CSCI4180 Tutorial-2

Hadoop Setup on OpenStack Windows Azure Guide

ZHANG, Mi

mzhang@cse.cuhk.edu.hk

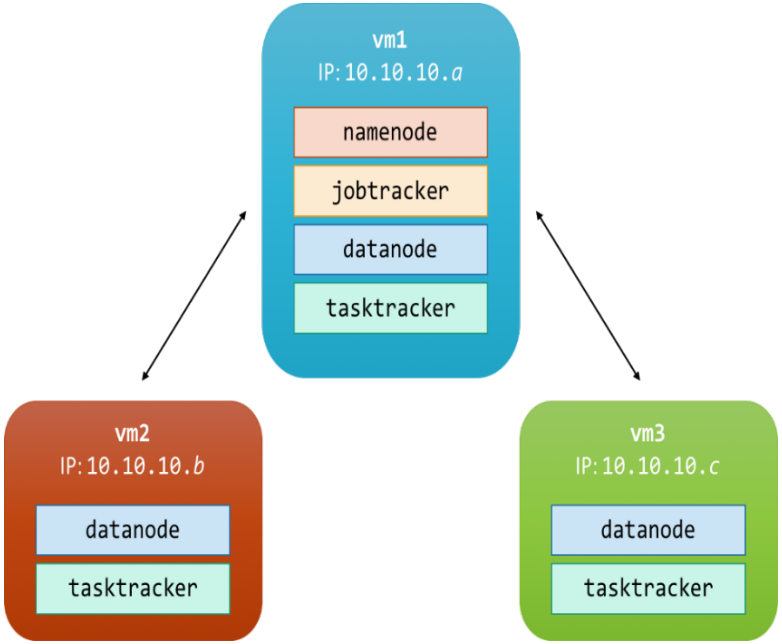
Sep. 24, 2015

Outline

- Hadoop setup on OpenStack
  - Set up Hadoop cluster
  - Manage Hadoop cluster
  - WordCount Example
- Windows Azure guide
  - Access Azure
  - Create VMs
  - Install Hadoop

Set up Hadoop Cluster

- We’ve created three VM instances of our own.
  - ➢Architecture



Set up Hadoop Cluster

- We'll set up small-scale Hadoop cluster using these VM instances.

- What you've done in tutorial-1:
  - Setting up HTTP proxy.
  - Installing Java.
  - Configuring /etc/hosts.

Set up Hadoop Cluster

- Switch to normal user “hadoop”
  - su - hadoop
- If you do not have user “hadoop”
  - adduser hadoop
  - ➢enter your password when necessary…
  - su - hadoop

Set up Hadoop Cluster

- Download Hadoop on EACH node
  - wget http://archive.apache.org/dist/hadoop/core/hadoop-0.20.203.0/hadoop-0.20.203.0rc1.tar.gz
- Place Hadoop in home directory on EACH node
  - tar xzf hadoop-0.*.*.tar.gz
  - mv hadoop-0.*.* hadoop

Set up Hadoop Cluster

- Set environment variable on EACH node.
  - I recommend you put them in ~/.bashrc
  - export HADOOP_HOME=~/hadoop
  - export PATH=$PATH:$HADOOP_HOME/bin
- Set hadoop environment on EACH node.
  - Append the following lines to ~/hadoop/conf/hadoop-env.sh
  - export JAVA_HOME=/usr/lib/jvm/java-7-oracle
  - #depends on where you put the jvm
  - export HADOOP_OPTS=-Djava.net.preferIPv4Stack=true
- Set path for HDFS storage on EACH node.
  - #under HOME directory
  - mkdir hadoop/tmp

Set up Hadoop Cluster

- Configure SSH on EACH node
  - ssh-keygen -t rsa -P ""
  - cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys
- Configure SSH on namenode only
  - ssh-copy-id -i $HOME/.ssh/id_rsa.pub hadoop@vm1
  - ssh-copy-id -i $HOME/.ssh/id_rsa.pub hadoop@vm2
  - ssh-copy-id -i $HOME/.ssh/id_rsa.pub hadoop@vm3
- Check SSH configuration
  - ➢ whether namenode can ssh all the datanodes without typing password. E.g.,
  - ssh vm2

Set up Hadoop Cluster

- Set hadoop core on EACH node
  - Add property in ~/hadoop/conf/core-site.xml

```
<property>
  <name>hadoop.tmp.dir</name>
  <value>/home/hadoop/hadoop/tmp</value>
</property>
<property>
  <name>fs.default.name</name>
  <value>hdfs://vm1:54310</value>
</property>
```

Set up Hadoop Cluster

- Set hadoop mapreduce on EACH node
  - Add property in ~/hadoop/conf/mapred-site.xml

```
<property>
  <name>mapred.job.tracker</name>
  <value>vm1:54311</value>
</property>
```

Set up Hadoop Cluster

- Set hadoop HDFS on EACH node
  - Add property in ~/hadoop/conf/hdfs-site.xml

```
<property>
  <name>dfs.replication</name>
  <value>3</value>
</property>
```

Set up Hadoop Cluster

- Set hadoop master on namenode
  - Add hostname which is supposed to run *NameNode* and *JobTracker* in ~/hadoop/conf/masters
  - vm1
- Set hadoop slaves on namenode
  - Add hostname which is supposed to run *DataNode* and *TaskTracker* in ~/hadoop/conf/slaves
  - vm1
  - vm2
  - vm3

Set up Hadoop Cluster

- Format namenode on namenode
  - hadoop namenode –format
- Start hadoop on namenode
  - start-dfs.sh
  - start-mapred.sh
  - # you can type “jps” to see whether the startup is successful.
- Stop hadoop on namenode
  - stop-mapred.sh
  - stop-dfs.sh

Set up Hadoop Cluster

- Some operations related to HDFS
  - ➢From Local to HDFS
    - hadoop dfs -copyFromLocal <local dir/file> <hdfs URI>

  (for user home URI: /home/hadoop)
  - ➢From HDFS to Local
    - hadoop dfs -copyToLocal <hdfs URI> <local dir/file>
  - ➢List files in HDFS
    - hadoop dfs -ls <hdfs URI>
  - ➢Cat files in HDFS
    - hadoop dfs -cat <hdfs URI>

Manage Hadoop Cluster

- Add one more instance into cluster
  - Stop Hadoop services on namenode
  - For the new instance, repeat steps from slide 4 to slide 11
  - Add IP of new instance in ~/hadoop/conf/slaves on namenode
  - Format namenode and start Hadoop

Manage Hadoop Cluster

- Remove one instance from cluster
  - Stop Hadoop services on namenode
  - Remove IP of the instance from ~/hadoop/conf/slaves
  - Format namenode and start Hadoop

WordCount Example

- Download the java source code from course website, say, WordCount.java, to your namenode, home directory
- Compile and run the program
  - mkdir wordcount
  - javac -classpath $HADOOP_HOME/hadoop-core-0.20.203.0.jar WordCount.java -d wordcount
  - jar -cvf wordcount.jar -C wordcount/ .
  - hadoop jar wordcount.jar org.myorg.WordCount */HDFS URI/to/input/file* */HDFS URI/to/output/directory*
  - ➢ Note that the part-r-00000 is the actual output.

