**Assignment(21-12-2023)**

**Installation of Apache Spark and Setup…**

**Apache PySpark..**

Apache Spark is an open-source distributed computing system that provides a fast and

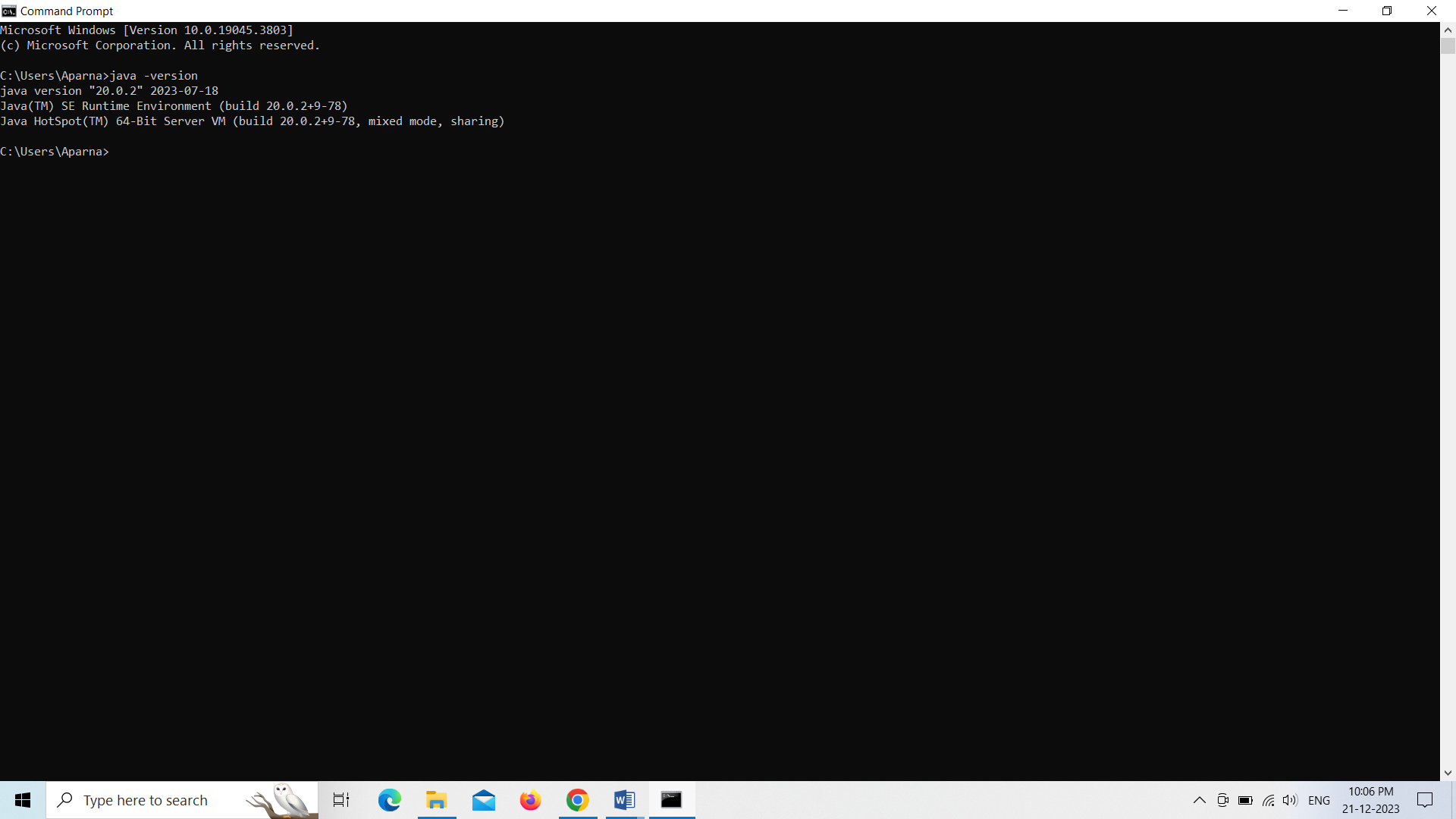
general-purpose cluster-computing framework for big data processing. PySpark is the Python

API for Apache Spark, allowing you to write Spark applications using Python.

