**PYSPARK**

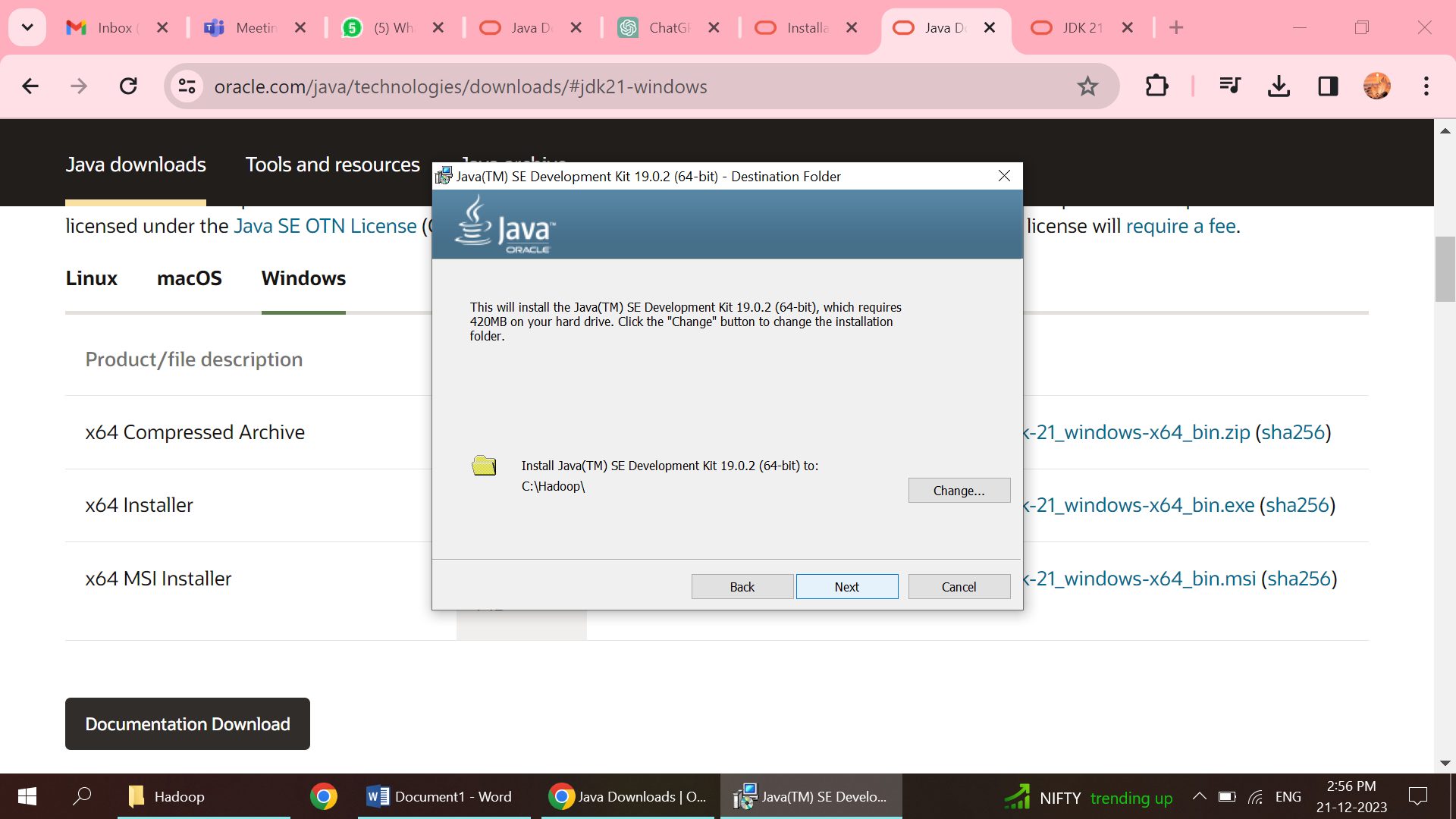