Windows Azure platform

- Windows Azure guide
  - Access Azure
  - Create VMs
  - Install Hadoop

Overview

- Get the 14-character code before you start.
- Redeem your Windows Azure at
  https://www.microsoftazurepass.com/azureu

Redeem the pass

Microsoft Azure

All Microsoft Sites

Search Microsoft Azure

Try Microsoft Azure

The Microsoft Educator Grant Program provides access to Azure for use in the classroom by university students and their professor.

ELIGIBILITY REQUIREMENTS:

- Requester must be a faculty member of an accredited university
- Passes must be used for a specific class
  a. Multiple classes should have multiple requests
  b. Multiple sections of same class may use one request
- Passes must be used by students
- Faculty must provide all information requested
  a. If you are unable to provide any of the requests, please attach a document explaining reason for omission, and another method to verify,

APPLICATION IS FOR EDUCATORS/FACULTY ONLY.

If you do not meet the eligibility requirements, the free trial program is also available for 1-month at no cost. (Free Trial Link)

Faculty will receive a 12 month, $250/month account
Students will receive a 6 month, $100/month account

All fields are required.

Full Name:

University Faculty Email:

Country: ---Select Country---

University (full name):

Course Name:

Course Description: (Please specifically indicate how you intend to leverage Microsoft Azure; 4000 character limit)

Course Start Date:

# of Student Passes Requested:

Course URL:

Faculty URL:

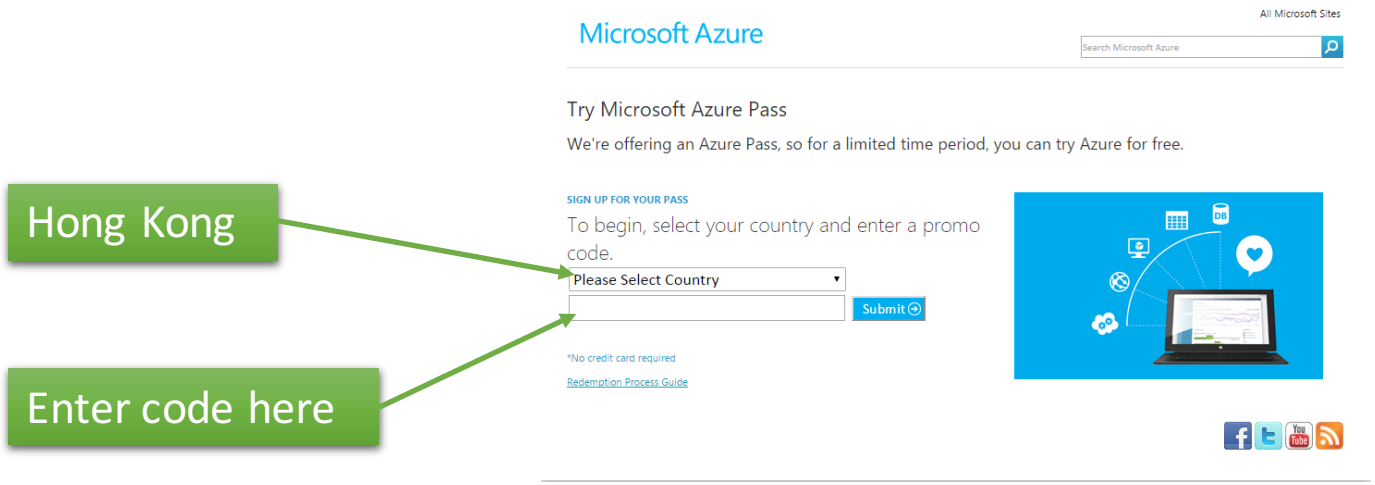
Course Syllabus: Choose File No file chosen