**Python+Spark=PySpark..**

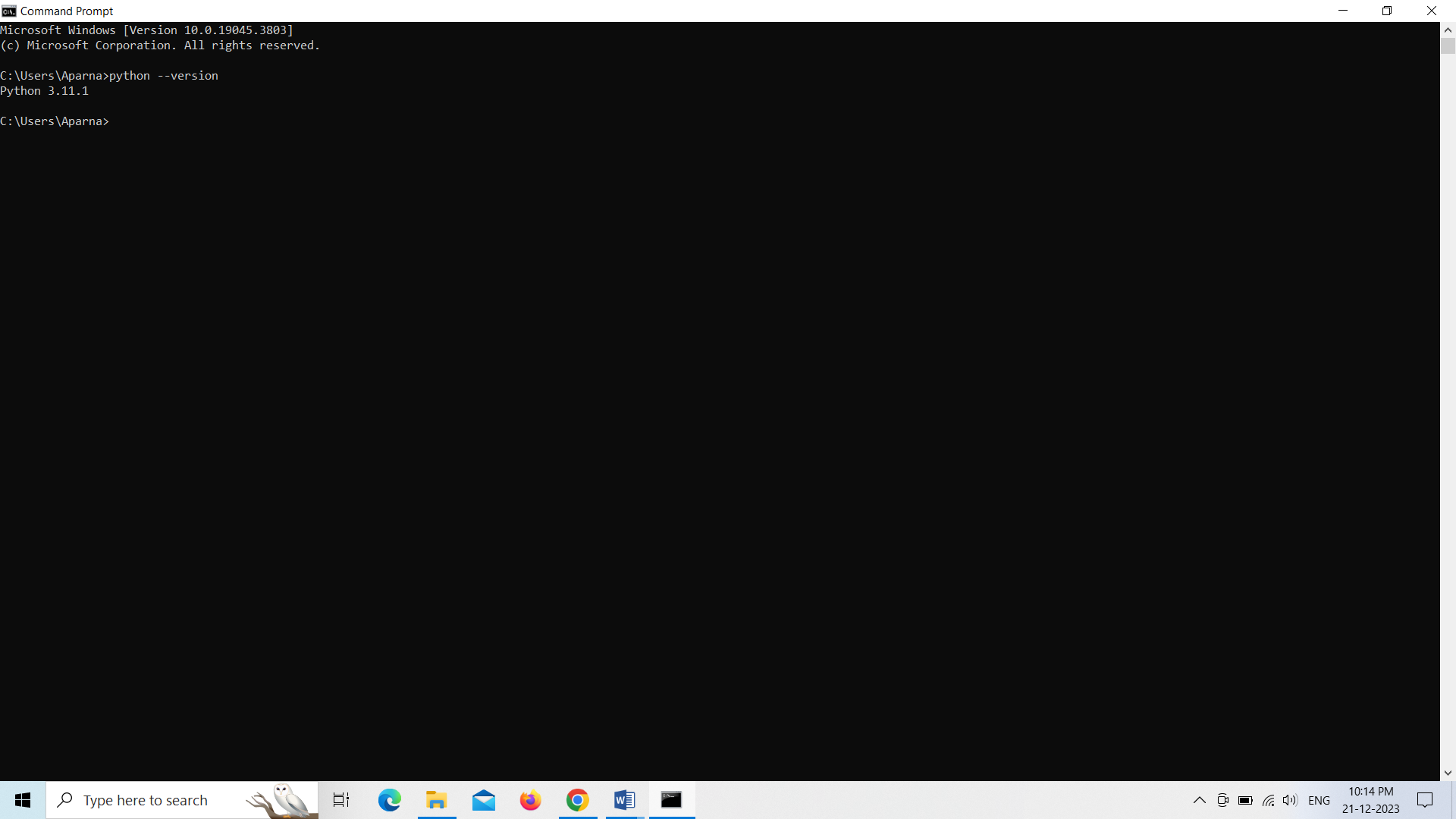
**Install Java**

* Firstly, we have to install java jdk version which is compatable to your system. Its good to download and install latest standard version of java**.**
* After that we have to check in command prompt by typing **java –version** which displays the version of java which you downloaded.
* In means that you have successfully downloaded and installed in your system.



