Install Docker and Hortonworks HDP

Install Docker Community Edition

For reference, we’re going to essentially follow directions from here: https://docs.docker.com/engine/installation/linux/docker-ce/ubuntu/#install-using-the-repository

- Start VM
- SSH to server
- Run the following commands (hit “Y” to confirm apt-get commands if prompted)
    - `sudo apt-get update`
    - `sudo apt-get install apt-transport-https curl software-properties-common`
    - `curl -fsSL https://download.docker.com/linux/ubuntu/gpg | sudo apt-key add -`

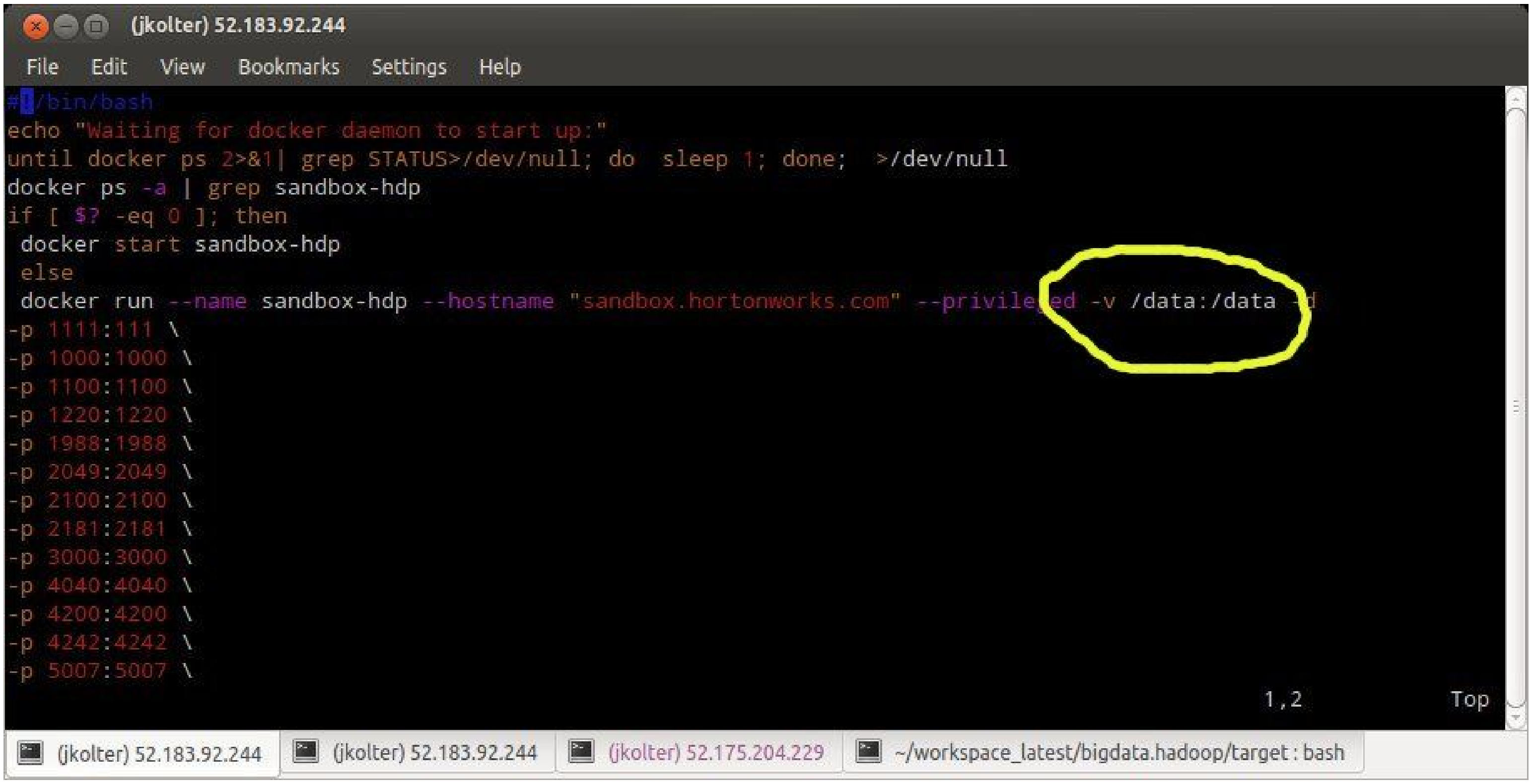
    -
      ```
      sudo add-apt-repository \
         "deb [arch=amd64]
      https://download.docker.com/linux/ubuntu \
         $(lsb_release -cs) \
         stable"
      ```
    - `sudo apt-get update`
    - `sudo apt-get install docker-ce`
- Docker should be running
    - Test installation by running “`sudo docker run hello-world`”
    - Output should include text like
        - Hello from Docker!
        - This message shows that your installation appears to be working correctly.