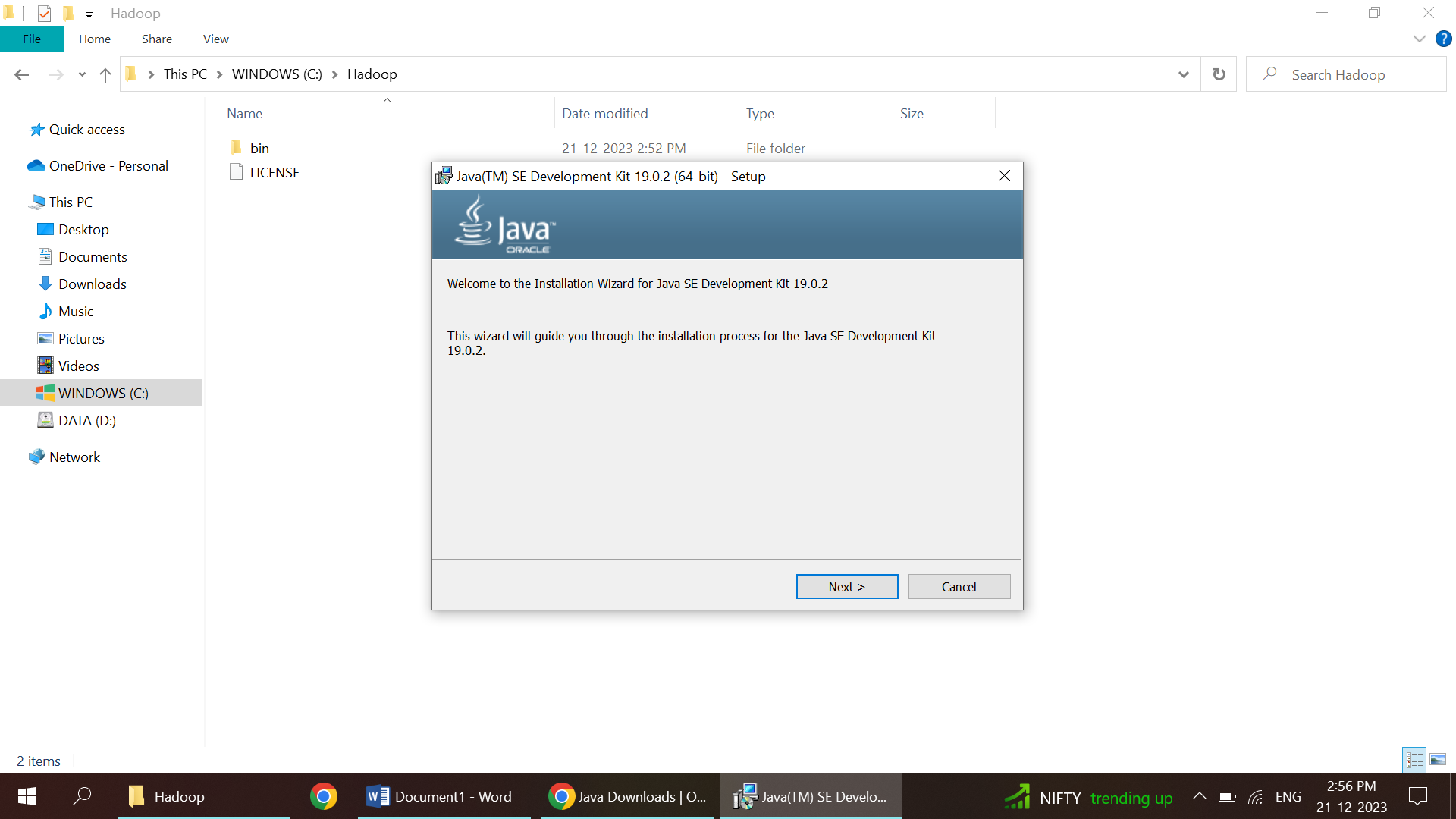
**INSTALLATION**