University ID: Choose File No file chosen

Click here

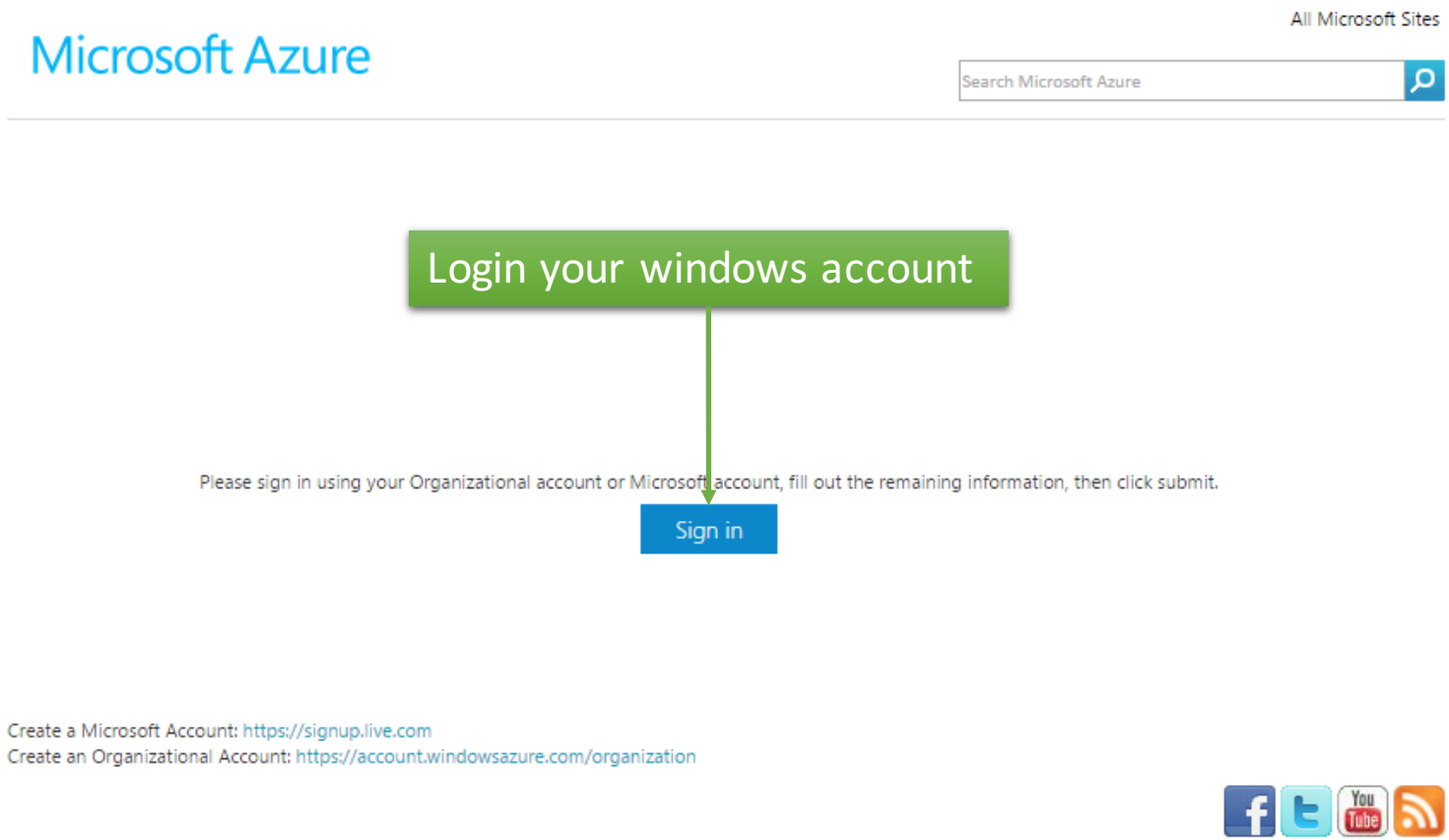
Have a code already? Redeem it here.

Redeem the pass

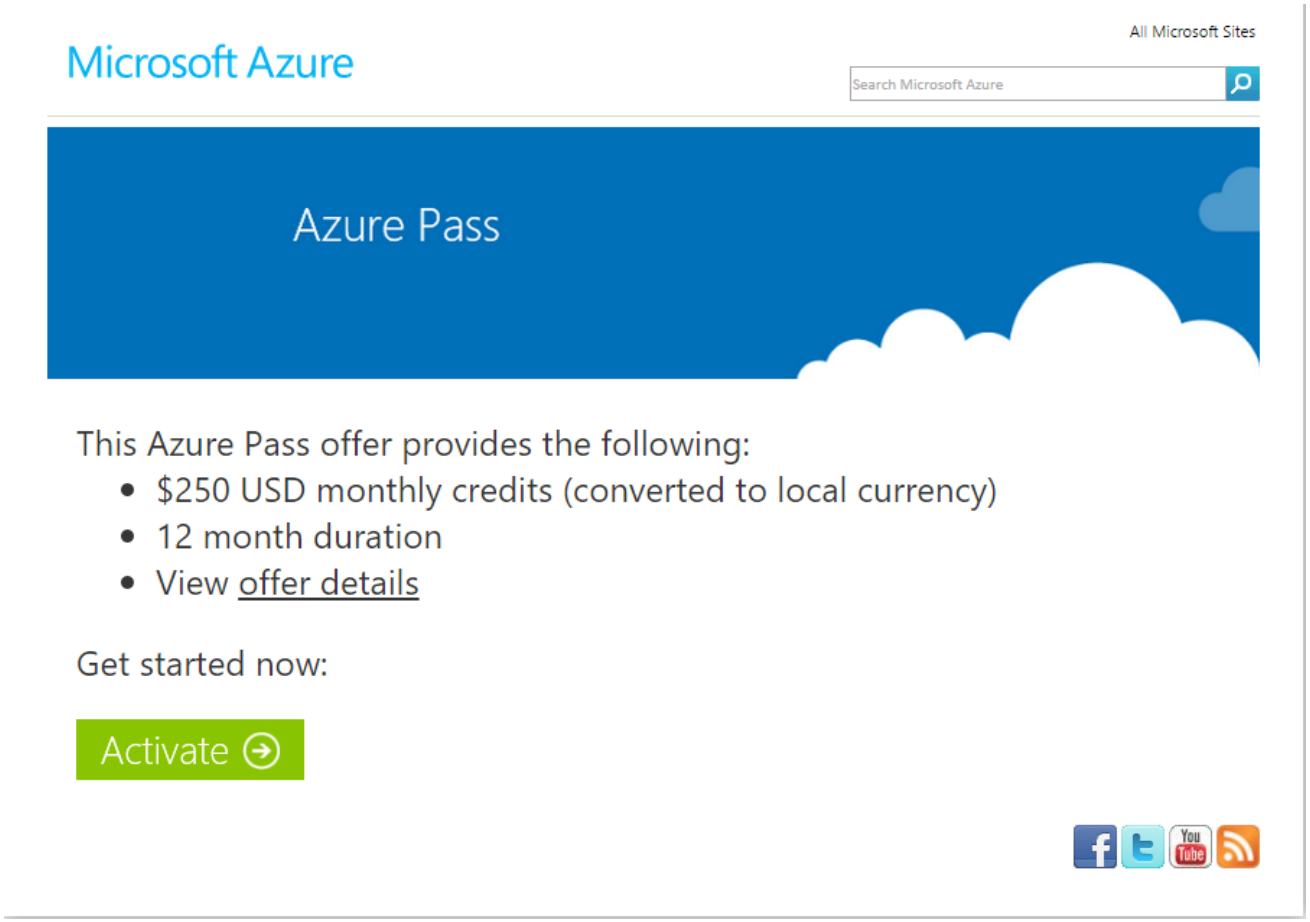


- Followed with some register information.
- To redeem the pass, you also need a windows live account.

