Mark as done

1. You can do your assignment in Python or Java. For python, you can install PySpark as below
    1. `pip install pyspark`
2. For Java, first ensure you have a recent java installation. For linux you can download the following https://builds.openlogic.com/downloadJDK/openlogic-openjdk/21.0.4+7/openlogic-openjdk-21.0.4+7-linux-x64-deb.deb

   and then run

   ```
   sudo apt install ~/Downloads/openlogic-openjdk-21.0.4+7-linux-x64-deb.deb
   ```

   to install java 21.
3. Download spark-hadoop jars using the following link and expand to a folder: https://www.apache.org/dyn/closer.lua/spark/spark-3.5.2/spark-3.5.2-bin-hadoop3.tgz
4. Setup instructions for Eclipse
    - Create a new eclipse Java project
    - 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 folder with hadoop jars and select all the jars in it
    - 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 or later version of Java.
            - Make sure to check the box for java-21-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.
            - Go to Run Configurations, and go to Classpath tab
            - Choose Advanced > JRE System Library and click on Next
            - Then choose java-21-openjdk

1. Create your required Java files and build them
2. You can run your spark program as follows:
    1. 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
    2. You can run spark-submit from the command line or run the same command from your IDE:
       export JAVA_HOME=/usr/lib/jvm/java-21-openjdk-amd64
        - Note that the JAVA_HOME above can be set from your .bashrc, so you don't need to do it each time
        - Run
          spark-3.5.2-bin-hadoop3/bin/spark-submit --class WordCount --master local[4] ~/workspace2/simple-project-1.0.jar

- WHERE WordCount 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
- **NOTE:** *Depending on how you created the jar file, you may need to add a folder path to SimpleApp in the command above. For example, if your project is lab6, you may need to use lab6.SimpleApp*

- Some of the Spark sample files require an input file. Preferably give a full path, or put it in the directory from where you run the spark-submit command
- If your program has any output files, it will put them in a directory. Make sure that the directory does not exist, by deleting it between each run. You can set options to overwrite but by default it will give an error.

Last modified: Sunday, 22 September 2024, 1:36 PM