1) PA5 teragen and terasort

Hi, I'm trying to run terasort on my EC2 instance for PA5. First, I run teragen:

It seems like teragen gets stuck on the "running job" line and never completes the job. Any clue on how to fix this? Thanks.

Solution:

restarting all the processes (yarn, dfs, jobhistory) worked for me.

2) NameNode doesn't appear when using JPS command?

I was going through PA5 and finished configuring the master and slave nodes. I tried starting Hadoop, but when using JPS to see what daemons were running, NameNode did not appear. I don't get any errors when running start-hdfs, but it just doesn't seem to start the NameNode.

What kind of error is going on? None of the online solutions have helped, and I've checked my configuration files.

Solution:

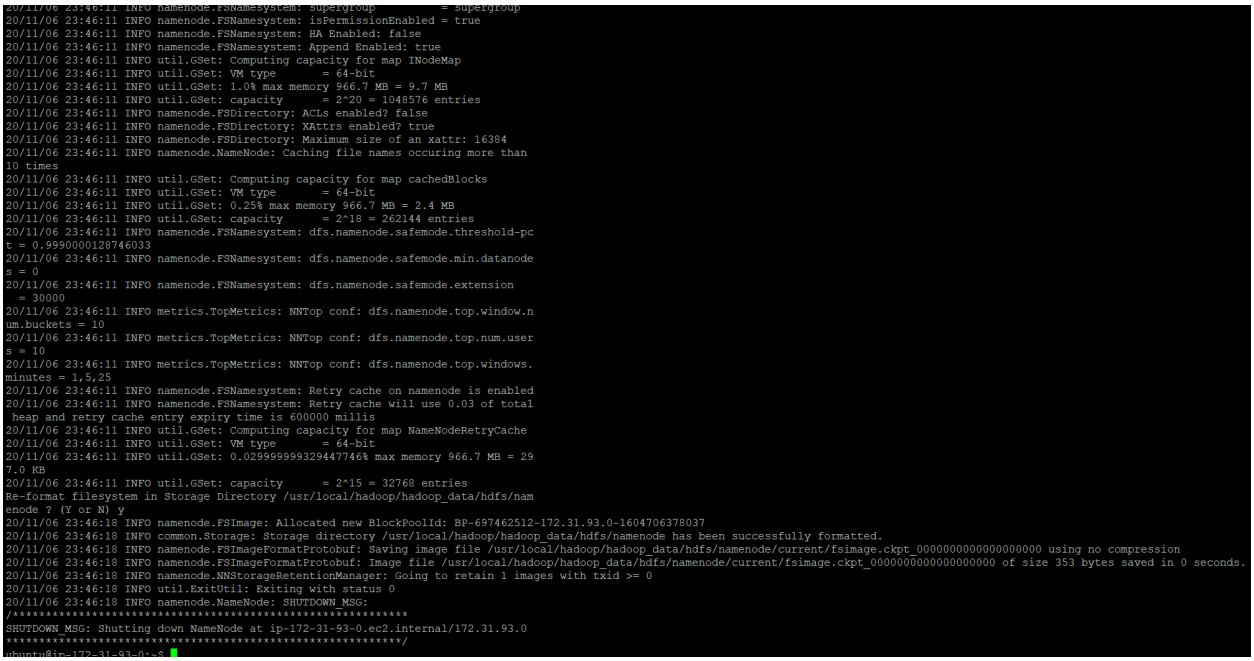
Instructor: Try these steps:
1. $HADOOP_HOME/sbin/stop-all.sh
2. sudo rm -rf /app/hadoop/tmp/
3. sudo mkdir -p /app/hadoop/tmp
4. sudo chown userName (i.e., ec2-user) /app/hadoop/tmp
5. sudo chmod 750 /app/hadoop/tmp
6. hdfs namenode -format
7. $HADOOP_HOME/sbin/start-all.sh
Check jps command to see if namenode is started.

Student: I ran the commands and namenode hasn’t started. Does it just take a while to start up?

Instructor: It should start straight away. Did start-all.sh start secondary namenode?

Student: Yes, it starts with Jps, ResourceManager, and SecondaryNameNode, but not NameNode.

