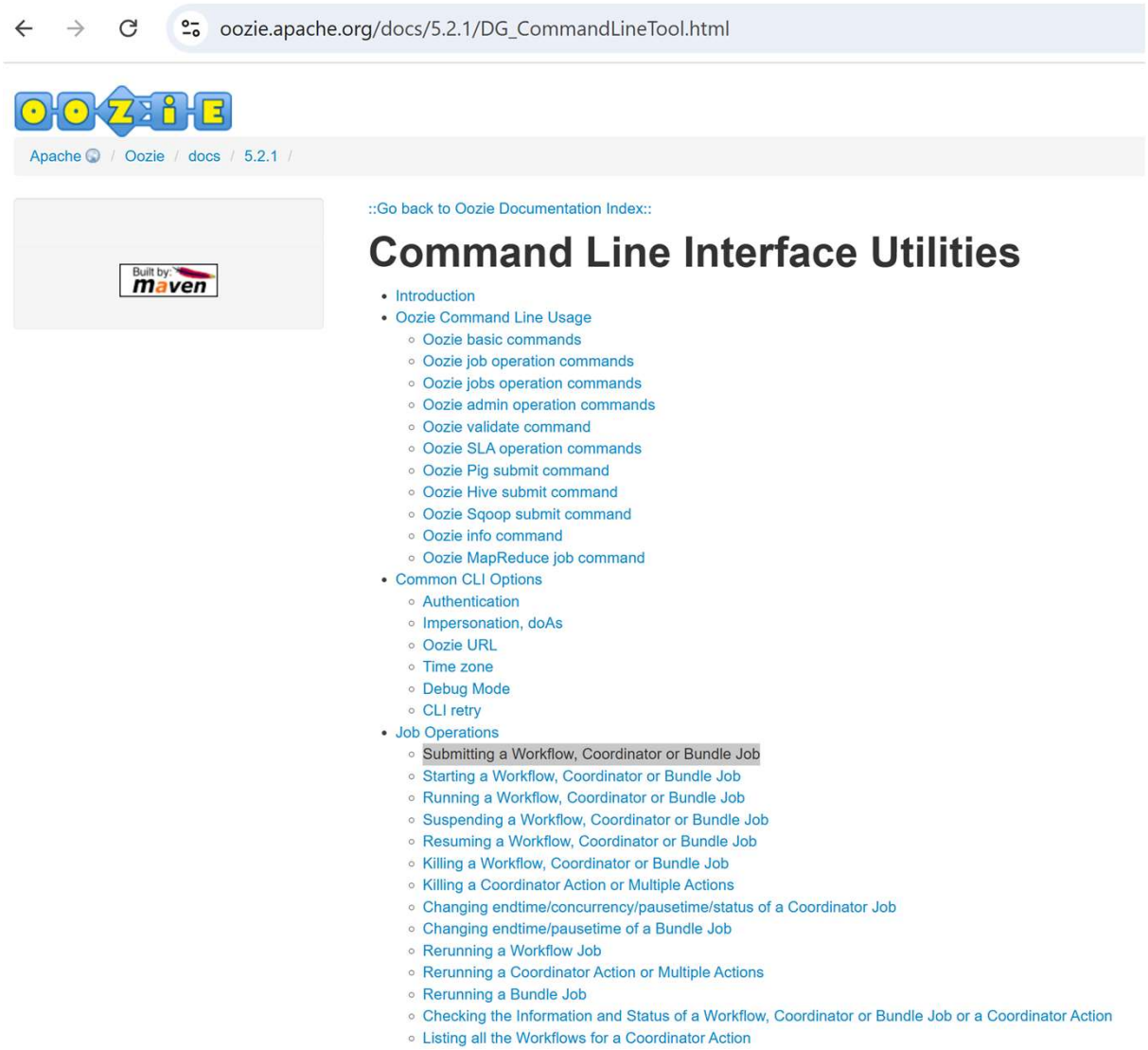
Quick Start

Enough reading already? Follow the steps in Oozie Quick Start to get Oozie up and running.

Developer Documentation

- Overview
- Oozie Quick Start
- Running the Examples
- Workflow Functional Specification
- Coordinator Functional Specification
- Bundle Functional Specification
- EL Expression Language Quick Reference
- Command Line Tool
- Workflow Re-runs Explained
- HCatalog Integration Explained
- Oozie Client Javadocs
- Oozie Core Javadocs
- Oozie Web Services API
- Action Authentication
- Fluent Job API

oozie Doc: Command Line Tool



doc : $ oozie job -config job.properties -submit

Submitting a Workflow, Coordinator or Bundle Job

- Submitting bundle feature is only supported in Oozie 3.0 or later. Similarly, all bundle operation features below are only supported in Oozie 3.0 or later.

