Tutorial 1 - Cluster and HDFS

Access the Cluster

In order to access the DIMA cluster, open a shell and establish a SSH session with the following command (instead of <user> insert the username you received in class):

```
ssh <user>@ibm-power-1.dima.tu-berlin.de -L 50070:ibm-power-1.dima.tu-berlin.de:50070
```

Input the password you received in class to establish a SSH connection and the port 50070 will be forwarded to your localhost (we will shortly see why this is important).
ibm-power-1.dima.tu-berlin.de is the DNS name of the first host in the cluster, containing 10 nodes. This node is running the HDFS namenode and the secondary namenode.

To check the status and other information regarding the HDFS, open your browser and go to http://localhost:50070. Information including DFS properties such as capacity, journal status, startup progress and data nodes.

For each user an empty HDFS home directory is created. The content can be listed by:

```
hdfs dfs -ls
```

The DFS can't be accessed directly such as the local file system. *hdfs dfs* therefore acts as a proxy accepting most of the known Unix file system commands.

HDFS Command Overview

Create a directory in your HDFS home directory

```
hdfs dfs -mkdir aim3/
```

Create a file on your local file system and upload it to the HDFS

```
echo "Hello HDFS, how are you?" > data
hdfs dfs -put data aim3/
```

View the content of /aim3

```
hdfs dfs -ls aim3/
```

View the content of /aim3/data

```
hdf dfs -cat aim3/data
```

View the size of /aim3/data

```
hdfs dfs -du -h aim3/data
```

HDFS commands can be combined with other Unix tools such as *less* or *head*

```
hdfs dfs -cat aim3/data | less
```

Copy /aim3/data to /aim3/data_copy

```
hdfs dfs -cp aim3/data aim3/data_copy
```

Download /aim3/data_copy to your local filesystem and name it data_localcopy

```
hdfs dfs -get aim3/data_copy data_localcopy
```

Remove /aim3/data from the HDFS

```
hdfs dfs -rm aim3/data
```

Remove /aim3 from the HDFS

```
hdfs dfs -rm -r aim3/
```