Instructor: Could you post the log (especially the last part) printed on the console after "namenode -format" command?



above is the end of the log, here is the beginning of the log:

```
ubuntu@ip-172-31-93-0:~$ hdfs namenode -format
20/11/06 23:46:10 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = ip-172-31-93-0.ec2.internal/172.31.93.0
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 2.7.3
STARTUP_MSG:   classpath = /usr/local/hadoop/etc/hadoop:/usr/local/hadoop/share/
hadoop/common/lib/jsr305-3.0.0.jar:/usr/local/hadoop/share/hadoop/common/lib/par
anamer-2.3.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-collections-3.2
.2.jar:/usr/local/hadoop/share/hadoop/common/lib/protobuf-java-2.5.0.jar:/usr/lo
cal/hadoop/share/hadoop/common/lib/zookeeper-3.4.6.jar:/usr/local/hadoop/share/h
adoop/common/lib/jackson-core-asl-1.9.13.jar:/usr/local/hadoop/share/hadoop/comm
on/lib/commons-lang-2.6.jar:/usr/local/hadoop/share/hadoop/common/lib/jackson-ja
xrs-1.9.13.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-compress-1.4.1.
```

Instructor: Try this command "hadoop dfsadmin -safemode leave".

```
ubuntu@ip-172-31-93-0:~$ hadoop dfsadmin -safemode leave
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

safemode: Call From ip-172-31-93-0.ec2.internal/172.31.93.0 to ec2-34-201-59-55.
compute-1.amazonaws.com:9000 failed on connection exception: java.net.ConnectExc
eption: Connection refused; For more details see:  http://wiki.apache.org/hadoop
/ConnectionRefused
```

Student: this is the result from running the command

Instructor: Start all the hadoop daemons first and then try that command.

Student: Yes, I ran it after $HADOOP_HOME/sbin/start-all.sh and it gave me the same error. It recognizes there is a NameNode, but it just doesn't start. When I stop-all, it says "no namenode to stop".

Instructor: Check if there is any process running at port 50070 by using the command "sudo netstat -tulpn | grep :50070". Use "sudo kill -9 'process_id'; to kill the process. Then try formatting namenode and restarting all the daemons. If the problem persists, I think it will be less time consuming to reinstall hadoop than to debug the problem.

Student: There was no output when running the grep :50070 command; the problem still appears when formatting and restarting. Should I just go through the tutorials again?

Instructor: Yes. Probably during installation, something went wrong. Before you reinstall, can you try my first suggestion again? This time, in step 5, use 777 instead 750.

3) Hey, im getting an error when trying to start up the hadoop cluster:

I'm not sure what is causing it. Also there are a few steps that are in the commands.txt and powerpoint that aren't in the tutorials, specifically the yarn-site.xml and mapred-site.xml file configuration, are we not supposed them, because I did.

Solution:
Check whether the environment variables are set properly.
One solution can be as follows:

Try to export the HADOOP_PREFIX environment variable.

Add the following line to your ~/.bashrc file:

export HADOOP_PREFIX=/path_to_hadoop_location

for example:
export HADOOP_PREFIX=/home/XX/hadoop-2.7.1
Then do . ~/.bashrc in your terminal and try again, hope this will fix the error.

4) Step 3 for Mac

When setting up ssh for Mac I didn't have to download WinSCP. How would i do step 3 on mac?

It doesn't have to be in the .ssh folder. You can just put the .pem and config file where you want locally and then specify the path when you do scp

I used the format below for step 3 on mac and filled in my own path and the right username and public dns name for the for each node

scp -i /path/my-key-pair.pem /path/my-key-pair.pem /path/config my-instance-user-name@my-instance-public-dns-name:~/.ssh

Would the user name be 'ubuntu' in that command?

Solution:

Instructor: yes

Student: This is the command I put in and I got an error saying 'permission denied(public key). '

scp -i /Downloads/ssh/hadoop-key.pem /Downloads/ssh/hadoop-key.pem /Downloads/ssh/config ubuntu@ec2-18-212-31-22.compute-1.amazonaws.com:~/.ssh

Instructor: ssh' is the folder with my key. Did u get this error as well?

Student: No I didn't. If both your files are actually in Download/ssh I'm not sure why that doesn't work. Did you do 'sudo chmod 600 /Downloads/ssh/hadoop-key.pem' before trying it

It seems like there isn't a .ssh folder when I ssh to ubuntu. Which might be why I'm getting this error. When I type 'ls', there is nothing.

