

Setting Up Your Microsoft Azure with PHP Account

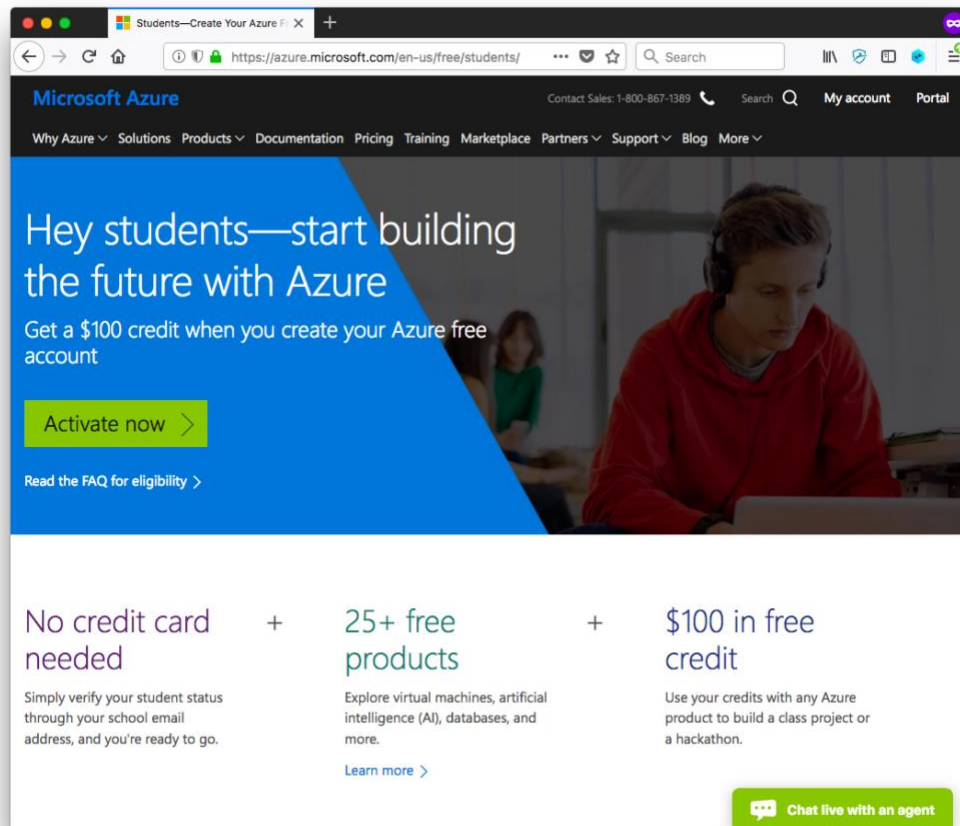
This semester we are allowing all students to explore cloud computing as offered by the Microsoft Azure cloud service.

1. Sign up for Microsoft Azure for Students

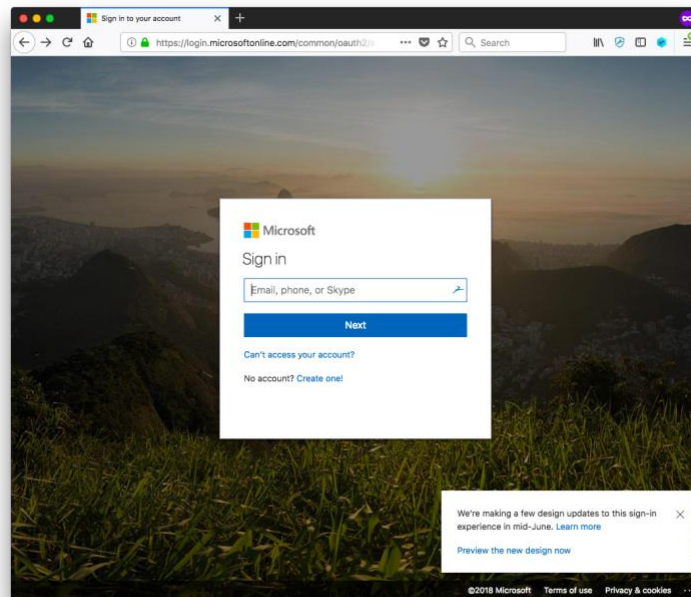
To sign up for the Free Trial, with a \$100 credit, you will not need a credit card. First of all, use a browser in **Incognito or In Private mode** (such as Chrome – File New Incognito Window or Firefox – File – New Private Window).