Install Hortonworks Sandbox into Docker

https://hortonworks.com/tutorial/sandbox-deployment-and-install-guide/section/3/

- Download Docker image
    - `cd /data`
        - Docker images are installed onto root filesystem so won’t have enough disk space for both raw image file and installed image
    - `sudo wget https://downloads-hortonworks.akamaized.net/sandbox-hdp-2.6.1/HDP_2_6_1_docker_image_28_07_2017_14_42_40.tar`
        - Will take about 20 min to download

- Load image file into Docker
    - `sudo docker load -i HDP_2_6_1_docker_image_28_07_2017_14_42_40.tar`
        - This will also take a while with little to no output to let you know it is actually doing anything
- Download startup script
    - `cd`
        - Return to home directory
    - `wget` https://raw.githubusercontent.com/hortonworks/data-tutorials/master/tutorials/hdp/sandbox-deployment-and-install-guide/assets/start_sandbox-hdp.sh
- Edit `start_sandbox-hdp.sh` using favorite text editor (vi, nano, emacs, etc…)
    - Add "`-v /data:/data`" to docker run command between "`--priviledged`" and "`-d`" (should be line 8)



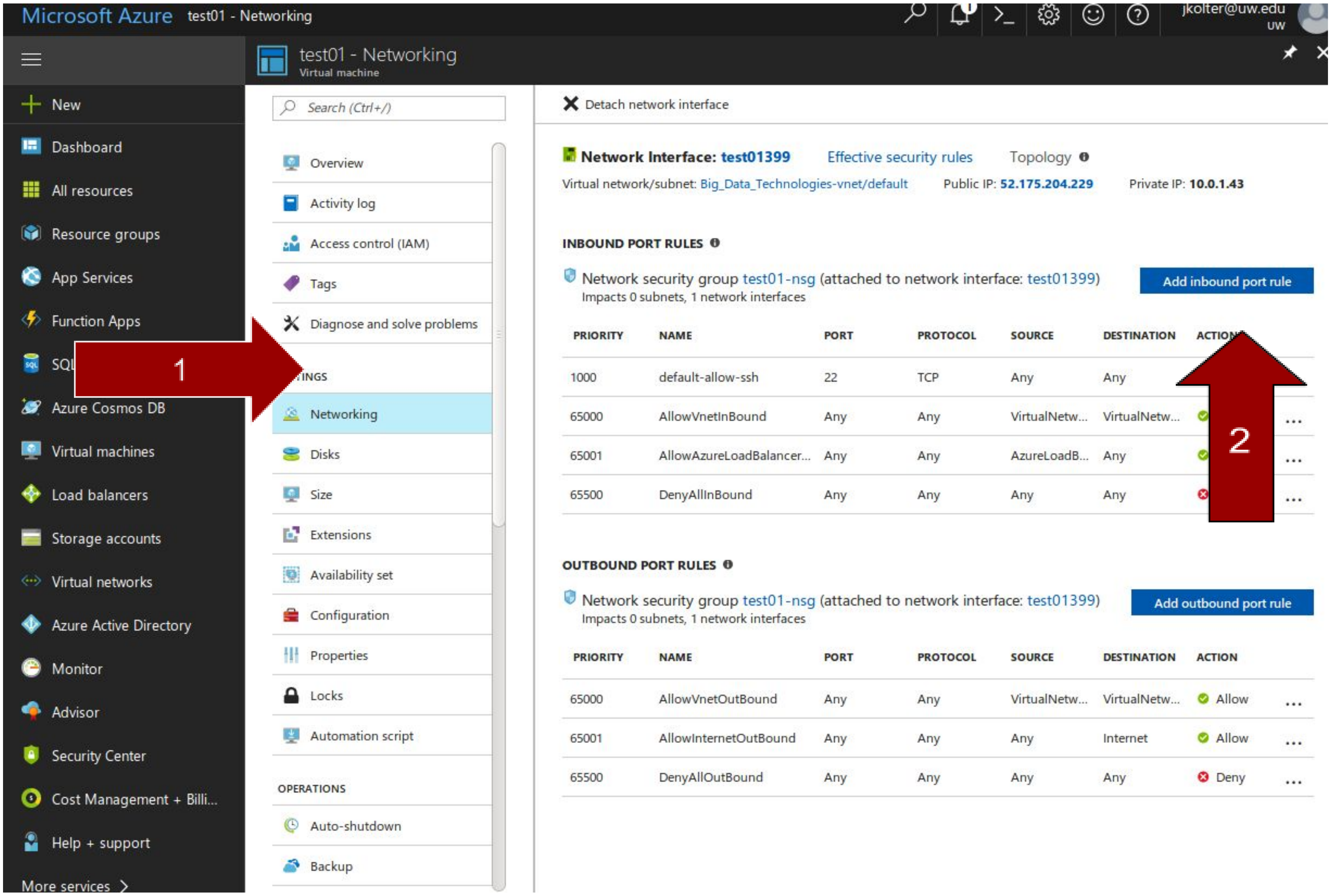
```
#!/bin/bash
echo "Waiting for docker daemon to start up:"
until docker ps 2>&1| grep STATUS>/dev/null; do  sleep 1; done;  >/dev/null
