G. Rupa Manasa 21-12-2023

Assessment – 16

**Installation and Setup of PySpark:**

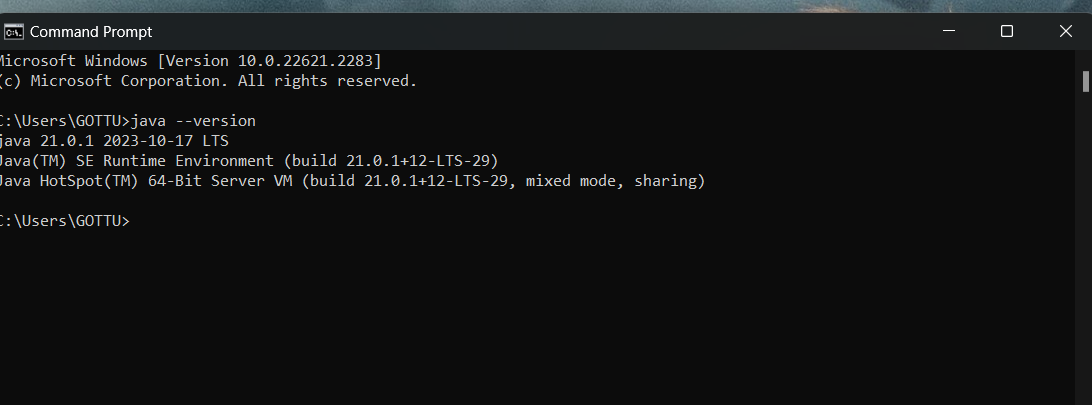
Pyspark = Python + Apache Spark

Apache Spark is a new and open-source framework used in the big data industry for real-time processing and batch processing.

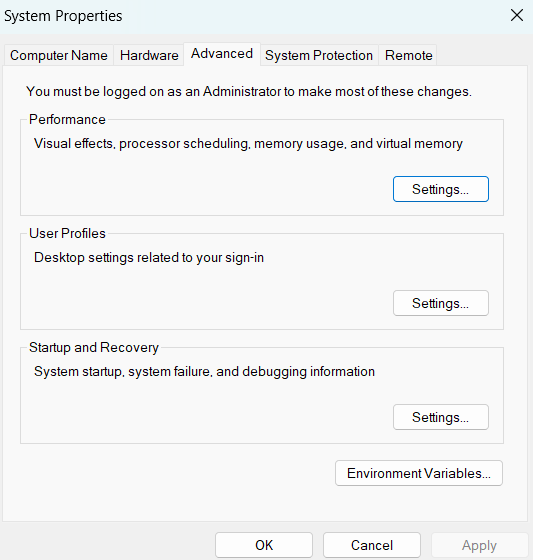
**Java Installation:**

Apache Spark requires Java 8+ to be installed. You can download and install Java from the official Oracle website or use OpenJDK.

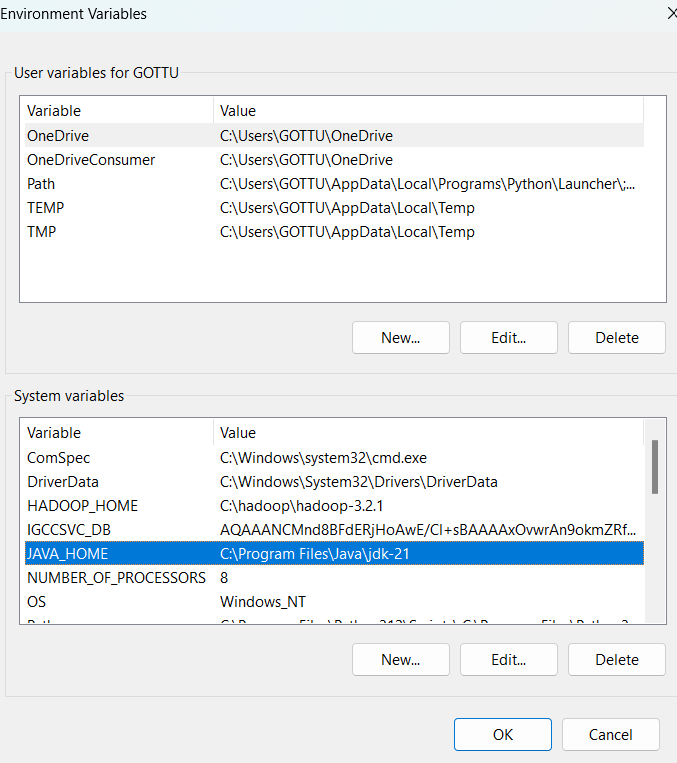
After installing, you can check in command prompt which version is installed.



