

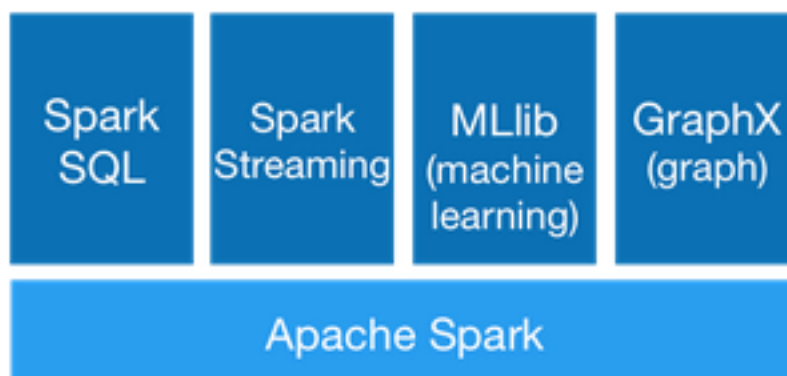
# MIDS W205

Lab #	6	Lab Title	Apache Spark and Spark SQL – An Introduction
Related Module(s)	6	Goal	Get you started on spark, pyspark and Spark SQL
Last Updated	9/27/15	Expected duration	60 minutes

## Introduction

Apache Spark is an open-source distributed computing. Spark uses in-memory processing and can in some situations be up to 100 times faster than regular Hadoop Map Reduce. The core of the Spark architecture is the concept of RDD's. You should read about them and why they enable scale and resiliency in a distributed environment with unreliable nodes.

Apache Spark has a basic computing substrate and several frameworks built on top of it. There is a framework for querying structured data (Spark SQL), a analytics framework for micro batching (they call it streaming) a machine-learning framework etc. In this Lab you will be learning about basic RDD's as well as about Spark SQL.



Spark SQL assumes that data is structured according to a relational model, which enables us to use SQL to query the data. It provides a programming abstraction based on what is called DataFrames. Spark SQL acts as a distributed SQL query engine.

Spark can be used from several different programming languages. We will be using Python as our preferred way of interacting with Spark. Spark is commonly utilized in a programmatic way. Sometimes there are command line interfaces (CLI's) that provide convenient ways of interacting using RDDs. In this Lab we will be using both.

## Instructions, resources and prerequisites

You can do this lab on one of the course provided AMI's, but you need to have Hadoop set up properly. Follow the instructions provided in other sections of the course. You can also do it on another computer such as you laptop by installing from the link in the table below.

Since the AMI's uses CDH 5.4.5 and Spark SQL is not officially support on that version you may encounter problems with the CLI portion of the lab. If you have problems you can skip the Spark SQL CLI portion (Step-4 and Step-5). But you still need to try the programmatic way of using Spark SQL shown in Step-6.

Resource	What
Spark download for your own install.	<a href="http://spark.apache.org/downloads.html">http://spark.apache.org/downloads.html</a>  <b>Set up Spark 1.5</b> Open <a href="https://spark.apache.org/downloads.html">https://spark.apache.org/downloads.html</a> 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, <pre>wget &lt;url for spark&gt; 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</pre> You can start pyspark as follows: <pre>\$SPARK_HOME/bin/pyspark --master yarn</pre>
<a href="http://spark.apache.org/docs/latest/programming-guide.html">http://spark.apache.org/docs/latest/programming-guide.html</a>	This guide shows each of these features in each of Spark's supported languages. It is easiest to follow along with if you launch Spark's interactive shell.  This includes reference to the basic commands you can perform on RDD's. Such as filter records, count records, join of data sets.
<a href="https://spark.apache.org/docs/1.1.0/sql-programming-guide.html">https://spark.apache.org/docs/1.1.0/sql-programming-guide.html</a>	Guide for using Spark SQL.
<a href="https://spark.apache.org/docs/latest/sql-programming-guide.html#running-the-spark-sql-cli">https://spark.apache.org/docs/latest/sql-programming-guide.html#running-the-spark-sql-cli</a>	Guide to Spark SQL CLI Shell.
<a href="https://spark.apache.org/docs/0.9.0/python-programming-guide.html">https://spark.apache.org/docs/0.9.0/python-programming-guide.html</a>	Python spark programming guide.
<a href="http://spark.apache.org/docs/">http://spark.apache.org/docs/</a>	Programming guide for the Spark Context object. Here you can

```
latest/api/scala/index.html#org.apache.spark.SparkContext
```

find actions available on the Spark Contexts.