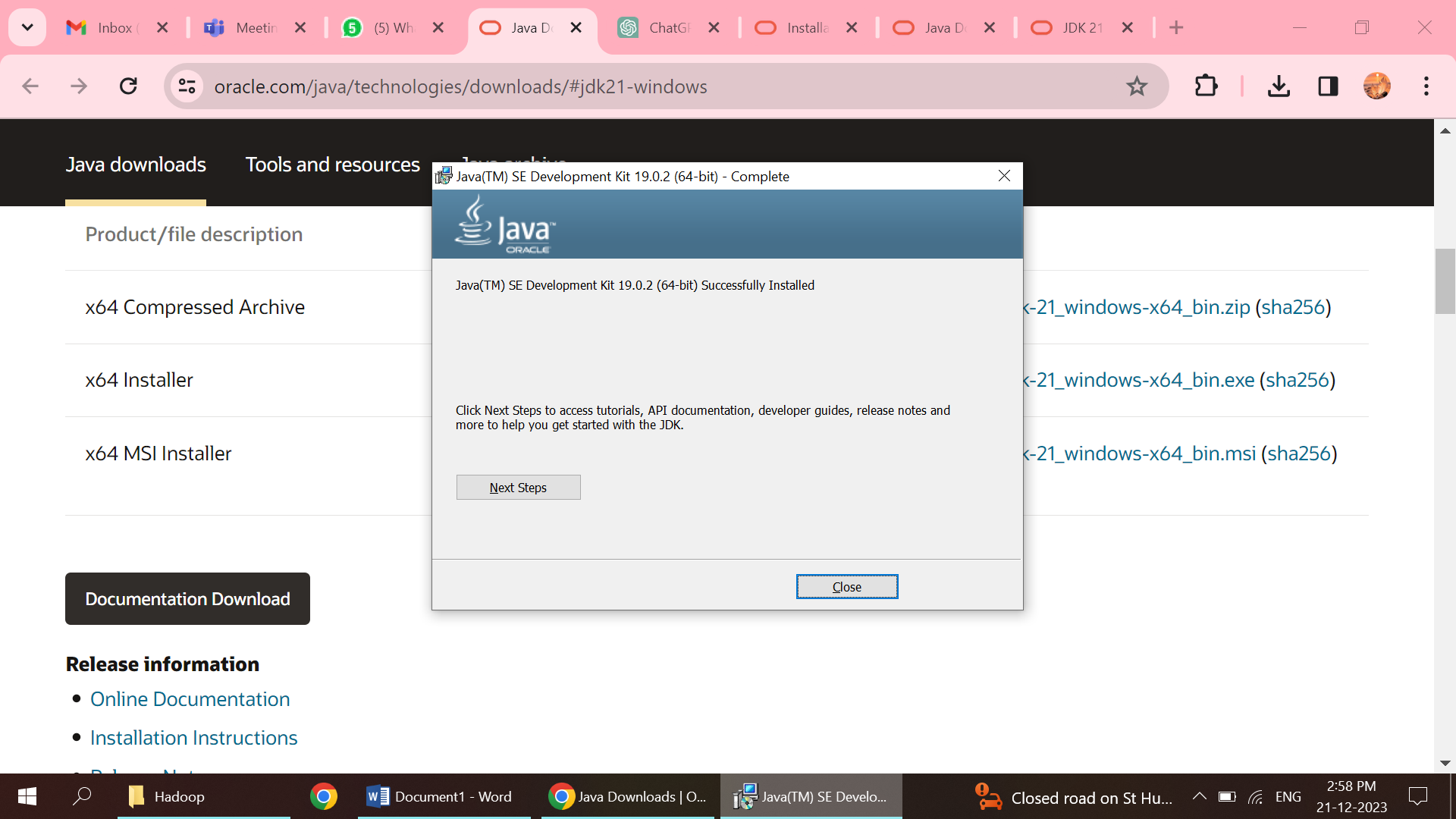
S.R.TAANUSRI

21.12.2023

Step 1: Install Java

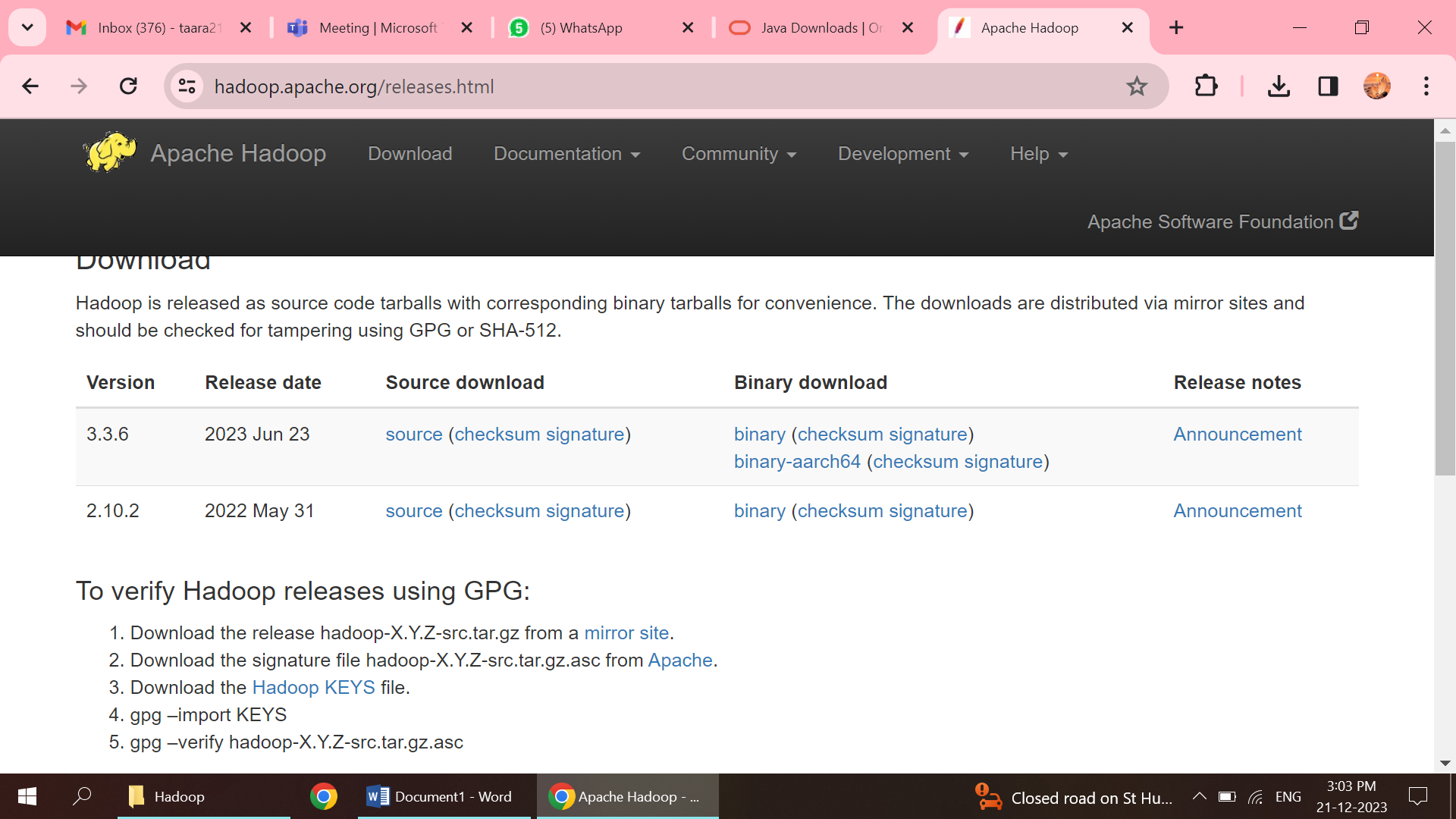
* PySpark relies on Java, Java need to be installed.
* Install the Java Development Kit (JDK) from the official Oracle website.