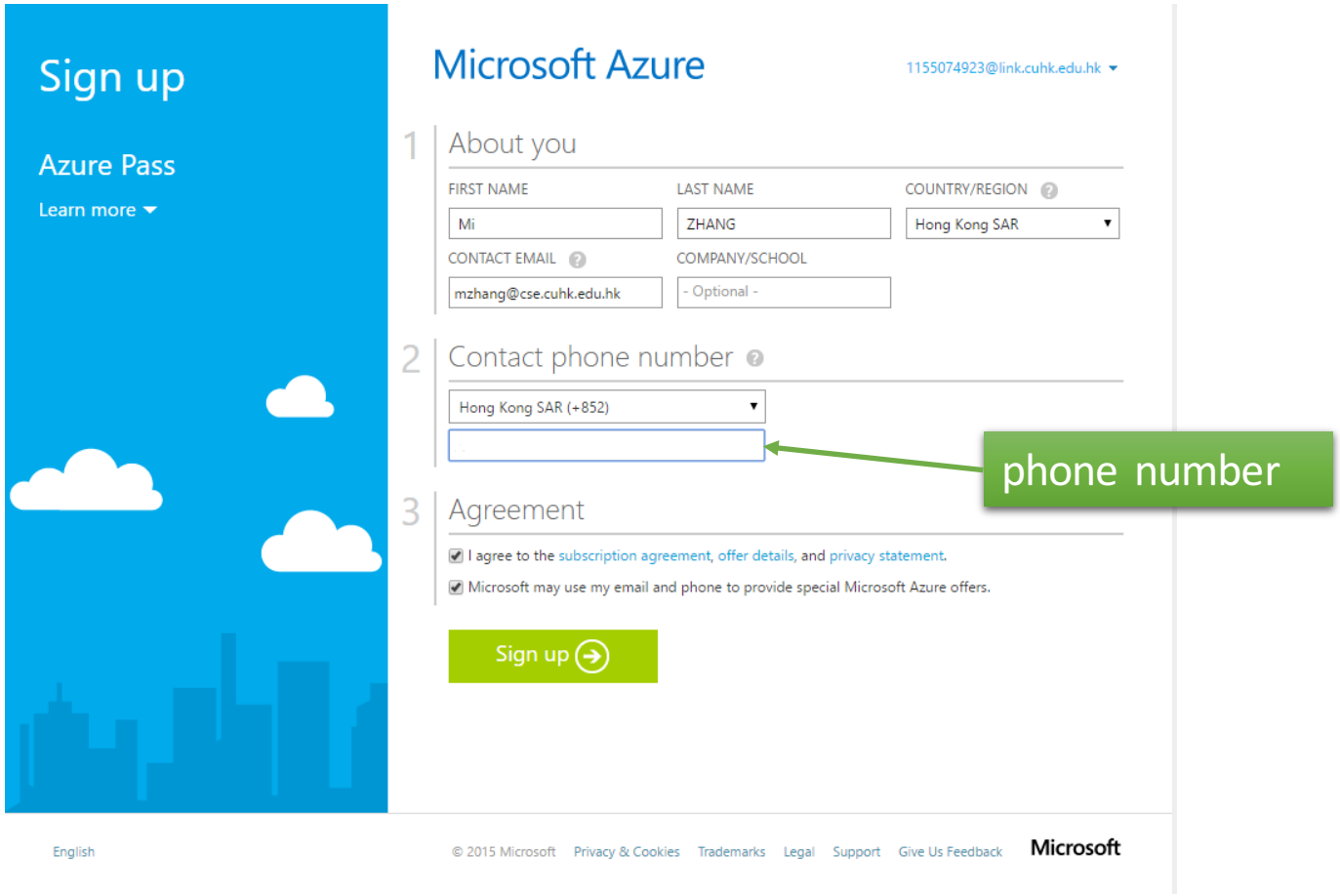
Redeem the pass



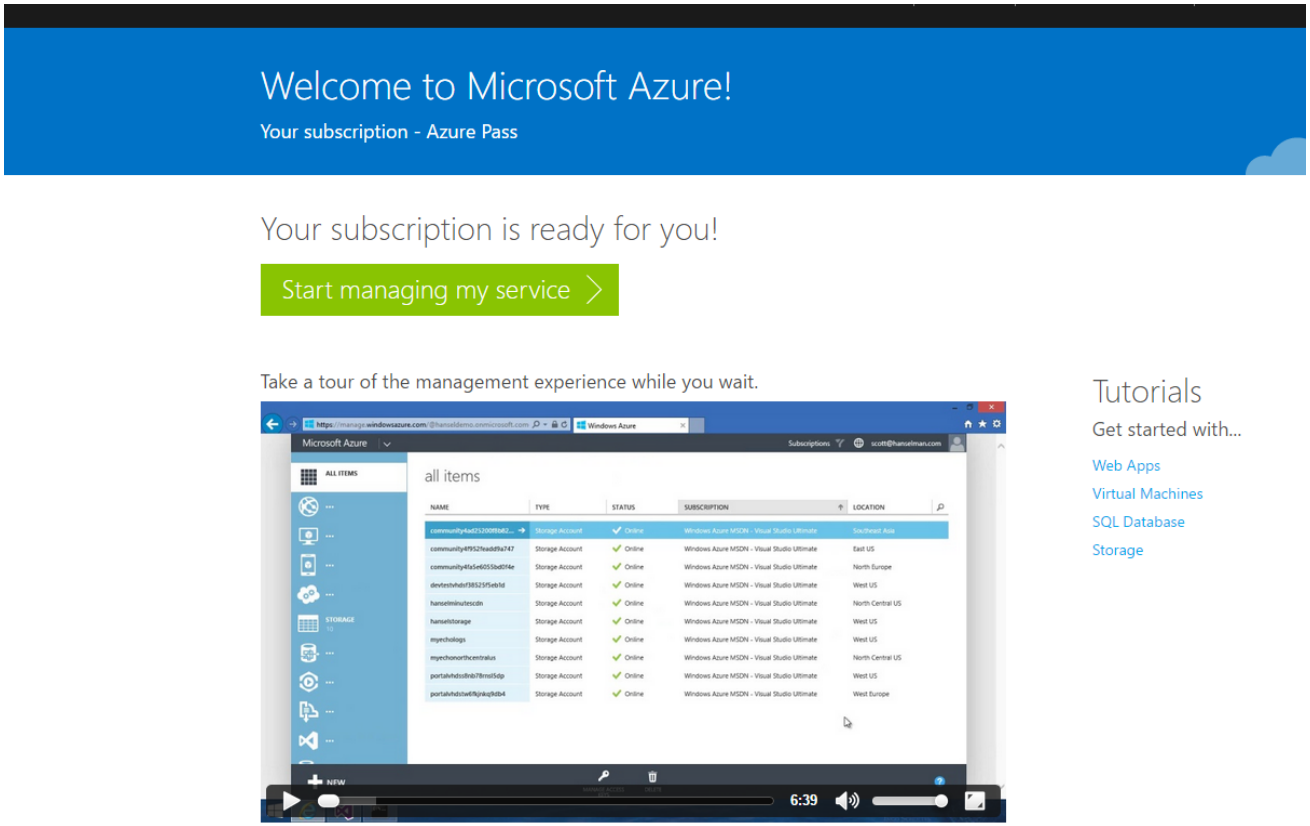
Redeem the pass



Redeem the pass

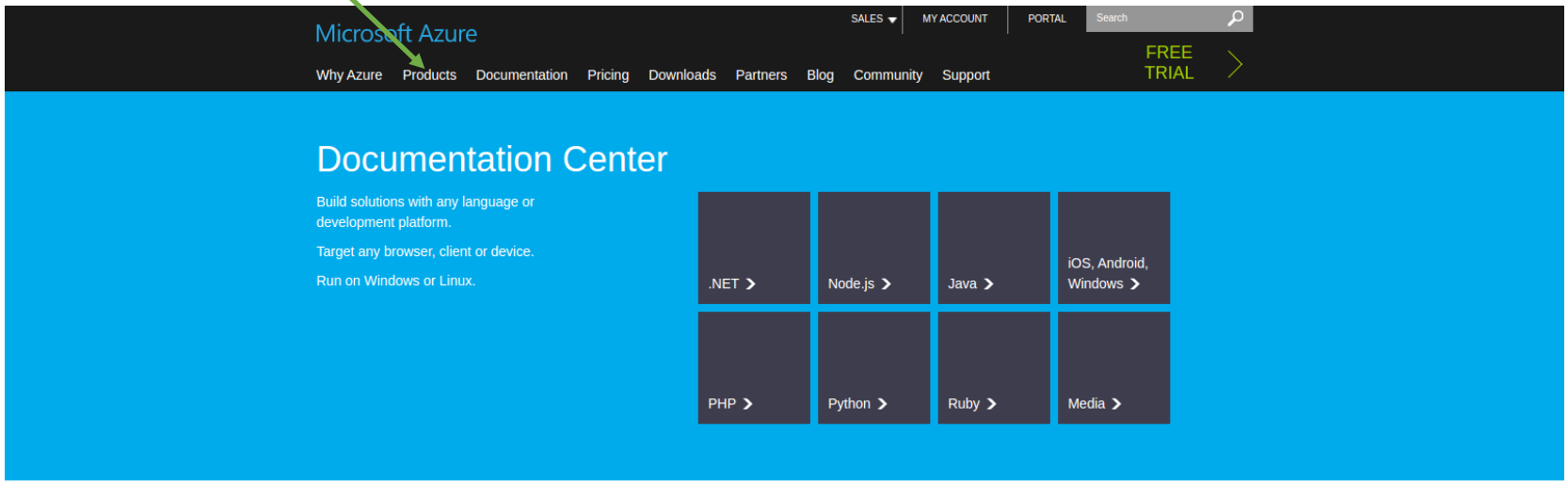


Redeem the pass



Create VMs

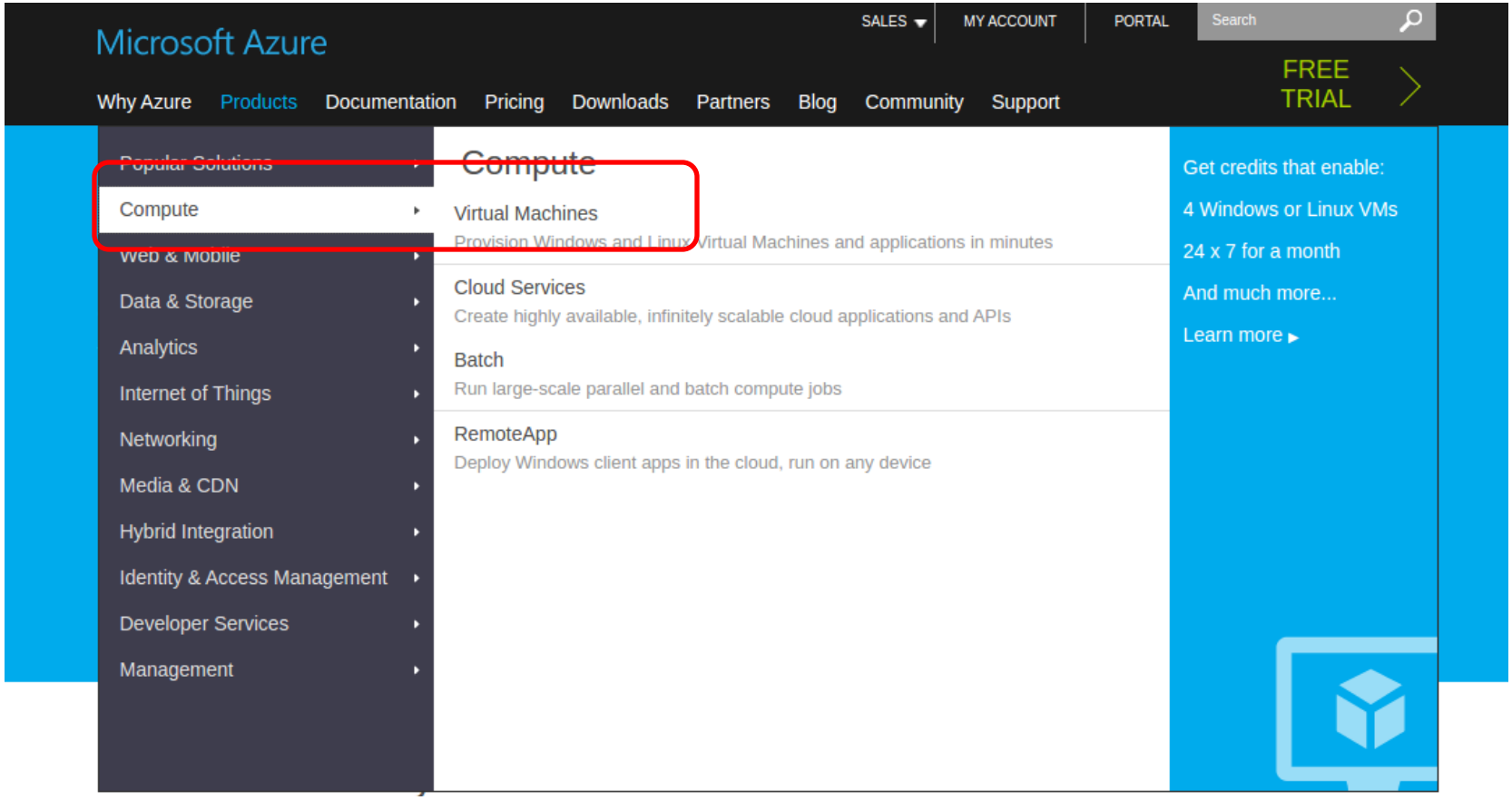
Click Products