## Step-0. Check installation and preparing data

In this section we will help you understand the basic of your Spark installation. We will also help you download and assemble a data set we like to use later in the Lab. We have to go through some trouble to get the dataset downloaded since it is reasonably big and stored in Github. Github has limitations on files size so we needed to split in manageable chunks. Your instructor may provide you with an alternative way of getting the same data set.

Firstly, if an example starts with the “\$” prompt, it is run in the Linux shell. If it starts with the “>>>” prompt it is run in pyspark shell.

Echo your SPARK\_HOME environment variable to see where you Spark is installed.

```
%echo $SPARK_HOME
```

You should see something like:

```
$ echo $SPARK_HOME
/Applications/devtools/spark
```

Got to that directory and look what is in it.

```
$ cd $SPARK_HOME
$ ls
CHANGES.txt NOTICE      README.md   bin    data  ec2    lib    python
LICENSE      R          RELEASE    conf   derby.log examples old
sbin
```

If you look in bin you will see spark-shell, pyspark and other tools for running and managing spark.

In your shell profile (often .bash\_profile) you may have something along the following lines so that the Spark commands are in your shell execution path.

```
export SPARK=/usr/lib/spark
export SPARK_HOME=$SPARK
export PATH=$SPARK/bin:$PATH
```

You can test to see that your shell can find them by running the Unix/Linux that command.

```
$ which spark-shell
$ which pyspark
```

```
$ which spark-sql
```

If your shell cannot find any of those programs you will likely need to check your installation. If which can find the commands it will return the location of the programs.

We will be using two different datasets in this lab. One is a web log data set. The other one is a data set with historic crime data from 2001 for the Chicago area. If you have time play around with the data set, it is interesting to explore types of crimes, density of crimes and so forth. For this lab you can use any substantial data file as replacement for "Crimes\_-\_2001\_to\_present.csv" but you need to modify the commands accordingly. You can download the "Crimes\_-\_2001\_to\_present.csv" file from github. Because it is a big file and we uploaded it to github we needed to compress and split the file. So you will need merge the parts and un-compress the result.

The instructors may have the files more easily accessible for you if you have problems with the below.

Get the files in this directory on github: [https://github.com/UC-Berkeley-I-School/w205-labs-exercises/tree/master/data/Crimes\\_-\\_2001\\_to\\_present\\_data](https://github.com/UC-Berkeley-I-School/w205-labs-exercises/tree/master/data/Crimes_-_2001_to_present_data)

You can get them by cloning the exercise repository if you have not already done that. If you done that can just get a `git pull`.

```
git clone https://github.com/UC-Berkeley-I-School/w205-labs-exercises.git
```

The files are in the `data/Crimes_-_2001_to_present_data` directory.

Once you have the files (should be seven of them all starting with the letter "x") run the following commands (make sure you do not have other files starting with x in the directory). The first one will just concatenate the split files into one file. The original file was a compressed csv file, so we name it appropriately. Next we un-compress it to get the original csv file.

```
$ cat x* > Crimes_-_2001_to_present.csv.gz
$ gunzip Crimes_-_2001_to_present.csv.gz
```

The result should be that you have the Crime data csv file in your directory. You also still have the original split files. If you run `ls` it should look something like this:

```
$ ls
Crimes_-_2001_to_present.csv  xac          xaf
xaa                          xad          xag
xab                          xae
```

You can check the correctness of the resulting files by checking the size of the file or number of rows in the file. The size may be different in different computers, but the number of lines should be the same.

```
$ du -s Crimes_-_2001_to_present.csv
2688712 Crimes_-_2001_to_present.csv
$ wc -l Crimes_-_2001_to_present.csv
5862796 Crimes_-_2001_to_present.csv
```

If you like you can remove the split files use; `rm`, but use the `-i` option so that you do not accidentally remove other files in the directory that you like to save.

```
$ rm -i x*
```

At this point you should be able to see the Crime data file in your directory.

```
$ ls
Crimes_-_2001_to_present.csv
```

You are good to go. In addition, also tried some useful Linux/Unix commands such as `du`, `wc` and `cat`.

## What you should have learnt

You should have checked the basic installation of Spark. And you have checked that the Spark programs can be found by your interactive Linux (Unix) shell.

## Step-1. Start pyspark

First we will start a spark shell so that we can access spark and interactively process Spark commands. We will be using `pyspark`, which is a python based shell for spark.