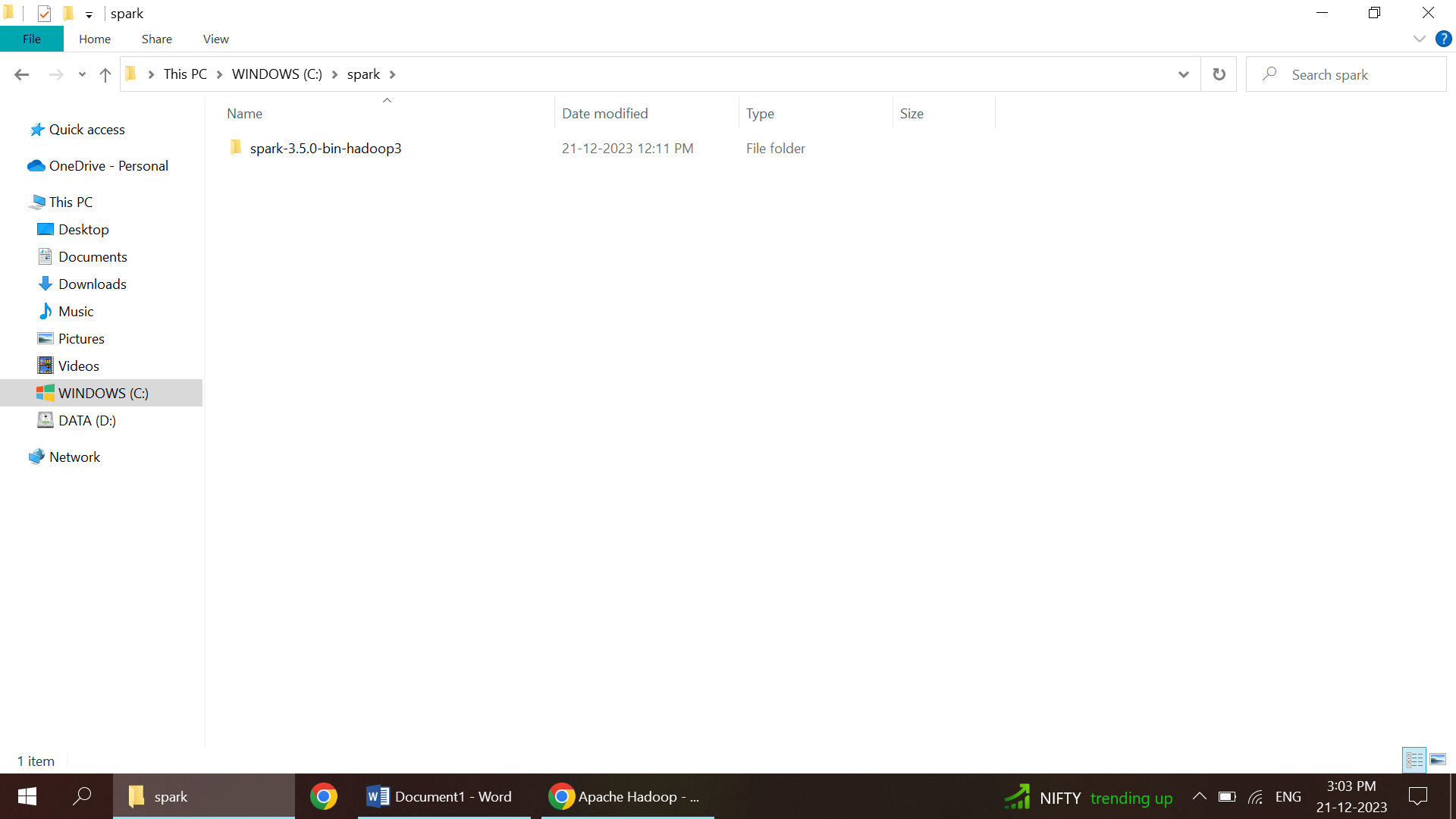
Step 2: Install Apache Spark

1. Visit the Apache Spark downloads page.
2. Choose the latest version of Spark.
3. Select the package type: "Pre-built for Apache Hadoop"
4. Download the compressed archive