To sign up go to: <https://azure.microsoft.com/en-us/free/students/>

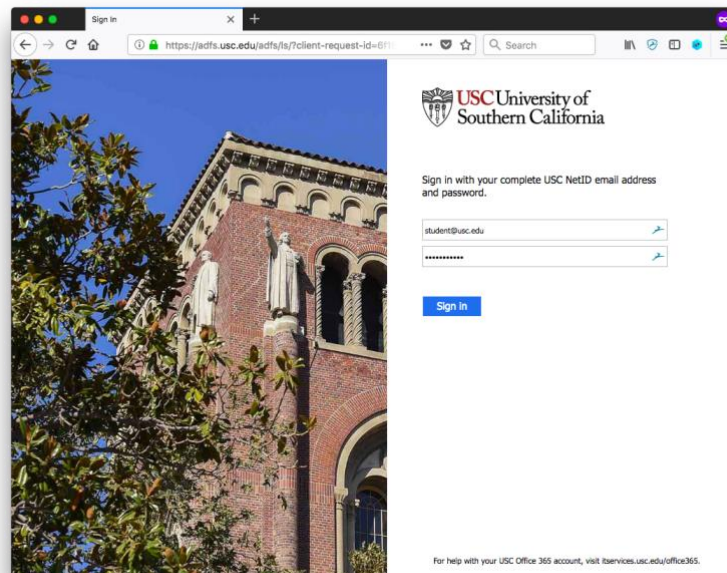
In the Microsoft Azure for Students page, click on **Activate now >**:



The Microsoft**Sign in** page will be displayed.

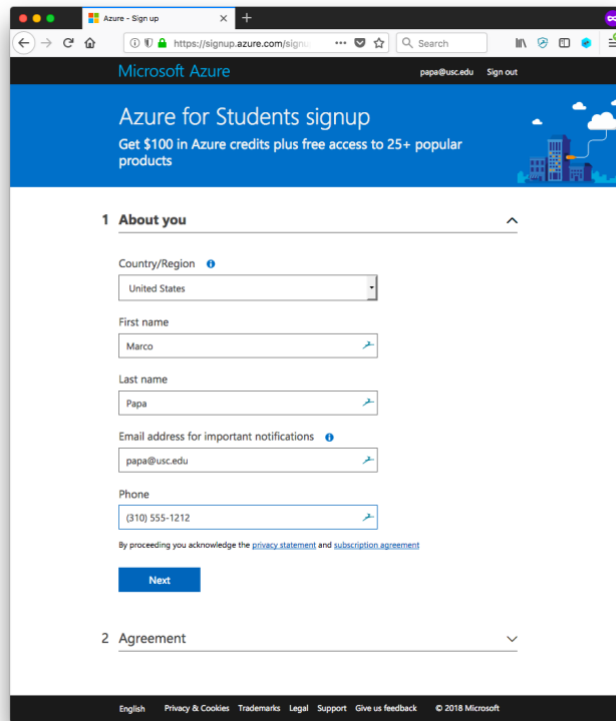


Enter your**USC e-mail address**. The USC Shibboleth page will be displayed and will be used to verify your academic status and receive the \$100 credit.



