

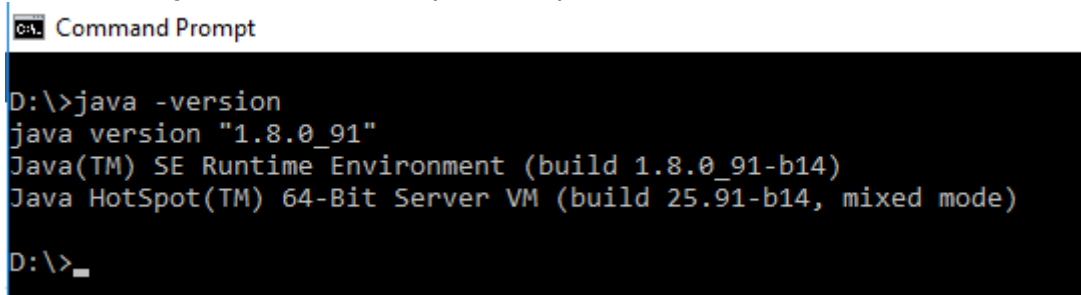
## CSC 3355

### Database Management

#### Instruction for installing Spark, Python in Windows

You need to have Administrator rights on your laptop/desktop where you are trying to install this softwares. All the following commands must be executed in a command-line window (*cmd*) ran as Administrator, i.e. using **Run as administrator** option while executing *cmd*. You can find *cmd* by browsing *All Programs -> Accessories -> Command Prompt*.

1. Look for the java installation in your computer.



```
Command Prompt
D:\>java -version
java version "1.8.0_91"
Java(TM) SE Runtime Environment (build 1.8.0_91-b14)
Java HotSpot(TM) 64-Bit Server VM (build 25.91-b14, mixed mode)
D:\>_
```

2. If the java version returns anything else than 1.9....., download and install Java 8 and place it in C:\Java. **Do not install JDK 9**; Spark is currently incompatible with JDK 9. If you need to download the JDK, please visit Oracle's download site: <http://www.oracle.com/technetwork/java/javase/downloads/index.html>. Please remember that while installing, you need to change the default folder to C:\java. (there is a problem in windows with setting PATH with folder names that has spaces in them, therefore the default folder "Program Files" won't work with Spark)
3. Download and install Spark 2.2.1 on your machine:  
<https://www.apache.org/dyn/closer.lua/spark/spark-2.2.1/spark-2.2.1-bin-hadoop2.7.tgz>
  - a. Use windows 7 zip to untar, place it in D:\spark-2.2.1-bin-hadoop2.7
4. Set environment variable (for Java)
  - a. System variable JAVA\_HOME = C:\ Java
  - b. Edit system variable PATH = C:\Java\bin
5. Set environment variable (for Spark)
  - a. System variable SPARK\_HOME = D:\spark-2.2.1-bin-hadoop2.7
  - b. Edit system variable PATH = D:\spark-2.2.1-bin-hadoop2.7\bin
6. Set System variable HADOOP\_HOME = D:\spark-2.2.1-bin-hadoop2.7 and download winutils.exe from (<https://github.com/steveloughran/winutils/blob/master/hadoop-2.7.1/bin/winutils.exe>) and copy it in D:\spark-2.2.1-bin-hadoop2.7\bin
7. Install Anaconda with python 2.7 from <https://www.anaconda.com/download/> and install it.
8. Once installed search for "anaconda prompt" and type "conda list" or jupyter notebook" to check the installation.

- In Anaconda prompt(looks like cmd) type "where conda", "where python" to get the folders where they are installed add these locations to system PATH variable.
- Open a fresh cmd, and type pyspark, you should see something as follows

```

Administrator: Command Prompt - pyspark
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

H:\>pyspark
Python 2.7.14 [Anaconda, Inc.] <default, Nov  8 2017, 13:40:45> [MSC v.1500 64 b
it <AMD64>] on win32
Type "help", "copyright", "credits" or "license" for more information.
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel
(newLevel).
18/04/10 10:49:24 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
18/04/10 10:49:36 WARN ObjectStore: Failed to get database global_temp, returnin
g NoSuchObjectException
Welcome to

          Spark
    version 2.2.1

Using Python version 2.7.14 <default, Nov  8 2017 13:40:45>
SparkSession available as 'spark'.
>>>

```

- Setup jupyter by adding the two below system variables

PYSPARK_DRIVER_PYTHON	jupyter
PYSPARK_DRIVER_PYTHON_OPTS	notebook

- Open a fresh cmd, and type pyspark, you should now see the Jupyter notebook running in your browser.
- Try the following code in jupyter notebook. You need to scroll to the correct folder where you download the attached employee.json input datafile and then start a python notebook and paste the content of employee.py in a cell. Make sure to open employee.py with notepad. If you see the following output, you are good with necessary installation.

```

Jupyter Untitled16 Last Checkpoint: 3 minutes ago (autosaved)
File Edit View Insert Cell Kernel Widgets Help Trusted Python 2.0
In [1]:
from pyspark.sql import SparkSession

df = spark.read.json("employee.json")
# Displays the content of the DataFrame to stdout
df.show()

+-----+
|age| name|salary|
+-----+
| 29|Michael| 3000|
| 25| Andy | 4500|
| 34| Justin| 3500|
| 36| Bertal| 4000|
+-----+

```