### Upcoming Events

- Looking for one-on-one HP C help? Email us to book an HP C Consultation.

### New Pages

- Spotted any errors or omissions in our new Wiki? Tell us about it!
| Transferri ng data to/from Mercer cluster using Globus |
| Writing and submittin g jobs with qsub |
| Available software |
| Building Software |
| Tutorials |
| Research Gallery |
| FAQs |
| NYU HPC Cheat Sheet |
| Wiki Map |

(Tip: click "<<" at bottom left to close Confluence sidebar)

---

**Dumbo - Hadoop cluster**

**Dumbo** is our 48-node Hadoop cluster, running Cloudera CDH 5.8.0 (Hadoop 2.6.0 with Yarn).

All HPC users also have an account on Dumbo. (See [Getting or renewing an HPC account](#) for instructions to get an account. If you are enrolled in a class using the clusters you may already have an account, try logging in first).

Your `$HOME` directory on Dumbo is different than the `$HOME` on the HPC cluster, Mercer. See [FAQ](#) for a way to set the command prompt to show which cluster you are on.

- Dumbo - Hadoop cluster
- Hardware specifications
- Logging In
  - Dynamic port forwarding
  - Hadoop Web UIs
- Storage on Dumbo
  - Moving files between your workstation and HDFS
  - Huge Data Transfer between your workstation and Dumbo
Hardware specifications

<table>
<thead>
<tr>
<th>System Name</th>
<th>Dumbo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Master nodes</td>
<td>&quot;babar&quot; and &quot;hathi&quot;</td>
</tr>
<tr>
<td></td>
<td>2×12-core Intel &quot;Haswell&quot; (c 2014) CPUs</td>
</tr>
<tr>
<td></td>
<td>128GB memory</td>
</tr>
<tr>
<td></td>
<td>8TB RAID1 disk</td>
</tr>
<tr>
<td>Login and Compute nodes</td>
<td>2 login/edge nodes: &quot;dumbo0&quot; and &quot;dumbo1&quot;. Logging in to &quot;dumbo&quot; gets a randomly selected login node.</td>
</tr>
<tr>
<td></td>
<td>44 compute nodes: &quot;compute-#-#&quot;.</td>
</tr>
<tr>
<td></td>
<td>2×8-core Intel &quot;Haswell&quot; (c 2014) CPUs</td>
</tr>
<tr>
<td></td>
<td>128GB memory</td>
</tr>
<tr>
<td></td>
<td>16×2 TB disk for HDFS</td>
</tr>
<tr>
<td>Network</td>
<td>10 Gb Ethernet</td>
</tr>
<tr>
<td>Operating System</td>
<td>Linux (Centos 6.5)</td>
</tr>
<tr>
<td>File systems</td>
<td>/home - 20GB quota per user, common $HOME</td>
</tr>
<tr>
<td></td>
<td>HDFS - 1.4 PB configured. 128MB block size</td>
</tr>
</tbody>
</table>

Logging In

If you wish to use Hadoop Web UIs, you will need to setup dynamic port forwarding before logging in.

Dynamic port forwarding

HDFS and other Hadoop components have Web-based user interfaces you can access with a web browser. However, all HPC clusters are behind the HPC bastion host and not visible to the NYU network or the outside world.

To make them visible, first configure Firefox to use a SOCKS proxy server.

1. Start Firefox
2. Under "Preferences", select "Advanced" and then "Network":
3. Next click "Settings" for "How Firefox connects to the Internet"
Select "Manual proxy configuration", and set SOCKS Host to "localhost" on port 8118 (you can choose any port you like, but you must use the same port in the ssh command below)

5. Check "SOCKSv5" and "Remote DNS"

6. Set up your SSH tunnel as per the instructions here (Windows / Mac)
   For Mac/Linux, add dumbo0 and dumbo1 as hosts, with a "DynamicForward" option as follows:
Host hpctunnel
   HostName hpc.nyu.edu
   User NetID
   LocalForward 8023 mercer.es.its.nyu.edu:22
   # you can add other hosts to the tunnel too - for example, to specifically use
   login node 1 on mercer:
   LocalForward 8024 mercer1.es.its.nyu.edu:22
   # and for dumbo
   LocalForward 8025 dumbo.es.its.nyu.edu:22
   # this line gives you WebUI access:
   DynamicForward 8118

Host dumbo
   HostName localhost
   Port 8025
   ForwardX11 yes
   User NetID

For Windows:

Please follow the instructions on the link and then move on to the next step: Logging in from a Windows workstation

In the "Tunnels" dialog of your hpctunnel, add a new forwarded port, but this time the Source port is "8118" and check "Dynamic" instead of "Local"
7. Now, after you have logged in to hpc.nyu.edu, you can log in to dumbo in the same way you log in to Mercer.

**Hadoop Web UIs**

Two very useful UIs are the Cloudera Manager (http://babar.es.its.nyu.edu:7180/) and Hue Interface (http://babar.es.its.nyu.edu:8888/). Please enter either link on the firefox web browser to access the UI.

Cloudera Manager is an administrator tool, but you can view the status of the cluster by logging in as "nyu" with the password "nyu".

The hue interface login is your NetId and Password.

**Storage on Dumbo**

You have two places for files on Dumbo: your $HOME and your space on HDFS.

Both are somewhat limited: $HOME has about 2.3TB of space shared by all users, and HDFS has about 16 TB, by 3 replicas, shared by all users. **We are currently not setting quotas on either $HOME of HDFS, but please be mindful that the disk space is limited and shared.** Also, it is not backed up.

