MSc Data Analytics Research Project Configuration Manual

Erin Gilheany X14106671 Supervisor: Vikas Sahni

# Table of Contents

Settin	ng up the Virtual Machines	4
1.	Download and Install Virtual Machine Software	4
2.	Download Ubuntu Operating System	7
3.	Set up our initial Virtual Machine – Hadoop 2.6	8
4.	Booting up our Initial Virtual Machine	13
5.	Change Screen Resolution on Virtual Box	20
6.	Clone your initial VM to be used for testing of Hadoop 2.7 and Spark 2.0	21
Hadoo	op 2.6: Set Up, Configuration and Process Implementation	24
1.	Set up Required Software and Environment	24
2.	Download, Set Up and Configure Hadoop 2.6.4	26
3.	Download and Set Up Mahout 0.12.2	29
4.	Download Dataset and move to HDFS	30
5.	Run Mahout Algorithms	32
1	1. TFIDF on Hadoop	33
2	2. Naïve Bayes on Hadoop	34
Hadoo	op 2.7: Set Up, Configuration and Process Implementation	35
1.	Updates to Process Needed for Hadoop 2.7.3	35
Spark	2.0: Set Up, Configuration and Process Implementation	36
1.	Download, Set Up and Configure Spark 2.0.0	36
2.	Install Scala and check Spark Set Up	37
3.	Run MLlib Algorithms	41
3	3. TFIDF on Spark	42
4	4. Naïve Bayes on Spark	43

# Setting up the Virtual Machines

### 1. Download and Install Virtual Machine Software

Firstly the software for setting up a virtual machine needs to be downloaded. I used "Oracle VM VirtualBox" which can be found <u>here</u>. The computer I am running this on is a Windows 64-bit so I downloaded the relevant release.

# VirtualBox binaries

By downloading, you agree to the terms and conditions of the respective license.

- VirtualBox platform packages. The binaries are released under the terms of the GPL version 2.
  - ✓ VirtualBox 5.1.4 for Windows hosts ⇒x86/amd64
    - VirtualBox 5.1.4 for OS X hosts ⇒ amd64
    - VirtualBox 5.1.4 for Linux hosts
    - VirtualBox 5.1.4 for Solaris hosts ⇒amd64

Go to the location where the above has been downloaded to and double-click to run the .exe file.



Click next on the below screen.



Click next on the below screen.

岃 Oracle VM VirtualBox 5.1.4 Setup	×
Custom Setup	
Select the way you want features to be installed.	
Click on the icons in the tree below to change the way features will be installed.	
VirtualBox Application         Oracle VM VirtualBox 5.1.4           VirtualBox USB Support         VirtualBox Networking	
VirtualBox Bridge VirtualBox Host-C VirtualBox Python 2.x Su VirtualBox Python 2.x Su	
Location: C:\Program Files\Oracle\VirtualBox\ Browse	
Version 5.1.4 Disk Usage < Back Next > Cancel	

Click next on the below screen.

🛃 Oracle VM VirtualBox 5.1.4 Setup			×
Custom Setup			
Select the way you want features to be	installed.		
Please choose from the options below:			
Create a shortcut on the desktop			
Create a shortcut in the Quick Laund	h Bar		
Register file associations			
Version 5.1.4	< Back	Next >	Cancel

Click Yes on the below screen.



Click Install on the below screen.

😸 Oracle VM VirtualBox 5.1.4 Setup		×
Ready to Install		
The Setup Wizard is ready to begin the C	ustom installation.	
Click Install to begin the installation. If yo installation settings, dick Back. Click Canc	u want to review or change any el to exit the wizard.	r of your

The software will install and there should be an icon on your desktop. Please double click on this icon to run the newly installed application.



You should be presented with the below screen.

Note: You may need to restart the computer to proceed to this step after installation process has completed.



# 2. Download Ubuntu Operating System

Download an Ubuntu Desktop .iso file so that we can boot on Virtual Machine from here.

Preferably copy this file to an external disk (to be used when setting up a new VM).

📭 2% complete	-		×
Copying 1 item from Downloads to Install Ubuntu (F:) 2% complete		II	×
✓ More details			

## 3. Set up our initial Virtual Machine – Hadoop 2.6

We return to our VM Ware application previously downloaded and click on "New" as per screen below.



We will start off setting up the virtual machine which we run Hadoop version 2.6 on. Therefore we will name it as per screen below.

Please ensure to enter Linux in the Type dropdown and Ubuntu in the Version dropdown with the bit that matches the host computer.

Create Virtual Machine

## Name and operating system

Please choose a descriptive name for the new virtual machine and select the type of operating system you intend to install on it. The name you choose will be used throughout VirtualBox to identify this machine.

Name:	Hadoop 2.6	~
Type:	Linux 👻	64
Version:	Ubuntu (64-bit) 🔻	

Expert Mode	Next	Cancel	

Set the VM memory size to be 4GB or 4096MB.

Note: The host machine in this instance has an overall memory of 6GB so this is close to the maximum memory that we can use for the virtual machines.

i bity		
	?	×
M) in megabytes to	be allocate	ed
1024 MB.		
	4096 韋	мв
6144 MB		
Next	Cance	4
	M) in megabytes to 1024 MB. 6144 MB	Provide the second s

For this instance we have chosen "Create a virtual hard disk now". Click Create.