If we assume you have the spark bin directory in your `PATH` environment variable you can start `pyspark` by just issuing the command.

```
$pyspark
```

Otherwise go to the `/bin` directory in the installation (`$SPARK_HOME`) folder and type.

```
$ ./pyspark
```

In the `pyspark` shell you can use python instructions and you can create RDDs. Consequently you can apply operations on RDD's as well. Create a Python variable with some value using this command. This is plain old Python.

```
>>> x = [1,2,3,4,5,6,7,8,9];
>>> print x
[1, 2, 3, 4, 5, 6, 7, 8, 9]
>>> len(x)
9
```

There is something call a Spark Context. The spark context is the “object” you use to refer to the spark cluster. Type `sc` to verify you has one already in you Python shell:

```
>>> sc
<pyspark.context.SparkContext object at 0x1063b3410>
```

So far you only used Python statements and you checked you have a Spark Context. You can use the Spark context to create a Spark RDD from Python data using the command `parallelize`

```
>>> distData = sc.parallelize(x);
>>> print distData;
ParallelCollectionRDD[0] at parallelize at PythonRDD.scala:391
```

The Spark Context `parallelize` action is used to take local programming collections and create RDDs from them. In this case we created a RDD from a Python array.

The resulting value of the `distData` variable is an RDD representation. Try doing `len(distData)` ; what happens and why? To count elements in an RDD you need to use RDD actions. You can find a list of actions in the programming guide. To count the elements you use the `count()` action. Try the following:

```
>>> nx=distData.count()
>>> print nx
9
```

As you probably noticed, the default level of logging can be distracting. To reduce the logging information go to the `$SPARK_HOME/conf` directory. Some times this is stored in another location. If you are using one of our AMI's, look in the `/usr/lib/spark/conf/` directory. Create a `log4j.properties` file. If you do not already have one you can get one by copying the `log4j.properties.template` file. Change the logging level to warnings (WARN) by change `INFO` to `WARN` in the property `log4j.rootCategory=WARN, console`

```
# Set everything to be logged to the console
log4j.rootCategory=WARN, console
log4j.appender.console=org.apache.log4j.ConsoleAppender
log4j.appender.console.target=System.err
log4j.appender.console.layout=org.apache.log4j.PatternLayout
```

Restarts pyspark and rerun the commands above. You should see much less output in the Spark shell.

## What you should have learnt

You should understand how to start `pyspark` and have learnt that `pyspark` is a shell with access to the Python language and to the underlying Spark cluster capabilities. We also show you how to make `pyspark` less verbose by reducing the amount of logging displayed. If you have problems you may consider increasing logging again to understand what is going on in your execution.

## Step-2. Load a file and count the rows

Spark is commonly used to process large sets of data, and naturally we often read these data files from disc. The action `textFile` is an operation on Spark Contexts that creates a RDD from a file that resides in HDFS or the local files system. Make sure you cloned the data file `Crimes_-_2001_to_present.csv` in the lab directory (assuming you are using directory `"/data/mylab"`). Now create an RDD from that file using this action:

```
crimedata =sc.textFile("file:///data/mylab/Crimes_-_2001_to_present.csv")
```

If you run the AIM you can copy the data to HDFS and have `pyspark` get the file from there.

Print the number of lines this RDD using the following command:

```
>>> print crimedata.count()
5862796
>>>
```

As you can see there are almost 6 million records, so it took a few seconds to count them.

You can get the first element of the RDD with the operation first:

```
>>> crimedata.first()
```

You can get the n first elements with the operation `take(n)`

```
>>> crimedata.take(10)
```

As you see the data include the header information. Remove the header from the RDD is not straightforward. The reason is that RDD is immutable by design. One way to remove it is to create a new RDD and to filter the first row out. There are many ways to do that, here is an example:

```
>>> noHeaderCrimedata = crimedata.zipWithIndex().filter(lambda
(row,index): index > 0).keys()
```

It is a quite a processing heavy way of doing it, but as we mentioned RDD's are immutable you cannot just go in and remove a record. It is an inherent assumption to Spark which allows it to certain things more efficiently.

Print the first line to check that the header if gone. And count the lines to make sure the number seems correct.

```
>>> noHeaderCrimedata.first()
>>> noHeaderCrimedata.count()
```

Here is a more Python like way of doing the same that may be easier to understand. We first define a Python function in our Python spark shell using the following line.

```
>>> def remove_header(itr_index, itr): return iter(list(itr)[1:]) if
itr_index == 0 else itr
```

We then execute a `mapPartitionWithIndex` operations passing that function as an argument.

```
>>> noHeaderCrimeData2 = crimedata.mapPartitionsWithIndex(remove_header)
```

Print the first line and the count to make sure it is correct.

