CS199 Lecture 3

Applied Cloud Computing - MapReduce pt.2

Prof. Robert J. Brunner

Quinn Jarrell

These Slides

Go to goo.gl/LaeG3w

Copy short URL

Lab 1 Discussion

How did it go?

Make an NCSA server account

- This is the same form we sent out over slack yesterday
- Don’t do it twice

http://bit.ly/2ktdWo7

Remount your folder

- Unfortunately auto mount only works with Ubuntu. We're using Centos instead
- So if you shutdown your VM when you load it back up you need to run the mount command
- Should be something like
  - `sudo mount -t vboxsf  SHAREDFOLDERONHOST ~/SHAREDFOLDERONVM`
  - `Example:`
  - `sudo mount -t vboxsf  shared ~/shared`

- Hadoop lets us do map reduce on a cluster of computers
- Our cluster is currently 20 nodes each with 8 VCPUs and 16 gb of ram
- This week's lab will have you testing out the cluster
  - Things may break
  - DON'T WAIT TILL WEDNESDAY NIGHT
  - Say something in slack if something is wrong

Running Hadoop on your VM

- There are too many of you
- So let’s try it in the VM before running it on the cluster
- SSH to your VM
- Log in as the hadoop user
  - sudo dhclient
  - sudo -u hadoop -s
  - cd ~
  - wget https://transfer.sh/Avc21/archive.tar.gz
  - tar -zxvf archive.tar.gz
  - cd ~/VMHadoop/
  - source ./hadoop.env
  - hdfs dfs -ls

Hadoop Streaming

- By default Hadoop only runs java programs
  - Hadoop is written in Java so it is only natural
- We want to write python programs instead
  - Each python program is run on each computer
  - Read input from STDIN
  - Write output to STDOUT

Map.py

```
#!/usr/bin/env python

import sys

# input comes from STDIN (standard input)
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # split the line into words
    words = line.split()
    # increase counters
    for word in words:
        # tab-delimited; the trivial word count is 1
        print '%s\t%s' % (word, 1)
```

Sort

- One thing we've skipped over so far is the sort operation of MapReduce
- After the map operation completes, hadoop will sort the map outputs by their values
- So for word counts
- Sample Map Output
  - ‘Hello’ 1 ‘goodbye’ 1 ‘Hello’ 1 ‘goodbye’ 1 ‘Hello’ 1 ‘goodbye’ 1 ‘Hello’ 1
  - After sorting
  - ‘goodbye’ 1 ‘goodbye’ 1 ‘goodbye’ 1 ‘Hello’ 1 ‘Hello’ 1 ‘Hello’ 1 ‘Hello’ 1
- Why does it sort?

Reduce.py

```
current_word = None
current_count = 0
word = None
for line in sys.stdin:
    word, count = line.strip().split('\t', 1)
    if current_word == word: # This works since it’s sorted
        current_count += count
    else:
        if current_word: # Initialize current_word
            # write result to STDOUT
            print '%s\t%s' % (current_word, current_count)
        current_count = count
        current_word = word
```

Hadoop Distributed File System (HDFS)

- Hadoop MapReduce needs data to process
- Can’t just keep the data on one computer
- Instead store it on MULTIPLE computers
- So HDFS

Hadoop Distributed File System (HDFS)

- On your VM if you are not in the hadoop user yet
  - sudo -u hadoop -s # This logs you in as the user hadoop
  - cd ~/
- Now
  - source ~/VMHadoop/hadoop.env

- If your hadoop@localhost is now green on your shell, it worked

Hadoop Distributed File System (HDFS)

- Test it out
  - Run hdfs dfs -ls
  - Run ~/VMHadoop/hdfs_shell.sh
    - This is a REPL you can use to explore HDFS interactively
    - Ls, rm, mkdir etc

NCSA Account

http://bit.ly/2ktdWo7

Accessing NCSA Server

Security Violations - things you should not do

- We are trusting you here, don’t take advantage
- Do not use the machine you to run other classes’ work
- Do not share these machines with students outside of this class

- It is fairly easy to get kicked out of this class, don’t do anything stupid

SSH to it

- If you entered a username and password before 4 PM today you should have an account. You cannot do the next steps if you did not send it before 4 PM

- SSH from either your VM or from your local machine
- ssh USERNAME@141.142.210.245
- source ~/hadoop.env
- If your username is red, it worked

How to get files on a server?

- SSHFS

Run on your VM not the ncsa server!

sudo su
rpm -Uvh http://dl.fedoraproject.org/pub/epel/7/x86_64/e/epel-release-7-9.noarch.rpm
yum install fuse-sshfs -y
mkdir ~/ncsaHadoop
sshfs YOURUSERNAME@141.142.210.245:/home/YOURUSERNAME ~/ncsaHadoop

How to run stuff on the server?

- Use the mapreduce command in your home directory
- Like so

~/mapreduce mapper.py reducer.py /tmp/helloworld.txt /user/quinnjarr

- The lab will have more detail

Source

- When you log into either your VM or the cluster, remember to source the right file
- Source sets up a bunch of environmental variables for you

For VM

source ~/VMHadoop/hadoop.env

For cluster

source ~/hadoop.env

So many SSH terminals

- You've probably seen how weird it is jumping between your VM and the cluster
- Use the color of your username to guide you
    - Green == VM
    - Red == Cluster

Editing files between VM and NCSA

- Edit on your host OS like you would normally
- Keep the files you're editing in the shared folder between host OS and VM
- cp from within the vm the files to the sshfs folder
    - Or scp if you want

Tips and Tricks Doc

https://goo.gl/bmZCBa

If you run into a problem and solve it, post the solution there

If you have a problem, check the doc before asking TAs

Lab 2

- Due in one week
- Run MapReduce using Hadoop on your VM/ the cluster
- This lab is simpler, we want problems to be from messed up settings rather than difficult code
- If something goes wrong unexpected POST IN THE SLACK CHANNEL

Project Ideas

- Start thinking of projects you would like to use the cluster for
- The technical report(s) subjects are very open