	?	×
<ul> <li>Create Virtual Machine</li> </ul>		
Hard disk		
If you wish you can add a virtual hard disk to the new can either create a new hard disk file or select one fro from another location using the folder icon.	machin m the li	e. You st or
If you need a more complex storage set-up you can s and make the changes to the machine settings once t created.	kip this : he mach	step ine is
The recommended size of the hard disk is <b>8.00 GB</b> .		
🔿 Do not add a virtual hard disk		
Oreate a virtual hard disk now		
$\bigcirc$ Use an existing virtual hard disk file		
testNCI-disk1.vmdk (Normal, 20.00 GB)		
Create	Can	icel

We have chosen to select "VDI (VirtualBox Disk Image)" as per screen below. Click Next.

	?	×
Create Virtual Hard Disk		
Hard disk file type		
Please choose the type of file that you would like to use for th hard disk. If you do not need to use it with other virtualization can leave this setting unchanged.	ne new virtu n software y	ial vou
VDI (VirtualBox Disk Image)		
○ VHD (Virtual Hard Disk)		
O VMDK (Virtual Machine Disk)		
Expert Mode Next	Can	cel

Choose "Fixed Size". Click Next.

? X

Create Virtual Hard Disk

Storage on physical hard disk

Please choose whether the new virtual hard disk file should grow as it is used (dynamically allocated) or if it should be created at its maximum size (fixed size).

A **dynamically allocated** hard disk file will only use space on your physical hard disk as it fills up (up to a maximum **fixed size**), although it will not shrink again automatically when space on it is freed.

A **fixed size** hard disk file may take longer to create on some systems but is often faster to use.

- O Dynamically allocated
- Fixed size

Marsh Garagel		_	
Next Cancel	Next Cancel		

Change hard disk size to 20.00 Gb. Click Create.

		?	×
~	Create Virtual Hard Disk		
I	File location and size		
	Please type the name of the new virtual hard disk file into the on the folder icon to select a different folder to create the file	box below in.	or click
	Hadoop 2.6		
	Select the size of the virtual hard disk in megabytes. This size i amount of file data that a virtual machine will be able to store of	is the limit on the har	on the d disk.
-		20	0,00 GB
	4.00 MB 2.00 TB		
-			
	Create	Car	ncel

The screen below will appear and make take a few minutes to complete.



When the above process you will have your Virtual Machine in the screen as per below and ready to boot/start up.



## 4. Booting up our Initial Virtual Machine

Double click on your "Hadoop 2.6" icon as per above screen.

In the screen below, press on the folder icon and find the location where you moved the Ubuntu .iso file if it does not automatically appear here. Click Start.



It may take a short time for the next screen to load.

Select language – English in this case – and click "Install Ubuntu".

*	Hadoop 2.6_	[Running	g] - Ora	cle VM Vir	tualBox	_		×
File	Machine	View	Input	Devices	Help			
								<u>^</u>
		tall						
	wei	.com	ie					
	Banas	a Indor	nesia					
	Catal	15KI à						
	Češtir	na						
	Cymra	aeg						
4	Dansk	(						
	Deuts	ch						
	Eesti							
	Englis	h						
	Españ	iol			Try Ubuntu Install Ubuntu			
	Esper	anto						
	Euska	ra		Y	ou can try Ubuntu without making any changes to your computer, directly from this CD.			
	Gaeile			_	r if you're ready, you can install Ubyoty algogside (er instead of) your system to getating system. This			
	Galeo	0		st	nouldn't take too long.			
	Hrvat	ski						
	Íslens	ka		Y	pu may wish to read the release notes.			
								v
					2 () # 2 () = #	' 🛄 🔇	🛃 Righ	t Ctrl

Follow instructions as per screen below by allowing 6.5GB available drive space, plug in your computer to a power source and keep it plugged in, and get and remain connected to the internet.

Click on "Download updates while installing".

Note: Following instructions above will allow for a smoother install if any issues were to arise.

Click Continue.

8 Install
Preparing to install Ubuntu
For best results, please ensure that this computer:
Ans at least 6.5 GB available drive space
is plugged in to a power source
✓ is connected to the Internet
🛛 Download updates while installing
Ubuntu uses third-party software to play Flash, MP3 and other media, and to work with some graphics and wi-fi hardware. Some of this software is proprietary. The software is subject to license terms included with its documentation.
Install this third-party software Fluendo MP3 plugin includes MPEG Layer-3 audio decoding technology licensed from Fraunhofer IIS and Technicolor SA.
Quit Back Continue
• • • • • •

Select "Erase and install Ubuntu" and click "Install Now".

😣 Install	
Installation type	
This computer currently has no detected operating systems. What would you like to do?	
Erase disk and install Ubuntu Warning: This will delete any files on the disk.	
<ul> <li>Encrypt the new Ubuntu installation for security</li> <li>You will choose a security key in the next step.</li> </ul>	
<ul> <li>Use LVM with the new Ubuntu installation</li> <li>This will set up Logical Volume Management. It allows taking snapshots and easier partition resizing.</li> </ul>	
<ul> <li>Something else</li> <li>You can create or resize partitions yourself, or choose multiple partitions for Ubuntu.</li> </ul>	
	Quit Back Install Now
••••	

Input your location if this is not automatically picked up – in my case it's Dublin.

Click Continue.



Select your desired Keyboard layout. I have chosen Irish as per below. Click Continue.

Note: This does not refer to Irish as a language, all language returns in English.

⊗ Install	
Keyboard layout	
Choose your keyboard layout: Hungarian Icelandic Indian Iraqi Irish Italian Japanese Japanese (PC-98xx Series) Kazakh	Irish Irish - CloGaelach Irish - Irish (UnicodeExpert) Irish - Ogham Irish - Ogham (IS434)
Type here to test your keyboard         Detect Keyboard Layout	Back Continue

Fill in the fields in screen below with "Require my password to log in". Click Continue.