**Install Python**

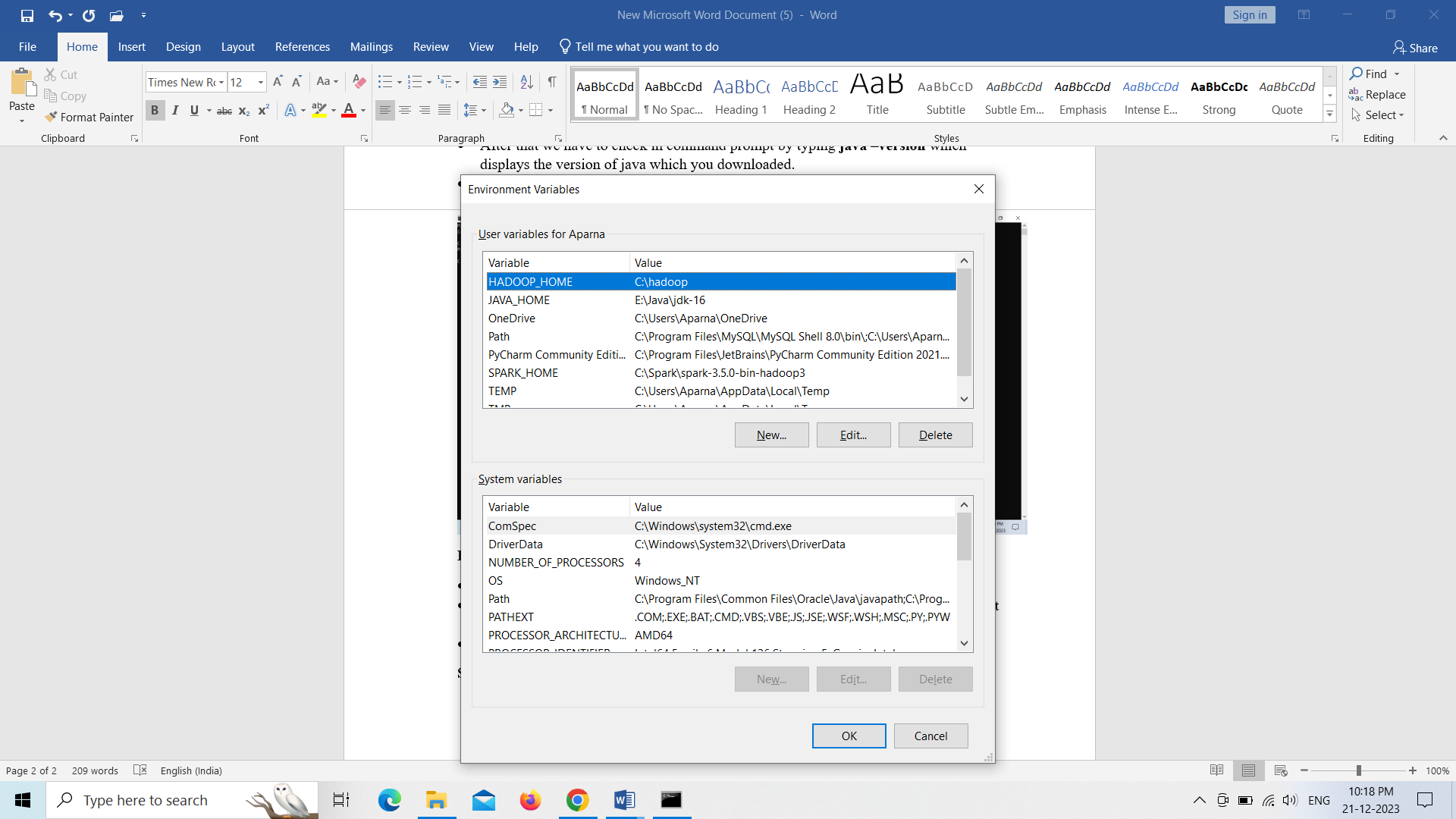
* Along with java, we have to install python environment into our system. Its good to download and install latest and standard version of python.
* After that we have to check in command prompt by typing **java –version** which displays the version of java which you downloaded.
* In means that you have successfully downloaded and installed in your system.

