Java –version
Install version 8 —> sudo apt install openjdk-8-jdk
version Change —> sudo update-alternatives --config java

sudo addgroup hadoop
sudo adduser --ingroup hadoop hduser
sudo adduser hduser sudo

sudo apt-get install openssh-server
which ssh
which sshd

su hduser
cd
ssh-keygen -t rsa -P ""
cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys
ssh localhost
exit

Install Hadoop
https://archive.apache.org/dist/hadoop/core/hadoop-2.9.0/
tar -xvf hadoop-2.9.0.tar.gz
sudo mv hadoop-2.9.0 /usr/local/hadoop
#Now, change the owner of hadoop folder using command:
sudo chown -R hduser /usr/local/

sudo gedit ~/.bashrc
#HADOOP VARIABLES START
export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64
export HADOOP_HOME=/usr/local/hadoop
export PATH=$PATH:$HADOOP_HOME/bin
export PATH=$PATH:$HADOOP_HOME/sbin
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib"
#HADOOP VARIABLES END

sudo gedit /usr/local/hadoop/etc/hadoop/hadoop-env.sh
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

sudo gedit /usr/local/hadoop/etc/hadoop/core-site.xml
<property>
<name>fs.default.name</name>
<value>hdfs://localhost:9000</value>
</property>

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

sudo gedit /usr/local/hadoop/etc/hadoop/mapred-site.xml
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property

sudo mkdir -p /usr/local/hadoop_tmp/hdfs/namenode
sudo mkdir -p /usr/local/hadoop_tmp/hdfs/datanode
sudo chown -R hduser:hadoop /usr/local/hadoop_tmp

sudo gedit /usr/local/hadoop/etc/hadoop/hdfs-site.xml
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.namenode.name.dir</name>
<value>file:/usr/local/hadoop_tmp/hdfs/namenode</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>file:/usr/local/hadoop_tmp/hdfs/datanode</value>
</property>

sudo gedit /usr/local/hadoop/etc/hadoop/yarn-site.xml
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
<property>
<name>yarn.nodemanager.aux-
services.mapreduce.shuffle.class</name>

<value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>

hadoop namenode -format
Start-all.sh
http://localhost:50070/

Additional Command
hdfs namenode -format
sudo chmod 777 -R filename/

Questions :
What is Hadoop?
Hadoop is an open source framework from Apache and is used to store process and analyze data which are very huge in volume.

Modules of Hadoop:-
1. **HDFS**: Hadoop Distributed File System. Google published its paper GFS and on the basis of that HDFS was developed. It states that the files will be broken into blocks and stored in nodes over the distributed architecture.
2. **Yarn**: Yet another Resource Negotiator is used for job scheduling and manage the cluster.
3. **Map Reduce**: This is a framework which helps Java programs to do the parallel computation on data using key value pair. The Map task takes input data and converts it into a data set which can be computed in Key value pair. The output of Map task is consumed by reduce task and then the out of reducer gives the desired result.
4. **Hadoop Common**: These Java libraries are used to start Hadoop and are used by other Hadoop modules.

Advantages of Hadoop:
Fast: In HDFS the data distributed over the cluster and are mapped which helps in faster retrieval. Even the tools to process the data are often on the same servers, thus reducing the processing time. It is able to process terabytes of data in minutes and Peta bytes in hours.
Scalable: Hadoop cluster can be extended by just adding nodes in the cluster.
Cost Effective: Hadoop is open source and uses commodity hardware to store data so it really cost effective as compared to traditional relational database management system.
Resilient to failure: HDFS has the property with which it can replicate data over the network, so if one node is down or some other network failure happens, then Hadoop takes the other copy of data and use it. Normally, data are replicated thrice but the replication factor is configurable.
Disadvantages
Still rough - means software under active development

Programming model is very Restrictive
Cluster Management is High

ssh has two main components:
3.1 ssh : The command we use to connect to remote machines - the client.
3.2 sshd : The daemon that is running on the server and allows clients to connect to the server.
What is ssh ?
Secure Shell (SSH) is a protocol for cryptographic network for operating network services securely over and unsecurely network

Rivest, Shamir, and Adleman (RSA)
The RSA algorithm is an asymmetric cryptography algorithm; this means that it uses a public key and a private key

Single Node Hadoop Cluster has only a single machine whereas a Multi-Node Hadoop Cluster will have more than one machine. In a single node hadoop cluster, all the daemons i.e. DataNode, NameNode, TaskTracker and JobTracker run on the same

A framework is used for building and deploying an application quickly. When we use a framework, we can use resources to facilitate faster development and a greater user experience. A library is used to enhance the functionality of an application.