Crucible
DATA SCIENCE ACCELERATED

SoM Azure HPC Onboarding Guide

Preparation

Join the *public-hardac* Crucible Slack channel

You will receive an email invitation or other direct guidance (if you're not sure, contact us). This will be the main place to go for communication and technical support for HPC.

Make sure you are on the Duke network

You must be on the Duke medicine network to connect to the Azure HPC cluster.
Use the VPN `dmvpn.duhs.duke.edu`

Connect to the login node

Using your CLI, SSH into the login node

```
$ ssh [NetID]@somhpc-scheduler.azure.dhe.duke.edu
...
...password: [xxxxx]
...
@ip-0A260C0B:~$
```

Migrate data to Azure HPC cluster

Data migration currently involves two steps:
First, from Source Server to Temporary Azure Blob Store (`dhpsomhpchardacs01`)
Second, from Temporary Azure Blob Store (`dhpsomhpchardacs01`) to Azure HPC cluster scratch space

1) From Source Server to Temporary Azure Blob Store (`dhpsomhpchardacs01`)

Create a personal container under the Azure storage container called `sharedcontainer`. For direct access to `sharedcontainer`, use this link.

If you do not have permission to access the container, reach out to your lab's PI to be added to your lab's grouper group.

Once the personal container is created, you can upload data to the temporary Azure blob store using **AzCopy**.
Follow this link for instructions on downloading and using AzCopy.
You will need to download and install AzCopy on a device that has access to the Duke medicine network/VPN or DHTS approved Duke public IP range

Sign in with AzCopy on a browser

Using your CLI, enter `$ azcopy login` and follow the instructions displayed:

```
To sign in, use a web browser to open the page https://microsoft.com/devicelogin and enter the code [XXXXXX] to authenticate.
INFO: Logging in under the "Common" tenant. This will log the account in under its home tenant.
INFO: If you plan to use AzCopy with a B2B account (where the account's home tenant is separate from the tenant of the target storage account), please sign in under the target tenant with --tenant-id
```

Open the url in a browser and put in the code generated in the terminal, then sign in with your Duke credentials.

Once you have signed in successfully, you will see this in the terminal:

```
INFO: Login succeeded.
```

Next, run the following command to transfer files:

```
azcopy copy [source-data-path] https://dhpsomhpchardacsa01.dfs.core.windows.net/sharedcontainer/[personal-directory]  --recursive
```

If data is transferred successfully you will see a response in the terminal:

```
INFO: Scanning...
INFO: Authenticating to destination using Azure AD
INFO: azcopy: A newer version 10.14.0 is available to download

INFO: Any empty folders will be processed, because source and destination both support folders

Job 4d9a081f-04d9-b347-775a-49d25acd71e1 has started
Log file is located at: /Users/jimmyhung/.azcopy/4d9a081f-04d9-b347-775a-49d25acd71e1.log

100.0 %, 0 Done, 0 Failed, 0 Pending, 0 Skipped, 0 Total (scanning...),

Job 4d9a081f-04d9-b347-775a-49d25acd71e1 summary
Elapsed Time (Minutes): 0.0334
Number of File Transfers: 5
Number of Folder Property Transfers: 1
Total Number of Transfers: 6
Number of Transfers Completed: 6
Number of Transfers Failed: 0
Number of Transfers Skipped: 0
TotalBytesTransferred: 37
Final Job Status: Completed
```

2) From Temporary Azure Blob Store (`dhpsomhpchardacsa01`) to Azure HPC cluster scratch space

Azure HPC scheduler node has the AzCopy CLI tool preinstalled. Once you ssh into the scheduler node and login with `azcopy login`, you can run the following command to transfer the files from temporary storage to your personal scratch space:

```
azcopy copy https://dhpsomhpchardacsa01.dfs.core.windows.net/sharedcontainer/[personal-directory] /data/[lab-directory]/[personal-directory] --recursive
```

NOTE: `dhpsomhpchardacsa01` is a temporary data placeholder. It is recommended that you remove data from the personal container once you finish the second step of data migration. Azure HPC cluster `/data` directory should be the permanent final destination.

Get code onto the cluster

Port 22 is disabled on Azure HPC cluster scheduler node (SSH goes over port 22). Downloading repository is only viable over HTTPs:

Download Gitlab Repos:

Refer to Gitlab's documentation on how to generate a personal access token.

When selecting scopes for the token, `read_repository` and `write_repository` scopes are sufficient in most scenarios.

Once you obtain your personal access token, you can download the repository in your scheduler terminal:

```
git clone https://[token-name]:[token]@[repository-url]
```

Example:

```
git clone https://hpctoken:long-token-string@gitlab.oit.duke.edu/myrepository
```

Download Github Repos:

Refer to Github's documentation on how to generate a personal access token and download a repository.

Frequently Asked Questions

Q: Can I use `scp` to copy files to Azure HPC node?

A: Due to the limitation of the ExpressRoute network capacity (the network that connects Duke Health network to Azure) `scp` is not recommended for transferring large dataset (hundreds of GB or TB).

Use `azcopy CLI` to transfer data over to the temporary blob store (`dhpsomhpchardacsa01`) over public internet with a higher network throughput. `azcopy` also has built-in parallel data transferring feature to expedite large data transfers.

Q: Why I can't ssh to x machine from scheduler node?

A: Outbound port 22 is closed by design

Need help?

Email questions to the Crucible team or post on the Slack channel `public-hardac`