**ASSESSMENT-16 Vinutha s**

21/12/2023

**Installation of PySpark (Windows)**

Apache Spark is a new and open-source framework used in the big data industry for real-time processing and batch processing. It supports different languages, like Python, Scala, Java, and R.

To set up pyspark I installed four major things they are Java, Python, Spark, Hadoop.

1.Java installation

* Gone to **Download Java JDK**.  
  Visited Oracle's website to download the Java Development Kit(JDK).
* Set up the installation process of JDK and checked in command prompt whether it is installed or not.

2. Similarly all the remaining three Python, Spark, Hadoop are installed and runned in command prompt.

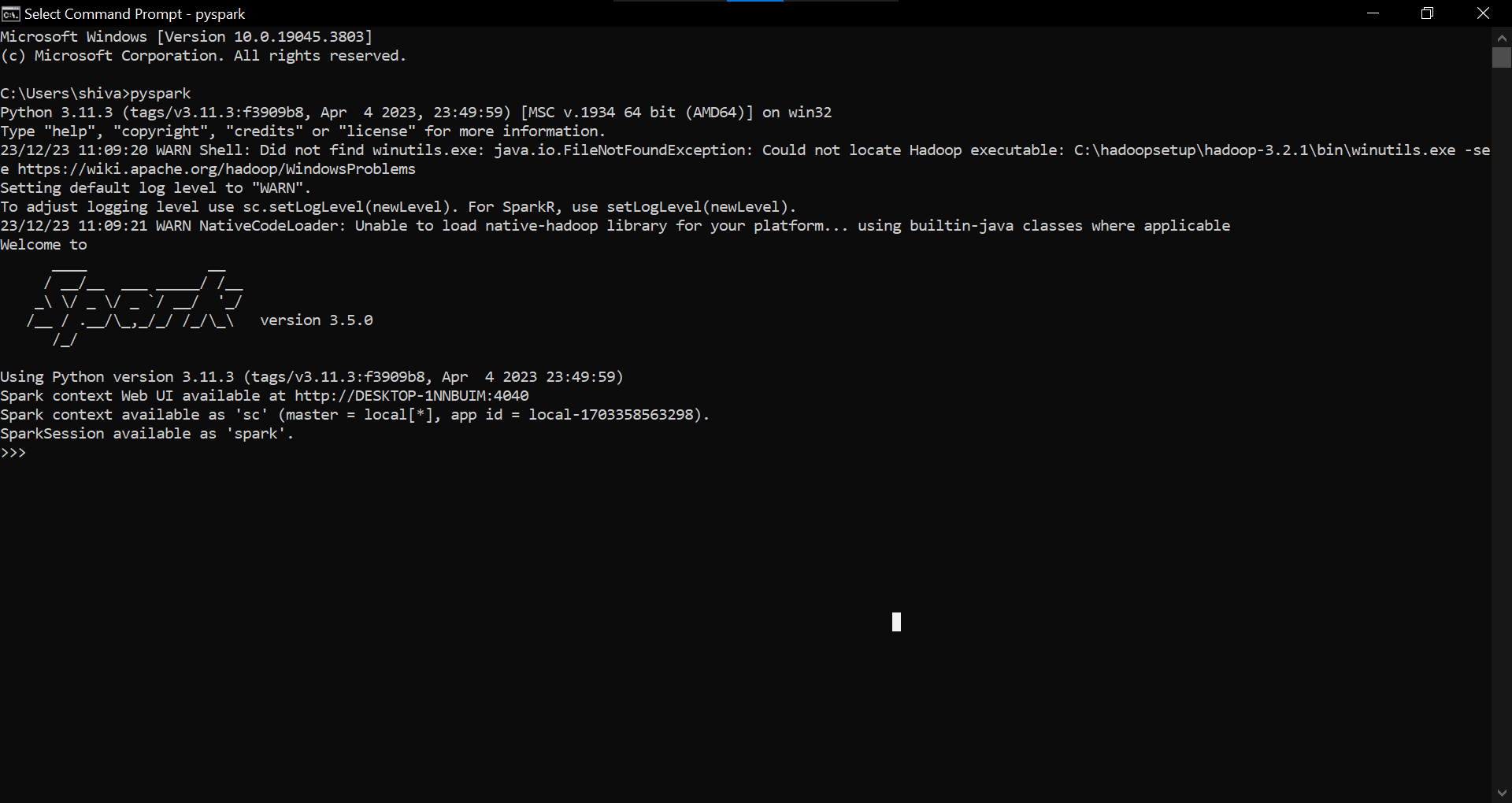
**Environment variables**

* created a new environment where variable name as "JAVA\_home" and variable value to be the location of JDK-19.
* created a new environment where variable name as "PYTHON\_home" and variable value to be the location of python 3.11.
* created a new environment where variable name as "SPARK\_home" and variable value to be the location of spark.
* created a new environment where variable name as "hadoop\_home" and variable value to be the location of winutils.

**Finalizing Pyspark Installation**

Opened Command Prompt and typed the following command “Pyspark”.

Finally I successfully completed the installation process of pyspark.

****