**Moving files between your workstation and HDFS**

There are three ways to do this:

1. Use the SSH tunnel to scp between your workstation and $HOME, and use 'hadoop fs -put' and 'hadoop fs -get' to move between $HOME and HDFS. If you do this, please remove extraneous files for $HOME afterwards:
Transfer from workstation to HDFS

my_workstation$ scp my_file dumbo:
my_workstation$ ssh dumbo
dumbo$ hadoop fs -put my_file
dumbo$ rm my_file

Transfer from HDFS to workstation

my_workstation$ ssh dumbo
dumbo$ hadoop fs -get my_file
dumbo$ exit
my_workstation$ scp dumbo:my_file .
my_workstation$ ssh dumbo rm my_file

2. Use the SSH tunnel and NFSGateway to scp files directly between your workstation and HDFS:

NFSGateway is currently down - please use method 1 or 3 instead

Transfer from workstation to HDFS

my_workstation$ scp my_file dumbo:/mnt/hdfs/user/<my_netid>/

Transfer from HDFS to workstation

my_workstation$ scp dumbo:/mnt/hdfs/user/<my_netid>/my_file .

3. Use the Hue Interface file browser option which is located in the top dashboard. On the right right hand side you will have an option to upload and download files to and from your workstation.

Huge Data Transfer between your workstation and Dumbo

Using Globus to transfer files to and from NYU HPC storage. Use Globus to transfer the data to /scratch on mercer first. The same /scratch folder is also available on dumbo login nodes.

The Globus project aims at providing powerful tools for scientific data management, to help researchers to focus on their domain subjects and solve data intensive research problems. Globus has been grown maturely to enable grid computing by connecting computing resources distributed globally across organizational boundary. Universities, national laboratories and computing facilities are using services of Globus.

Using Hive

Hive defaults to using the "default" database. This means that tables you create in this database might clash with tables others have created. Therefore, we highly recommend creating and using your own database, eg:
Hive CLI is deprecated and migration to Beeline is recommended.

bash-4.1$ beeline

beeline> !connect jdbc:hive2://babar.es.its.nyu.edu:10000/
Enter username for jdbc:hive2://babar.es.its.nyu.edu:10000/: <net_id>
Enter password for jdbc:hive2://babar.es.its.nyu.edu:10000/: ***********

0: jdbc:hive2://babar.es.its.nyu.edu:10000/> create database <netid> ;
0: jdbc:hive2://babar.es.its.nyu.edu:10000/> use <netid> ;

Please make sure to enter the Hive Heap size with your query for a MapReduce job.

Currently we do not have a way to automatically choose a suitable compute node to connect to - you should choose one at random. The Cloudera Manager WebUI can help, on the front page there are charts showing host CPU usage and host memory usage. It is wise to avoid hosts which already show high CPU or memory usage!

You can see a list of all compute hosts by clicking the "Hosts" link on the top dashboard of the Cloudera Manager WebUI main page.

Using Impala

To connect to impala use hostname where impala daemon is running. In fact, impala daemons are running on all slaves/datanodes on compute nodes of dumbo. So, connecting to any compute nodes, impala will work.

Here is the process to connect to Impala-shell.

bash-4.1$ impala-shell
Starting Impala Shell without Kerberos authentication
Error connecting: TTransportException, Could not connect to login-1-1.local:21000
********************************************************************
Welcome to the Impala shell. Copyright (c) 2015 Cloudera, Inc. All rights reserved.
(Impala Shell v2.6.0-cdh5.8.0 (8d8652f) built on Tue Jul 12 15:43:17 PDT 2016)

[Not connected] > connect compute-1-1;
Connected to compute-1-1:21000
Server version: impalad version 2.6.0-cdh5.8.0 RELEASE (build 8d8652f69461f0dd8d5f474573fb5de7ce0ee6b)
[compute-1-1:21000] > use <netid> ;
[compute-1-1:21000] >

Using Python

Python defaults to version 2.6 on dumbo. Incase there is need to work on python with libraries i.e., NumPy, SciPy, Pandas, NLTK, Gensim, Geopandas, follow below steps which points to Python 2.7.11 with required libraries. Contact hpc@nyu.edu for installing new libraries.

NOTE: If you want to work with Hadoop Streaming, Refer Big Data Tutorial 1: MapReduce for MapReduce Hadoop Streaming execution process.
Using Spark

Spark is a data-processing framework that operates on the distributed data collections. Hadoop and Apache Spark are both big-data frameworks, which really serve the different purposes. Hadoop is essentially a distributed data infrastructure, it distributes massive data collections across multiple nodes within a cluster of commodity servers whereas Spark is a data-processing tool.

Refer Big Data Tutorial 3 to learn more about Spark.

Here is the process to work with Spark on Dumbo:

```
-bash-4.1$ spark-shell
Welcome to
    / __/|__ _ _ _|__ /__
   \
   /._\ / _ _\ | / / / /\ _ _ _ _\ _
  /___/_/\_\_/\_/\_\_\_\_\_\_\_\_\_\_\_\ version 1.6.0
/\Using Scala version 2.10.5 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0_79)
Type in expressions to have them evaluated.
Type :help for more information.
Spark context available as sc (master = yarn-client, app id = application_1472492286365_0001).
SQL context available as sqlContext.
scala>
```

Please refer back to this page in the near future for instructions on how to use Oozie, Flume and Kafka.