This is the error i get for chmod:

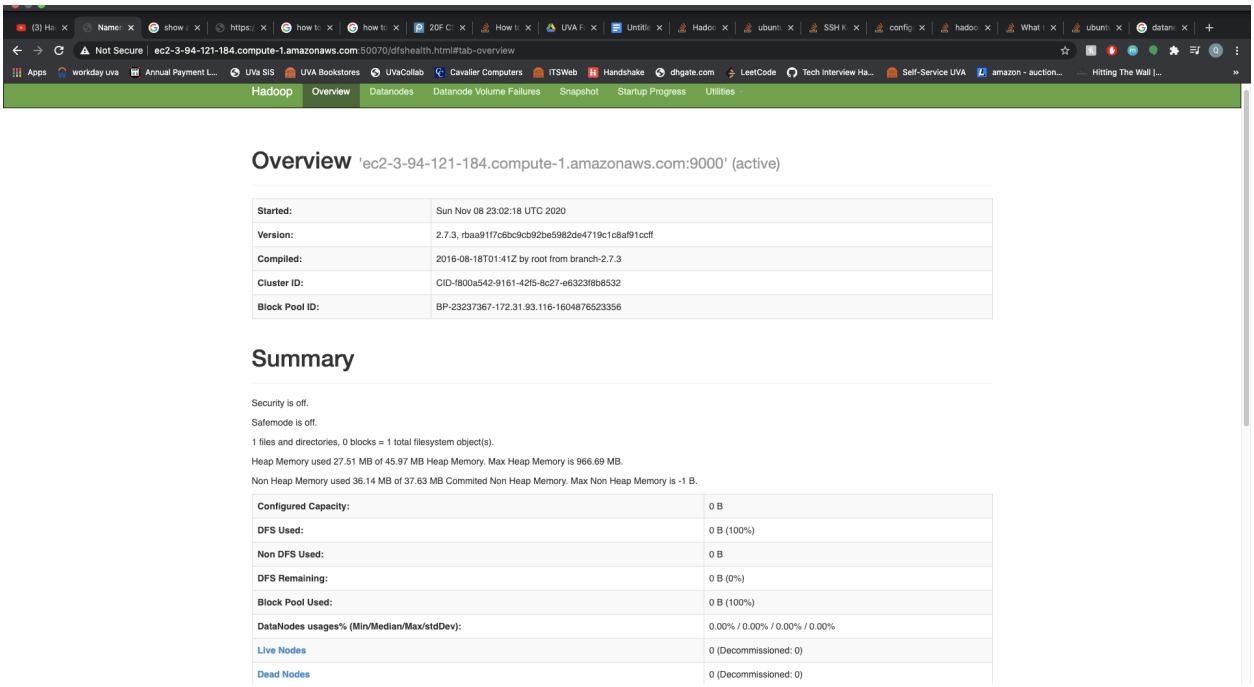
chmod: cannot access '/home/ubuntu/.ssh/hadoop-key.pem': No such file or directory

Instructor: well if you haven't copied hadoop-key.pem onto the node yet it's not going to be there. do 'sudo chmod 600 /Downloads/ssh/hadoop-key.pem' in your terminal on your mac before you do any scp commands to copy the pem and config files onto all 4 nodes. And you shouldn't be able to see the .ssh folder with ls. Anything with a '.' before it is hidden

Student: Got it thanks!

5) Issue with Datanode

When I run start-dfs.sh, my datanodes aren't starting but my namenode is, what should I do?



I had gotten this error but I restarted the steps from part 5 making sure the master and slave files in namenode corresponded to the config aliases

Solution:
Instructor: stop-all the running processes by typing the stop-all.sh command. Then remove all the folders inside your hadoop directory by typing the command 'rm -Rf /tmp/hadoop-your-username/*'
Finally, restart the namenode with the command 'bin/hadoop namenode -format' .
Then, restart the process with start-dfs.sh

Student: Same error! Were u able to figure it out?

Unfortunately no I wasn't :(

I also have the same error. I tried debugging in @340 but was not successful.

I don't seem to have a hadoop folder in my bin folder

I am not sure regarding your folder organization. You just need to format the namenode using the hadoop namenode -format command
oh okay that worked but it's still showing 0 live data nodes

Can you please double check the config, hosts and the configuration files in the data node whether they have the correct ip address?