After that, we need to provide the environment variables. For that go to the search bar and “EDIT THE ENVIRONMENT VARIABLES”. Click on environment variables option.

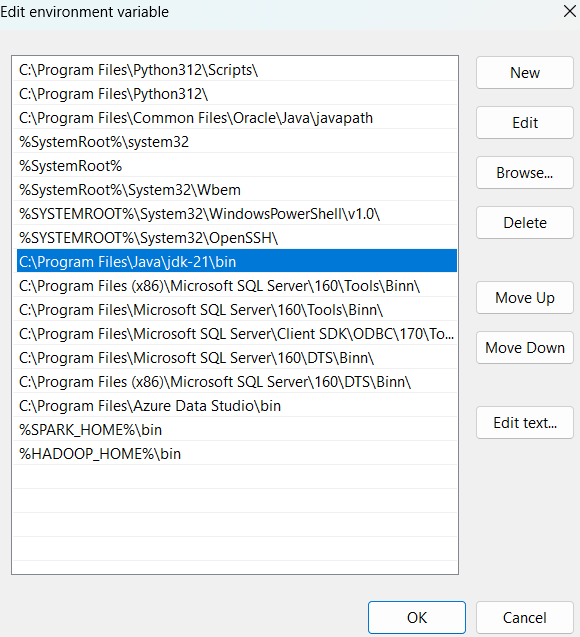


In the System Variables, Use Variable Name as "JAVA\_HOME' and your Variable Value as location of your java folder.



Then edit the path in System Variables, give the location of the folder.

Then click ok.



**Python Installation:**

Visit the python official website, go to downloads and download the python latest version then install it.

Open the command prompt, give the command “python –version”, to check whether the python is installed or not.

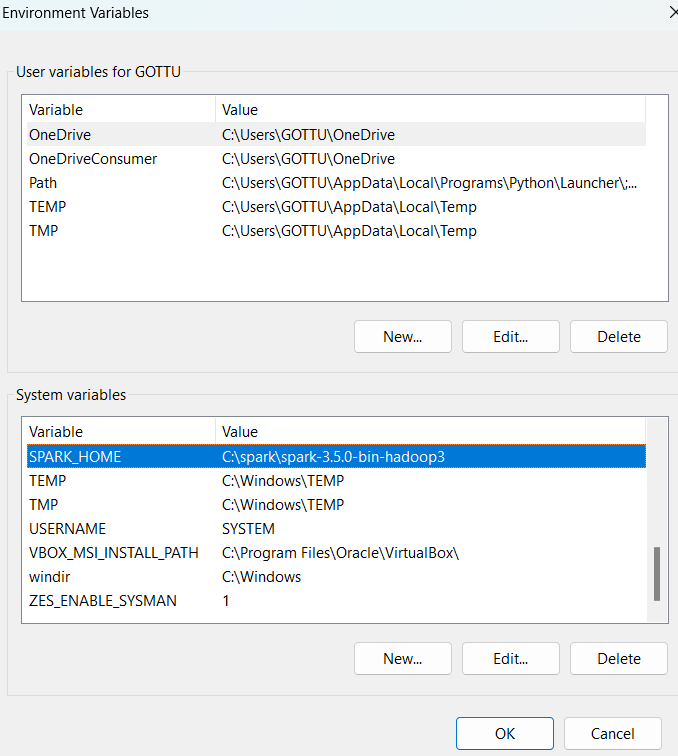
Set the Environment variables as well.

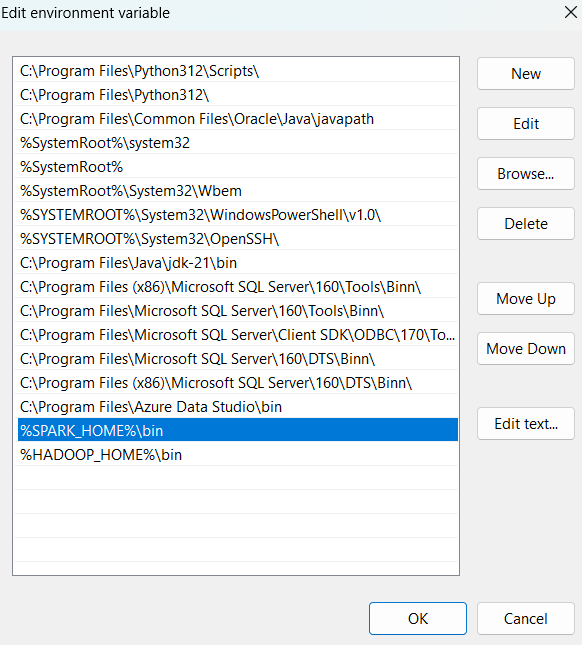
**Apache Spark Installation:**

Download Spark: Go to the official Apache Spark website and navigate to the "Downloads" section. Choose the latest version and download the pre-built binary.

