EASY BUTTON

Note: THIS WILL DELETE ALL DATA ON YOUR EBS VOLUME.

This document is put together from various sources from the w205 course materials

1.0 Install Environment

Preliminaries

1. If you have not already, create an EBS volume. Note the region in which the volume was created.

2. Start an instance ofucbw205_complete_plus_postgres_PY2.7 (ami-558fc730) in the **same region** as your EBS volume.

3. Attach the volume to the instance.

4. Login to the instance via ssh. At login, you will be the root user.

Setup and install

1. Determine which device is your EBS drive by running fdisk –l

2. The last entry is typically your EBS volume. For example /dev/xvdf

3. Copy the device path (/dev/xvdf) to your clipboard

4. Download the setup script like this:

```
wget https://s3.amazonaws.com/ucbdatasciencew205/setup_ucb_complete_plus_postgres.sh
```

5. Run the script like this:

```
bash setup_ucb_complete_plus_postres.sh <paste your device path>
```

6. Follow the onscreen instructions

After the script runs

1. Hadoop will be installed and started

2. Postgres will be installed and started

3. Hive and SparkSQL will use Postgres as a metastore

4. A w205 user will exist. This is the user you should work as

5. A script called setup_zeppelin.sh has been created

a. If you want to setup zeppelin, type: ./setup_zeppelin.sh

b. You can start zeppelin by typing /data/zeppelin/bin/zeppelin.sh

You only need to go through this process **ONCE**. After the install, you should interact with your instance like this:

2.0 Set up Spark 1.5

Open https://spark.apache.org/downloads.html in your browser
Select a release as follows:

Spark 1.5.0
Pre-built for Hadoop 2.6 or later
Direct download

Copy the URL to download spark

As your personal user,

```
wget <url for spark>
tar xvzf spark-1.5.0-bin-hadoop2.6.tgz
mv spark-1.5.0-bin-hadoop2.6 spark15
export SPARK_HOME=$HOME/spark15
export HADOOP_CONF_DIR=/etc/hadoop/conf
```

You can start pyspark as follows:

```
$SPARK_HOME/bin/pyspark --master yarn
```

3.0 Integrating SparkSQL and the Hive Metastore

Integrating SparkSQL with the Hive Metastore is straightforward. However, we need to make sure that SparkSQL knows where our Hive metadata is.

Place a Hive configuration in Spark

```
mv spark15 /data

ln -s /data/spark15 $HOME/spark15

cp /data/hadoop/hive/conf/hive-site.xml /data/spark15/conf

export SPARK_HOME=$data/spark15
```

Edit /data/spark15/conf/hive-site.xml and change the following:

```
<!-- <property>
  <name>hive.metastore.uris</name>
  <value>thrift://localhost:9083</value>
  <description>IP address (or fully-qualified domain name) and port of the metastore host</description>
</property>
-->
```

To

```
<property>
  <name>hive.metastore.uris</name>
  <value>thrift://localhost:9083</value>
  <description>IP address (or fully-qualified domain name) and port of the metastore host</description>
</property>
```

Set up a Hive Metastore Service Script

In a file called /data/start_metastore.sh place the following:

```
 #! /bin/bash
nohup hive --service metastore &
```

In a file called /data/stop_metastore.sh place the following:

```
#! /bin/bash
ps aux|grep org.apache.hadoop.hive.metastore.HiveMetaStore|awk '{print $2}'|xargs kill -9
```

From now on, when you want to use Hive Data in Spark, you must do:

```
/data/start_postgres.sh

/data/start_metastore.sh
```

Then start spark using one of the following

```
/data/spark15/bin/pyspark
```

OR

```
/data/spark15/bin/spark-sql
```

OR

```
/data/spark15/bin/spark
```

OR

```
/data/spark15/bin/spark-submit
```

Zeppelin

Configuring Zeppelin as a primary interface for Pyspark, Spark and Spark-SQL (picks up Hive metadata)

As root, do the following:

- Copy the hive-site.xml from /data/spark15/conf to /data/zeppelin/conf
- Copy the Hadoop configurations (*-site.xml) from /etc/hadoop/conf to /data/zeppelin/conf

Before starting zeppelin, **make sure your metastore is started!**