PySpark – Report

NAME: SIVA DHARSHANA G

ROLL NO: 23ADR151

DEPARTMENT: AI & DS

SUBJECT: BIG DATA ANALYTICS

OBJECTIVE:

The goal is to install and set up PySpark on a Windows 10 system to allow the development and execution of Spark applications using Python

1. **Prerequisites Installation**
   - **1.1. Java Installation**

- ✓ **Step 1: Download the Java JDK (version 8 or above) from the official oracle website**
- ✓ **Step 2: Run the installer and complete the installation process.**
- ✓ **Step 3: During installation, ensure to check:**
  - ✅ **Set JAVA_HOME environment variable**
  - ✅ **Add Java to PATH**
- ✓ **Step 4: Verify installation by running:**
- ✓ **java –version**

2. **Apache Spark Installation**
   - **2.1 Download and Setup**

- ✓ **Step 1: Download a pre-built version of Apache Spark from spark.apache.org.**
- ✓ **Step 2: Extract the ZIP archive to a known location (e.g., C:\spark).**
- ✓ **Step 3: Set the environment variable SPARK_HOME to the Spark directory.**
- ✓ **Step 4: Add the bin directory of Spark to the system PATH.**

Explanation:

This step is crucial for correctly setting up Apache Spark. Carefully following the instructions is essential to prevent any installation problems.

2.2. WinUtils Configuration

- ✓ **Step 1: Download winutils.exe from a trusted source that matches your Spark/Hadoopversion.**
- ✓ **Step 2: Create a Hadoop directory structure such as: C:\hadoop\bin**
- ✓ **Step 3: Place winutils.exe inside the bin folder.**
- ✓ **Step 4: Set the environment variable:HADOOP_HOME = C:\hadoop**

3.Python and pip –Installation

3.1. Python Installation

- ✓ **Download the latest version of Python 3.x from python.org.**
- ✓ **Ensure the checkbox 'Add Python to PATH' is selected during installation.**
- ✓ **Verify using: python –version pip –version**

4. Install PySpark

Using pip

- ✓ **Open a command prompt or PowerShell and run: pip install pyspark**
- ✓ **This installs PySpark and its dependencies.**

5. Running PySpark Shell

- ✓ **Open command prompt or PowerShell and run: pyspark**
- ✓ **A Spark session should start with a prompt, confirming successful installation.**