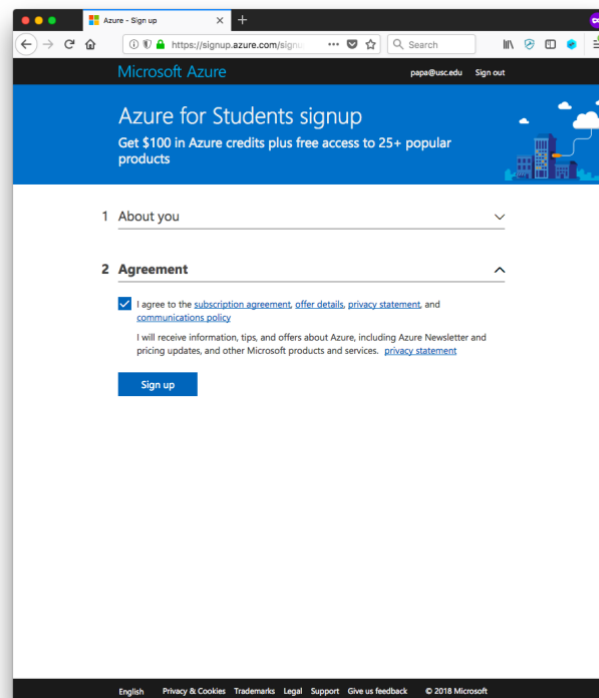
Enter your**USC account password**. Click**Sign in**.

The Azure for Students sign up page is displayed.



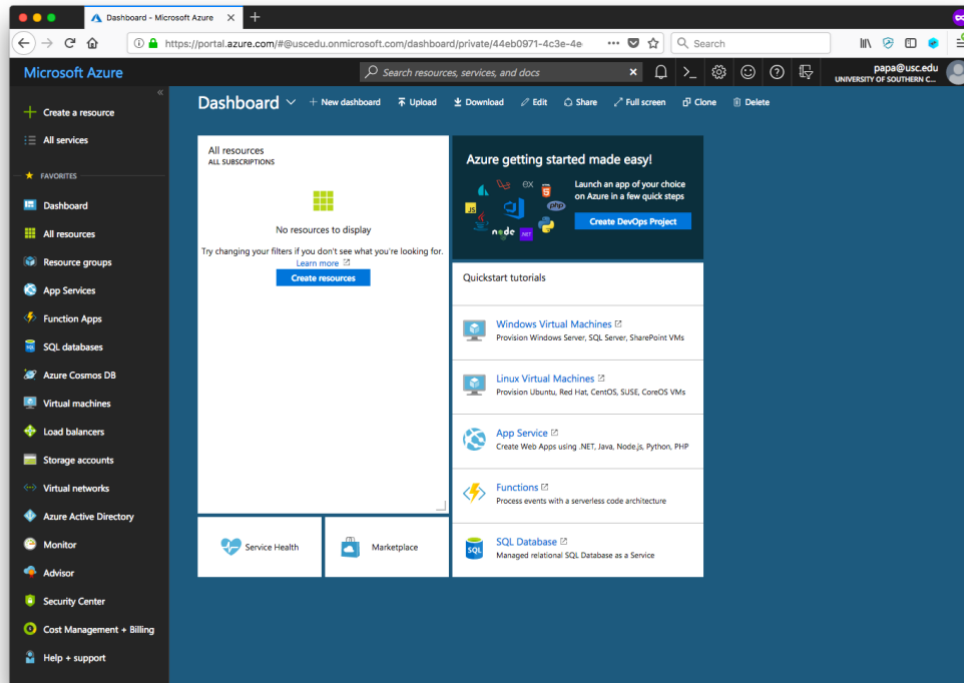
A screenshot of a web browser showing the 'Azure for Students sign up' page. The browser's address bar shows 'https://signup.azure.com/signup'. The page has a blue header with the Microsoft Azure logo and a 'Sign out' link. Below the header, the main heading is 'Azure for Students sign up' with a subtext 'Get \$100 in Azure credits plus free access to 25+ popular products'. The page is divided into two sections: '1 About you' and '2 Agreement'. The 'About you' section contains several form fields: 'Country/Region' (a dropdown menu showing 'United States'), 'First name' (text input with 'Marco'), 'Last name' (text input with 'Papa'), 'Email address for important notifications' (text input with 'papa@usc.edu'), and 'Phone' (text input with '(310) 555-1212'). Each text input field has a small blue icon to its right. Below these fields is a link to 'privacy statement' and 'subscription agreement'. At the bottom of the 'About you' section is a blue 'Next' button. The '2 Agreement' section is currently collapsed.