### What you should have learnt

You should understand how to create an RDD from a file, how apply operations on RDD. You used examples such as `first` and `count`. We also illustrated that RDD's are immutable and that to even remove one row (the header) you need to create a new RDD.

## Step-3. Filter records, structures

One obvious operation for Spark is to filter the data. Lets filter out all crimes that seem to be related to "NARCOTICS".

We can do this using the filter operations and a lambda function that checks for each row if the word "NARCOTICS" appears in that specific row. We will only return rows that included that word.

```
narcoticsCrimes = noHeaderCrimedata.filter(lambda x: "NARCOTICS" in x)
>>> narcoticsCrimes.count()
663712
```

It appears that 663712 crimes are related to narcotics. Use `take(n)` to check that the data seems ok. For example:

```
>>> narcoticsCrimes.take(20)
```



The RDD we have are just long strings, the fact that fields are comma separated does not mean anything to the Spark RDD. If we want to create a structure form more advanced computations and manipulations we need to parse the rows and create the appropriate structure or our data. The operations below splits each record up as an array using the Python operation split. It then create a new RDD were each row is an array of strings as opposed to one long string.

```
>>> narcoticsCrimeRecords = narcoticsCrimes.map(lambda r :  
r.split(","))
```

You can see the first array record using:

```
>>> narcoticsCrimeRecords.first()
```

You can check that you still have the same number of rows using:

```
>>> narcoticsCrimeRecords.count()
```

### **What you should have learnt**

You should understand that RDD's are immutable. You can filter RDD's but it creates a new RDD. You should also understand that RDD's do not understand anything about the structure of the records (except for key-value structures which we discuss in the next section). But you can store any Python structure that seems useful in an RDD.

### **Step-3. Key-values**

An important structure in spark is called Key Value pairs. In Python those are represented as Python tuples. A tuples is an immutable sequence of elements of various types.

You can create a new RDD consisting of tuples using the following operation:

```
>>> narcoticsCrimeTuples = narcoticsCrimes.map(lambda x :  
(x.split(",")[0], x))
```

You can check that the number of tuples is the same as the number of records in the data.

```
>>> narcoticsCrimeTuples.count()
```

It takes the first element and makes it a key and the rest if the row becomes the value part of the tuple. You can examine the tuple using the following operations:

```
>>> narcoticsCrimeTuples.first()
```

And you can check it out using these RDD and Python functions.  
Get the first tuple:

```
>>> firstTuple=narcoticsCrimeTuples.first()
```

How many elements do you have in the tuple?

```
>>> len(firstTuple)
```

What is the key of the first tuple?

```
>>> firstTuple[0]
```

What is the value of the first tuple?

```
>>> firstTuple[1]
```

There is one little problem with the tuple. Can you spot it? How should we change the map and lambda functions above to address that?

There are many operations you can do once you have a key-value tuple. You can join, reduce, map etc. You can read about the operations in the RDD Spark programming guide. One operation you can do is to sort by key:

```
>>> sorted=narcoticsCrimeTuples.sortByKey()
```

If print the first element in the sorted RDD and the original RDD you will see they are different.

```
>>> sorted.first()
```

```
>>> narcoticsCrimeTuples.first()
```

[SUBMISSION 1: Submit the first 10 rows of the unsorted and the sorted RDD to show that you successfully created both. Also explain the issue with the tuple and how to possibly correct it in the map/lambda function that was used to create the tuples.](#)

## What you should have learnt

Now you should understand the concept of Key-Value tuples, and understand how you can create them. You have also tried one operation on RDD's using the key-value structure.

## Step-4. Start Spark-SQL

**Important:** You can skip this step if you run on an AMI that does not have appropriate support for spark-sql. But if you have a later Spark with Hadoop installed on your computer you should be able to run this. Proceed to Step-6 to learn about Spark SQL.

Spark SQL can be used directly from `pyspark` or a `scala` shell. But there is also a Spark SQL CLI called the Beeline client. If you use `pyspark` or use Spark SQL programmatically

you need create a special Spark SQL contexts. With the Spark SQL CLI the context is already there for you and you can use SQL commands.

You start Beeline with the command

```
$spark-sql
```

Once started you can run some commands to see that it works. Show tables, create a table and drop the table by running the commands below.

```
spark-sql> show tables;

spark-sql> create table dummy (somedata varchar(500));
OK
Time taken: 0.369 seconds

spark-sql> show tables;
dummy false
Time taken: 0.053 seconds, Fetched 1 row(s)

spark-sql> drop table dummy;

spark-sql> show tables;
```

### What you should have learnt

You should understand the difference between using the Spark SQL CLI and using Spark SQL programmatically. You are able to start Spark SQL CLI and issue some basic commands to see that it works.