docker ps -a | grep sandbox-hdp
if [ $? -eq 0 ]; then
 docker start sandbox-hdp
 else
 docker run --name sandbox-hdp --hostname "sandbox.hortonworks.com" --privileged -v /data:/data -d
-p 1111:111 \
-p 1000:1000 \
-p 1100:1100 \
-p 1220:1220 \
-p 1988:1988 \
-p 2049:2049 \
-p 2100:2100 \
-p 2181:2181 \
-p 3000:3000 \
-p 4040:4040 \
-p 4200:4200 \
-p 4242:4242 \
-p 5007:5007 \
                                                                                1,2           Top
```

- Run startup script
    - `chmod a+x ./start_sandbox-hdp.sh`
    - `sudo ./start_sandbox-hdp.sh`
    - This will also take a while the first time

HDP Sandbox is now running in a Docker container on your VM

- First time setup
    - SSH into docker container “`ssh -p 2222 root@localhost`”
        - Default password is “hadoop” will make you change it first time

    - Unblock port 8080 for Ambari Web UI Access
        - 1) On Azure portal go to “Networking” tab for your VM
        - 2) Select “Add inbound port Rule



Network Interface: test01399

Virtual network/subnet: Big_Data_Technologies-vnet/default   Public IP: **52.175.204.229**   Private IP: **10.0.1.43**

INBOUND PORT RULES

Network security group test01-nsg (attached to network interface: test01399)
Impacts 0 subnets, 1 network interfaces