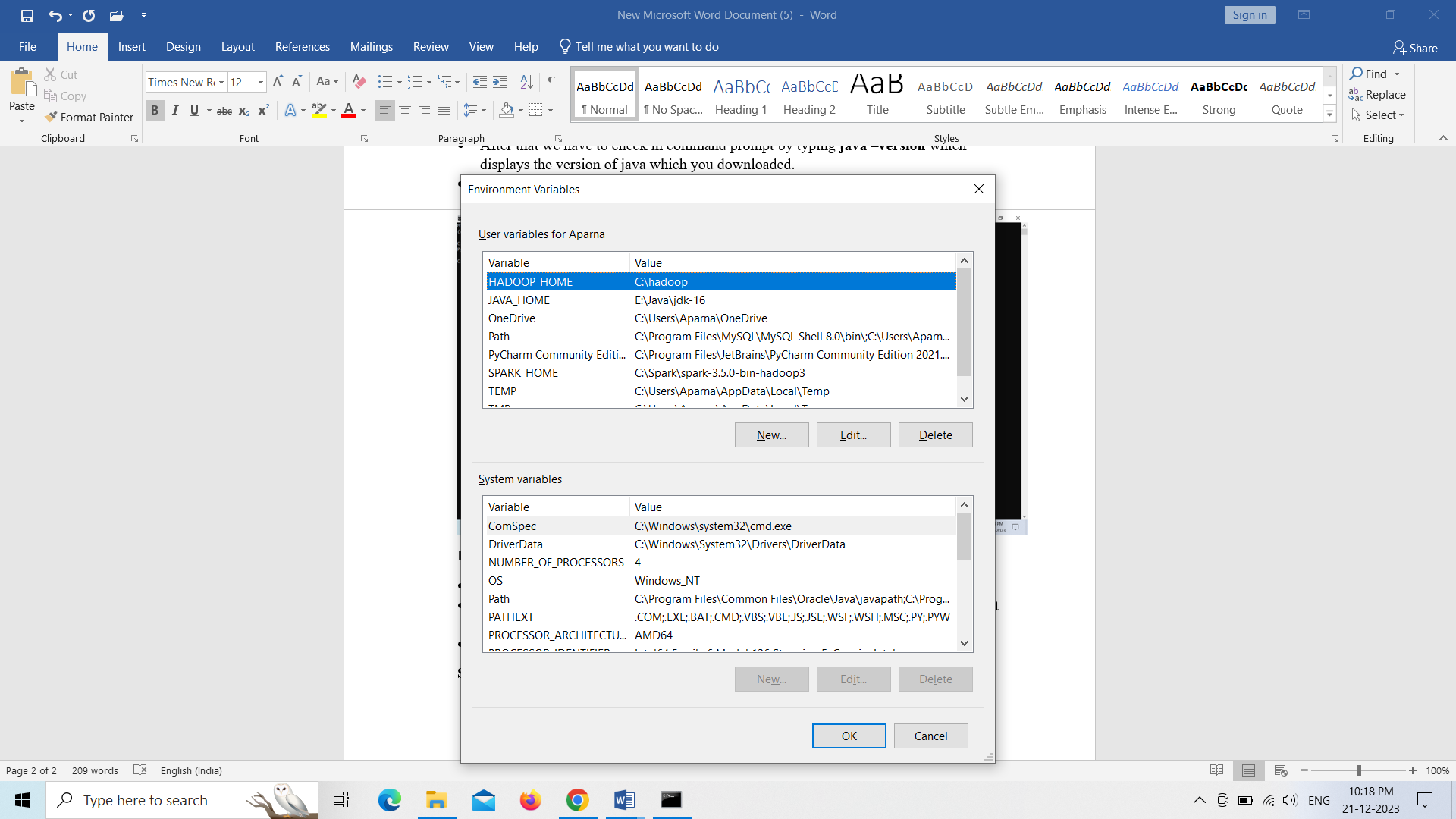


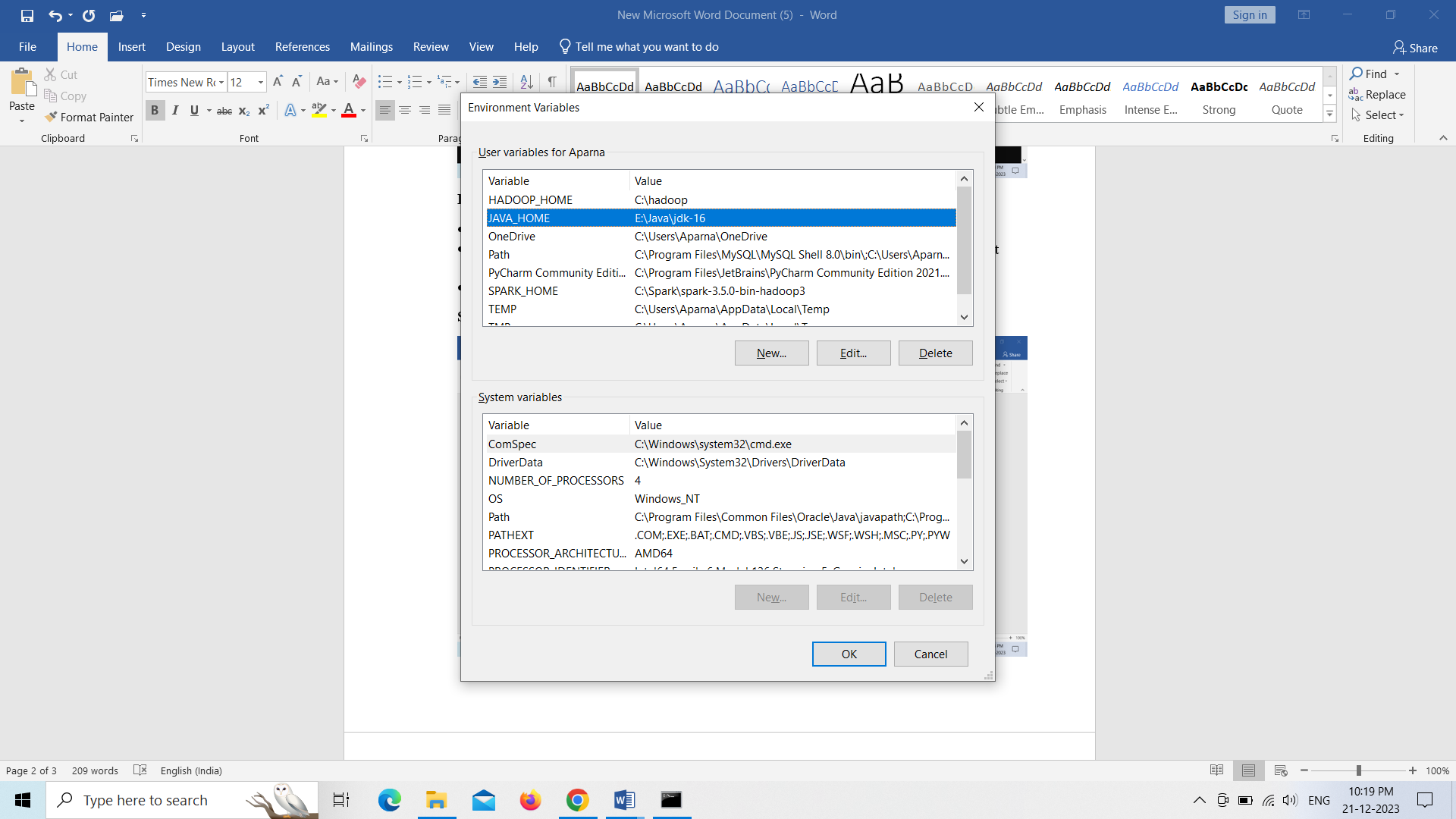
**Install Apache