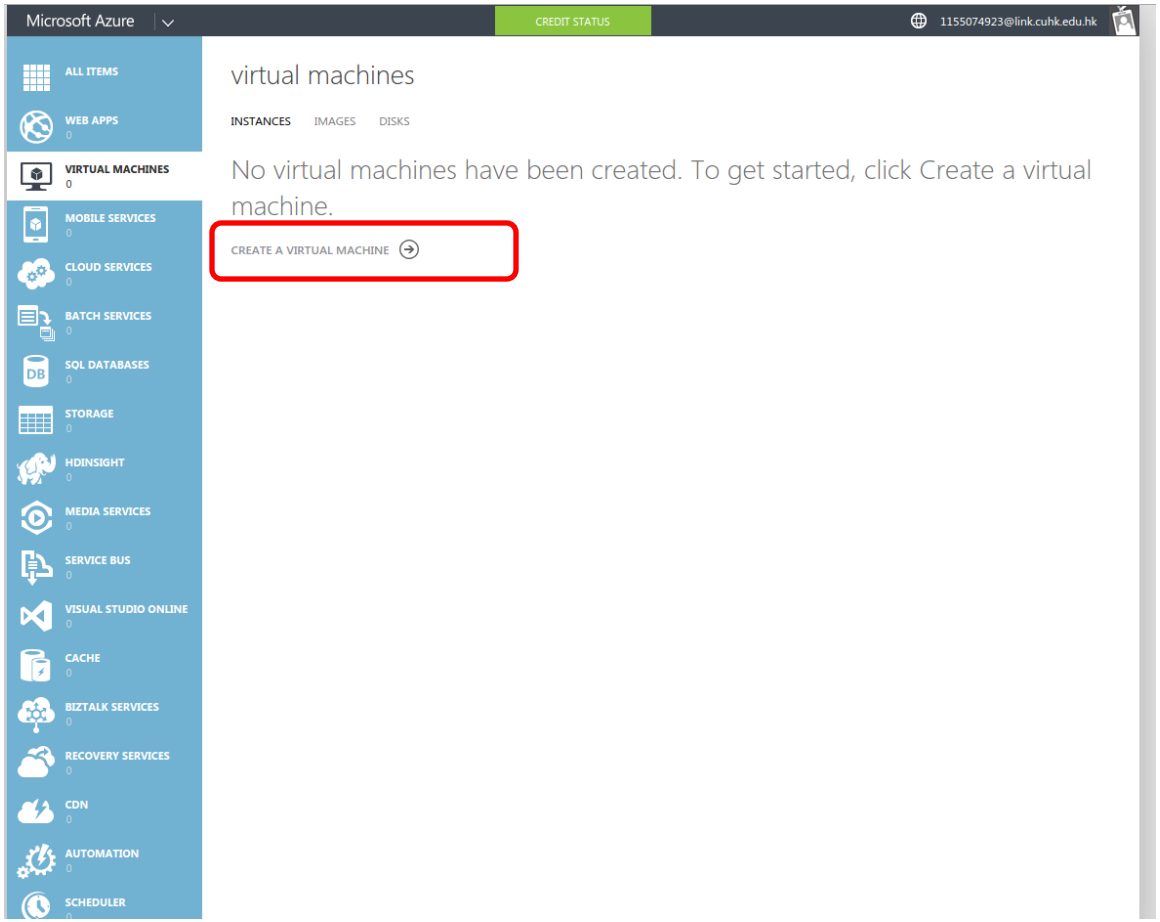
Documentation by service

Compute	Web & Mobile	Data & Storage	Analytics	Networking
Virtual Machines	App Service	DocumentDB	HDInsight	Virtual Network
Cloud Services	Websites	SQL Database	Machine Learning	ExpressRoute
Batch	Mobile Services	Redis Cache	Stream Analytics	Application Gateway
RemoteApp	API Management	Storage	Data Factory	Traffic Manager
Service Fabric	Push Notifications	StorSimple	Event Hubs	DNS
	Mobile Engagement	Azure Search	Data Catalog	Load Balancer
		SQL Data Warehouse		VPN Gateway