Enter your personal data in the form. Click **Next**.

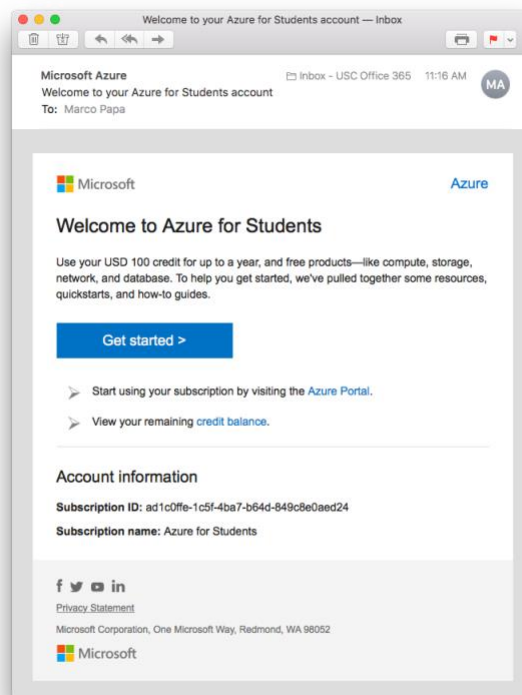


A screenshot of the same 'Azure for Students sign up' page, but now the '2 Agreement' section is expanded. The '1 About you' section is collapsed. The '2 Agreement' section contains a checkbox that is checked, with the text 'I agree to the [subscription agreement](#), [offer details](#), [privacy statement](#), and [communications policy](#)'. Below this is a paragraph: 'I will receive information, tips, and offers about Azure, including Azure Newsletter and pricing updates, and other Microsoft products and services. [privacy statement](#)'. At the bottom of the 'Agreement' section is a blue 'Sign up' button. The footer of the page contains links for 'English', 'Privacy & Cookies', 'Trademarks', 'Legal', 'Support', 'Give us feedback', and '© 2018 Microsoft'.

Agree to the free subscription offer by selecting the checkbox. Click **Sign Up**. You will be taken to the Microsoft Azure Dashboard.



You should also receive a confirmation **“Welcome to Azure for Students”** e-mail.



2. Create a HDInsight cluster.

Azure offers services to perform all the big data activities on the cloud. For that, Azure provide users a machine compatible of with all the big data services such as Hadoop, Pig, Hive, Sqoop, Kafka etc. Earlier, Azure used to provide both windows and linux virtual machines, but eventually they deprecated windows and thus now only linux machines are remaining. Below is a way to perform map-reduce operation on Azure.

(**Note:** Windows being Microsoft's own product, they have provided many different tools for windows clients, but the process below is useful for everyone (Windows, Linux, Mac).)

You can follow the steps below to create a new HDInsight cluster :

1. Go to portal.azure.com and log in using your USC ID (with which, you have accepted the 100\$ credits).
2. Go to 'Create a Resource' and search for **Azure HDInsight**. Open the service and click on **Create**. This will take you to the configurations window.
3. For configurations, there are 5 tabs. Let us start with **Basics** :

Subscription : Azure for Students (If you have crated student's account properly)

Resource Group : Create one or use the existing one (It is just for project grouping)

Cluster Name : Give a good cluster name

Region : Choose the corresponding region to your place. (For USC, West US 2)

Cluster Type : Choose Hadoop. (This will enable version tab)

Version : Choose the Hadoop version which you have used to write the code.

