LAB 1 – GETTING FAMILIAR WITH UTD CLUSTER

Pre-requisites:
You will need an account with the UTD CS department.
You should be on-campus or be using the VPN if off-campus.

1. Logging in:
ssh YourNetID@cs6360.utdallas.edu

2. Adding path to your environment variables. Download the file to your account using the command:

wget http://www.utdallas.edu/~axn112530/cs6350/lab1/EnvVariables

Add this to your bash profile so that you don't have to worry about it every time.

Use the following commands:
vim ~/.bashrc

Add the following line at the end of the file:
source EnvVariables

Save and close the file.

Log out and log back in and you should be ready to go.

2. Check the version of HDFS

hadoop version

3. Let's check some configuration settings:
Go to following location:

cd $HADOOP_CONF_DIR

4. Here is a good explanation of various config files:
http://www.edureka.co/blog/hadoop-cluster-configuration-files/

http://www.edureka.co/blog/explaining-hadoop-configuration/

Answer following questions:

- What is the replication factor?
  Look at file hdfs-site.xml
- What is the name of the master node?
- How many slave nodes are there?
- What is the name of the cluster?
  Look at file core-site.xml
- What is the maximum amount of memory a DataNode will use for caching?
  Look at the parameter dfs.datanode.max.locked.memory in file hdfs-site.xml

5. Create a directory for yourself on HDFS (if it doesn't exist already)

hdfs dfs –mkdir /user/<YourNetID>

6. Create a local file:

echo "Hello World" > test.text

7. Upload it to HDFS:

hdfs dfs –copyFromLocal test.txt /user/<YourNetID>

8. Check that it exists:

hdfs dfs –ls

9. Run following command to see blocks:

hadoop fsck /path/to/file -files -blocks

10. Generate a report of the cluster by using the following command:
hdfs dfsadmin -report