Example:

```
$ oozie job -oozie http://localhost:11000/oozie -config job.properties -submit
.
job: 14-20090525161321-oozie-joe
```

The parameters for the job must be provided in a file, either a Java Properties file (.properties) or a Hadoop XML Configuration file (.xml). This file must be specified with the `-config` option.

The workflow application path must be specified in the file with the `oozie.wf.application.path` property. The coordinator application path must be specified in the file with the `oozie.coord.application.path` property.The bundle application path must be specified in the file with the `oozie.bundle.application.path` property. Specified path must be an HDFS path.

The job will be created, but it will not be started, it will be in `PREP` status.

Submit Oozie workflow using command line
$ oozie help

```
hdfs@hn1-an-tes:~$ oozie help
usage:
      the env variable 'OOZIE_URL' is used as default value for the '-oozie' option
      the env variable 'OOZIE_TIMEZONE' is used as default value for the '-timezone' option
      the env variable 'OOZIE_AUTH' is used as default value for the '-auth' option
      custom headers for Oozie web services can be specified using '-Dheader:NAME=VALUE'

      oozie help : display usage for all commands or specified command

      oozie version : show client version

      oozie job <OPTIONS> : job operations
                -action <arg>          coordinator rerun/kill on action ids (requires -rerun/-kill);
                                       coordinator log retrieval on action ids (requires -log)
                -allruns               Get workflow jobs corresponding to a coordinator action
                                       including all the reruns
                -auditlog <arg>        job audit log
                -auth <arg>            select authentication type [SIMPLE|BASIC|KERBEROS]
                -change <arg>          change a coordinator or bundle job
                -config <arg>          job configuration file '.xml' or '.properties'
                -configcontent <arg>   job configuration
                -coordinator <arg>     bundle rerun on coordinator names (requires -rerun)
                -D <property=value>    set/override value for given property
```

workflow.xml

```
<workflow-app name="useooziewf" xmlns="uri:oozie:workflow:0.2">
   <start to="my-shell"/>

   <action name='my-shell'>
      <shell xmlns="uri:oozie:shell-action:0.1">
         <job-tracker>${jobTracker}</job-tracker>
         <name-node>${nameNode}</name-node>
         <configuration>
            <property>
              <name>mapred.job.queue.name</name>
              <value>${queueName}</value>
            </property>
         </configuration>
         <exec>my-shell-script.sh</exec>
         <argument>arg1</argument>
         <argument>arg2</argument>
         <file>my-shell-script.sh</file>
      </shell>
      <ok to="end" />
      <error to="fail" />
   </action>

      <kill name="fail">
       <message>Job failed, error message
           [${wf:errorMessage(wf:lastErrorNode())}] </message>
      </kill>
      <end name="end"/>
</workflow-app>
```

job.properties

```
# cf parameter fs.defaultFS (contains oozie jars)
nameNode=abfss://an-test-hdi4-2024-11-11t14-01-25-666z@teststoragehdifrance.dfs.core.windows.net

jobTracker=headnodehost:8050
queueName=default
oozie.use.system.libpath=true
user.name=admin

# dir containing workflow.xml + referenced <file>
oozie.wf.application.path=\
    abfss://an-test-hdi4-2024-11-11t14-01-25-666z@teststoragehdifrance.dfs.core.windows.net/user/test1/test-oozie-shell
```