Cluster Credentials : These are the credentials, which will be used to access the new cluster machine. There are two ways to access the machine, one through portal and other through SSH.

Cluster login username : Choose any username you want.

Cluster login password : Choose a good password according to Azure's policy. (1 Uppercase Letter, 1 Lowercase Letter, 1 Special Symbol, 1 number, length > 10)

Secure Shell Username : This username is going used to SSH into our system. It should be different than 'Cluster login username'.

You can check the checkbox saying 'Use cluster login password for SSH' to keep the same password, else you can create a new password as well.

Microsoft Azure

Home > New > Marketplace > Azure HDInsight > Create HDInsight cluster

Create HDInsight cluster

Go to classic create experience

your resources

Subscription * Azure for Students

Resource group * AjinkyaChaturHW3CSC571

Cluster details

Name your cluster, pick a region, and choose a cluster type and version. [Learn more](#)

Cluster name * hw3test

Region * West US 2

Cluster type * Hadoop

Version * Hadoop 2.7.3 (HDI 3.6)

Cluster credentials

Enter new credentials that will be used to administer or access the cluster.

Cluster login username * admin

Cluster login password *

Confirm cluster login password *

Secure Shell (SSH) username * sshuser

Use cluster login password for SSH ☒

Review > create

Previous

Next: Storage >

Now click **Next : Storage**. This will take you to the storage tab. Inside that :

Primary Storage Type : Azure Storage

Selection Method : Select from List

Primary Storage Account : Create New (Give a good name.)

Keep everything else as it is. This will create a new primary storage for us for all the HDFS storage during map-reduce job.

Microsoft Azure

Home > New > Marketplace > Azure HDInsight > Create HDInsight cluster

Create HDInsight cluster

Go to classic create experience

Basics Storage Security + networking Configuration + pricing Review + create

Select or create storage accounts that will be used for the cluster's logs, job input, and job output. Configure the cluster's access to these accounts, if needed.

Primary storage

Select or create a storage account that will be the default location for cluster logs and other output.

Primary storage type * Azure Storage

Selection method * ☒ Select from list ☐ Use access key

Primary storage account * (New) hw3testhdstoragea

Container * hw3test-2019-11-26t14-46-30-200z

Data Lake Storage Gen1

Provide details for the cluster to access Data Lake Storage Gen1. The cluster will be able to access any Data Lake Storage Gen1 accounts that the chosen service principal has access to.

Data Lake Storage Gen1 access [Configure access settings](#)

Additional Azure storage

Link additional Azure storage accounts to the cluster.

Account name

Add Azure storage

Metastore settings

Review > create

Previous

Next: Security + networking >

We will not change any any **Security + Networking**. In **Configuration and pricing** tab, you can change the systems type and number for the operation. The cheapest one is **D12 v2 (4 Cores, 28 GB RAM)** and that works fine with us.

Microsoft Azure

Home > New > Marketplace > Azure HDInsight > Create HDInsight cluster

Create HDInsight cluster

Go to classic create experience

Basics Storage Security + networking **Configuration + pricing** Review + create

Configure cluster performance and pricing. [Learn more](#)

Node configuration

Configure your cluster's size and performance, and view estimated cost information.

The cost estimate represented in the table does not include subscription discounts or costs related to storage, networking, or data transfer.