## Step-5. Spark SQL table loaded with data from a CSV file

**Important:** You can skip this step if you run on an AMI that does not have appropriate support for spark-sql. But if you have a later Spark with Hadoop installed on your computer you should be able to run this. Proceed to Step-6 to learn about Spark SQL.

You can create a Spark SQL table. The following create statement creates a table that has a schema that corresponds to the `web_session_log` data. Run the following create statement directly on the `spark-sql` shell prompt.

```
create table web_session_log
(DATETIME varchar(500),
USERID varchar(500),
SESSIONID varchar(500),
PRODUCTID varchar(500),
REFERERURL varchar(500))
row format delimited fields terminated by '\t'
stored as textfile;
```

You can load files from the local files system or from HDFS. Lets load a web\_log data available on github.

Run the describe command to see that it was created correctly. Otherwise you may need to drop it and try again correcting any mistakes.

```
spark-sql> describe web_session_log;
datetime      varchar(500)      NULL
userid        varchar(500)      NULL
sessionid     varchar(500)      NULL
productid     varchar(500)      NULL
refererurl    varchar(500)      NULL
Time taken: 0.083 seconds, Fetched 5 row(s)
```

Assuming you have the weblog data in the directory were you are running spark-sql shell, you can load the file from the files system into the table using the command below. If the file is located somewhere else you need to modify the path of the file.

```
spark-sql> LOAD DATA LOCAL INPATH "./weblog_lab.csv" INTO TABLE
web_session_log;
```

Once the data is loaded you can count the number of rows. You can count the number of rows with the following select statement.

```
spark-sql> select count(*) from web_session_log;
```

You can check that this seems reasonable by comparing with the number of rows in the original file. You can get the number of files using the Unix/Linux command wc.

```
$ wc -l weblog_lab.csv
```

Using Spark SQL you can use a number of SQL command so query your data. Lets select all rows in the web log that are related eBay.

```
spark-sql> select * from web_session_log where refererurl =
"http://www.ebay.com" ;
```

And now lets count the number of rows that are related to eBay.

```
spark-sql> select count(*) from web_session_log where refererurl =
"http://www.ebay.com" ;
```

## What you should have learnt

In this section you should have learnt how to create a table in the Spark SQL CLI and how to load data into the empty table. You also practiced some simple SQL commands on the loaded data set.

## Step-6. Accessing Spark-SQL in Python Code

Spark SQL can also be used from a program or directly from a shell such as pyspark. It will require that you create the appropriate spark context and programmatically define schemas and such. A table is ultimately represented as a Spark DataFrame. Below we will go through step by step how what you need to imports, how you read the data, how you create an object that represents the schema, and finally how you combine the schema definition and the data to create a table DataFrame.

We also show some simple queries using SQL in the resulting SQL DataFrame.

First make sure to import all necessary types etc. to you Python environment. You can try the below example interactively in a pyspark shell. Make sure you run it in the directory were you having the data file, or adapt the path in the example accordingly.

```
$ pyspark

>>> from pyspark.sql import SQLContext
>>> from pyspark.sql.types import *
```

Create the Spark SQL Context.

```
>>> sqlContext = SQLContext(sc)
```

Read the web log data into an RDD. You may need to adjust the path to the data based on were you stored it.

```
>>> lines = sc.textFile('file:///data/labs/w205-labs-
exercises/data/weblog_lab.csv ')
```

Create a map of the data so that it can be structured into a table.

```
>>> parts = lines.map(lambda l: l.split('\t'))

>>> Web_Session_Log = parts.map(lambda p: (p[0], p[1],p[2], p[3],p[4]))
```

Create string with the name of the columns of your table.

```
>>> schemaString = 'DATETIME USERID SESSIONID PRODUCTID REFERERURL'
```

Create a data structure of StructFields that can be used to create a table.

```
>>> fields = [StructField(field_name, StringType(), True) for  
field_name in schemaString.split()]
```

Combine the fields into a Schema object.

```
>>> schema = StructType(fields)
```

Create a table based on a DataFrame using the data that was read and the structure representing the Schema.

```
>>> schemaWebData = sqlContext.createDataFrame(Web_Session_Log, schema)
```

Register the object as a table with a table name.

```
>>> schemaWebData.registerTempTable('Web_Session_Log')
```

Query the table.

```
>>> results = sqlContext.sql('SELECT count(*) FROM Web_Session_Log')
```

Use the DataFrame operation show to print the content of the result of the query.

```
>>> results.show()
```

The following query can be used to query the number of rows related to ebay.

```
select count(*) from web_session_log where REFERERURL'=  
"http://www.ebay.com" ;
```

Note: the case sensitivity is sometime different depending on where you run the command. Keep that in mind if you see certain errors.

