Submission by:

Shubham Gupta (202318052)

Big Data Processing

(Assignment 4)

Title: Implementation of PySpark RDD and Dataframe using given dataset.

Objective:

• To configure PySpark in Windows

• learn about PySpark RDD and DataFrame API and how to use them to manipulate the data.

Step 1: Install Java 8

Start > type cmd> click Command Prompt. Check java -version.

If you don't have Java installed:
1. Open a browser window, and navigate to https://java.com/en/download/.
2. Click the **Java Download** button and save the file to a location of your choice.
3. Once the download finishes double-click the file to install Java.

Step 2: Check Python -version.

In my case its already installed with anaconda.

Step 3: Download Apache Spark

1. Open a browser and navigate to https://spark.apache.org/downloads.html.
2. Under the ***Download Apache Spark*** heading, there are two drop-down menus. Use the current non-preview version.

- *Choose a Spark release* -> 3.5.0
- *Choose a package type* -> Pre-built for Apache Hadoop 3.

3. Click the *spark-3.5.0-bin-hadoop3.tgz* link

4. A page with a list of mirrors loads where you can see different servers to download from. Pick any from the list and save the file to your Downloads folder.

Step 4: Verify Spark Software File

1. Verify the integrity of your download by checking the checksum of the file. This ensures you are working with unaltered, uncorrupted software.
2. Navigate back to the *Spark Download* page and open the **Checksum** link, preferably in a new tab.
3. Next, open a command line and enter the following command:

```
certutil -hashfile C:\Users\<username>\Downloads\spark-3.5.0-bin-hadoop3.tgz SHA512
```

4. Change the username to your username. The system displays a long alphanumeric code, along with the message "Certutil: -hashfile completed successfully".
5. Compare the code to the one you opened in a new browser tab. If they match, your download file is uncorrupted.

Step 5: Install Apache Spark

Installing Apache Spark involves extracting the downloaded file to the desired location.

1. Create a new folder named Spark in the root of your C: drive. From a command line, enter the following:

   ```
   cd \
   mkdir Spark
   ```

2. In Explorer, locate the Spark file you downloaded.
3. Right-click the file and extract it to *C:\Spark* using the tool you have on your system (e.g., 7-Zip).
4. Now, your *C:\Spark* folder has a new folder *spark-3.5.0-bin-hadoop3* with the necessary files inside.

Step 6: Add winutils.exe File

Download the winutils.exe file for the underlying Hadoop version for the Spark installation you downloaded.

1. Navigate to this URL *https://github.com/cdarlint/winutils* and inside the bin folder, locate winutils.exe, and click it.

2. Find the Download button on the right side to download the file.

3. Now, create new folders Hadoop and bin on *C:* using Windows Explorer or the Command Prompt.

4. Copy the winutils.exe file from the Downloads folder to *C:\Hadoop\bin*

Step 7: Configure Environment Variables

Configuring environment variables in Windows adds the Spark and Hadoop locations to your system PATH. It allows you to run the Spark shell directly from a command prompt window.

1. Click Start and type environment.

2. Select the result labelled Edit the system environment variables.

3. A System Properties dialog box appears. In the lower-right corner, click Environment Variables and then click New in the next window.

4. For Variable Name type *SPARK_HOME*.

5. For Variable Value type *C:\Spark\spark-3.5.0-bin-hadoop3* and click OK. If you changed the folder path, use that one instead.

6. In the top box, click the Path entry, then click Edit. Be careful with editing the system path. Avoid deleting any entries already on the list.

7. You should see a box with entries on the left. On the right, click New.

8. The system highlights a new line. Enter the path to the Spark folder *C:\Spark\spark-3.5.0-bin-hadoop3\bin*. We recommend using *%SPARK_HOME%\bin* to avoid possible issues with the path.

9. Repeat this process for Hadoop and Java.

For Hadoop, the variable name is *HADOOP_HOME* and for the value use the path of the folder you created earlier: *C:\Hadoop. Add C:\Hadoop\bin* to the Path variable field, but we recommend using *%HADOOP_HOME%\bin*.

For Java, the variable name is JAVA_HOME and for the value use the path to your Java JDK directory (example, C:\Program Files\Java\<jdk_version>).

10. Click OK to close all open windows.

Step 8: Launch Spark

1. Open a new command prompt Window using the right-click and Run as administrator:

2. To start Spark, enter:

C:\Spark\spark-3.5.0-bin-hadoop3\bin\spark-shell

If you set the environment path correctly, you can type spark-shell to launch Spark.

3. The system should display several lines indicating the status of the application. You may get a Java pop-up. Select Allow access to continue.

Finally, the Spark logo appears, and the prompt displays the Scala shell.

4., Open a web browser and navigate to *http://localhost:4040/*.

5. You can replace localhost with the name of your system.

6. You should see an Apache Spark shell Web UI. The example below shows the Executors page.

7. To exit Spark and close the Scala shell, press ctrl-d in the command-prompt window or Exit using quit()

Open a command prompt Window using the right-click and Run as administrator, If you installed Python, you can run Spark using Python with this command:

C:\Spark\spark-3.5.0-bin-hadoop3\bin\pyspark

Thus Spark is successfully configured with windows.