ⓘ This configuration will use 24 of 230 available cores in the West US 2 region. [View cores usage](#)

Add application

Node type	Node size	Number of ...	Estimated cost/hour
Head node	D12 v2 (4 Cores, 28 GB RAM), 0.37 USD/...	2	0.75 USD
Worker node	D12 v2 (4 Cores, 28 GB RAM), 0.37 USD/...	4	1.50 USD

☐ Enable autoscale [Learn more](#)

Total estimated cost/hour 2.24 USD

[Review + create](#) [Previous](#) [Next: Review + create >](#)

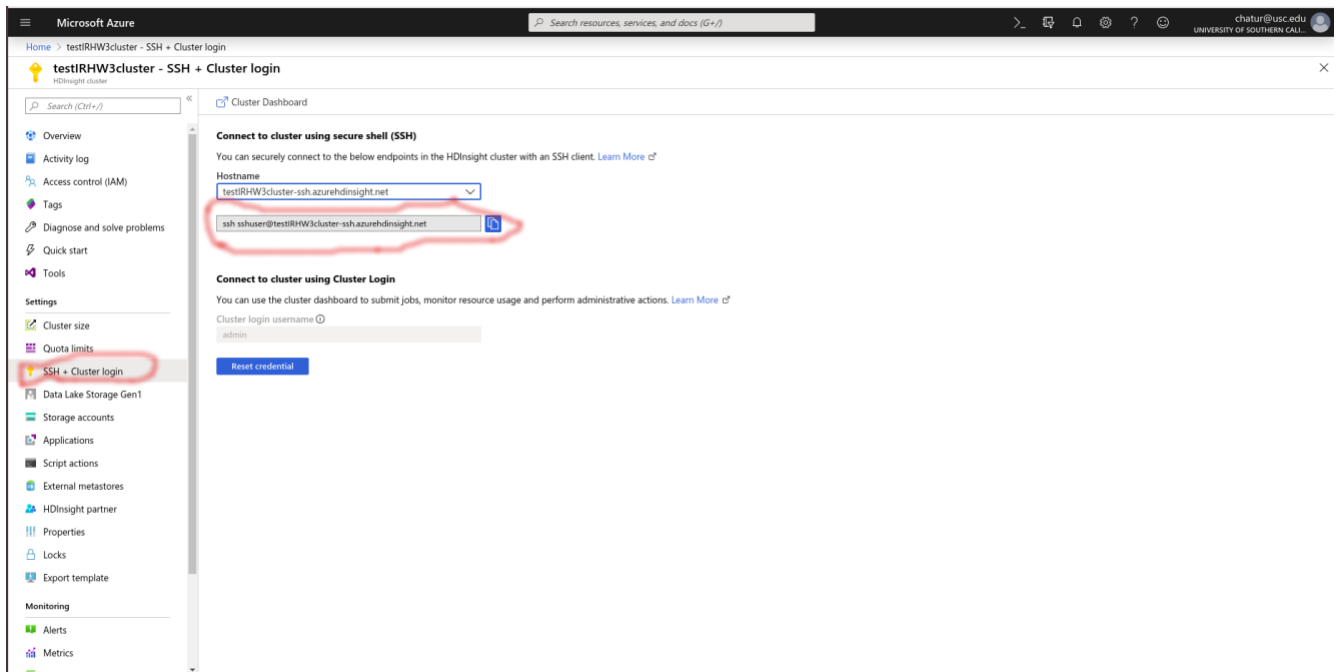
Go to **Review + Create**. This might take some time as Azure is creating the system with specified configurations. This will give a review of all the configurations. If you want, you can note down the usernames here. After this click on **Create** to provision the cluster. It will redirect to a screen showing **Your deployment is underway**. (Note: This might take some time.)

Issues :

If the **Status** section shows anything other than **OK** then change the resource group of your cluster.

3. Run map-reduce job

Once the cluster is provisioned, go to dashboard by clicking **Microsoft Azure** on top left corner. There search for **HDInsight Cluster**. It will showcase all the HDInsight clusters you have. Choose the one you want to work on. In the left menu go to **SSH + Cluster login**. Inside **Hostname** dropdown, choose your cluster name. This will autofill the textbox below. Copy everything in that textbox. We will need it to do SSH into this cluster.