Same error but and tried instruction from the instructor and no luck still, any one else able to fix this

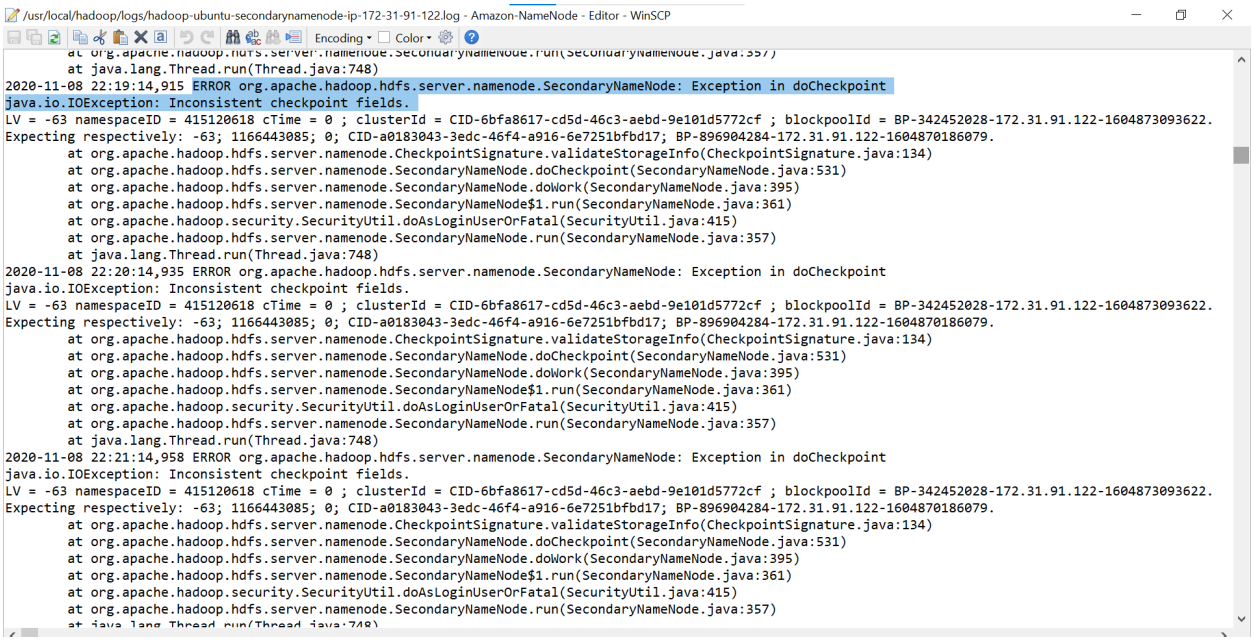
6) Datanodes won't start

Much like @339 my datanodes won't start, but my namenode will. When I ssh to the data notes and cat the log files located at /usr/local/hadoop/logs I find the following error.

java.io.IOException: Incorrect configuration: namenode address dfs.namenode.servicerpc-address or dfs.namenode.rpc-address is not configured.

I tried some commands from stackoverflow, but no luck. Any idea why this could be occuring?

Thanks!



Solution:
I found the fix. I copied the configuration from nodename in the core-site.xml file and pated it to all the datanode's core-site.xml files. This file is in the following directory path: /usr/local/hadoop/etc/hadoop/

Please check your core-site.xml whether it is configured properly.
Then restart the node using process mentioned in @339

7) Error: Temporary failure in name resolution

After running this command:

scp ~/.ssh/my-hadoop-key.pem ~/.ssh/config Datanode1:~/.ssh

I got the following error:

ssh: Could not resolve hostname datanode1: Temporary failure in name resolution
lost connection

I setup my config files based on the 3rd video in the tutorial, and I used the same naming conventions as were used in the video, so I don't know why I'm getting this error. My partner got the same error. How could I fix this?

Solution:

1. Please check whether the folder organization is correct
2. Please whether the name is (e.g., Datanode1 or datanode1) are same in both config and in the command.
3. Please check whether the public-dns-name is correct for the namenode and datanodes

8) step 3 on Mac with ssh file

I can connect to all four of my nodes, but I cannot start step 3 because I do not have a ssh file made. How do you go about doing that without WIN SCP?