Note: Please ensure that a memorable password is selected as we will need to use it later in the process.

8 Install		
Who are you?		
	(	0
Your name:	hduser	×.
Your computer's name:	hduser-VirtualBox 🖌	
	The name it uses when it talks to other computers.	
Pick a username:	hduser 🖌	
Choose a password:	Fair password	
Confirm your password:	••••••••	
	Log in automatically	
	Require my password to log in	
	Encrypt my home folder	
		Back Continue
	•••••	

The following screen will appear. We just have to wait until the install finishes. This can take a few minutes.



Click on Restart Now when the screen below appears.



The below screen will appear. We do not have to press anything here.

wait–for–state stop/waiting * Stopping rsync daemon rsync		ОК	
* speech-dispatcher disabled; edit /etc/default/speech-dispatcher	nø	al	l n
emaining processes to terminate	[	OK	
∗ All processes ended within 3 seconds ModemManager[1260]: <info≻ caught="" down<="" shutting="" signal,="" td=""><td></td><td>OK</td><td></td></info≻>		OK	
ModemManager[1260]: <warn>  Could not acquire the 'org.freedesktop.Modem ' service name</warn>	nMai	nag	er1
* Unmounting temporary filesystems * Deactivating swap * Stopping early crypto disks		0K 0K 0K	

You may need to restart the VM to ensure everything is working correctly. Click on the X at top right hand side of the VM screen.



On the below select "Power off the machine" and click ok.



Reopen your newly created Virtual Machine and you will be presented with the following screen.

1	Hadoo	p 2.6_	[Runnin	ng] - Ora	cle VM Vir	tualBo	ох					<u></u>		×
File	Mac	hine	View	Input	Devices	Hel	р							1
hdu	ser-V	irtua	lBox						Ø	4	le		23:25	<u>ب</u>
k														
8														
			hdus	er										2
			Pass	word										
			Gues	t Sess	sion									
	Tik		1110											
	UL	, ui	icų.	14.0										
								0	0 7	0		🗗 🔘 🔇	🕑 Right	Ctrl

5. Change Screen Resolution on Virtual Box

Log into machine.

Open Terminal, type in the following:

sudo apt-get install virtualbox-guest-utils virtualbox-guest-x11 virutalbox-guest-dkms

Enter your login password

Enter Y and press enter when it displays: [Do you want to continue]

Restart computer and the screen resolution should fit your host machine's screen resolution and will adapt as your maximize or minimize the screen.

This is the last step in setting up the virtual machine. Before we dive into setting up software applications on this virtual machine, we close out of the above because we have to make an identical clone of this VM so that we can separately set up the other version of Hadoop we are going to test, Hadoop 2.7 and subsequently Spark 2.0.0. We will not set up a separate machine for Hadoop 2.7 and Spark 2.0 as Spark will sit on top of this version of Hadoop. It may be possible to set these all up on the one machine but this strategy avoids any confusion and possible mistakes when connecting Spark to HDFS.

## 6. Clone your initial VM to be used for testing of Hadoop 2.7 and Spark 2.0

Back in the Oracle VM VirtualBox Manager screen we right click on the Hadoop 2.6 VM icon.

From menu that appears Click Clone.



We will name our new VM in line with the application to be installed on it "Hadoop 2.7 and Spark 2.0". Click Next



Select "Full clone". Click Clone.



The following will appear. We just wait until this reaches 100% and our VM is fully cloned.



When complete open up the VM "Hadoop 2.7 and Spark 2.0" to check that the system boots and you can log in. And you will be presented with the below. You click on X on the two banners at the top of the page. Close out of this VM.



Note: This virtual machine will be identical at this point to VM "Hadoop 2.6" including the users and passwords. Anything done in either of the environments will only take effect within that VM, they are not synced or linked to one another.

# Hadoop 2.6: Set Up, Configuration and Process Implementation

## 1. Set up Required Software and Environment

All of below commands will require your log-in password.

Note: At any stage of above commands you are asked "Do you want to continue? [Y/n]". Enter "Y" and hit return.

### Open terminal and run the following commands:

sudo apt-get update	
sudo apt-get install openjdk-7-jre	
sudo apt-get install openjdk-7-jdk	
sudo apt-get install ssh	
sudo apt-get install rsync	

sudo addgroup hadoop

If you have not already set up hduser account:

sudo adduser --ingroup hadoop hduser

If you set up the machine as hduser account:

sudo usermod -a -G hadoop hduser

Run this if you are not already logged in as hduser

su - hduser

ssh-keygen -t rsa -P ""

Enter "/home/hduser/.ssh/authorized\_keys" when it asks for a location as per screenshot below and hit return.



ssh localhost

If you are presented with the following message, enter "yes" and hit return.

The authenticity of host 'localhost (127.0.0.1)' can't be established. ECDSA key fingerprint is 67:2c:13:da:40:f1:28:bd:1a:47:2f:ce:bc:0a:b1:32. Are you sure you want to continue connecting (yes/no)? yes

You may then have to enter your password and the below will return.

Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts. hduser@localhost's password: Welcome to Ubuntu 14.04.1 LTS (GNU/Linux 3.13.0-32-generic x86\_64) \* Documentation: https://help.ubuntu.com/ New release '16.04.1 LTS' available. Run 'do-release-upgrade' to upgrade to it. The programs included with the Ubuntu system are free software; the exact distribution terms for each program are described in the 'individual files in /usr/share/doc/\*/copyright. Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by applicable law.