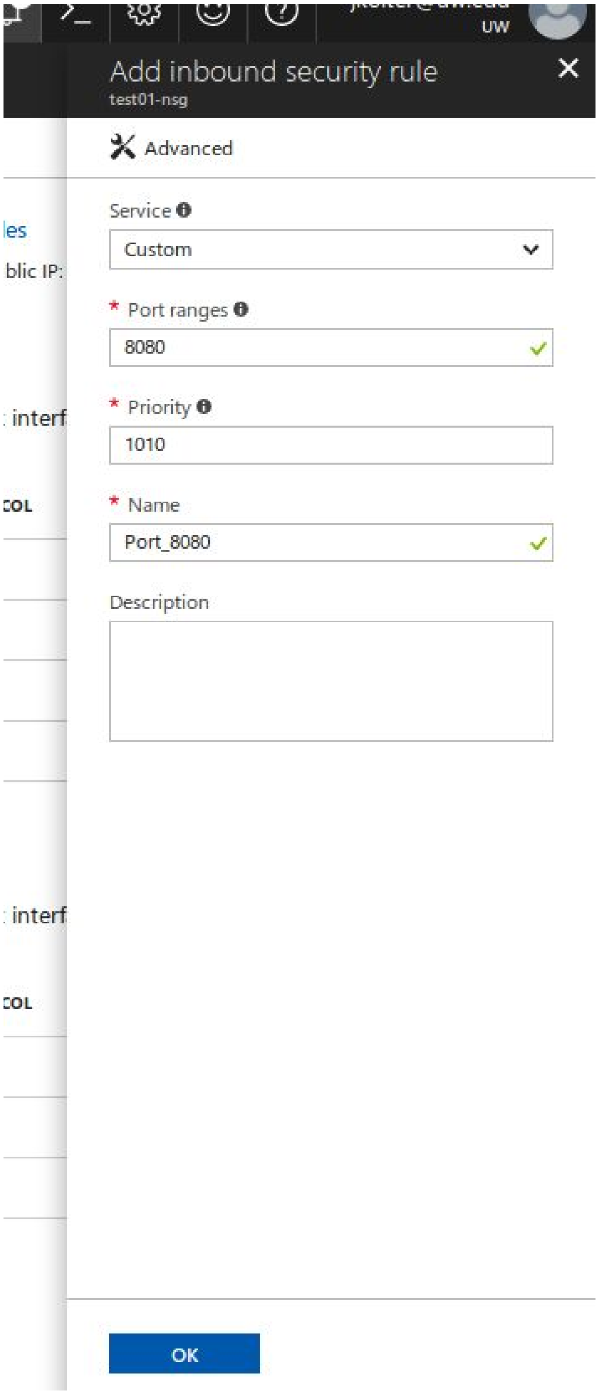
PRIORITY	NAME	PORT	PROTOCOL	SOURCE	DESTINATION	ACTION
1000	default-allow-ssh	22	TCP	Any	Any	
65000	AllowVnetInBound	Any	Any	VirtualNetw...	VirtualNetw...	✓
65001	AllowAzureLoadBalancer...	Any	Any	AzureLoadB...	Any	✓
65500	DenyAllInBound	Any	Any	Any	Any	✗

OUTBOUND PORT RULES

Network security group test01-nsg (attached to network interface: test01399)
Impacts 0 subnets, 1 network interfaces

PRIORITY	NAME	PORT	PROTOCOL	SOURCE	DESTINATION	ACTION
65000	AllowVnetOutBound	Any	Any	VirtualNetw...	VirtualNetw...	✓ Allow
65001	AllowInternetOutBound	Any	Any	Any	Internet	✓ Allow
65500	DenyAllOutBound	Any	Any	Any	Any	✗ Deny

Add Port 8080 rule



Add inbound security rule

test01-nsg

Service: Custom

* Port ranges: 8080

* Priority: 1010

* Name: Port_8080

Description:

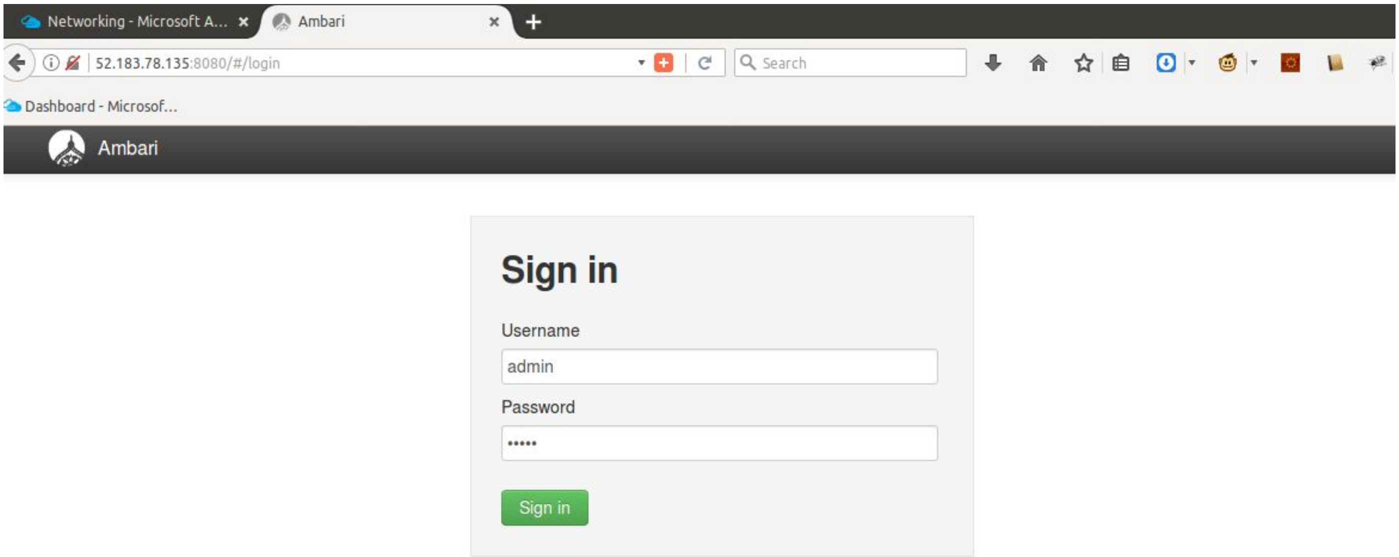
Hortonworks HDP Installation and Setup Completed!!!

Tutorial of getting to know sandbox available at:

https://hortonworks.com/tutorial/learning-the-ropes-of-the-hortonworks-sandbox

Fun With Hive Lab

Go to http://IP.ADD.RE.SS:8080 to see Ambari Web UI
Log in as “admin” using password ”4o12t0n” (without quotes)