Extract Spark: Extract the downloaded Spark archive to a directory of your choice.

Set the Environment variables as well.





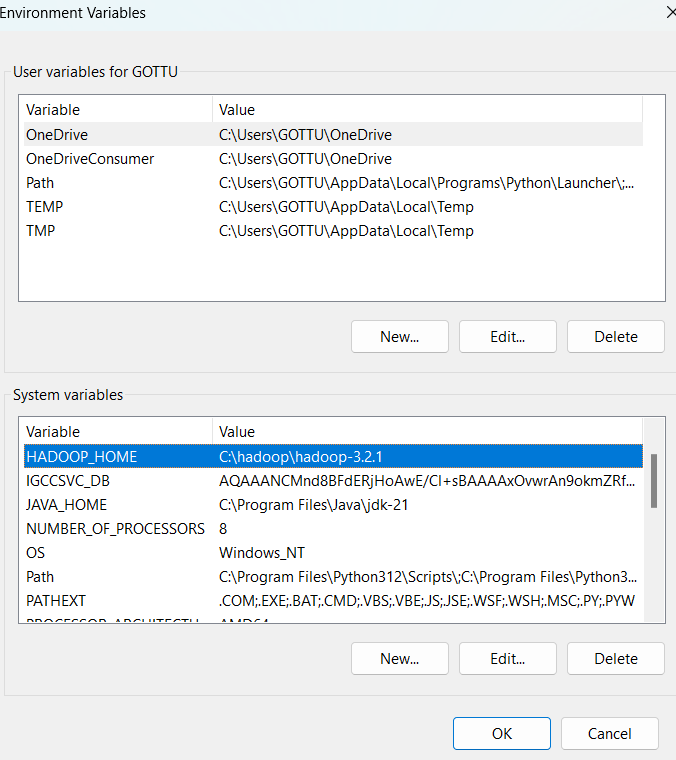
**Hadoop Installation:**

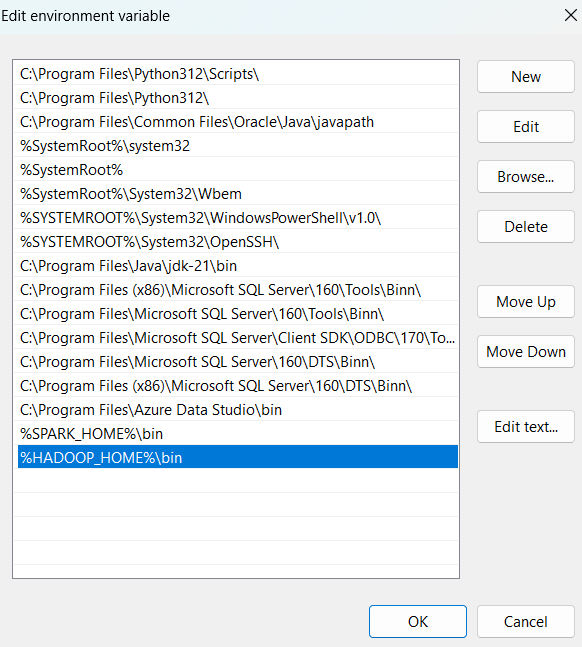
Download Hadoop: Visit the official Apache Hadoop website and navigate to the "Downloads" section.

Download the binary distribution of Hadoop. Choose the version that fits your requirements.

Extract Hadoop: Extract the downloaded Hadoop archive to a directory of your choice