We exit out of ssh and log out of the VM OS by running the following commands:

exit

### hduser@hduser-VirtualBox:~\$ exit logout Connection to localhost closed.

If you have not set up as hduser run the exit command again to logout out of hduser. (Note: not the case for my set up as I only have hduser set up)

#### exit

Disable IPv6 as Hadoop is not supported for these networks.

sudo gedit /etc/sysct1.conf

Enter password if necessary. An empty document will appear. Enter the below text. Save and close the document.

# disable ipv6
net.ipv6.conf.all.disable\_ipv6 = 1
net.ipv6.conf.default.disable\_ipv6 = 1
net.ipv6.conf.lo.disable\_ipv6 = 1

## 2. Download, Set Up and Configure Hadoop 2.6.4

We will download Hadoop 2.6.4 itself.

sudo wget http://ftp.heanet.ie/mirrors/www.apache.org/dist/hadoop/common/hadoop-

2.6.4/hadoop-2.6.4.tar.gz

sudo cp hadoop-2.6.4.tar.gz /usr/local/

cd /usr/local

sudo tar xvf hadoop-2.6.4.tar.gz

sudo In -s hadoop-2.6.4 hadoop

sudo chown -R hduser:hadoop hadoop-2.6.4

sudo rm hadoop-2.6.4.tar.gz

sudo update-alternatives --config java

You will see similar text to the below return, please take note of the location where Java is installed, this will be needed later in the process. In my case it is "/usr/lib/jvm/java-7-openjdk-amd64" as per screen below:

hduser@hduser-VirtualBox:/usr/local\$ sudo update-alternatives --config java There is only one alternative in link group java (providing /usr/bin/java): /usr/lib/jvm/java-7-op enjdk-amd64/jre/bin/java Nothing to configure.

Note: the location is not the full text as printed out but up until just before jre directory.

If not currently logged in as hduser do so:

su - hduser cd /usr/local/hadoop/etc/hadoop cp mapred-site.xml.template mapred-site.xml

In our hadoop-env.sh we enter the java location we took note of just a few steps ago.

vi hadoop-env.sh

#### export JAVA\_HOME=/usr/lib/jvm/java-7-openjdk-amd64

vi mapred-site.xml

Insert the following configuration properties to the file.

#### <configuration>

<property>

<name>mapreduce.jobtracker.address</name> <value>local</value> </property>

</configuration>

mkdir ~/tmp

mkdir ~/hdfs

chmod 750 ~/hdfs

vi core-site.xml

Enter the below properties to this file.

#### <configuration>

<property> <name>hadoop.tmp.dir</name> <value>/home/hduser/tmp</value> </property> <property> <name>fs.defaultFS</name> <value>hdfs://localhost:54310</value> </property> </configuration>

vi hdfs-site.xml

<configuration> <property>

cd /usr/local/hadoop

bin/hdfs namenode -format

sbin/start-dfs.sh

\*Enter password multiple times and type "yes" and return if asked to continue.

sbin/start-yarn.sh

jps

You should see the following return:

hduser	r@hduser-VirtualBox:/usr/local/hadoop\$	jps
13154	DataNode	
13327	SecondaryNameNode	
13867	Jps	
13466	ResourceManager	
13773	NodeManager	
13013	NameNode	_

A useful resource for all information on Hadoop Environment and Administration is to open a web browser/tab in web browser and type in the following location:

localhost:50070

If you click on "Utilities" tab at top right and the "Browse the file system" you will see a page like screenshot below. Click Go to browse through file system or go directly to a file or directory.

Note: This is empty at the moment because we have created any directories or added any files to hdfs.

e 🖲 lo	localhost:50070/explorer.html#/								▼ C <sup>r</sup> ] 🚼 ▼ Google				÷	A	8
	Hadoop														
	Brow	se Dire	ectory										Go	-	
	Permissio	n	Owner	Gr	oup S	ize	Replication		Block Size		Na	me	L		

Verify All Applications for Cluster Browser is running as expected. Enter the following address into the web browser and you will see the screen below return:





## 3. Download and Set Up Mahout 0.12.2

#### In terminal enter

#### cd ~

sudo wget http://mirrors.whoishostingthis.com/apache/mahout/0.12.2/apache-mahout-

distribution-0.12.2.tar.gz

\*Enter log-in password

sudo cp apache-mahout-distribution-0.12.2.tar.gz /usr/local

cd /usr/local

sudo tar xvf apache-mahout-distribution-0.12.2.tar.gz

sudo In -s apache-mahout-distribution-0.12.2 mahout

sudo chown -R hduser:hadoop apache-mahout-distribution-0.12.2

*sudo rm apache-mahout-distribution-0.12.2.tar.gz* 

Update bashrc file to ensure all of the following variables are included:

vi ~/.bashrc

#### #HADOOP VARIABLES START

export JAVA\_HOME=/usr/lib/jvm/java-7-openjdk-amd64 export HADOOP\_INSTALL=/usr/local/hadoop export PATH=\$PATH:\$HADOOP\_INSTALL/bin export PATH=\$PATH:\$HADOOP\_INSTALL/sbin export HADOOP\_MAPRED\_HOME=\$HADOOP\_INSTALL export HADOOP\_COMMON\_HOME=\$HADOOP\_INSTALL export HADOOP\_HDFS\_HOME=\$HADOOP\_INSTALL export YARN\_HOME=\$HADOOP\_INSTALL export YARN\_HOME=\$HADOOP\_INSTALL export HADOOP\_COMMON\_LIB\_NATIVE\_DIR=\$HADOOP\_INSTALL/lib/native export HADOOP\_OPTS="-Djava.library.path=\$HADOOP\_INSTALL/lib" #HADOOP VARIABLES END

export MAHOUT\_HOME=/usr/local/mahout

After you have updated the bashrc file as above you will need to restart the computer for changes to take effect.

