CLOUDERA

Welcome To

Cloudera Developer Training for Spark & Hadoop

Instructor: Dr. Mary Myers Email: mmyers@cloudera.com

Google Doc: http://tiny.cloudera.com/DevSHWestpac

Training Environments

***The Instructor will let you know when to log into your assigned Virtual Machine with specific instructions. Please remember that the passwords to use consistently within the Virtual Machine (VM) should be ID* = training, *password* = training.**

Johnson Wang	https://cloud.skytap.com/vms/60ff5cca3351335d3558576ec7cf00f9/desktops
Amna Hassan	https://cloud.skytap.com/vms/27a32fc5af09b94a93f382ef9cf381a6/desktops
Dominique MacKenzie	https://cloud.skytap.com/vms/e8356793ced204190f75bf46333ad410/desktops
Riki Mitchell	https://cloud.skytap.com/vms/cee4d7baf71a22ffeb83961d00c0d9d0/desktops
Yifan Zhang	https://cloud.skytap.com/vms/5341897a695a758dfa09b45c04e40dfb/desktops
Steve Manion	https://cloud.skytap.com/vms/2f546bf9fa75babcb23c794cc4396ebc/desktops
*** Peter Bowman**	https://cloud.skytap.com/vms/a81a7e7bc0a17c7020146278d8f60869/desktops
Sam Cox	https://cloud.skytap.com/vms/a3fa502e02cb4b66128e9242fa1b7c85/desktops
Kalyan Emani	https://cloud.skytap.com/vms/d4b7d80badd25d8a9c8692ba9b6f3826/desktops
Neetika Srivastava	https://cloud.skytap.com/vms/4d242357045a95e59175ef38cb6a7758/desktops
Laks Arunachalam	https://cloud.skytap.com/vms/b3cd8ef5a622eaa36d526c6d082c14b0/desktops
Sam (Yujia Liu)	https://cloud.skytap.com/vms/4c5644eff22da370b41423c7dd8d142b/desktops
Huan	https://cloud.skytap.com/vms/42237af09c7ca0d86b647e8f8aa31730/desktops

Extra:
https://cloud.skytap.com/vms/fb1a8a2972ea23ff5217fa7ec7082bba/desktops
https://cloud.skytap.com/vms/d31edbb94a7700d872f0502811d234c0/desktops
https://cloud.skytap.com/vms/ed0d19e330df939108cfc0e47b8b6af8/desktops

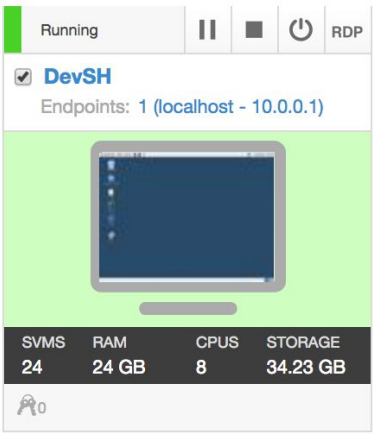
Instructor: https://cloud.skytap.com/vms/ed0d19e330df939108cfc0e47b8b6af8/desktops

Please ensure you are selecting the link assigned to you.

In case you need to modify your keyboard in Skytap:

You can modify their own keyboard from the default English(US) by following the instructions here: https://help.skytap.com/Setting_an_International_Keyboard_Layout.html

This course provides a single-host exercise environment running a pseudo-distributed Cloudera cluster (that is, a cluster running on a single host that emulates a multi-host environment) to complete the exercises.

