CS4830 - Big Data Lab

Big Data Lab

Outline:

This course will introduce the students to practical aspects of analytics at large scale, i.e. big data. The course will start with a basic introduction to big data concepts spanning hardware, systems and software, and then delve into details of algorithm design and execution at large scale.

Goals:

1. Introduction to Big Data concepts: divide-and-conquer, parallel algorithms, distributed virtualized storage, distributed resource management, orchestration and scheduling, lambda architecture, data flow paradigm, real-time event processing.
2. Technology deep-dive: Map-Reduce using Java and Python, Spark for Batch processing, Spark SQL, data flow processing libraries (Beam, Spark Streaming, Flink).
3. Hardware deep-dive: Shared-nothing MPP architecture, Cloud architecture, GPU-based acceleration and processing.
4. Analytics at Large Scale: Libraries of algorithms including Spark MLlib, H20 Sparkling Water; integrations with TensorFlow and PyTorch; ML on cloud; use of Zeppelin and Databricks Notebooks.

Reference:

- Mining of massive datasets: - http://infolab.stanford.edu/~ullman/mmds/book.pdf
- Hortonworks website - https://hortonworks.com

Lab - 1

29th January 2019

Objectives:

1. Introduction to MapReduce. What is Hadoop and Spark?
2. Setup Microsoft Azure account and spawn a Virtual Machine (VM)
3. Setup Hortonworks HDP: https://hortonworks.com/tutorial/sandbox-deployment-and-install-guide/
4. Go through: https://hortonworks.com/tutorial/learning-the-ropes-of-the-hortonworks-sandbox/
5. Advanced resource: https://hortonworks.com/tutorial/sandbox-architecture/

1-page Report:

Answer the following questions:

1. What is a Virtual Machine? How is it different from a normal machine?
2. Read up on docker (https://docs.docker.com/engine/docker-overview/). Why do we need a service like docker?
3. What is Ambari? Name 3 services that ambari provides an interface to (vertical bar on the left of the Ambari interface).

Instructions:

Setting Up Azure:

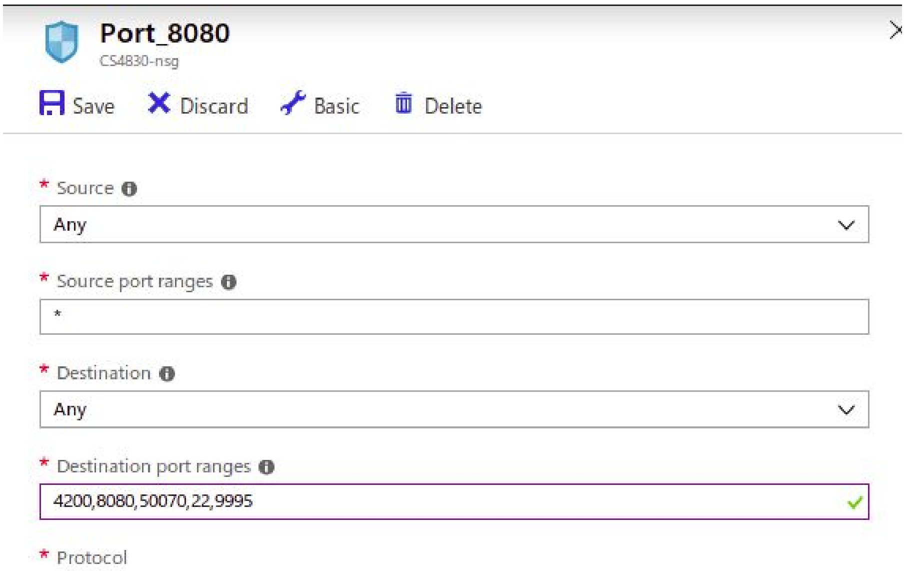
1. Create a Microsoft account: https://account.microsoft.com/account?lang=en-us
2. Go to https://aka.ms/JoinEdu and activate enter the code (443BC).
3. Now navigate to https://portal.azure.com/#blade/Microsoft_Azure_Education/ClassroomMenuBlade/assignments/classroomId/400
4. Now setup lab the lab (HDP-lab) which is part of the course "Big Data" https://portal.azure.com/#blade/Microsoft_Azure_Education/EducationMenuBlade/quickstart/section/quickstart
5. Click setup lab which should give you $50.00 in Azure credits (expires Jun 29, 2019)
6. Azure students FAQ: https://portal.azure.com/#blade/Microsoft_Azure_Education/EducationMenuBlade/support