Once you have restarted the computer re-enter the following commands to start up hadoop services again.

cd /usr/local/hadoop	
sbin/start-dfs.sh	
sbin/start-yarn.sh	

## 4. Download Dataset and move to HDFS

sudo mkdir ~/20newsgroups

cd 20newsgroups

sudo wget http://people.csail.mit.edu/jrennie/20Newsgroups/20news-bydate.tar.gz

sudo tar xvf 20news-bydate.tar.gz

cd ~

sudo chown -R hduser 20newsgroups

cd /usr/local/hadoop

bin/hdfs dfs -mkdir /20newsgroups

bin/hdfs dfs -copyFromLocal ~/20newsgroups/\*/\* /20newsgroups

If we look back in our Hadoop Administration @ localhost:50070 and go to Utilities and Browse the file system, you will now see the directory as follows:

<b>())</b>	localhost:50	070/explorer.l	ntml#/				v	C Soogle	(	<b>Q</b>	î 🏚
	Hadoop										
	Brow	se Dir	ectory								
	Permission Owner		Group S		Size Replication		Block Size	Name			
	drwxr-xr-x		hduser	supergrou	ip 0 B	0		0 B	20newsgro	ups	

And when you click on 20newsgroups you will see subdirectories (20 news group categories) as follows (only some categories/subdirectories show because all cannot fit on screenshot, scroll down to view all of them):

alhost:50070/explo	orer.html#/20new	sgroups			▼ C	👌 🕶 Google 🔍 🐧
Browse	rowse Directory					
/20newsgroups						
Permission	Owner	Group	Size	Replication	Block Size	Name
drwxr-xr-x	hduser	supergroup	0 B	0	0 B	alt.atheism
drwxr-xr-x	hduser	supergroup	0 B	0	0 B	comp.graphics
drwxr-xr-x	hduser	supergroup	0 B	0	0 B	comp.os.ms-windows.misc
drwxr-xr-x	hduser	supergroup	0 B	0	0 B	comp.sys.ibm.pc.hardware
drwxr-xr-x	hduser	supergroup	0 B	0	0 B	comp.sys.mac.hardware
drwxr-xr-x	hduser	supergroup	0 B	0	0 B	comp.windows.x
drwxr-xr-x	hduser	supergroup	0 B	0	0 B	misc.forsale
drwxr-xr-x	hduser	supergroup	0 B	0	0 B	rec.autos
drwxr-xr-x	hduser	supergroup	0 B	0	0 B	rec.motorcycles
drwxr-xr-x	hduser	supergroup	0 B	0	0 B	rec.sport.baseball
drwxr-xr-x	hduser	supergroup	0 B	0	0 B	rec.sport.hockey

## Each subdirectory contains a large number of documents within as per below:

localhost:50070/explore	t:50070/explorer.html#/20newsgroups/alt.atheism ▼ C 🔀 ▼ Google						
Browse D	Directory						
/20newsgroups/alt.al	theism					G	<b>o!</b>
Permission	Owner	Group	Size	Replication	Block Size	Name	
-rw-rr	hduser	supergroup	11.61 KB	1	128 MB	49960	
-rw-rr	hduser	supergroup	31.3 KB	1	128 MB	51060	
-FW-FF	hduser	supergroup	4 KB	1	128 MB	51119	
-rw-rr	hduser	supergroup	1.59 KB	1	128 MB	51120	
-rw-rr	hduser	supergroup	773 B	1	128 MB	51121	
-rw-rr	hduser	supergroup	4.8 KB	1	128 MB	51122	
-rw-rr	hduser	supergroup	618 B	1	128 MB	51123	
-rw-rr	hduser	supergroup	1.42 KB	1	128 MB	51124	
-rw-rr	hduser	supergroup	2.7 KB	1	128 MB	51125	
-rw-rr	hduser	supergroup	427 B	1	128 MB	51126	
-rw-rr	hduser	supergroup	742 B	1	128 MB	51127	

You can manually download these files if you wish to see text within by clicking on Name and then download on the pop up that appears.

# 5. Run Mahout Algorithms

## Create folder to store results/output of process

mkdir ~/mahout

Run transformation to sequence process

cd /usr/local/mahout

bin/mahout seqdirectory -i/20newsgroups -o/20newsgroups-seq -ow

After this is successfully run we will see a new directory in HDFS files system (go to localhost:50070 -> Utilities -> Browse the file system)

## Browse Directory

/							Go!
Permission	Owner	Group	Size	Replication	Block Size	Name	
drwxr-xr-x	hduser	supergroup	0 B	0	0 B	20newsgroups	
drwxr-xr-x	hduser	supergroup	0 B	0	0 B	20newsgroups-seq	

And within this directory two files:

Browse I	Directory	/					
/20newsgroups-seq							Go!
Permission	Owner	Group	Size	Replication	Block Size	Name	
-FW-FF	hduser	supergroup	0 B	1	128 MB	_SUCCESS	
-rw-rr	hduser	supergroup	18.31 MB	1	128 MB	part-m-00000	

Return to the terminal take note of the output of previous command.

Copy and paste this into a file as we will use this for analysis later on.