Enhance the script to execute the above query and answer Submission 2.

[SUBMISSION 2: submit the number of rows returned by the select on ebay entries.](#)

It should print that the result has 1 cell with some value.

[SUBMISSION 3: submit the number returned in the DataFrame of the result.show\(\) command above.](#)

Another query, with a screen shot.

```
>>> results = sqlContext.sql('SELECT * FROM Web_Session_Log')  
>>> results.show()
```

See below for screenshot of what you should see.

```
>>> results = sqlContext.sql('SELECT * FROM Web_Session_Log')
>>> results.show()
```

DATE	TIME	USERID	SESSIONID	PRODUCTID	REFERERURL
2008-01-31	15:54:25	__RequestVerifica...	+.ASPXAUTH=C31HD...	/product/YJ29I0CVQI	http://www.abc.com
2005-12-08	02:36:30	__RequestVerifica...	+.ASPXAUTH=H7HTS...	/product/MVI9HHP8A	http://www.ebay.com
2015-06-07	23:27:58	__RequestVerifica...	+.ASPXAUTH=58SZL...	/search/P5XK03AC9I	http://www.abc.com
2009-03-12	03:16:27	__RequestVerifica...	+.ASPXAUTH=VBWZJ...	/product/A13025WBT	http://www.shoppe...
2014-07-23	08:36:03	__RequestVerifica...	+.ASPXAUTH=VXBLE...	/search/SPI9XD6LZ	http://www.facebo...
2002-12-30	08:42:09	__RequestVerifica...	+.ASPXAUTH=YABJB...	/product/WS80XJFW2	http://www.xyz.com
2004-11-03	20:29:10	__RequestVerifica...	+.ASPXAUTH=2F90N...	/product/OJ201IBUN	http://www.homes...
2012-01-26	12:39:57	__RequestVerifica...	+.ASPXAUTH=SEWRR...	/product/OA3QGXF1U	http://www.xyz.com
2008-04-30	02:01:34	__RequestVerifica...	+.ASPXAUTH=60B10...	/search/K1IRBE1DU	http://www.abc.com
2003-08-23	09:44:43	__RequestVerifica...	+.ASPXAUTH=1NRGS...	/product/ANGEKDMKM	http://www.shoppe...
2008-04-09	01:24:24	__RequestVerifica...	+.ASPXAUTH=2Y8NA...	/product/LC94NBS9A	http://www.facebo...
2000-08-07	06:45:19	__RequestVerifica...	+.ASPXAUTH=KS9LL...	/search/HDKNJ50RV	http://www.facebo...
2013-10-09	05:22:31	__RequestVerifica...	+.ASPXAUTH=UA1WH...	/search/5LPS3BTJI	http://www.facebo...
2006-07-31	08:12:44	__RequestVerifica...	+.ASPXAUTH=GDMV0...	/search/BW80TIDQP	http://www.xyz.com
2014-07-27	13:23:18	__RequestVerifica...	+.ASPXAUTH=OY5S5...	/search/D5S8HFH9D	http://www.facebo...
2001-01-10	18:23:03	__RequestVerifica...	+.ASPXAUTH=VMOYI...	/product/I8VLXARNQ	http://www.xyz.com
2011-09-24	21:28:13	__RequestVerifica...	+.ASPXAUTH=SDVEA...	/search/S44PIHRYX	http://www.shoppe...
2008-09-19	02:52:53	__RequestVerifica...	+.ASPXAUTH=7NEBV...	/search/CX28DBZYW	http://www.shoppe...
2006-03-01	20:10:27	__RequestVerifica...	+.ASPXAUTH=K58W1...	/product/GG8EXER8K	http://www.amazon...