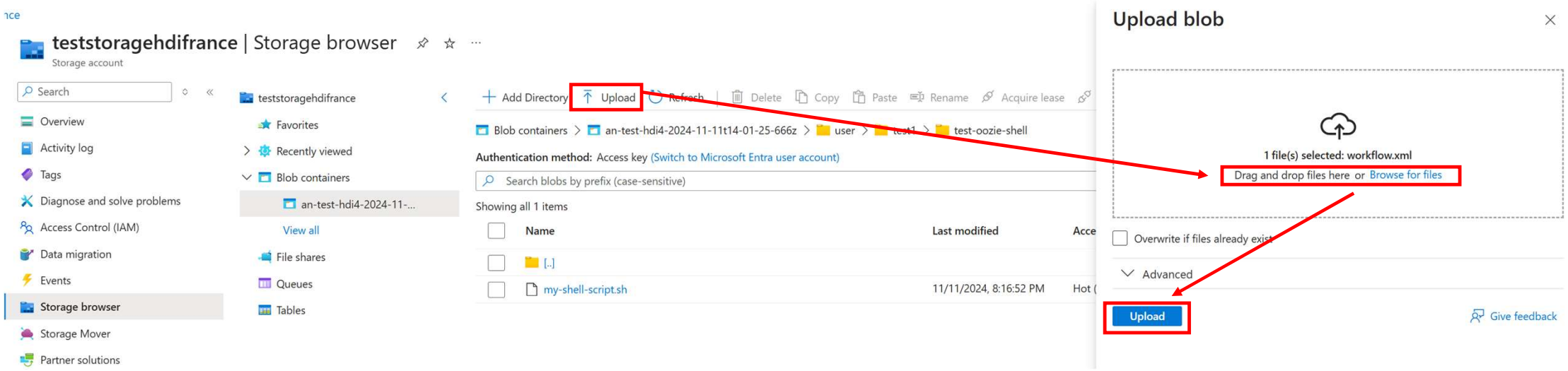
my-shell-script.sh

```
#!/bin/bash
# file my-shell-script.sh

echo "executing shell script $0"
echo "using shell arguments: $*"

for i in $(seq 1 6);
do
        echo "step [$i/6] sleep 10"
        sleep 10
done

echo "Finish shell script ==> exit 0"
exit 0
```

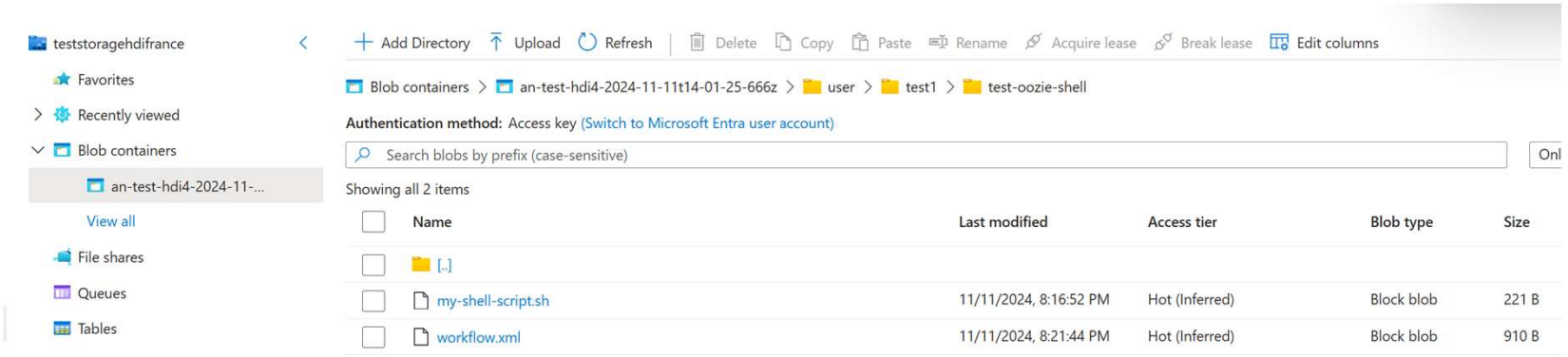
Testing locally "my-shell-script.sh"

```
arnaud@DesktopArnaud /cygdrive/c/apps/hadoop/hd-insight
$ ./my-shell-script.sh a b c
executing shell script ./my-shell-script.sh
using shell arguments: a b c
step [1/6] sleep 10
step [2/6] sleep 10
step [3/6] sleep 10
step [4/6] sleep 10
step [5/6] sleep 10
step [6/6] sleep 10
Finish shell script ==> exit 0
```

Deploying workflow.xml + my-shell-script.sh to abfss://...



Checking before deployment



$ hdfs dfs -ls /user/test1/test-oozie-shell
Found 2 items
-rwxrwxrwx 1 hdfs hadoop 221 2024-11-11 19:16 /user/test1/test-oozie-shell/my-shell-script.sh
-rwxrwxrwx 1 hdfs hadoop 910 2024-11-11 19:21 /user/test1/test-oozie-shell/workflow.xml

Testing before submit ...

```
$ hdfs dfs -cat /user/test1/test-oozie-shell/workflow.xml
<workflow-app name="useooziewf" xmlns="uri:oozie:workflow:0.2">
   <start to="my-shell"/>
      <action name='my-shell'>
      <shell xmlns="uri:oozie:shell-action:0.1">
...

$ hdfs dfs -get /user/test1/test-oozie-shell/my-shell-script.sh

$ chmod u+x ./my-shell-script.sh

$ ./my-shell-script.sh a b c
executing shell script ./my-shell-script.sh
using shell arguments: a b c
step [1/6] sleep 10
```