Creating a virtual machine:

1. In the nav-bar on the left, click on **Virtual Machines**.
2. Click on **Add.**
3. In the **Basics** tab, specify the following settings:

    a. **Subscription** – Your subscription name
    b. **Resource Group** – Create a new resource group with your name
    c. **VM name** – CS4830
    d. **Region** – East US, East US 2
    e. No Infrastructure redundancy required
    f. **Image** – Ubuntu 18.04 LTS
    g. **VM Size** – Standard B4ms
    h. Authentication type: Password
    i. Set your username and password. (**Remember the credentials, you will use this throughout the semester. If you lose this password the VM data will be lost permanently)**
    j. In **Inbound port rules**, select **Allow Selected Ports** and select all four – HTTP, HTTPS, RDP, SSH.

4. Click on **Next: Disks**.
5. Select **OS disk type** as **Standard HDD** and click on **Next: Networking**.
6. Set **NIC security group** to be **Advanced** and make the following changes:
    a. Click on on **Create New** in **Configure Network Security Group**.
    b. Add **Inbound Rule**:
        i. Source=Any
        ii. Destination=Any
        iii. Source Port range=*
        iv. Destination port ranges=50070, 4200, 9995, 8080, 22,
        v. Priority=110
    c. Add **Outbound Rule**: (change name [to anything] since it won't allow same name as inbound rule)
        i. Source=Any
        ii. Destination=Any
        iii. Source Port range=*
        iv. Destination port ranges=80
    d. Above two commands allow access to select ports. It is generally unsafe but is okay for the purpose of this lab.



Port_8080
CS4830-nsg

* Source: Any

* Source port ranges: *

* Destination: Any

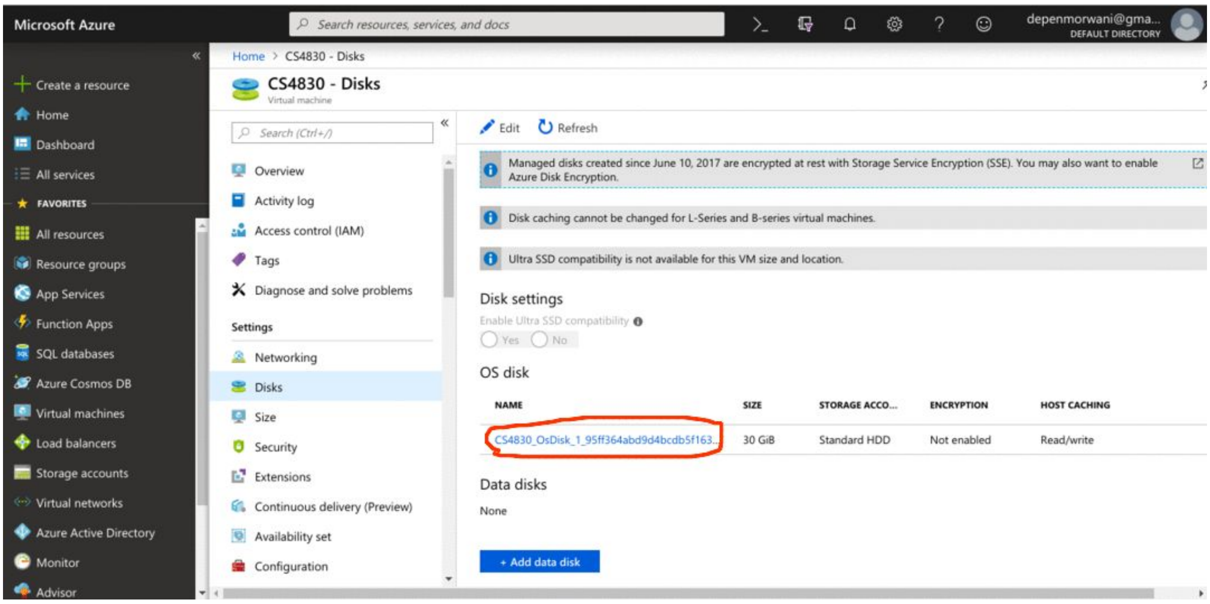
* Destination port ranges: 4200,8080,50070,22,9995