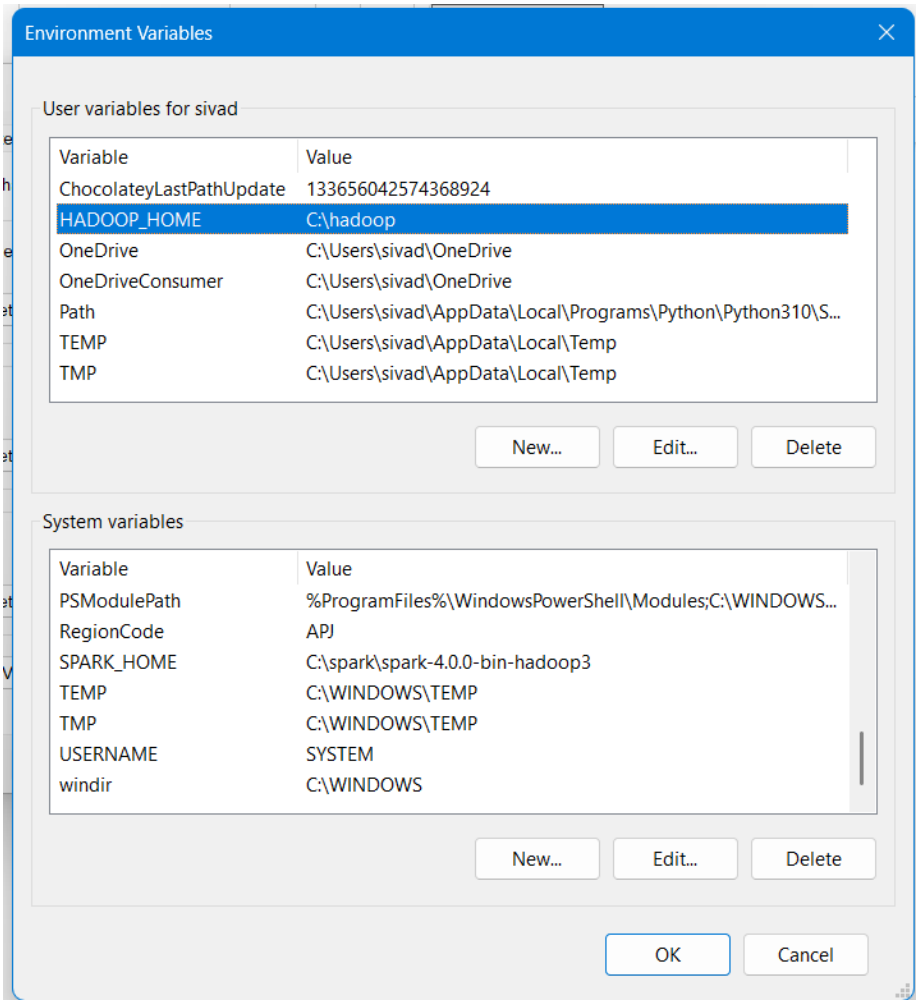
6. Troubleshooting Notes

- ✓ **If Spark does not launch:**
- ✓ **- Double-check that all environment variables are set correctly.**
- ✓ **- Ensure compatible versions of Java, Spark, and Hadoop are used.**
- ✓ **- Make sure winutils.exe is present in the correct path.**
- ✓ **- Use absolute paths when setting JAVA_HOME, SPARK_HOME, and HADOOP_HOME.**

OUTPUT:

```
Microsoft Windows [Version 10.0.26100.4061]
(c) Microsoft Corporation. All rights reserved.

C:\Users\sivad>java --version
openjdk 21.0.4 2024-07-16 LTS
OpenJDK Runtime Environment Temurin-21.0.4+7 (build 21.0.4+7-LTS)
OpenJDK 64-Bit Server VM Temurin-21.0.4+7 (build 21.0.4+7-LTS, mixed mode, sharing)
```



```
(c) Microsoft Corporation. All rights reserved.

C:\>certutil -hashfile
Expected at least 1 args, received 0
CertUtil: Missing argument

Usage:
  CertUtil [Options] -hashfile InFile [HashAlgorithm]
  Generate and display cryptographic hash over a file

Options:
  -Unicode          -- Write redirected output in Unicode
  -gmt              -- Display times as GMT
  -seconds          -- Display times with seconds and milliseconds
  -v                -- Verbose operation
  -privatekey       -- Display password and private key data
  -pin PIN                  -- Smart Card PIN
  -sid WELL_KNOWN_SID_TYPE  -- Numeric SID
            22 -- Local System
            23 -- Local Service
            24 -- Network Service

Hash algorithms: MD2 MD4 MD5 SHA1 SHA256 SHA384 SHA512

CertUtil -?              -- Display a verb list (command list)
CertUtil -hashfile -?    -- Display help text for the "hashfile" verb
CertUtil -v -?           -- Display all help text for all verbs
```

```
C:\Users\sivad>spark-shell
WARNING: Using incubator modules: jdk.incubator.vector
Using Spark's default log4j profile: org/apache/spark/log4j
Setting default log level to "WARN"
To adjust logging level use sc.setLogLevel(newLevel).
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 0.0.0
      /_/

Using Scala version 2.13.16 (OpenJDK 64-Bit Server VM, Java 21.0)
Type in expressions to have them evaluated.
Type :help for more information.
25/07/31 12:23:15 WARN NativeCodeLoader: Unable to load native-
hadoop library for your platform... using builtin-java classes whero
Spark context Web UI available at http://DESKTOP-GATIOEL:4040
Spark context available as 'sc' (master = local[*], app id =
Spark session available as 'spark'.
```

Conclusion:

Once all the steps are completed successfully, PySpark will be fully functional on a Windows system, allowing you to run and test Spark applications locally using Python.