* Install EC2 instance
  + Select Ubuntu 20.04
  + Graphical user interface, application

    Description automatically generated
  + Create a key pair if you have not had one, download the \*.pem key (make sure you remember where you put it) Note the screenshot shows dsci2024 but you can use any other name you want (e.g., I am using dsci2023).

Graphical user interface, text, application, email

Description automatically generated

* 10-20GB is sufficient (minimum is 8GB)

Graphical user interface, text, application, email

Description automatically generated

* Click Launch instance!
* Select your instance, and go to Connect. Find tab for SSH client:
* Copy the Example command
* Open a terminal with access to the ssh client program
  + If you have Windows OS, install Cygwin or use Powershell (see note at the end)
  + If you have Mac, just open a terminal windows which already has access to ssh
* “cd” to the place where you have downloaded the \*.pem key.
  + Execute: chmod 400 <your pem key>
  + (see the AWS screenshot)
* Paste the command you copied. The command looks like this:
  + ssh -i <your pem key> [ubuntu@ec2-xxxxx.us-west-2.compute.amazonaws.com](mailto:ubuntu@ec2-xxxxx.us-west-2.compute.amazonaws.com)
    - again, see the screenshot for example
  + say yes to the question.
  + You should now be connected to EC2.

Graphical user interface, text, application, email

Description automatically generated

* + Note: when you restart the instance, its ip address changes. You need to recopy the ssh connection string from EC2 web site.
  + Text editor on EC2:
    - nano
    - vi
* Before installing the following software, please first update package database by executing:
  + sudo apt update
* Install MySQL:
  + sudo apt install mysql-server
  + sudo mysql
  + In MySQL prompt (mysql>):
    - (note: do not copy and paste the following command, since it might add a new line character after ‘-‘.)
    - alter user 'root'@'localhost' identified with mysql\_native\_password by 'Dsci-551';
    - exit
  + mysql -u root -p
    - on password prompt, type: Dsci-551 and hit enter
  + (note) MySQL server consumes a lot of main memory
    - Step the server first, please you run other program, e.g., hdfs, spark, …
    - Stop the server by executing:
      * sudo service mysql stop
    - You may start the server by executing:
      * sudo service mysql start
* Install Java SDK:
  + sudo apt install default-jdk
  + (there might be a configuration menu popping up, just hit the tab key to select OK, and hit enter).
  + (add this line to the end of your ~/.bashrc file on EC2)
    - export JAVA\_HOME=/usr/lib/jvm/default-java
  + log out and log in to EC2 again
  + note whenever you modify ~/.bashrc file, you can either execute “source ~/.bashrc” or exit EC2 and log in again so that the updated can be re-executed.
* Install Spark:
  + (note: please install Spark, you need to install Java SDK first, see previous step)
  + wget <https://dlcdn.apache.org/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz>
  + tar xvf spark-3.5.0-bin-hadoop3.tgz
  + (add this line to ~/.bashrc)
    - export PATH=$PATH:~/spark-3.5.0-bin-hadoop3/bin
  + pyspark
* Install Hadoop:
  + wget <https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz>
  + tar xvf hadoop-3.3.6.tar.gz
  + **(Skip this step if you are taking 351):** Follow the instructions in: <https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html> on “pseudo distributed operation”. In particular,
    - Follow the configuration steps
    - Follow the “set up passphraseless ssh” steps
    - Edit the file: :~/hadoop-3.3.6/etc/hadoop/hadoop-env.sh
      * add the following line (you can edit line 54):
      * export JAVA\_HOME=/usr/lib/jvm/default-java
    - follow the execution steps to format namenode, start dfs, etc.
  + add this line to ~/.bashrc file:
    - export PATH=$PATH:~/hadoop-3.3.6/bin:~/hadoop-3.3.6/sbin
  + note (**ignore this if you are 351 students**):
    - if namenode does not start, try to reformat the namenode
    - if datanode does not come up, try:
      * rm -rf /tmp/hadoop-ubuntu/dfs/data
        + (note) this will remove the directory where hdfs stores its data node content.
      * Restart the dfs
* Install MongoDB:
  + Follow the instructions in <https://www.mongodb.com/docs/manual/tutorial/install-mongodb-on-ubuntu/>
  + Make sure using the steps for Ubuntu 20.04
  + If you get the key missing error message like this:

“W: GPG error: https://repo.mongodb.org/apt/ubuntu focal/mongodb-org/7.0 Release: The following signatures couldn't be verified because the public key is not available: NO\_PUBKEY 160D26BB1785BA38”

Try: sudo apt-key adv --keyserver keyserver.ubuntu.com --recv-keys 160D26BB1785BA38

Note: replace 160D26BB1785BA38 with your missing key

* + To start server: sudo service mongod start
  + Run client: mongosh
* Install pip:
  + sudo apt install python3-pip
* Windows OS: If you are using Windows, please first download & install Cygwin.
  + If you want to use Cygwin
    - Please go to Cygwin.com
    - Download and execute setup-x86\_64.exe
    - Make sure you select openssh package when installing
    - Your Cygwin default installation directory will be “c:\cygwin64”
      * Note: your home directory will be in msy2 will be like:
        + c:\cygwin64\home\<your user id>
      * copy your \*.pem file downloaded from AWS to this directory