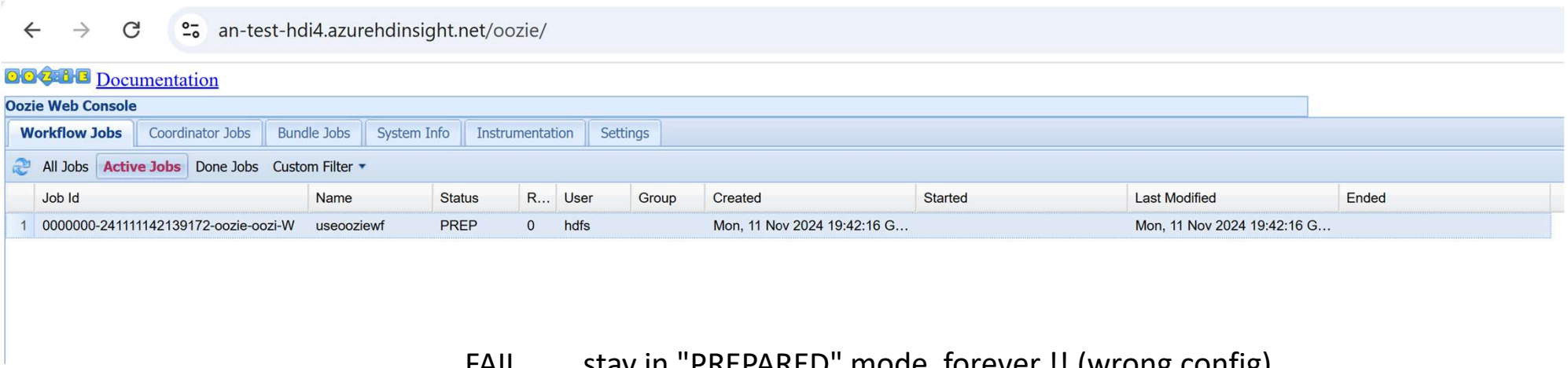
$ oozie job -config job.properties -submit

$ **oozie job -config job.properties -submit**

SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/hdp/5.1.6.7/oozie/libext/slf4j-reload4j-1.7.35.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/5.1.6.7/oozie/embedded-oozie-server/webapp/WEB-INF/lib/slf4j-reload4j-1.7.35.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Reload4jLoggerFactory]

job: 0000000-241111142139172-oozie-oozi-W

Checking running Job in Oozie UI

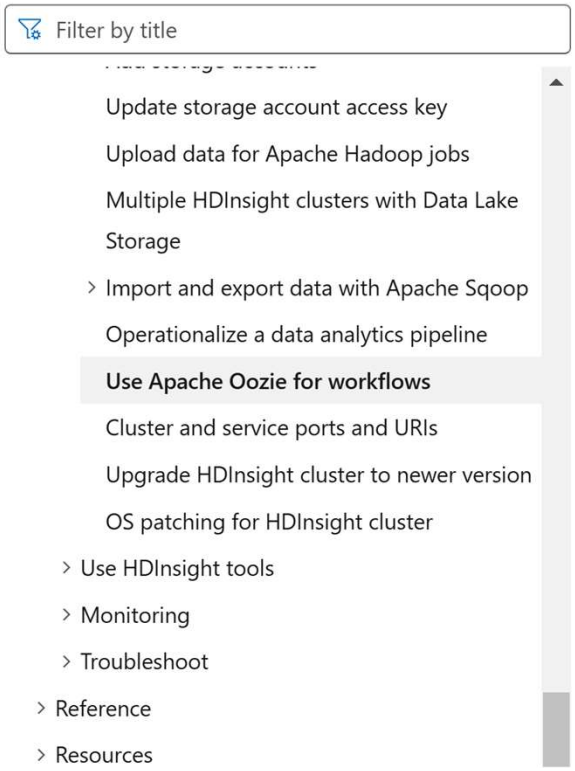


FAIL .... stay in "PREPARED" mode forever !! (wrong config)

sshuser ... can not be authenticated/impersonnated by oozie ?!

$ **sudo adduser sshuser users**
Adding user `sshuser' to group `users' ...
Adding user sshuser to group users
Done.

cf doc https://learn.microsoft.com/en-us/azure/hdinsight/hdinsight-use-oozie-linux-mac




```
hdfs dfs -mkdir -p /tutorials/useoozie/data
```

> **Note**
>
> The `-p` parameter causes the creation of all directories in the path. The `data` directory is used to hold the data used by the `useooziewf.hql` script.