only showing top 20 rows

You can also run the program as a script. Lets assume you create a script `mysql.py` which you placed in `/tmp` with the data file. The content of `mysql.py` is as follows. You may need to adjust the location of data file based on were you stored it and if you opted to have it on HDFS.

```
from pyspark import SparkContext
from pyspark.sql import SQLContext
from pyspark.sql.types import *
sc = SparkContext("local", "weblog app")
sqlContext = SQLContext(sc)
lines = sc.textFile('file:///data/labs/w205-labs-exercises/data/weblog_lab.csv')
parts = lines.map(lambda l: l.split('\t'))
Web_Session_Log = parts.map(lambda p: (p[0], p[1], p[2], p[3], p[4]))
schemaString = 'DATETIME USERID SESSIONID PRODUCTID REFERERURL'
fields = [StructField(field_name, StringType(), True) for field_name in schemaString.split()]
schema = StructType(fields)
schemaWebData = sqlContext.createDataFrame(Web_Session_Log, schema)
schemaWebData.registerTempTable('web_session_log')
results = sqlContext.sql('SELECT * FROM web_session_log')
```

```
results.show()
```

You can now run the python Spark SQL script using the command. You can use `pyspark` to run the command, but the recommended way is to use `spark-submit`.

```
$ spark-submit /tmp/mysql.py
```

The output may look something like the following screen shot.

```
jkoister-mac:lab_6 jkoister$ spark-submit /tmp/mysql.py
15/09/26 20:16:13 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using
builtin-java classes where applicable
15/09/26 20:16:13 WARN Utils: Your hostname, jkoister-mac resolves to a loopback address: 127.0.0.1; usin
g 10.0.0.13 instead (on interface en0)
15/09/26 20:16:13 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
15/09/26 20:16:15 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
15/09/26 20:16:15 WARN MetricsSystem: Using default name DAGScheduler for source because spark.app.id is
not set.
+-----+-----+-----+-----+-----+
|      DATETIME|      USERID|      SESSIONID|      PRODUCTID|      REFERERURL|
+-----+-----+-----+-----+-----+
|      date|      userid|      sessionid|      productid|      refererurl|
|2008-01-31 15:54:25|__RequestVerifica...|;+.ASPXAUTH=C31HD...|/product/YJ29I0CVQ| http://www.abc.com|
|2005-12-08 02:36:30|__RequestVerifica...|;+.ASPXAUTH=H7HTS...|/product/MVI9HHP8A| http://www.ebay.com|
|2015-06-07 23:27:58|__RequestVerifica...|;+.ASPXAUTH=58SZL...|/search/P5XK03AC9| http://www.abc.com|
|2009-03-12 03:16:27|__RequestVerifica...|;+.ASPXAUTH=VBWZJ...|/product/A13025WBT|http://www.shophe...|
|2014-07-23 08:36:03|__RequestVerifica...|;+.ASPXAUTH=VXBLE...|/search/5PI9XD6LZ|http://www.facebo...|
|2002-12-30 08:42:09|__RequestVerifica...|;+.ASPXAUTH=YABJB...|/product/WS80XJFW2| http://www.xyz.com|
|2004-11-03 20:29:10|__RequestVerifica...|;+.ASPXAUTH=2F90N...|/product/0J201IBUN|http://www.homesh...|
|2012-01-26 12:39:57|__RequestVerifica...|;+.ASPXAUTH=SEWRR...|/product/0A3QGXF1U| http://www.xyz.com|
|2008-04-30 02:01:34|__RequestVerifica...|;+.ASPXAUTH=60B10...|/search/K1IRBE1DU| http://www.abc.com|
|2003-08-23 09:44:43|__RequestVerifica...|;+.ASPXAUTH=1NRGS...|/product/ANGEKDMKM|http://www.shophe...|
|2008-04-09 01:24:24|__RequestVerifica...|;+.ASPXAUTH=2Y8NA...|/product/LC94NBS9A|http://www.facebo...|
|2000-08-07 06:45:19|__RequestVerifica...|;+.ASPXAUTH=KS9LL...|/search/HDKWJ50RV|http://www.facebo...|
|2013-10-09 05:22:31|__RequestVerifica...|;+.ASPXAUTH=UA1WH...|/search/5LPS3BTJI|http://www.facebo...|
|2006-07-31 08:12:44|__RequestVerifica...|;+.ASPXAUTH=GDVM0...|/search/BW80TIDQP| http://www.xyz.com|
|2014-07-27 13:23:18|__RequestVerifica...|;+.ASPXAUTH=0YSS5...|/search/D5S8HFH9D|http://www.facebo...|
|2001-01-10 18:23:03|__RequestVerifica...|;+.ASPXAUTH=VMOYI...|/product/I8VLXARNQ| http://www.xyz.com|
|2011-09-24 21:28:13|__RequestVerifica...|;+.ASPXAUTH=SDVEA...|/search/S44PIHRYX|http://www.shophe...|
|2008-09-19 02:52:53|__RequestVerifica...|;+.ASPXAUTH=7NEBV...|/search/CX28DBZYW|http://www.shophe...|
|2006-03-01 20:10:27|__RequestVerifica...|;+.ASPXAUTH=K58W1...|/product/GG8EXER8K|http://www.amazon...|
+-----+-----+-----+-----+-----+
only showing top 20 rows
```

## What you should have learnt

You should have learnt how to create a Python script that uses Spark SQL and how to run the script.