Now open **Terminal** if you have linux or MAC or **Command Prompt** in case of windows and paste the content copied. For the first time, it will ask you for username, where you will have to give username you filled in the field of **Secure Shell Username** previously. Then it will ask for password. Once that is done, you are inside your cluster system.

For Azure, this is a new Linux machine, with all the packages and some built in samples for you. So, once you are inside the system, you can test if everything has been done successfully by using :

```
yarn jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-mapreduce-examples.jar wordcount /example/data/gutenberg/davinci.txt /example/data/davinciwordcount
```

This command will run a sample word count map reduce job on existing data.

After this, for your map reduce job, first of all you will have to upload data from your local machine to the cluster machine. For that go to the folder where your data is using new terminal / command prompt. There use the following command :

```
scp devdata sshuser@testIRHW3-ssh.azurehdinsight.net:
```

Keep the command as it is. The devdata is folder name of input data. Next to it is address to our cluster machine. The format of address is

```
sshusername@clustername-ssh.azurehdInsight.net:
```

The **‘.’** in the end are important as they denotes the root directory.

With this, if you type **ls** inside the terminal / command prompt, which is used to do SSH into cluster system, you can see the folder.

Similarly upload the **JAR** file of code you have created.

Now we successfully uploaded the data and JAR file to a cloud machine, but now we need to upload this data to the Hadoop environment of that machine. For that, inside cloud machine type,

hadoop fs -copyFromLocal devData input

using this command, we are copying the local data folder **devData** to **input** folder on hadoop environment. To check if it has been successfully transferred, use following command :

hdfs dfs -ls /

This command lists all the files and folders of HDFS.

With this, we are ready to run our map-reduce job. For that, use command :

yarn jar jarname.jar /input /output

(Note: for each map reduce job, output folder should be unique. So, for each try, you will have to use a new name for output.)

Wait for job to complete and you will have your output inside output folder.

Issues :

One of the most common errors is **Java Heap Space error**. This is a hadoop error and not the Azure's. For that, inside your cluster machine type following commands :

cd /etc/hadoop/conf

this will take you to configurations directory. There type :

sudo nano mapred-site.xml

This will open the file in **nano** editor. You can use **Vim** if you want. Inside that, look for :

```
<property>
  <name>mapreduce.map.java.opts</name>
  <value>-Xmx756M -Xms756M -Djava.net.preferIPv4Stack=true -XX:NewRatio=8 -
XX:+UseNUMA -XX:+UseParallelGC</value>
</property>
```

field. There change 756 to 5048. Also, look for :

```
<property>
```

```
<name>mapreduce.map.memory.mb</name>
```

```
<value>756</value>
```

```
</property>
```

and change the value to 5048. The type **ctrl+X**. There it will ask you if you want to overwrite, press **Enter** , then it will ask for the file name, press **Enter** again. It will save the changes.

Here we have increased the heap space availed to map function in terms of Mbs.

4. Close all services

Once all the work is done, it is very important to close all the services, otherwise Azure will keep charging you for them. Before doing this, download the output folder. For that, we will first copy the output folder from Hadoop environment to cluster machine and then from there to out local machine. For that, use the following commands on cluster machine :

```
sudo hadoop fs -copyToLocal /hadoopOutputDir outputClusterDirName
```

This will copy the hadoop output to cluster machine output. Now to download it to local machine, use :

```
scp root@server.ip.adress:/root/file.txt ~/homecomputer/directory
```

This will copy output file from cluster machine to local machine. Now is the time to close all the services. For that, go to the dashboard at **portal.azure.com** .

Search for **HDInsight cluster**. Go to the cluster and click on **Delete** to delete the HDInsight cluster. Then again go back to dashboard and search for **Storage Accounts**. Click on the storage account you have created for the job and click on delete.

This will delete all the Azure services, which you were using.

Have fun and explore Azure Web Services!