Step 3: Extract Spark Archive

1. Extract the downloaded Spark archive to a C directory.
2. Create a specific folder named spark in the directory.



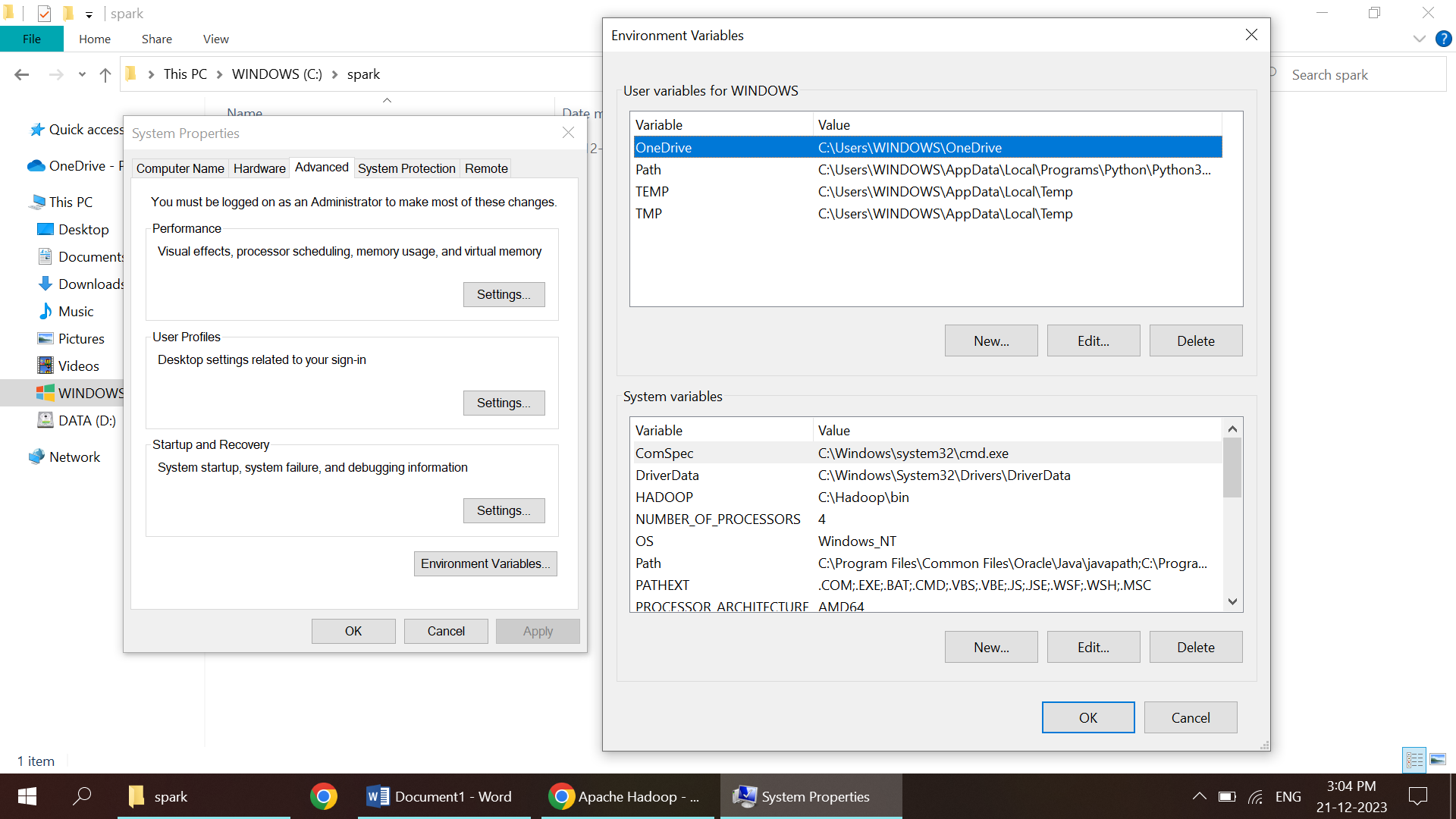
Step 4: Set Environment Variables

1. Set the SPARK\_HOME environment variable to the C directory
2. In the system variable for name give spark as name and in value give the path