8 🗖 🗊	hduser@hduser-VirtualBox: /usr/local/mahout
	FILE: Number of write operations=0
	HDFS: Number of bytes read=35855003
	HDFS: Number of bytes written=19202391
	HDFS: Number of read operations=75471
	HDFS: Number of large read operations=0
	HDFS: Number of write operations=3
	Map-Reduce Framework
	Map input records=18846
	Map output records=18846
	Input split bytes=1409104
	Spilled Records=0
	Falled Shuffles=0
	CC time alapsed (ms)-606
	CDU time clapsed (MS)=090
	Dhusical memory (hytes) spanshot-0
	Virtual memory (bytes) snapshot-0
	Total committed heap usage (bytes)=77549568
	File Input Format Counters
	Bytes Read=0
	File Output Format Counters
	Bvtes Written=19202391
16/09/0	1 19:52:30 INFO MahoutDriver: Program took 150807 ms (Minutes: 2.51345)
hduser@	nduser-VirtualBox:/usr/local/mahout\$

1. TFIDF on Hadoop

Run the following commands:

bin/mahout seq2sparse -i /20newsgroups-seq -o /20newsgroups-vectors -Inorm -nv -wt tfidf

# Browse Directory

1							Go!
Permission	Owner	Group	Size	Replication	Block Size	Name	
drwxr-xr-x	hduser	supergroup	0 B	0	0 B	20newsgroups	
drwxr-xr-x	hduser	supergroup	0 B	0	0 B	20newsgroups-seq	
drwxr-xr-x	hduser	supergroup	0 B	0	0 B	20newsgroups-vectors	

## Browse Directory

/20newsgroups-vectors	Go!
Permission Owner Group Size Replication Block Size Name	
drwxr-xr-x hduser supergroup 0 B 0 0 B df-count	
-rw-rr hduser supergroup 1.85 MB 1 128 MB dictionary.file-0	
-rw-rr hduser supergroup 1.8 MB 1 128 MB frequency.file-0	
drwxr-xr-x hduser supergroup 0 B 0 0 B tf-vectors	
drwxr-xr-x hduser supergroup 0 B 0 0 B tfidf-vectors	
drwxr-xr-x hduser supergroup 0 B 0 0 B tokenized-documer	its
drwxr-xr-x hduser supergroup 0 B 0 0 B wordcount	

# 2. Naïve Bayes on Hadoop

bin/mahout trainnb -i /20newsgroups-vectors/tfidf-vectors -o /model -li /labelindex -ow

## Browse Directory

/							Go!
Permission	Owner	Group	Size	Replication	Block Size	Name	
drwxr-xr-x	hduser	supergroup	0 B	0	0 B	20newsgroups	
drwxr-xr-x	hduser	supergroup	0 B	0	0 B	20newsgroups-seq	
drwxr-xr-x	hduser	supergroup	0 B	0	0 B	20newsgroups-vectors	
-rw-rr	hduser	supergroup	658 B	1	128 MB	labelindex	
drwxr-xr-x	hduser	supergroup	0 B	0	0 B	model	
drwxr-xr-x	hduser	supergroup	0 B	0	0 B	user	

Save output of command in terminal to a file for later analysis.

# Hadoop 2.7: Set Up, Configuration and Process Implementation

## 1. Updates to Process Needed for Hadoop 2.7.3

Shut down and close out of VM "Hadoop 2.6".

Open up VM "Hadoop 2.7 and Spark 2.0".

The process for this will be the exact same as the steps we toke for Hadoop 2.6 except for 1 thing – hadoop version references.

We just need to update the command:

sudo wget http://ftp.heanet.ie/mirrors/www.apache.org/dist/hadoop/common/hadoop-

2.6.4/hadoop-2.6.4.tar.gz

with

sudo wget http://www-eu.apache.org/dist/hadoop/common/hadoop-2.7.3/hadoop-

2.7.3.tar.gz

Also please ensure that any command which contain reference to Hadoop version 2.6.4, update this command to reference Hadoop 2.7.3.

# Spark 2.0: Set Up, Configuration and Process Implementation

## 1. Download, Set Up and Configure Spark 2.0.0

We are still using VM "Hadoop 2.7 and Spark 2.0", so ensure we are still logged in to this after completing Set Up, Configuration and Process Implementation for Hadoop 2.7.

Our implementation of Spark will connect directly into the same data we used for the Hadoop experiment, which sits in hdfs (hadoop file system).

So we download Spark from the following location

http://spark.apache.org/downloads.html

And ensure that we have selected the correct versions of both Spark (2.0.0) and Hadoop (2.7)



We then move the downloaded Spark file to a specific directory. In this case I have moved it to the same location as Hadoop: */usr/local/* 

Using the Linux commands below:

cd ~/Downloads sudo cp spark-2.0.0-bin-hadoop2.7.tgz /usr/local/

We then go to this directory, extract the file and set up a symbolic link (and name it Spark) for easier navigation to spark.

cd /usr/local

sudo tar -xvzf spark-2.0.0-bin-hadoop2.7.tgz

sudo In -s spark-2.0.0-bin-hadoop2.7 spark

(Optional) I then removed the original downloaded .tgz file from this directory just to keep the folder as clean as possible.

sudo rm spark-2.0.0-bin-hadoop2.7.tgz

The user needs to be given permission to be able to use spark.

sudo chown -R hduser:hadoop /usr/local/spark

## 2. Install Scala and check Spark Set Up

First of all we will test Spark in local mode to ensure it is running correctly.