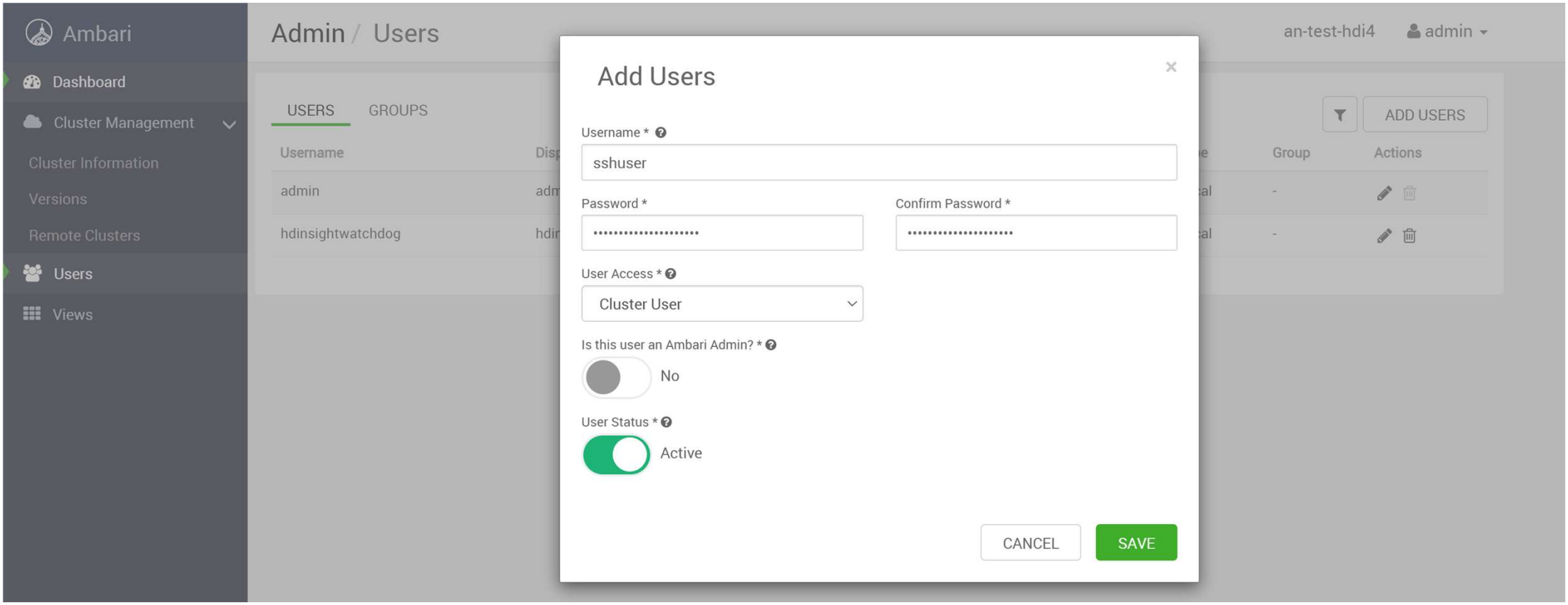
3. Edit the code below to replace `sshuser` with your SSH user name. To make sure that Oozie can impersonate your user account, use the following command:

Bash Copy

```
sudo adduser sshuser users
```

> **Note**
>
> You can ignore errors that indicate the user is already a member of the `users` group.

Trying to add user



login in Ambari for this user

using Oozie via Http Rest API + curl

$ export PASSWORD="..."

$ **curl -u admin:$PASSWORD** https://an-test-hdi4.azurehdinsight.net**/oozie/versions**
[0,1,2]

submit Oozie via Http curl -X POST

$ export PASSWORD="..."

$ **curl -u admin:$PASSWORD** https://an-test-hdi4.azurehdinsight.net**/oozie/v1/job -d config.xml**

WARN ... Oozie POST => PREPARE mode

https://oozie.apache.org/docs/5.0.0/WebServicesAPI.html#Job_Submission

Request:

```
POST /oozie/v1/jobs
Content-Type: application/xml;charset=UTF-8
.
<?xml version="1.0" encoding="UTF-8"?>
```

Looks simple ... BUT WARN !!!

```
}
```

A created job will be in `PREP` status. If the query string parameter 'action=start' is provided in the POST URL, the job will be started immediately and its status will be `RUNNING` .

You need explicitly to POST /api/v1/ jobs?action=start

Unfinished document

More to do:
- oozie
- livy
- spark-submit
...

Questions ?

arnaud.nauwynck@gmail.com