Sign in

Username

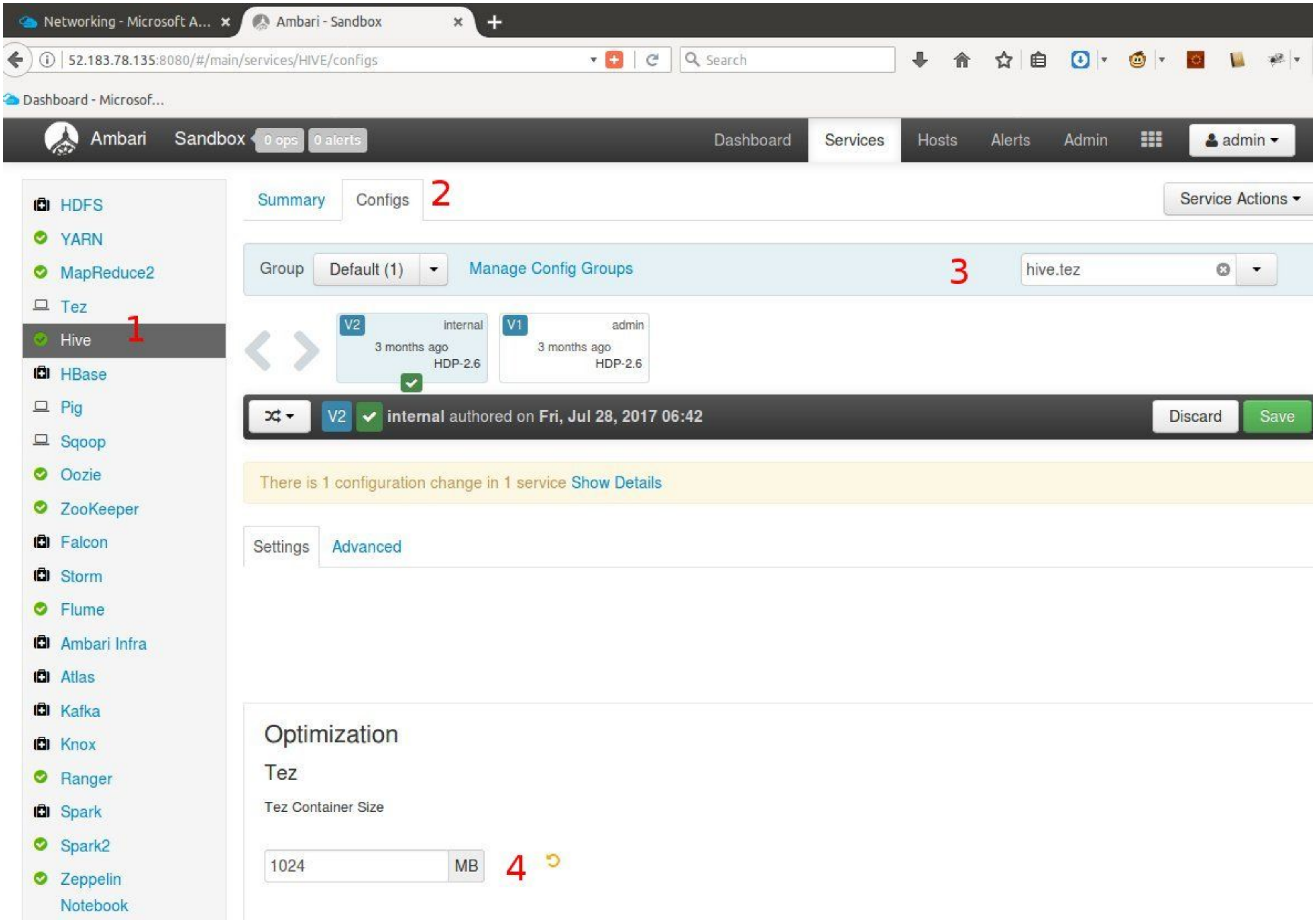
admin

Password

•••••

The default memory size for Hive is quite small on this sandbox so we will increase it by changing the Hive configuration

1) Select the Hive service from the service list on the left side of the page,
2) Choose “Configs” tab
3) In the filter box input “`hive.tez`” to filter the configs we want to change
4) Set `hive.text.container.size` to 1024 MB