Spark**

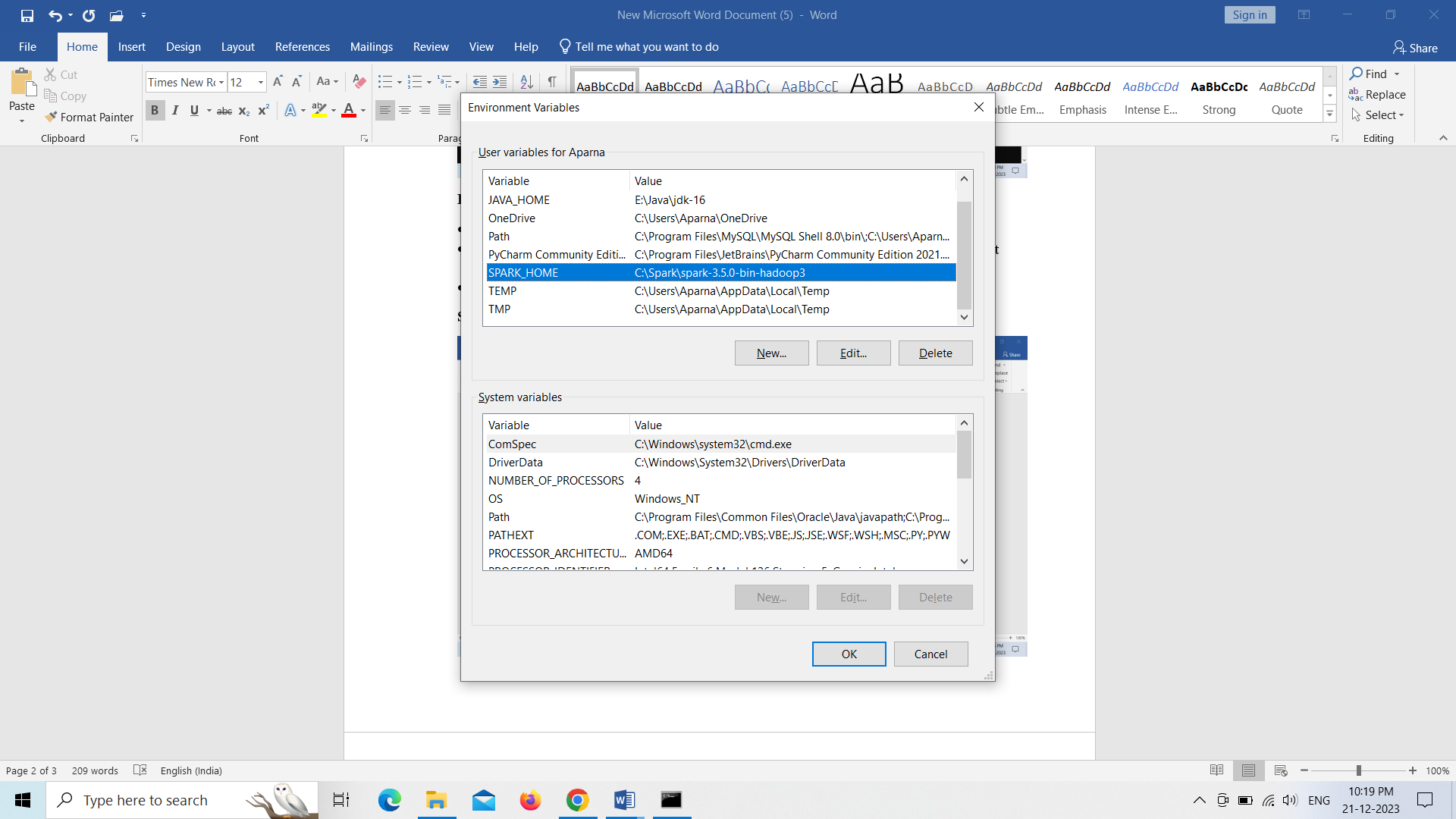
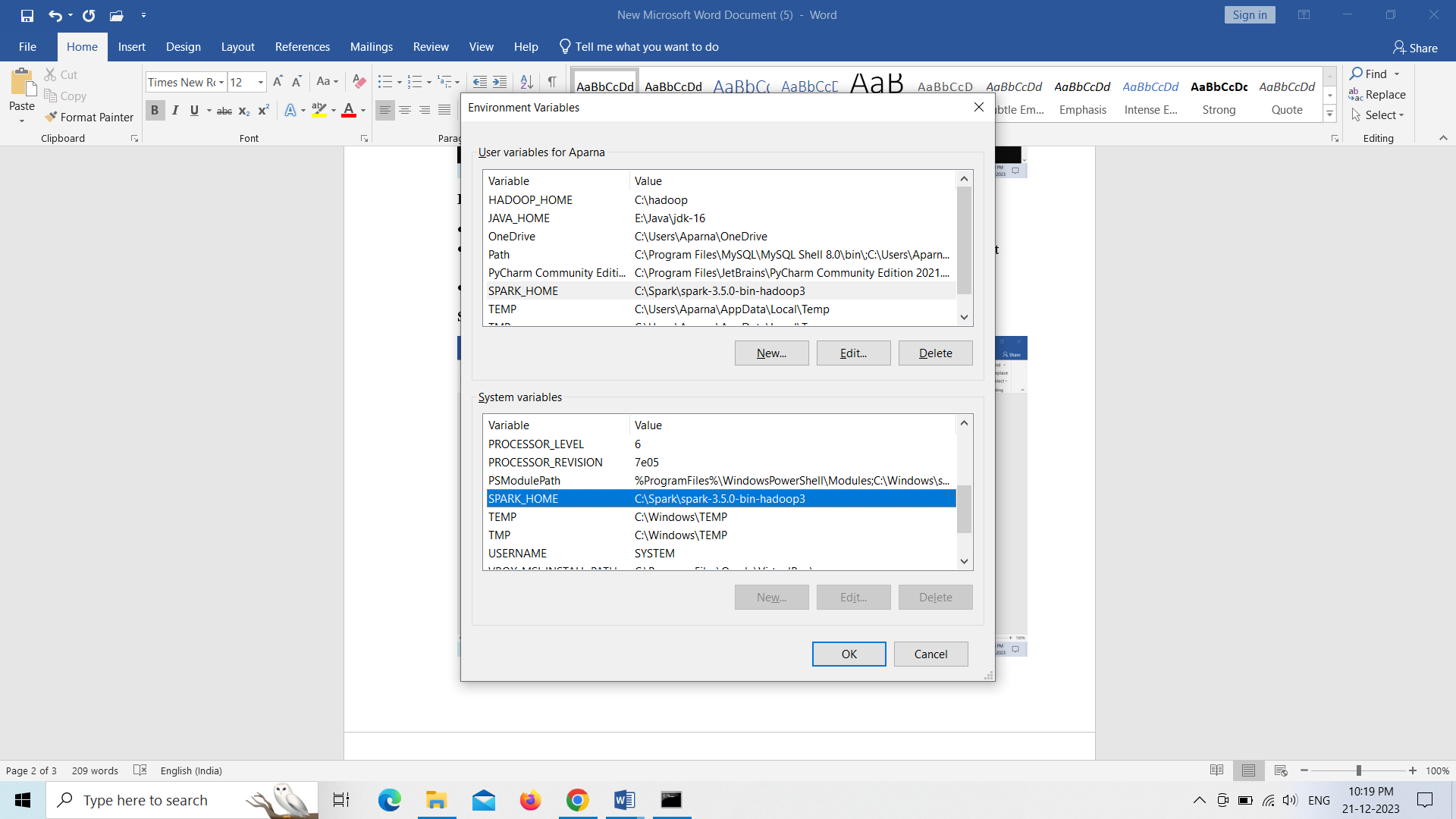
* Visit the Apache Spark download page.
* Choose the latest version of Spark and download the pre-built package for Hadoop. It will be a tarball (.tgz) file.
* Extract the downloaded tarball to a location on your machine.