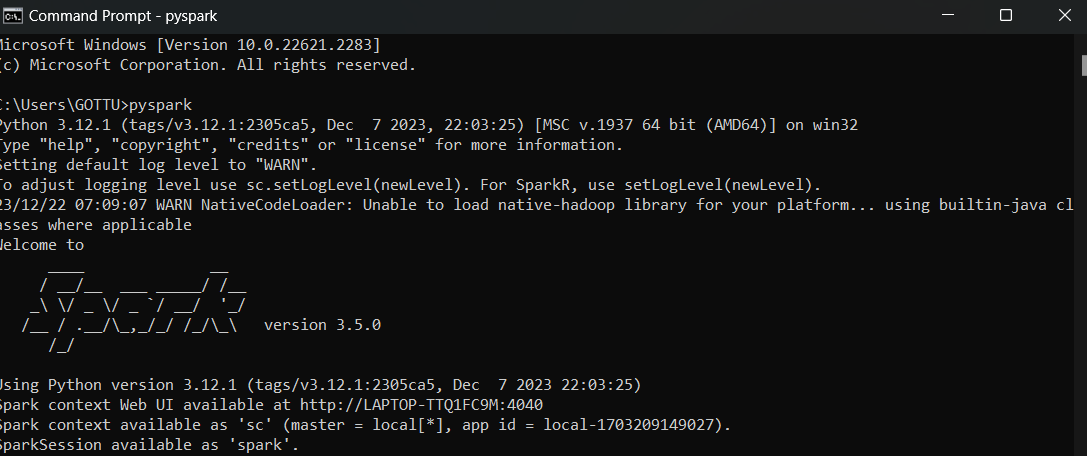
Set the environment variables as well.





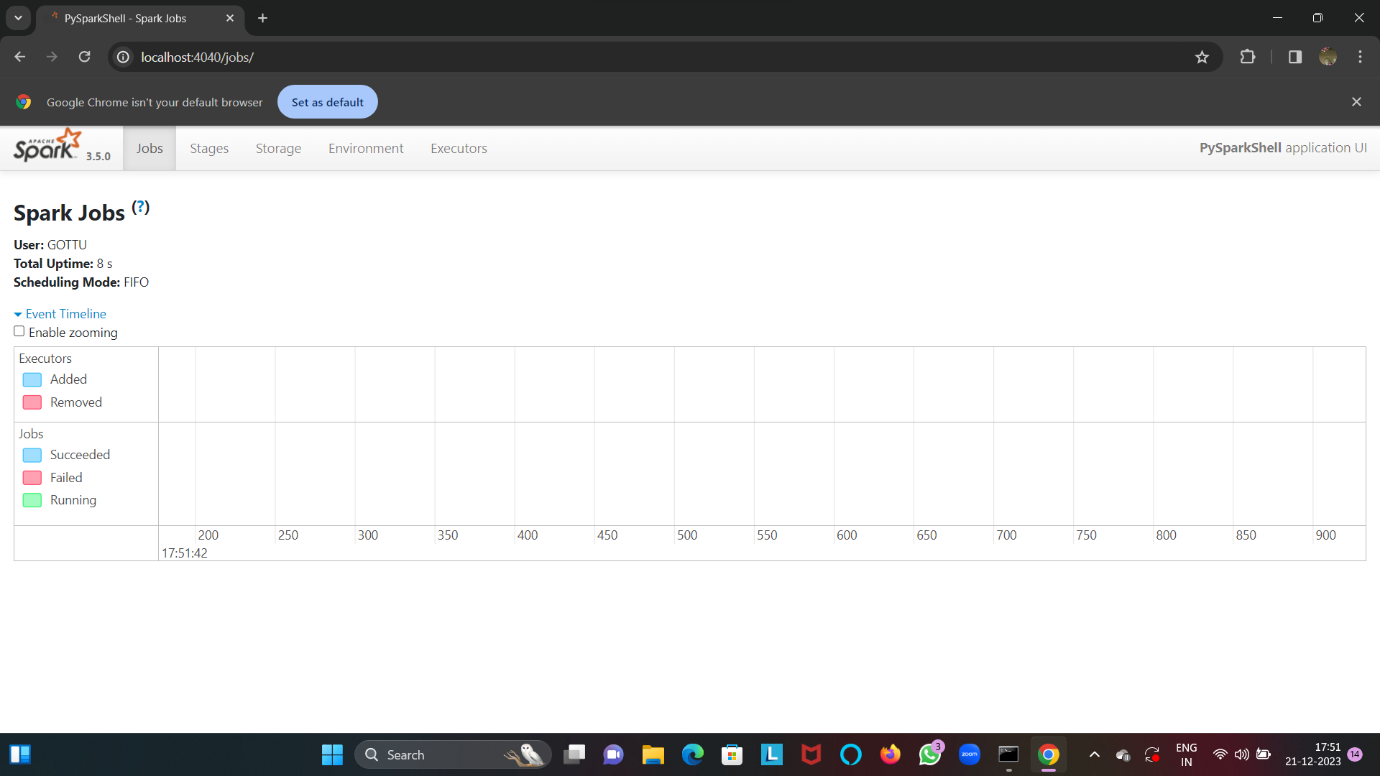
**Checking the PySpark is Installed or not:**

Open the command prompt and give the command “pyspark”.



**Check for web UI:**

Give “http//localhost:4040/jobs/” in web browser.

****