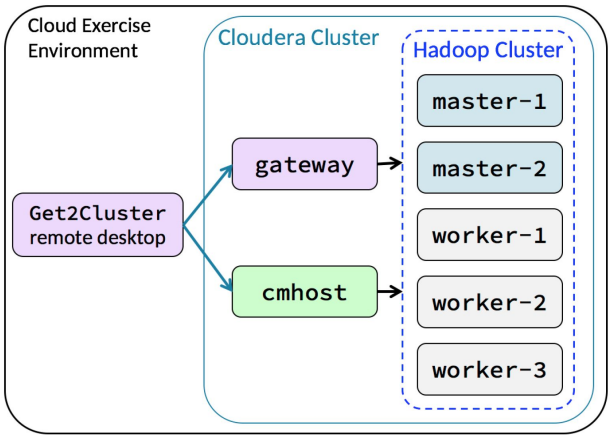


Although all the cluster services are running on the same remote machine, different services are configured with different host names. In order for the exercises be as realistic as possible, instructions will refer to services using the host names corresponding to the types of nodes that are part of in a typical cluster. For example, services that would usually run on a master node are configured with the host name master.

The table below shows the various host names in the pseudo-distributed cluster and the corresponding role that type of host would play in a full cluster.

Host Names	Role
`gateway`	A gateway node (sometimes referred to as an edge node) is a node outside the cluster that provides you with access to the services running on the cluster. Users and developers typically do their work on gateway nodes rather than cluster nodes.
`cmhost`	This node runs Cloudera Manager, which installs, configures, and monitors the services on the Hadoop cluster.
`master-1` `master-2`	Master nodes run the services that manage the Hadoop cluster.

`worker-1`
`worker-2`
`worker-3`
	Worker nodes execute the distributed tasks for applications that run on the Hadoop cluster.



NOTE: **The environment will NOT auto-suspend during class time. After the course ends, auto-shutdown is set on the environment (triggered by 30 minutes of inactivity), the environment is granted 10 hours of runtime, and terminates 30 days later.**

Troubleshooting Resources for the Skytap environment:

- Connecting to VMs Troubleshooting Guide: http://help.skytap.com#SmartClient_Help_Page.html
- Connectivity Checker - Use this tool to check your connectivity to Skytap: https://cloud.skytap.com/tools/connectivity
- Speedtest - Use this tool to check your network performance to Skytap: http://speedtest.skytap.com/

Hadoop Developer Class Tools Available

Tools:

Editing Tools -- from terminal command line:

Pluma: Click on the Gedit shortcut in the top toolbar. To invoke the graphical editor from the command line, type gedit followed by the path of the file you wish to edit. Appending & to the command allows you to type additional commands while the editor is still open. Here is an example of how to edit a file named myfile.txt:

pluma myfile.txt &

Nano: sudo nano *filepath&filename*
Ctrl + O and enter to save file
Ctrl + X to exit

Vl: sudo vi *filepath&filename*
Enter "I" o switch to Text mode – then type
Hit ESC to switch to Command mode – where you can enter"

:w	Write the file and continue editing
:q!	Quit (without saving)
:wq!	Write and quit (overwrite if necessary)

Emacs:

Web Browser: FireFox – link is at top of VM

Terminal Window:

File Browser: for local files: caja
Applications, System Tools, File Browser

Maven: mvn package *from the directory storing the files.*

Eclipse:

Linux Commands Used

Purpose	**Command**
Catchup Script	$DEVSH/scripts/catchup.sh
Shortcut to Data folder	$DEVDATA /home/training/training_materials/data
Shortcut to class files	$DEVSH /home/training/training_materials/devsh
Submit a Spark Program	Spark-submit (ex. Spark-submit --master yarn-client wordcount.py /loudacre/kb/*)
To view services running	**sudo -u hdfs jps**
To stop a service	sudo service *service name* stop
To start a service	sudo service *service name* start
History server	sudo service hadoop-mapreduce-historyserver restart
Spark History Server	sudo service spark-history-server restart
Zookeepers	**sudo service zookeeper-server restart**
Kafka	**sudo service kafka-server restart**
Copy a file	sudo cp hive-site.xml /etc/impala/conf
Delete a file	sudo rm /etc/impala/conf
Check for corrupt problems	sudo -u hdfs hdfs fsck /
Move corrupted files	sudo -u hdfs hdfs fsck / -move