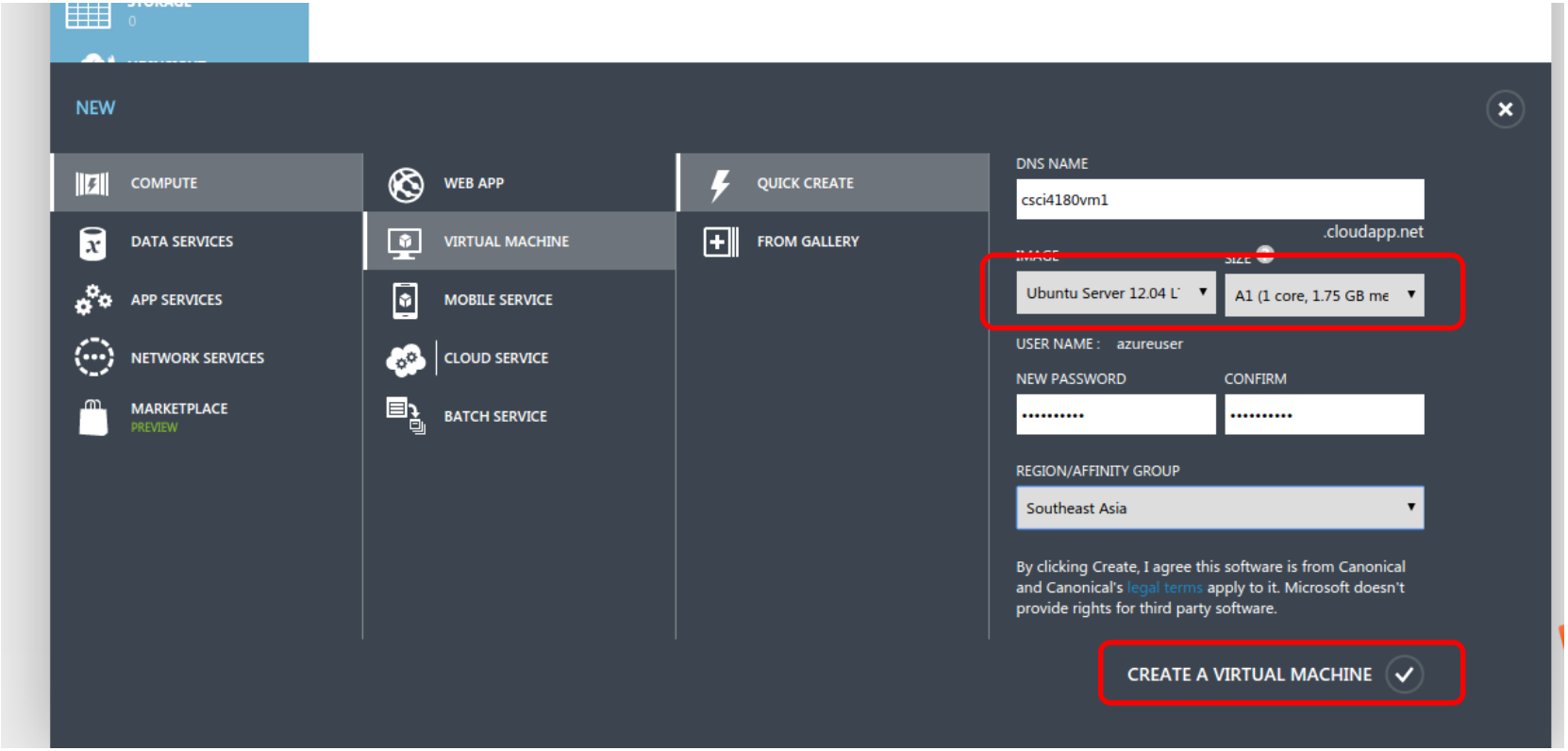
Create VMs


Microsoft Azure
SALES
MY ACCOUNT
PORTAL
Search
Why Azure
Products
Documentation
Pricing
Downloads
Partners
Blog
Community
Support
FREE TRIAL
Popular Solutions
Compute
Web & Mobile
Data & Storage
Analytics
Internet of Things
Networking
Media & CDN
Hybrid Integration
Identity & Access Management
Developer Services
Management
Compute
Virtual Machines
Provision Windows and Linux Virtual Machines and applications in minutes
Cloud Services
Create highly available, infinitely scalable cloud applications and APIs
Batch
Run large-scale parallel and batch compute jobs
RemoteApp
Deploy Windows client apps in the cloud, run on any device
Get credits that enable:
4 Windows or Linux VMs
24 x 7 for a month
And much more...
Learn more

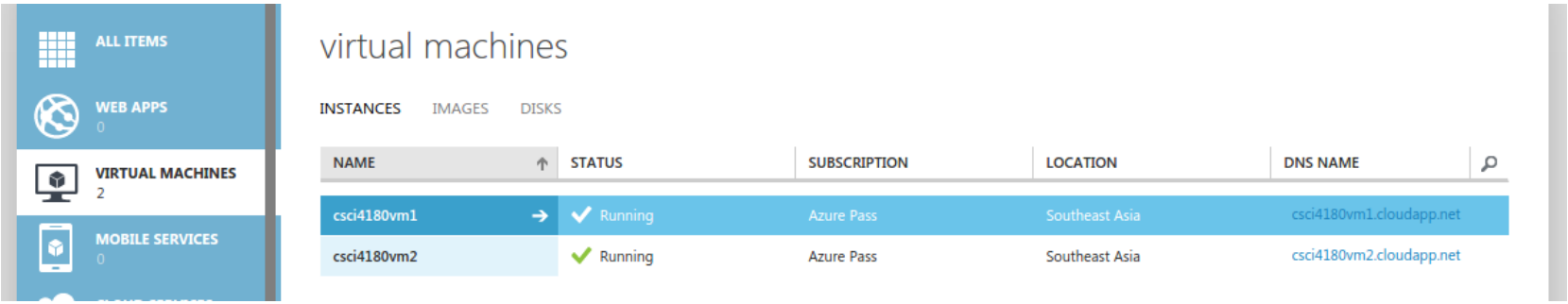
Create VMs


Microsoft Azure
CREDIT STATUS
1155074923@link.cuhk.edu.hk
ALL ITEMS
WEB APPS
VIRTUAL MACHINES
MOBILE SERVICES
CLOUD SERVICES
BATCH SERVICES
SQL DATABASES
STORAGE
HDINSIGHT
MEDIA SERVICES
SERVICE BUS
VISUAL STUDIO ONLINE
CACHE
BIZTALK SERVICES
RECOVERY SERVICES
CDN
AUTOMATION
SCHEDULER
virtual machines
INSTANCES
IMAGES
DISKS
No virtual machines have been created. To get started, click Create a virtual machine.
CREATE A VIRTUAL MACHINE