## Step-7. Caching tables and Un-caching tables

### Caching tables

This is extremely useful when you are joining tiny dataset with huge dataset.



CACHE TABLE and UNCACHE TABLE statements are available to do the above making it very easy.

To Cache a table:

```
CACHE TABLE logs_last_month;
```

To UnCache a table:

```
UNCACHE TABLE logs_last_month;
```

Once as table is cached, you can use in your spark queries.

## Submissions

There are three items that needs to be submitted from this lab to be approved.

In Step 3:

SUBMISSION 1: submit the first 10 rows of he unsorted and the sorted RDD to show that you successfully created both. Also explain the issue with the tuple and how to possibly correct it in the map/lambda function that was used to create the tuples.

In Step 5:

SUBMISSION 2: submit the number of rows returned by the select on ebay entries.

In Step 6:

SUBMISSION 3: submit the number return in he DataFrame of the result.show() command above.

## Troubleshooting

### Connection problems on Laptop or Macbook.

If you get an exception looking like "ERROR SparkContext: Error initializing SparkContext. java.net.UnknownHostException:..." make sure you have the make of our computer added to the /etc/hosts file.

The remedy is to add the line "127.0.0.1 <myhost>", where <myhost> is the name of our computer.

### Hadoop connection problem on EC2 instance.

If you get an error that look something like on the AMI:

```
py4j.protocol.Py4JJavaError: An error occurred while calling
o34.collect.
: java.net.ConnectException: Call From ip-10-61-206-
219.ec2.internal/10.61.206.219 to localhost:8020 failed on connection
exception: java.net.ConnectException: Connection refused; For more details
see: http://wiki.apache.org/hadoop/ConnectionRefused
```

Your hadoop instance is likely not running. With the “**ucb\_w205\_complete** - ami-71cdb014” use the `start-hadoop.sh` and `stop-hadoop.sh` scripts to start and stop the Hadoop service.

## AMI

If you are using the AMI you need to attach an EBS volume, create the file system and mount the file system.

Just follow these instructions on how to create a volume and how to attach it to your instance:

<http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ebs-attaching-volume.html>

When you create a Volume you give it a name, I used `/dev/sdh` (h for hadoop) just so that I remember it. When you check volumes on your EC2 instance you will see different name. In the example below my `sdh` got named `xvdh`. I recognized by the h and the size.

```
[root@ip-10-61-206-219 ~]# lsblk
NAME MAJ:MIN RM  SIZE RO TYPE MOUNTPOINT
xvda1 202:1    0   10G  0 disk /
xvdb  202:16   0    4G  0 disk
xvdh  202:112  0  100G  0 disk /data
```

Now you have an attached volume. At this point you need to create a file system on that raw disk volume and then mount the volume on your root file system so that it can be accessed by programs. A file system is like a tree of trees, and mounting essentially means you attach another tree to a branch of another tree. File systems are also a structure of information on your disk that tracks files, blocks of files, who owns the files etc. That is why you need to create the file system.

The guide below explains how to create and mount the file system. The only problem I noticed with the guide was that it left out that you should provide the type of file system to mount. If the example in the guide fails for you, try this:

```
sudo mount -t ext4 /dev/xvdh /data
```

<http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ebs-using-volumes.html>

## Error at ‘\t’

Sometimes cutting and pasting changes the representation of characters. Edit the command in `spark-sql` and make sure you have “ quotes.

## Local Spark Installation for spark-sql CLI

You create a new installation of Spark using the below instructions you will be able to run the CLI. The following is for the AMI, but you can do the corresponding steps on your own computer or in another environment.

```
wget http://www.us.apache.org/dist/spark/spark-1.5.0/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
```

I would also propose changing Log tracing to WARN in the conf/log4j file.  
If you check which spark-sql you use you should see the following.

```
# which spark-sql
/root/spark15/bin/spark-sql
```

You need to make sure that

Hadoop is started. Then start spark-sql and you should be able to do step-4 and step-5 of this lab.