In order to do so we need to have the latest version of Scala installed as the Spark Shell uses this programming language so we download Scala from the following location:

 sudo apt-get remove scala-library scala

 sudo wget http://www.scala-lang.org/files/archive/scala-2.11.8.deb

 sudo dpkg -i scala-2.11.8.deb

 sudo apt-get update

 sudo apt-get install scala

To test that the local install is working correctly we navigate to the spark directory and open the spark shell

cd /usr/local/spark bin/spark-shell

We see a screen that appears as follows:

Welcome to
$ ( \  \  \  \  \  \  \  \  \  \  \  \  \$
// versen 2.0.0
Using Scala version 2.11.8 (OpenJDK 64-Bit Server VM. Java 1.7.0 95)
Type in expressions to have them evaluated.
Type :help for more information.
scala>

Enter ":quit" and hit return to exit shell.

We now need to configure Spark so that it connects with the Hadoop set up that has already been set up which currently consists of one datanode as follows



We add the following lines to the bashrc file, opening it using the command:

vi ~/.bashrc

#### #SPARKVARIABLES

export SPARK\_HOME=/usr/local/spark export PATH=\$PATH:\$SPARK\_HOME/bin

Restart the VM so bashrc changes take effect. Log in to VM again.

We must configure the file "spark-env.sh" using the commands as follows:

cd /usr/local/spark/conf

Here there is a file named "spark-env.sh.template". We use this as the basis to our "spark-env.sh" as it does not yet exist. So we duplicate and remove "template" in the name and save it in the same location (/usr/local/spark/conf/)

sudo cp spark-env.sh.template spark-env.sh

## export HADOOP\_CONF\_DIR=/usr/local/hadoop/conf

### Set up a log folder and ensure permission is granted to user hduser:hadoop

cd /usr/local/spark	
sudo mkdir logs	
sudo chown hduser:hadoop logs	

#### Restart all services

cd /usr/local/hadoop	
sbin/stop-all.sh	
sbin/start-dfs.sh	
sbin/start-yarn.sh	

On top of the Hadoop services we also have to start our Spark Master and Worker

cd /usr/local/spark

sbin/start-all.sh

### jps

This is to ensure Master is running as in below:

hduser@ncistudent-VirtualBox:/usr/local/spark\$ jps
2899 NodeManager
11646 Jps
2763 ResourceManager
3473 SparkSubmit
2559 SecondaryNameNode
3241 Master
11606 Worker
2228 NameNode
2367 DataNode

You can view more in info relating to Spark master at address in web browser:

localhost:8080

					C Q Search		☆目	• +	î
Spark 2.0.0 Sp	oark Master	at <mark>spark</mark>	://ncistudent-Vir	tualBox:707	77				
URL: spark://ncistudent-)	VirtualBox:7077								
REST URL: spark://ncistu	udent-VirtualBox:6066	(cluster mode)							
Alive Workers: 0	llead								
Memory in use: 0.0 B To	otal, 0.0 B Used								
Annella stimmer O. Dunning	0. Openalista d								
Applications: 0 Hunning	g, u Completed								
Drivers: 0 Running, 0 Co	ompleted								
Drivers: 0 Running, 0 Co Status: ALIVE	ompleted								
Drivers: 0 Running, 0 Co Status: ALIVE Workers	pmpleted								
Applications: o Hunning Drivers: 0 Running, 0 Co Status: ALIVE Workers Worker Id	, o Completed	Address		State	Cores		Memory		
Applications of Authing Drivers: 0 Running, 0 Co Status: ALIVE Workers Worker Id Running Applicatio	ons	Address		State	Cores		Memory		
Applications of Aufming Drivers: 0 Running, 0 Co Status: ALIVE Workers Worker Id Running Application Application ID	ons Name	Address	Memory per Node	State	Cores ubmitted Time	User	Memory State	Dura	ition
Applications: 0 Hullming, 0 Co Drivers: 0 Hullming, 0 Co Status: ALIVE Workers Worker Id Running Application Application ID	ons Name	Address	Memory per Node	State State	Cores ubmitted Time ke Screenshot	User	Memory State	Dura	ation
Applications: 0 Halfming, 0 Co Drivers: 0 Halfming, 0 Co Status: ALIVE Workers Worker Id Running Application Application ID Completed Applica	ons Name	Address	Memory per Node	State State	Cores ubmitted Time ke Screenshot	User	Memory State	Dura	ation

Run spark-shell in order to use machine learning

bin/spark-shell --master yarn --driver-memory 1g --executor-memory 3g



Please take note of address as per above: Spark Context Web UI available at <u>http://10.0.2.15:4041</u> as this is where we will see our metrics for analysis.

This may be different in your implementation. Enter this address into web browser and you will see the following UI.