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

1. Add the bin directory to your PATH environment variable:

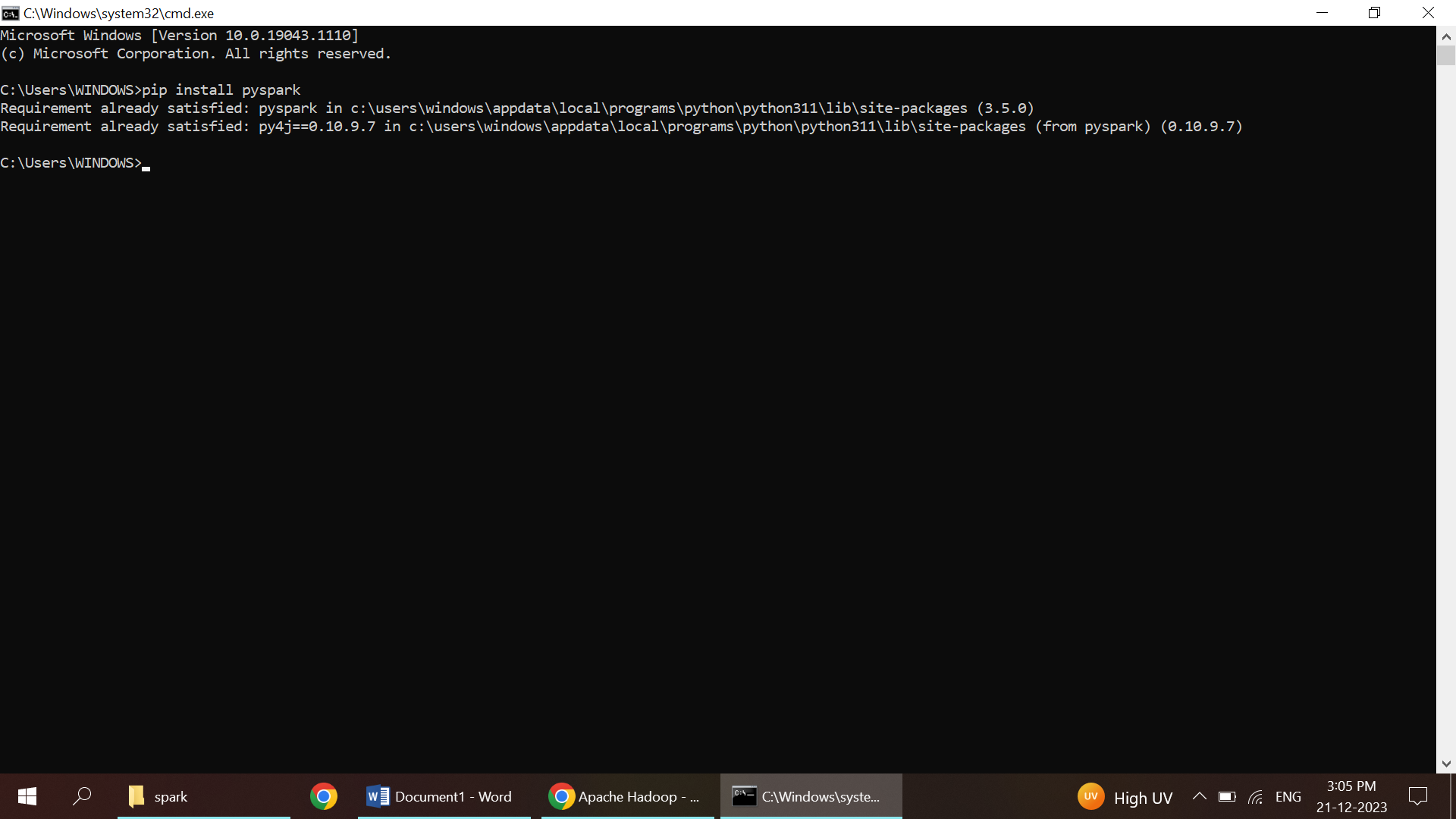
C:\spark\spark-3.5.0-bin-hadoop3



Step 5: Install PySpark Using pip

1. Open a new terminal or command prompt.
2. Install PySpark using pip:

pip install pyspark