Hadoop Commands Used

Purpose	**Command**
To list files	hdfs dfs –ls /*pathway/filename*
Remove file on the cluster	hdfs dfs –rm –r /*pathway/filename*
Copy local file to HDFS	hdfs dfs –put *localpath/filename HDFSpath/filename*
Copy HDFS to local	hdfs dfs –get *HDFSpath/filename localpath/filename*
Find a file on the cluster	locate filename
To make a directory on HDFS	hdfs dfs -mkdir /*directory*
View contents on screen	hdfs dfs -cat /*pathway/filename*
Help - HDFS DFS commands	hdfs dfs

Scala & Python Language Tutorials

Scala Help
General:

- http://docs.scala-lang.org/tutorials/
- Cheat sheet for Scala syntax: http://brenocon.com/scalacheat/

For Java Programmers:

- Scala for Java Programmers: http://docs.scala-lang.org/tutorials/scala-for-java-programmers.html
- Scala Tutorial: www.scala-lang.org/docu/files/ScalaTutorial.pdf
- Learning Scala: http://joelabrahamsson.com/learning-scala/

Scala Partial Functions (case functions) explained "without a PhD"
http://blog.bruchez.name/2011/10/scala-partial-functions-without-phd.html

Python Help
General:

- https://docs.python.org/2/tutorial/
- http://www.learnpython.org/ (interactive)
- https://developers.google.com/edu/python/

For Java Programmers:

- Python for Java Programmers: http://python4java.necaiseweb.org/Main/TableOfContents
- Python for the busy Java Developer: http://antrix.net/static/pages/python-for-java/online

Python Commands Used

Purpose	**Command**
To start Python pyspark shell	pyspark2
To view info on spark session object	spark
Exit shell	Ctrl + D or exit
Work with lines	(lambda line: “.jpg” in line)
Exit Python shell	Ctr + D or exit

Scala Commands Used

Purpose	**Command**
To start Scala shell	spark-shell2
To view info on spark session object	spark
To exit shell	Sys.exit or Ctrl + D
Work with lines	(line => line.contains(“.jpg”))
Exit Scala shell	sys.exit

Note: Scala must be compiled into a .jar file. Cloudera suggests use of Maven. If you are not familiar, this link goes to a blog post by Sandy Ryza that explains how to build and run a Spark program with Maven. https://blog.cloudera.com/blog/2014/04/how-to-run-a-simple-apache-spark-app-in-cdh-5/

Spark Commands Used

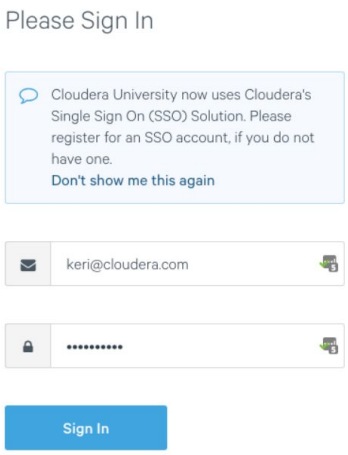
Purpose	**Command**
Read a file	read.*filetype*()
Display data	printSchema() or printSchema
Count number of items	count()
Display all data returned	collect()
Display specific number of items returned	take()
Break data on delimiter	split(*delimiter*)
To save as text file	saveAsTextFile("/pathway/dir")
Read the entire file	wholeTextFiles(files)
Flattens a value list into multiple rows	flatMap()
Return a new distributed dataset formed by passing each element of the source through a function *func*.	map()
A shortcut function - adds a key but leaves the whole string value in tact in results	keyBy()
A reduce function that merges the values for each key. The values it combines are all those associated with the same key—thus the "by key" part of the name.	reduceByKey()
PAIR RDD Operations:	
Returns a map with the count of occurrences	countByKey()
Groups all the values for each key in an RDD	groupByKey()
Sorts in ascending or descending order	sortByKey()
Returns an RDD containing all pairs with matching keys from two RDDs	join()
Returns an RDD of just the keys, without the values	keys()