😧 🔿 🛛 localhost:40	40/stages/s	tage/?id=0&at	tempt=0			C	Q Search	☆ 自 ♥ ♥	<b>^ 9</b>
Spork 2.0.0	Job	s Stages	Storage	Environme	nt Executors	SQL		Spark shell	application
Details for	Stage	0 (Attem	pt 0)						
Fotal Time Across Locality Level Sum Input Size / Record	All Tasks: 1 mary: Any: 1 s: 34.2 MB /	1 min I ; Process local 18846	:1						
<ul> <li>DAG Visualization</li> <li>Show Additional M</li> <li>Event Timeline</li> </ul>	letrics								
<ul> <li>DAG Visualization</li> <li>Show Additional M</li> <li>Event Timeline</li> <li>Summary Metri</li> </ul>	letrics cs for 2 C	completed T	asks						
DAG Visualization     Show Additional M     Event Timeline Summary Metri Metric	letrics cs for 2 C	Completed T	asks	25th pe	rcentile	Median	75th percentile	Max	
DAG Visualization     Show Additional M     Event Timeline     Summary Metri Metric Duration	letrics cs for 2 C	Completed T Min 4.8 min	asks	25th pe 4.8 min	rcentile	Median 5.8 min	75th percentile 5.8 min	Max 5.8 min	
DAG Visualization     Show Additional M     Event Timeline     Summary Metri     Metric     Duration     GC Time	letrics cs for 2 C	Completed T Min 4.8 min 4 s	asks	25th per 4.8 min 4 s	rcentile	Median 5.8 min 5 s	75th percentile 5.8 min 5 s	<b>Max</b> 5.8 min 5 s	
DAG Visualization     Show Additional M     Event Timeline     Summary Metri     Metric     Duration     GC Time     Input Size / Record:	cs for 2 C	Completed T Min 4.8 min 4 s 17.1 MB / 8113	asks	25th per 4.8 min 4 s 17.1 MB	/ 8113	Median 5.8 min 5 s 17.1 MB / 10733	75th percentile 5.8 min 5 s 17.1 MB / 10733	Max 5.8 min 5 s 17.1 MB / 10733	
DAG Visualization     Show Additional M     Event Timeline     Summary Metri     Duration     GC Time     Input Size / Record:     Aggregated Me	trics by E	Min 4.8 min 4 s 17.1 MB / 811: Executor	asks	25th per 4.8 min 4 s 17.1 MB	/ 8113	Median           5.8 min           5 s           17.1 MB / 10733	<b>75th percentile</b> 5.8 min 5 s 17.1 MB / 10733	Max 5.8 min 5 s 17.1 MB / 10733	
DAG Visualization     Show Additional M     Event Timeline     Summary Metri     Metric     Duration     GC Time     Input Size / Record:     Aggregated Me     Executor ID ▲	s Addre	Min 4.8 min 4 s 17.1 MB / 8111 Executor Ess	asks 3 Tasi	25th per 4.8 min 4 s 17.1 MB k Time	/ 8113 Total Tasks	Median 5.8 min 5 s 17.1 MB / 10733 Failed Tasks	75th percentile 5.8 min 5 s 17.1 MB / 10733 Succeeded Tasks	Max 5.8 min 5 s 17.1 MB / 10733 Input Size / Records	

# 3. Run MLlib Algorithms

Staying in the spark shell -> Import required packages.

import org.apache.spark.mllib.feature.{HashingTF, IDF}
import org.apache.spark.mllib.linalg.Vector
import org.apache.spark.rdd.RDD
import org.apache.spark.mllib.regression.LabeledPoint
import org.apache.spark.mllib.classification.{NaiveBayes, NaiveBayesModel}

Import data input files from HDFS into an RDD of (String, String)

val example = sc.wholeTextFiles("hdfs://127.0.0.1:54310/20newsgroups/\*")

Manipulate and transform this RDD so that it is in the correct format for running TD-IDF as well as Naïve Bayes

<pre>val example1 = example.map{case(directory, text) =&gt; (text, directory)}</pre>				
val example2 = example1.map{case(text, directory) =>				
(if(directory contains "alt.atheism"){1}				
else if(directory contains "comp.graphics"){2}				
else if(directory contains "comp.os.ms-windows.misc"){3}				
else if(directory contains "comp.sys.ibm.pc.hardware"){4}				
else if(directory contains "comp.sys.mac.hardware"){5}				

else if(directory contains "comp.windows.x"){6} else if(directory contains "misc.forsale"){7} else if (directory contains "rec.autos"){8} else if(directory contains "rec.motorcycles"){9} else if(directory contains "rec.sport.baseball"){10} else if(directory contains "rec.sport.hockey"){11} else if(directory contains "sci.crypt"){12} else if(directory contains "sci.electronics"){13} else if(directory contains "sci.med"){14} else if(directory contains "sci.space"){15} else if (directory contains "soc.religion.christian"){16} else if(directory contains "talk.politics.guns"){17} else if(directory contains "talk.politics.mideast"){18} else if(directory contains "talk.politics.misc"){19} else if(directory contains "talk.religion.misc"){20} *else* {0}, *text*)}

val labels = example2.map{case(labels,text) => labels}

val pretf = example2.map{case(labels,text) => text.split(" ").toSeq}

## 3. TFIDF on Spark

Run IDF and subsequently TFIDF on output from transformations on input data above.

val hashingTF = new HashingTF()	
pretf.cache()	
val tf: RDD[Vector] = hashingTF.transform(pretf)	
val idf = new IDF().fit(tf)	

The latest command will be where we measure our performance of TFIDF. From here we perform 4 runs of the commands which create spark tasks, ignoring the first one and record these results for our

findings. Our results will be taken from localhost:4040 (or respective address as per spark-shell set up).

4. Naïve Bayes on Spark

We will prepare the data further for input into Naïve Bayes

pretf.unpersist()	
val tfidf = idf.transform(tf)	
val training = labels.zip(tfidf)	

Make labeled points for input of Naïve Bayes and run multinomial Naïve Bayes.

<pre>val training1 = training.map{case(x,y) =&gt; LabeledPoint(x,y)}</pre>	
training1.cache()	
val model = NaiveBayes.train(training1, lambda = 1.0, modelType = "multinomial")	

From here we perform 4 runs of the commands which create spark tasks, ignoring the first one and record these results for our findings. Our results will be taken from localhost:4040 (or respective address as per spark-shell set up).