**Set Environment Variables**

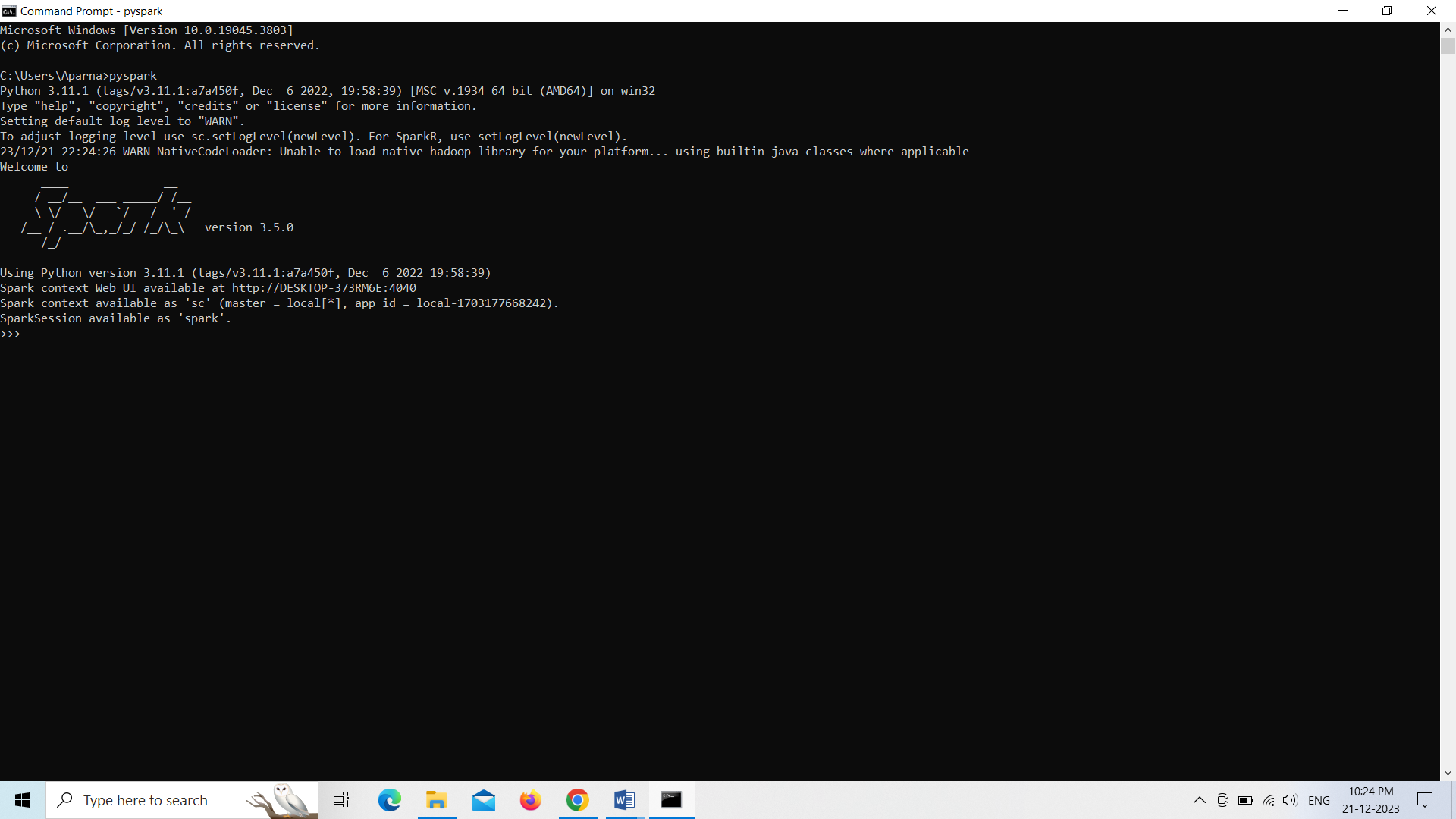






After setting up the environment variables,we need to save all of them and have to go to the command prompt and type **pyspark** as below.



* If it shows like these, you have successfully set up the Apache Spark setup.
* If you want to see pyspark web UI ,you have to type in chrome as like
* **<http://localhost:4040/jobs/>** you will get the web page as follows which indicates your pyspark is installed as follows

