Spark Resources

This page contains links to various Spark resources.

- Spark overview talk (somewhat outdated, from 2012)
    - https://spark.apache.org/talks/overview.pdf or https://spark.apache.org/talks/overview.pptx
    - And a newer and more detailed talk covering internals also (not required for you): http://www.slideshare.net/AGrishchenko/apache-spark-architecture
- Spark quick start: creating applications:
    - Read: http://spark.apache.org/docs/latest/quick-start.html#self-contained-applications
    - The page also has stuff on the Spark Shell, which we have skipped since it only works with Scala/Python.
    - Note: if you do this on your own machine, you should install Java 8 and Maven
- Spark programming guide: http://spark.apache.org/docs/latest/programming-guide.html
    - The guide has information for 3 languages: Scala, Java and R. Choose the Java tab in all cases, unless you wish to use Scala.
- Spark Java API docs: http://spark.apache.org/docs/latest/api/java/index.html
- Some Spark/Java8 examples
    - http://blog.cloudera.com/blog/2014/04/making-apache-spark-easier-to-use-in-java-with-java-8/
    - https://github.com/ypriverol/spark-java8

Setting up Spark on software lab machines

1. Create a new eclipse Java project
2. Import all the jars in the Spark jars folder into eclipse (select all the jar files) as follows:
    - Right click on the project and select: Properties > Build Path > Libraries : Add External Jars
    - Browse to the following folder and select all the jars in it
      ~sudarsha/spark-2.0.0-bin-hadoop2.7/jars
3. Right click on project and select:
    - Run As > Run Configurations > Java Application > New_configuration
    - then choose the JRE tab, click on the Alternate JRE button, and then select java 8. If it's not present,
    - then choose Add, and add /usr/lib/jvm/java-8-openjdk-amd64
    - Make sure to check the box for java-8-openjdk so it gets used for compilation.
    - Then go back to your project Run As > Run Configurations and make sure to choose New Configuration for it.
4. Go to Run Configurations, and go to Classpath tab
    - Choose Advanced > JRE System Library and click on Next
    - Then choose java-8-openjdk
5. Create your required Java files
6. Build them; but don't run (you can click on run, but it will give error messages and not actually run)
7. Export to a jar file with any name you choose The jar file gets created in the workspace folder of eclipse.
    - NOTE: you must export each time you update a file
8. Now run spark-submit from the command line:
    - export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64
    - ~sudarsha/spark-2.0.0-bin-hadoop2.7/bin/spark-submit --class SimpleApp --master local[4] ~/workspace2/simple-project-1.0.jar

      WHERE SimpleApp is the class you want to run, and simple-project-1.0.jar is the jar file you created when you exported to the jar file
    - Some of the Spark sample files require an input file. Make sure to create it in an appropriate directory (such as the one where you run the spark-submit commandpwd from)
    - Note that the JAVA_HOME above can be set from your .bashrc, so you don't need to do it each time