When I enter this code:

sudo chmod 600 ~/.ssh/my-hadoop-key.pem

I get the error:

chmod: cannot access '/home/ubuntu/.ssh/my-hadoop-key.pem': No such file or directory

Solution:

You can run scp command on the terminal of your computer to copy the my-hadoop-key.pem file from your computer to the EC2 instance. The format of the command is "scp -i /path/to/my-hadoop-key.pem /path/to/file_you_want_to_copy username@instance_public_dns_address: /path/to/dir/where_you_want_to_upload_the_file (~/.ssh in this case)".

9) PA5: How to create config file for Step3
I am having trouble with the config file, I saved the config file as config.cfg, but when I try to run the following line
scp ~/.ssh/my-hadoop-key.pem ~/.ssh/config.cfg datanode1:~/.ssh,
it shows this error ssh:
Could not resolve hostname datanode1: Name or service not known
lost connection

Solution:
scp wasn't working consistently for me so I switched it out for rsync. Replace "scp" with "rsync" and see if that works. I did not have the confit file with the .cfg type. I just saved it as config Also make sure that you dragged and dropped the files into the .ssh directory

10) Instance status checks - Instance reachability check failed

I had issues running teragen and terasort commands the first time, so i decided to stop and restart the nodes. When I tried to restart I got a connection timed out error. I checked my ec2 for my node's health, and my datanode1 failed one of the staus checks. It failed the second status check, with the following error:

Instance reachability check failed

According to aws it has been failing for about 5 hours.

How can I fix this?

Solution:

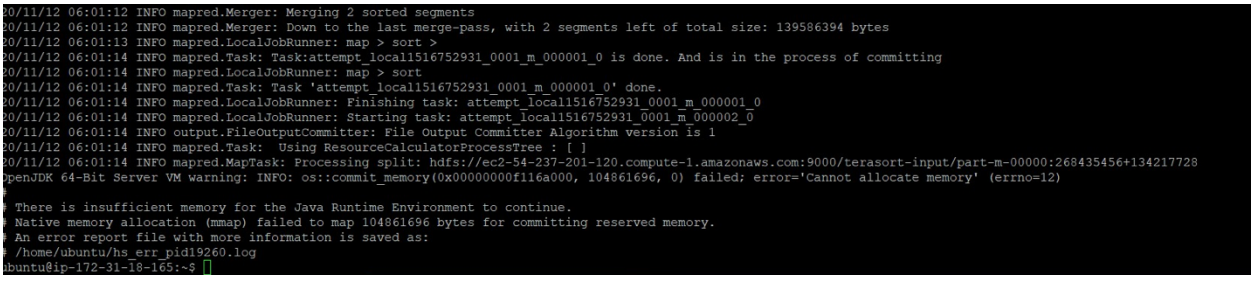
Instructor: Can you ssh to the datanode independently using putty? If the issue is occuring with AWS, it would be difficult to solve.
Are your teammates facing the same issue? You can consider submitting the last step with your teammates.

Student: Yes, I was able to ssh to the other datanodes, but I wasn't able to do so with datanode1, as I got a connection timed out error. This is the same node that is failing the status checks on aws. Does this mean I should submit the last part with a partner?

Instructor: Yes, you can submit the last part your partner.

11) PA-5 Terasort memory issues

When running Terasort I am running into memory issues that I assume is due to the server and out of my control. I have gotten everything to work correctly including Teragen however I have this error: Will I not get full credit for this?



Solution:
stop all the processes and restart them after formatting the namenode. Run teragen again, but generate less amount of data (probably drop 2 zeros from the size mentioned) and then run terasort on the generated data.

12) AWS DataTranfer cost

Anybody know how to stop the DataTransfer charges after the PA5? I shutdown all my EC2 instances but getting charged ~10 dollars. Not sure if there is anything else I need to terminate on AWS.

Solution:

Student: I'm glad to know I wasn't the only one with this problem... I contacted AWS support and they said that this happens if your security groups are open to all inbound traffic. So if you delete the inbound rules on your security groups that allow all traffic, it should stop the charges. It has worked for me so far. However, I was using my Educate account for PA5, so there were just 2 given-by-default security groups. I'm not sure if your groups are different, but if you're using Educate, they should be the same 2 default ones.