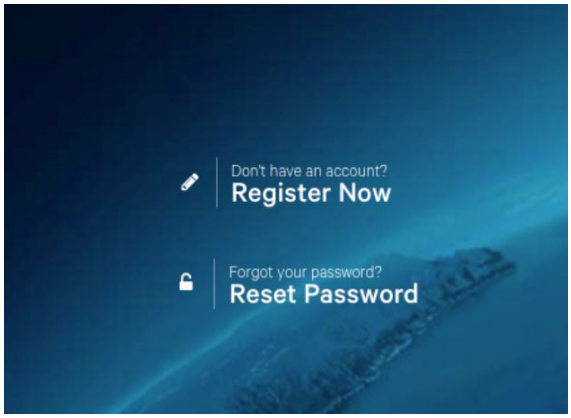
Returns an RDD of just the values, without the keys	values()
Returns the value(s) for a key	lookup(key)
Other joins	leftOuterJoin(), rightOuterJoin(), fullOuterJoin()
Execute a function on just the values, keeping the key the same	mapValues(), flatMapValues()
Change log level (to WARN)	setLogLevel("WARN")
Spark web user interface	localhost:4040
Read the schema of a parquet file	Parquet-tools schema *file.parquet*
Read the top contents of a parquet file	Parquet-tools head *file.parquet*

To access the course files:

Step 1: Go to: https://university.cloudera.com/user/learning/enrollments (Also available at university.cloudera.com, then select the Returning Students button , and then My Learning Dashboard.) If you have **not registered** for this class (class code: O9M7H0), go to this site and register: **https://university.cloudera.com/auth/class/register**

Step 2: If you have registered for this class: Click the **'Sign In'** button on the right. Use the ID and password that was used when you registered.





You may be prompted for any additional information needed.

There is a Reset Password link on the right if you are not certain of your password.

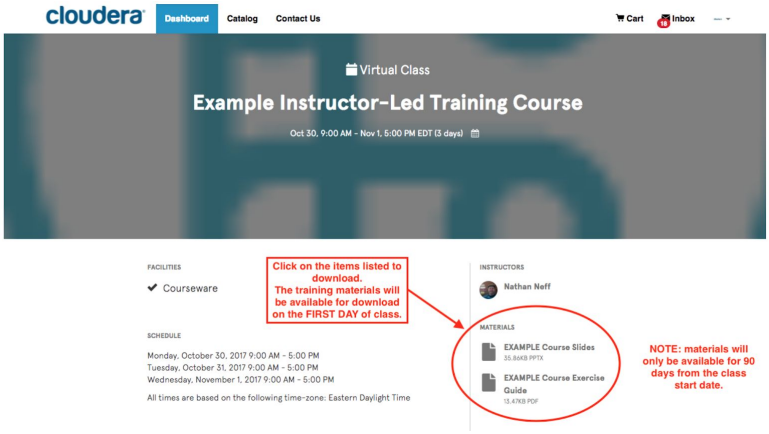
If your SSO account is "locked" You will need to contact SSO support website-login@cloudera.com

Step 3: Select the course you are taking - "**Cloudera Developer for Spark and Apache Hadoop**". It will look similar to the image below:

Step 4: You will then be able to view information regarding this course. **Download** the two PDF's listed on the *right side* of the course page under materials as shown in the list below.

Files to download are:

- **DevSH_190617a_Exercise_Manual.pdf**
- **DevSH_190617a_Student_Slides.pdf**



The files at training.cloudera.com will be available to you after class completes, and you may return any time within the next 90 days to download them again at your convenience. There is a maximum number of times you can download them though.

Questions and Resources during class:

If questions are asked during class and the answer is not in the course material, I will document the answer here for future reference. I also will include additional references periodically.

1. Cloudera courses: Scroll down for nine free ones: https://www.cloudera.com/about/training.html
2. Location of local Spark examples:
   /opt/cloudera/parcels/CDH/lib/spark/examples/lib
3. Examples of use cases for Cloudera software:
   Komatsu: Doubling equipment utilization:
   https://www.cloudera.com/more/customers/komatsu-mining.html
   Thomson Reuters: Separating Real News from Fake News on Twitter in 40 milliseconds:
   https://www.cloudera.com/more/customers/thomson-reuters.html
   Deutsche Telekom: 20% reduction in loss: https://player.vimeo.com/video/250886103 Select "watch on vimeo".
4.