52.183.78.135:8080/#/main/services/HIVE/configs

hive.tez

V2 internal authored on Fri, Jul 28, 2017 06:42

There is 1 configuration change in 1 service Show Details

Tez Container Size: 1024 MB

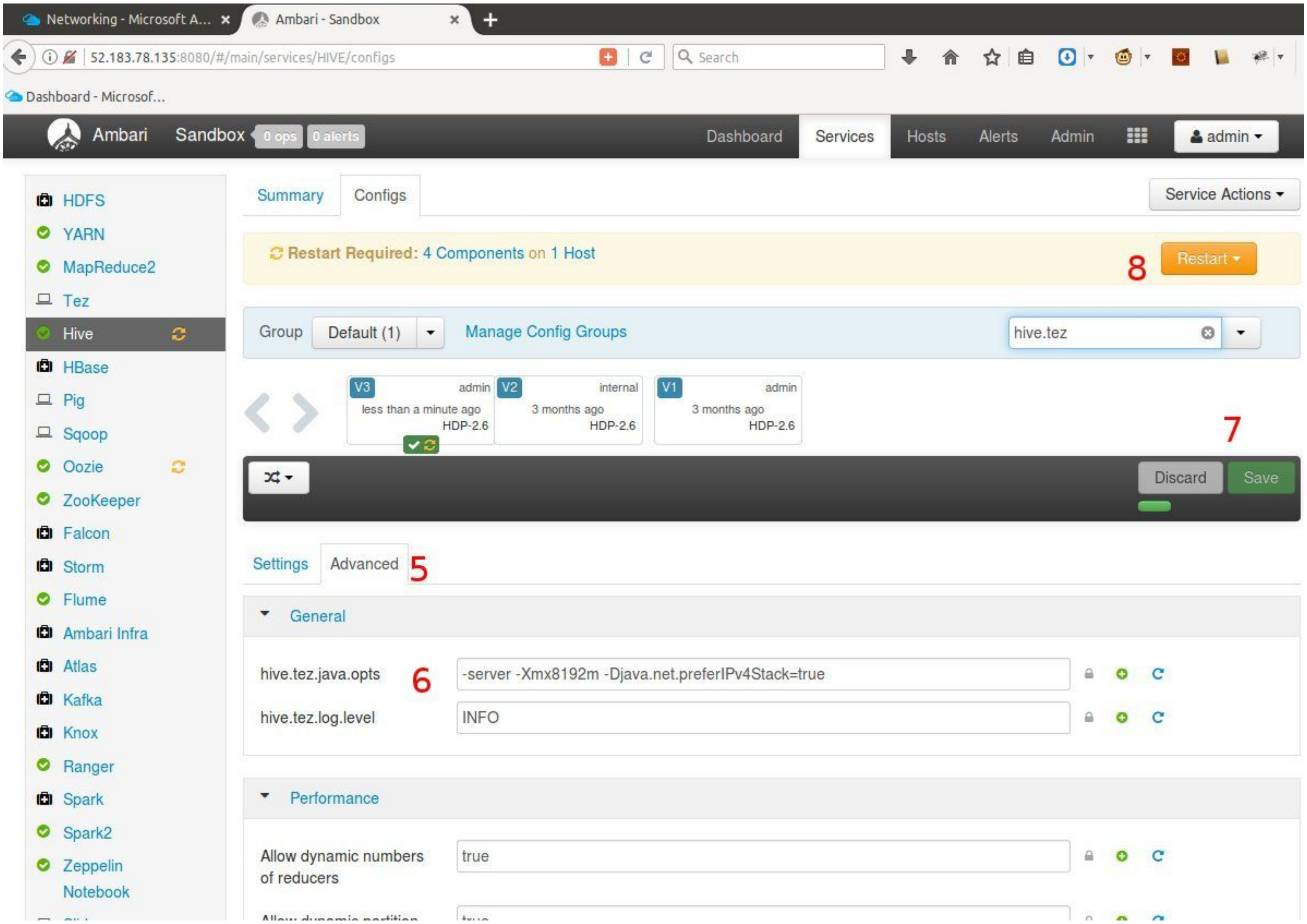
5) Choose the “Advanced” tab

6) Modify `hive.tez.java.opts` to set `-Xmx8192m`

7) Save the changes

8) Restart all affected services (any recommendations you get just accept or proceed)

Confirm restart if prompted



52.183.78.135:8080/#/main/services/HIVE/configs

Restart Required: 4 Components on 1 Host

Group Default (1)

hive.tez

General

hive.tez.java.opts: -server -Xmx8192m -Djava.net.preferIPv4Stack=true

hive.tez.log.level: INFO

Performance

Allow dynamic numbers of reducers: true

Once the restart is done, open the Hive2 View, this is what was demo'd at the end of class. Check the week 2 recording at 2:50:00 to see a demonstration of the rest, including how to load Hive view and create a table. Only instead of loading from HDFS you will simply upload from your local machine for this assignment, but please feel free to experiment with HDFS.

Download the sample data from here:
https://canvas.uw.edu/files/44270021/download?download_frd=1

Create a new Hive table, and choose upload table and provide the file you just downloaded (make sure the box to use "first row as header" is checked)

Hive will create and load the table for you based on the data you provided.
Now switch to the query view. Run the following query: `select count(*) from home_data`
You should get a result of **21613** if everything is working

Assignment 2

Use the sample data provided to answer the following questions. Please submit your answers to the following questions along with the query you executed to get that answer. For question 5 save the results file using the “Save as” option from the Ambari UI and attach to the assignment submission.

1) What is the most and least expensive houses in the data set?
2) What is the most expensive zipcode in the data set, defined as highest average sales price?
3) How many houses were built prior to 1979?

Using the same process as before, download and create a table from additional data file: https://canvas.uw.edu/files/44270416/download?download_frd=1

This is a zip code lookup table for WA, King County. Use this new table to join with the previous table to solve the next questions

4) How many homes were sold with a zipcode defined as being in “Seattle”?
5) Output a report showing the most expensive house in each city. Include at a minimum the price, zipcode and city.

Bonus Exercise 1:

Price per sq. ft. is a common metric for valuing housing. Calculate the avg. price per sq. ft for the houses sold in this data set. You do not need to get the answer using one query i.e. can use two or more queries to get the data you need to solve.

Bonus Exercise 2:

“date” is a reserved keyword in Hive. The sample data has a column named “date”. How can you use this column in a query without an error? Generate a list of all the unique date values represented in this data set.