Hope this helps! It was definitely frustrating to see.

13) Generate Config file for step 3

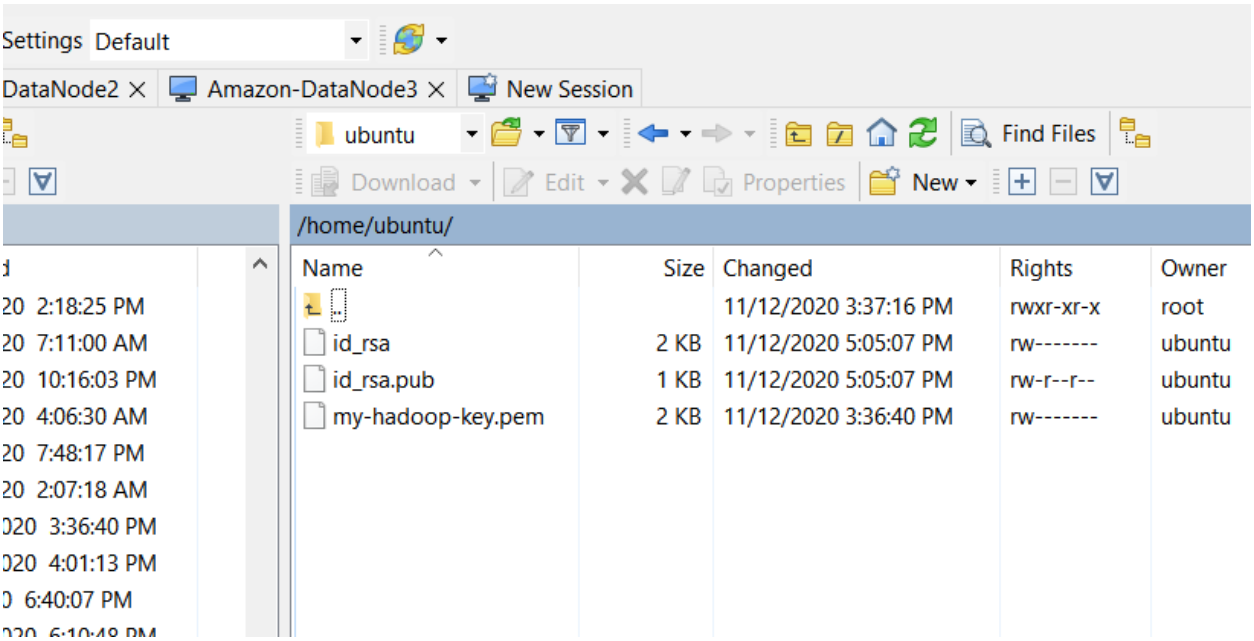
How do we generate the config file for step 3? He just seems to have on in the tutorial, and I cant see exactly what's on it from the video

Solution:
You need to create the config file by your own. Just create a file with the name 'config' without any extension. The contents of the file is basically providing aliases to the ec2 instances and the name of the key to access the instances.
A sample config file content can be as follows: You would need to change the public dns of the ec2 instances and the aliases accordingly.

Host Namenode
Hostname ec2-34-227-13-225.compute-1.amazonaws.com
User ubuntu
IdentityFile ~/.ssh/my-hadoop-key.pem
Host Datanode-1
Hostname ec2-34-205-127-60.compute-1.amazonaws.com
User ubuntu
IdentityFile ~/.ssh/my-hadoop-key.pem
Host Datanode-2
Hostname ec2-34-238-115-236.compute-1.amazonaws.com
User ubuntu
IdentityFile ~/.ssh/my-hadoop-key.pem
Host Datanode-3
Hostname ec2-54-197-16-131.compute-1.amazonaws.com
User ubuntu
IdentityFile ~/.ssh/my-hadoop-key.pem

14) Regarding the tutorial

Do I have to set a .ssh directory like the tutorial? Or can I just work on this directory in WinSCP for Step 3: Setup Passwordless SSH?



Solution:
When you first ssh to an instance, it automatically creates the ".ssh" directory and "authorized_keys" file in it. You can access the ".ssh" directory on WinScp by going to options->preferences->panel and checking "view hidden files". The files are not required to be stored in the ".ssh" directory but the "id_rsa.pub" key needs to be appended to "authorized_keys" which is located in the ".ssh" directory.