**Data visualization**

List of software need to be installed

1. Miniconda
2. Jupyter notebook
3. Pyspark
4. Pip
5. Package like matplotlib
6. **How to install miniconda**

**Step 1. Write the following in vm shell**

wget <https://repo.continuum.io/miniconda/Miniconda3-latest-Linux-x86_64.sh>

chmod -R 755 Miniconda3-latest-Linux-x86\_64.sh./Miniconda3-latest-Linux-x86\_64.sh

Step 2. Accept the license and to prepend the Miniconda3 install

Do you wish the installer to prepend the Miniconda3 install location

to PATH in your /home/username/.bashrc ? [yes|no]

[no] >>> yes

**Step 3. Refresh your environment variables**

. ~/.bashrc

**Step 4. Check your default python location**

which python

step 5. **Anaconda having currently no internal update to 3.7, for upgrading Python being under 3.6 for example, you will have to do something like**

conda install python==3.7

**step 6. Upgrade the packages to the newest version**

conda update --all

**step 7. Anaconda update doesn’t clean automatically older packages, for this we have to use the following command**

conda clean -p

1. **Install the Jupyter Notebook**

**Step 1: write this no the shell**

conda install -c conda-forge notebook

**step 2: You can now launch the Jupyter Notebook with**

jupyter notebook

1. **How to install Pyspark**
2. **Down load .tag file from the official cite of apache spark stable version (in my case : spark-3.0.0-preview2-bin-hadoop2.7 )**
3. **Un tar it using command: tar –xvzf ‘tag file’**
4. **Ste the home and path as follow**

**export SPARK\_HOME=** /home/cloudera/miniconda3/spark-3.0.0-preview2-bin-hadoop3.2

**export PATH=$SPARK\_HOME:$PATH:**/home/cloudera/lib/spark-3.0.0-preview2-bin-hadoop2.7/bin

1. **Installing Pip**

python -m pip install -U pip

1. **Installing matplotlib**

python -m pip install -U matplotlib

Section II

Connect PYSPARK to HIVE

Step 1: see the path where hive is installed

Ps -ef | grep hive (see the path after org.apache.hadoop.util.RunJar)

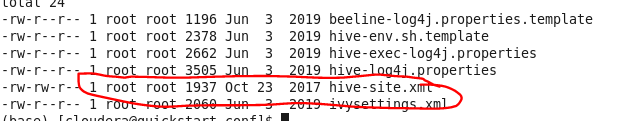


Step 2. Move to that directory (and see the configuration )

Cd /usr/lib/hive/conf

Step 3. List the configuration list and focus on hive -site-xml

Type: ll



Step 4. Get a path for it

Type: pwd



Step 5. Now, get the path for pyspark

Type: ps -ef | grep spark

Step 6. Chage the directory

Type: cd /usr/lib/spark/conf

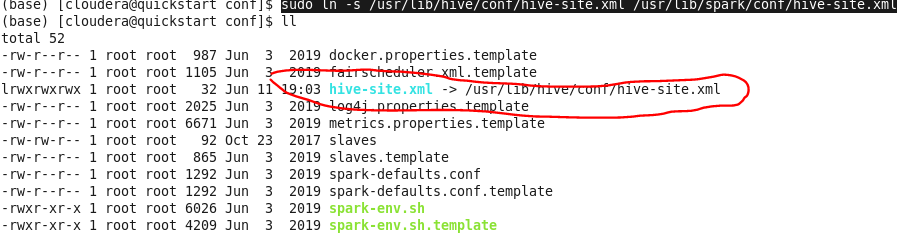
Step 7. List them

Type: ll

Note : still the link has not been created

Step 8: create symbolic link (softlink)

Type: sudo ln -s /usr/lib/hive/conf/hive-site.xml /usr/lib/spark/conf/hive-site.xml



Note : there are two way of excuting after connection, the first is on the shell and the second , more elegant was is to use spark submit followed by .py pyspark file

Type: ./bin/ spark-submit file.py

important : vim ~/.bash\_profile

export JAVA\_HOME=/usr/java/jdk1.8.0\_181

export PATH=$ JAVA\_HOME/bin:$PATH

**export SPARK\_HOME=** /home/cloudera/miniconda3/spark  
export PATH=$SPARK\_HOME/bin:$PATH  
export PYSPARK\_PYTHON=python3

export PYSPARK\_DRIVER\_PYTHON=jupyter

export PYSPARK\_DRIVER\_PYTHON\_OPTS='notebook'

**Reference**

1. <https://benjaminberhault.com/post/2018/07/11/install-python-jupyter-notebook-on-rhel-centos-7.html>
2. <https://html.developreference.com/article/12687980/Hive+on+windows+%3A+The+root+scratch+dir%3A++tmp+hive+on+HDFS+should+be+writable+%5Bduplicate%5D>
3. <https://kontext.tech/column/code-snippets/402/pandas-dataframe-plot-pie-chart>
4. <https://spark.apache.org/downloads.html>
5. <https://www.liquidweb.com/kb/how-to-install-pip-on-centos-7/>