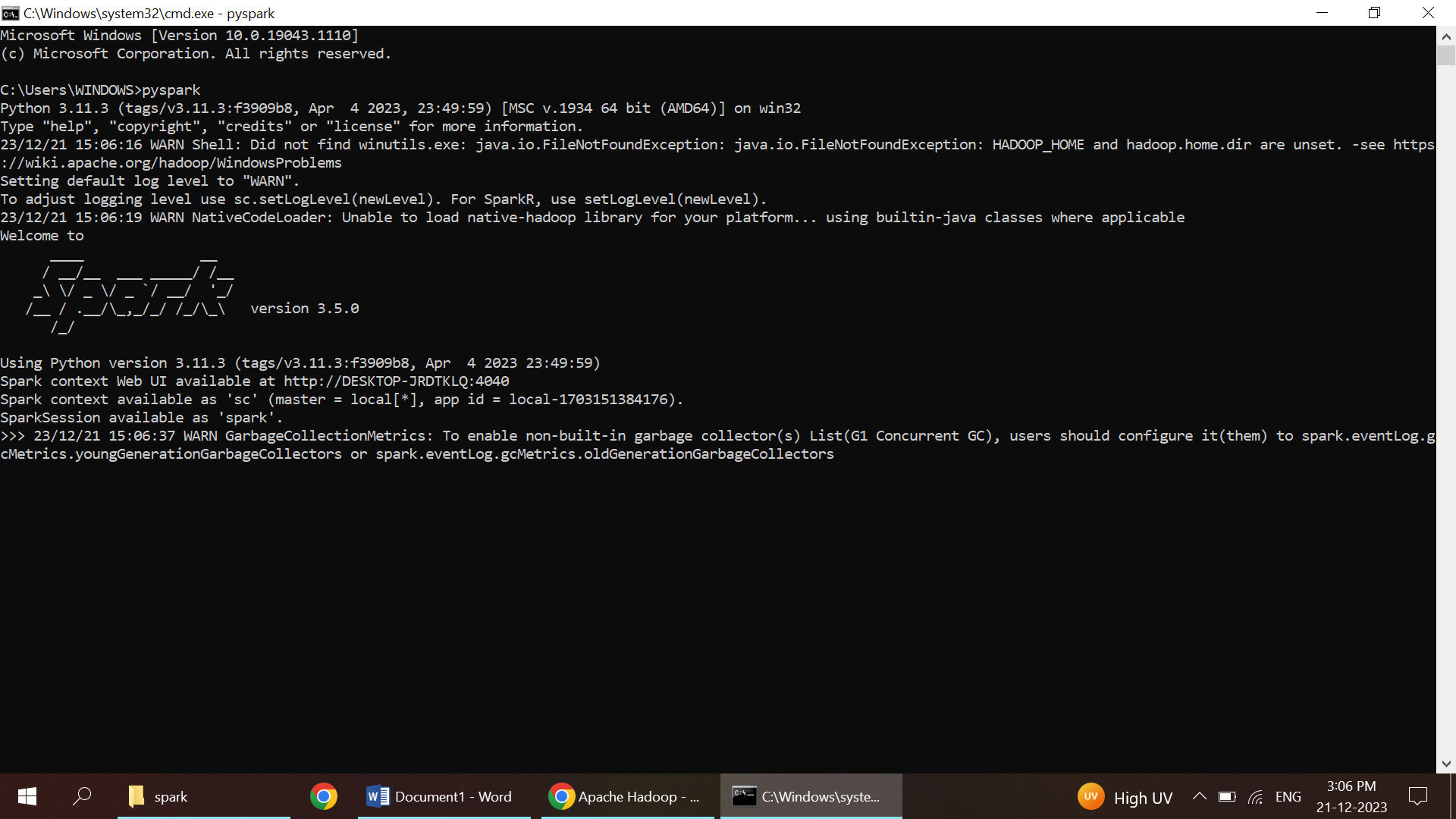


Step 6: Verify Installation

1. Open the command prompt and type:

Pyspark

(If everything is set up correctly, the Spark version will come without any errors.)



Step 7: Setup in webpage

1. Open the command promt and type pyspark
2. In browser type the following line:

<http://localhost:4040/jobs>