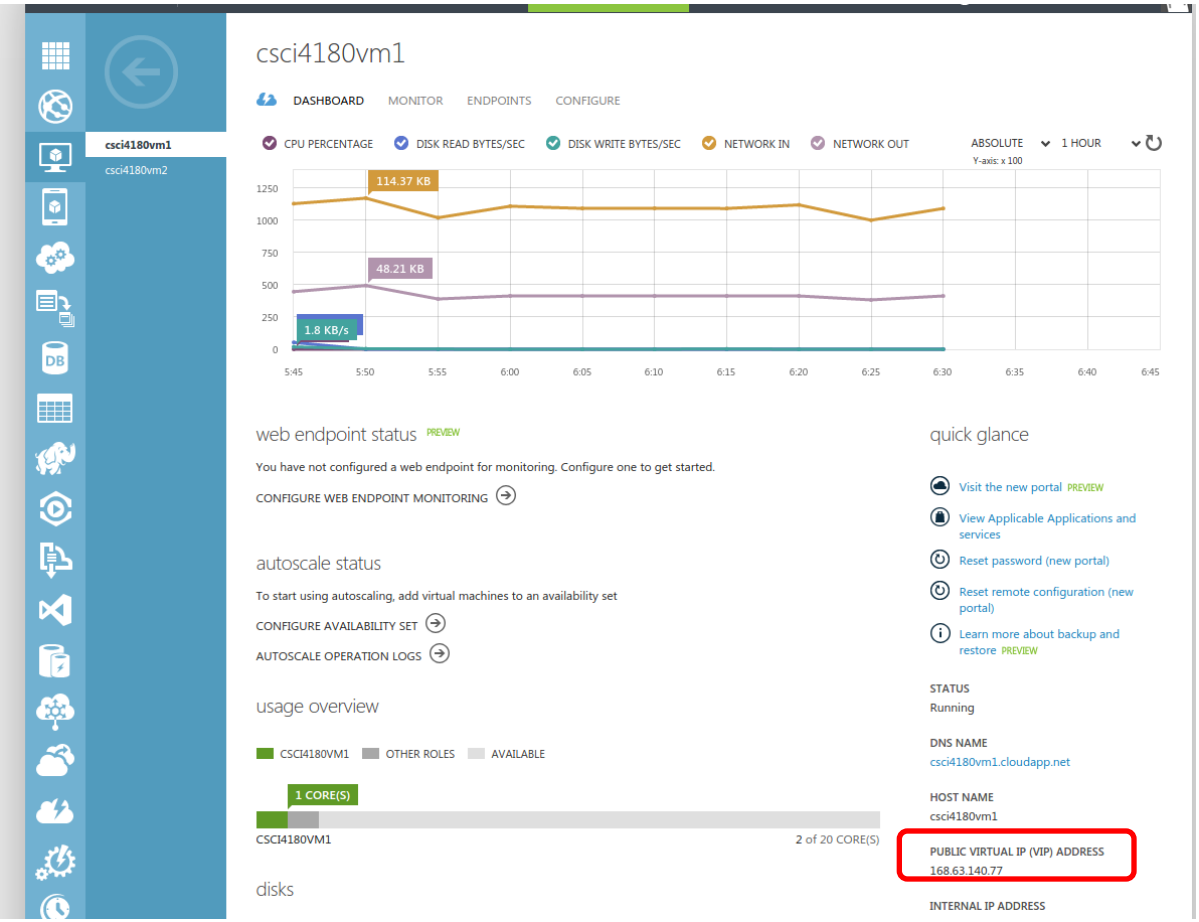
Create VMs


NEW
COMPUTE
DATA SERVICES
APP SERVICES
NETWORK SERVICES
MARKETPLACE
PREVIEW
WEB APP
VIRTUAL MACHINE
MOBILE SERVICE
CLOUD SERVICE
BATCH SERVICE
QUICK CREATE
FROM GALLERY
DNS NAME
csci4180vm1
.cloudapp.net
Ubuntu Server 12.04 L
A1 (1 core, 1.75 GB me
USER NAME : azureuser
NEW PASSWORD
CONFIRM
REGION/AFFINITY GROUP
Southeast Asia
By clicking Create, I agree this software is from Canonical and Canonical's legal terms apply to it. Microsoft doesn't provide rights for third party software.
CREATE A VIRTUAL MACHINE

Create VMs


ALL ITEMS
WEB APPS
0
VIRTUAL MACHINES
2
MOBILE SERVICES
0
virtual machines
INSTANCES
IMAGES
DISKS
NAME
STATUS
SUBSCRIPTION
LOCATION
DNS NAME
csci4180vm1
Running
Azure Pass
Southeast Asia
csci4180vm1.cloudapp.net
csci4180vm2
Running
Azure Pass
Southeast Asia
csci4180vm2.cloudapp.net

Create VMs



You can ssh to your VM using this IP.

Create VMs

- In your terminal, ssh azureuser@*your vm IP*

```
azureuser@csci4180vm1.cloudapp.net's password:
Welcome to Ubuntu 12.04.5 LTS (GNU/Linux 3.13.0-63-generic x86_64)

 * Documentation:  https://help.ubuntu.com/

  System information as of Wed Sep 16 11:27:08 UTC 2015

  System load:  0.04              Processes:           217
  Usage of /:   3.8% of 28.80GB   Users logged in:     0
  Memory usage: 7%                IP address for eth0: 10.62.162.156
  Swap usage:   0%

  Graph this data and manage this system at:
    https://landscape.canonical.com/

  Get cloud support with Ubuntu Advantage Cloud Guest:
    http://www.ubuntu.com/business/services/cloud

0 packages can be updated.
0 updates are security updates.

New release '14.04.3 LTS' available.
Run 'do-release-upgrade' to upgrade to it.
```

Create VMs

- If you SSH the VM via cse wired network, you may need to configure your ssh setting.
  - Append the following lines to ~/.ssh/config
    - Host *your hostname*
    - User azureuser
    - HostName *your hostname*
    - ProxyCommand nc -x socks.cse.cuhk.edu.hk:1080 %h %p
- Then you can login to the vm using ssh
  - ssh *your hostname*

Install Hadoop

- Install Java on EACH VM:
  - sudo apt-get update
  - sudo apt-get upgrade
  - sudo add-apt-repository ppa:webupd8team/java
  - sudo apt-get update
  - sudo apt-get install oracle-java7-installer
- You could follow the instruction in Tutorial 1.
  - http://mtyiu.github.io/csci4180-fall15/

Install Hadoop

- Repeat the process of installing hadoop on OpenStack from slide 5 to slide 11.
  - Slide 8: replace vm1, vm2 with their respective public IP.
  - Slide 9, 10: change "vm1" to "127.0.0.1" when editing .xml files.
- Set hadoop masters on namenode
  - Edit ~/hadoop/conf/masters
    - 127.0.0.1
- Set hadoop slaves on namenode
  - Edit ~/hadoop/conf/slaves
    - 127.0.0.1
    - *Another vm public IP *
- After starting HDFS and MapReduce, you can run the WordCount example.

Thank you!