* Protocol

    e. Click on **Review+Create**, followed by **Create.**

Increasing Disk Size of the VM:

- In the notifications tab, when the resource deployment is complete (could take about 5-10 mins), click on **Go To Resource** and **"Stop" running Virtual Machine** by clicking on the **Stop** button.
- On the left hand side of the window, click on **Disks** option and select your disk displayed on the right hand side of the window:



Managed disks created since June 10, 2017 are encrypted at rest with Storage Service Encryption (SSE). You may also want to enable Azure Disk Encryption.

Disk caching cannot be changed for L-Series and B-series virtual machines.

Ultra SSD compatibility is not available for this VM size and location.

Disk settings

Enable Ultra SSD compatibility

○ Yes ○ No

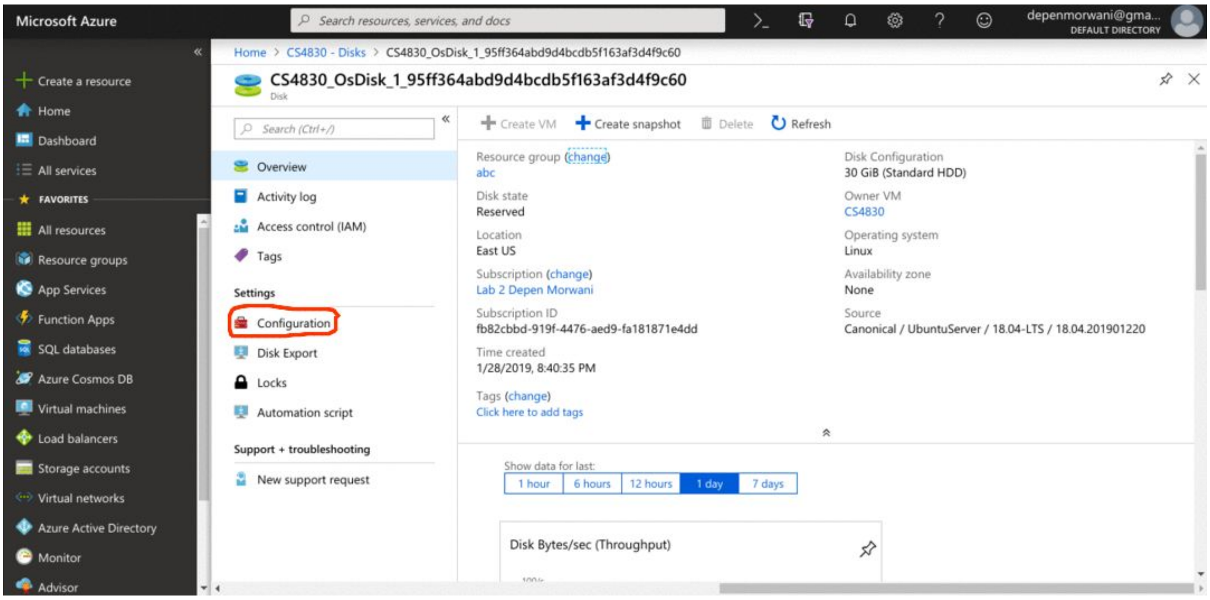
OS disk

NAME	SIZE	STORAGE ACCO...	ENCRYPTION	HOST CACHING
CS4830_OsDisk_1_95ff364abd9d4bcdb5f163...	30 GiB	Standard HDD	Not enabled	Read/write

Data disks

None

- Click on **Configuration** on the left hand side of the window:



Resource group (change)
abc

Disk state
Reserved

Location
East US

Subscription (change)
Lab 2 Depen Morwani

Subscription ID
fb82cbbd-919f-4476-aed9-fa181871e4dd

Time created
1/28/2019, 8:40:35 PM

Tags (change)
Click here to add tags

Disk Configuration
30 GiB (Standard HDD)

Owner VM
CS4830

Operating system
Linux

Availability zone
None

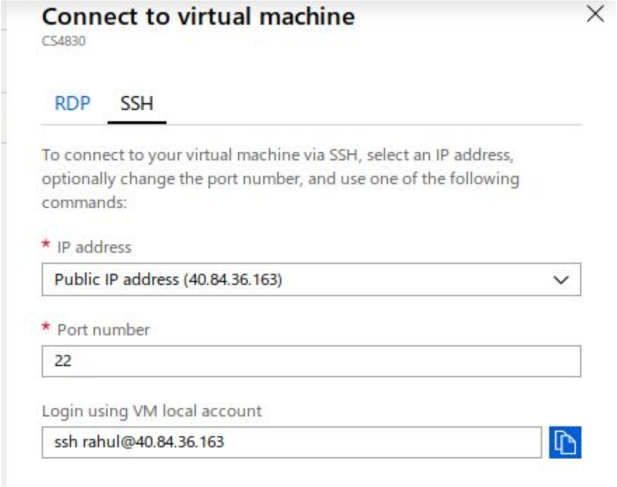
Source
Canonical / UbuntuServer / 18.04-LTS / 18.04.201901220

- Change the size of the disk to **100 GB** and click on **Save** option on the top of the window.
- After this, click on **CS4830-Disks** on the top of the window.
- Click on **Overview** tab on the left hand side.
- You can now **Start** and **Stop** the VM.
- **Do not "Delete" or "Refresh"** the VM since this will delete all the data.

Connecting to VM and downloading docker:

- Start up the VM using the **Start** button.

- Click on the **Connect** button and get the public IP (For our case it is 40.84.36.163):



Connect to virtual machine

CS4830

To connect to your virtual machine via SSH, select an IP address, optionally change the port number, and use one of the following commands:

* IP address

Public IP address (40.84.36.163)

* Port number

22

Login using VM local account

ssh rahul@40.84.36.163

- The virtual machine should be accessible from the **Dashboard** (found on left navigation bar).
- Use an ssh client (Download putty in windows or use the Terminal in Ubuntu/Mac).

```
# Connect to VM
ssh user@public-ip

# Once connected, execute the following
sudo apt update
sudo apt upgrade

# Install docker
sudo apt install docker.io
# Verify docker is running
sudo systemctl status docker
# Press q to exit status
```

- Download HDP and deploy it:

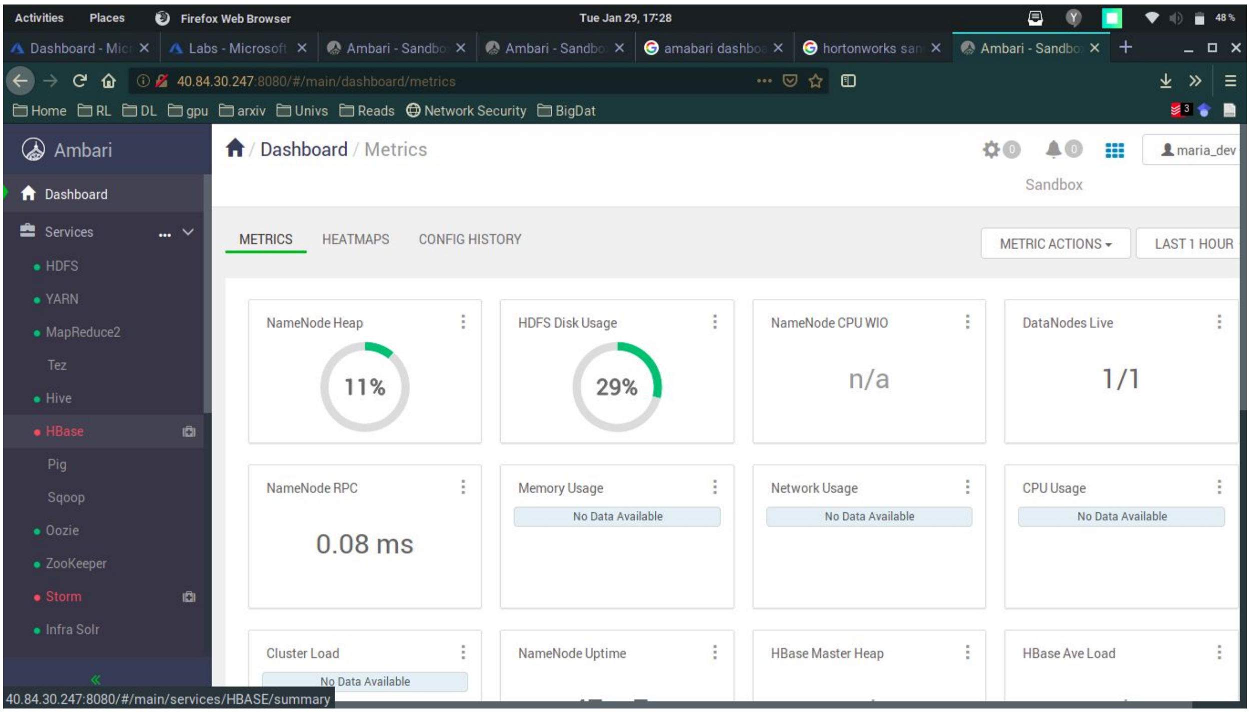
```
# Install unzip
sudo apt install unzip

# Download HDP
wget https://github.com/synchon/IITM-CS4830/raw/master/HDP_3.0.1.zip
unzip HDP_3.0.1.zip -d 'HDP_3.0.1'
cd HDP_3.0.1
cd HDP_3.0.1-test
# This command should take about an hour
sudo sh docker-deploy-hdp30.sh
```

- Now open your browser and access ***<public-ip>:4200*** (for our case it is http://40.84.36.163:4200); login using **root** as the username and **hadoop** as the password. Once done, enter **hadoop** as the UNIX password and set a new password for yourself. After that, run the following command in the shell:

```
ambari-admin-password-reset
```

When prompted for the admin password, enter **admin** and let the Ambari server restart.

- Now open your browser and try to access **<public-ip>:8080** (for our case it is http://40.84.36.163:8080). Try logging into ***admin*** account:
  - **Username**: admin
  - **Password**: admin
- Go through the Ambari interface.
- Check out: https://hortonworks.com/tutorial/learning-the-ropes-of-the-hortonworks-sandbox/
- **Stop** the machine when you are done (so that you don't bleed credits).



Starting a stopped VM:

- Start up the VM using the **Start** button.
- Enter the following two commands to get docker running:

```
# Start docker containers
sudo docker start sandbox-hdp
sudo docker start sandbox-proxy
```

Important:

- You can check your credits usage at: https://portal.azure.com/#blade/Microsoft_Azure_Education/ClassroomMenuBlade/assignments/classroomId/400

- Please **Stop** the VM when you are not using. The cost for a VM in usage is about 0.2 $/hour and a stopped VM has 0.003$ for an hour. Make sure that you do not unnecessarily waste credits by letting the VM run indefinitely.
- **Do not Delete/Refresh** the VM since that will delete all your data.
- If you are stuck don't hesitate to immediately contact one of the TAs or post a message on the Google group. This is especially true for the installation.
    - Rahul: rahul13ramesh@gmail.com
    - Synchon: synchonmandal@gmail.com
    - Depen: depenmorwani@gmail.com
    - Saurabh: saketd403@gmail.com
    - Pranshu: pranshumalviya2@gmail.com
    - Hafeez: a.hafeez123456@gmail.com
    - Akshay: cs14b038@smail.iitm.ac.in

- Get your Laptops to the Lab, do not disconnect any power-sockets for any of the DCF machines when in the lab.