

How to install Hadoop

A. Install Hadoop on Windows Machine:

Install Java machine:

using command line type: **java -version**

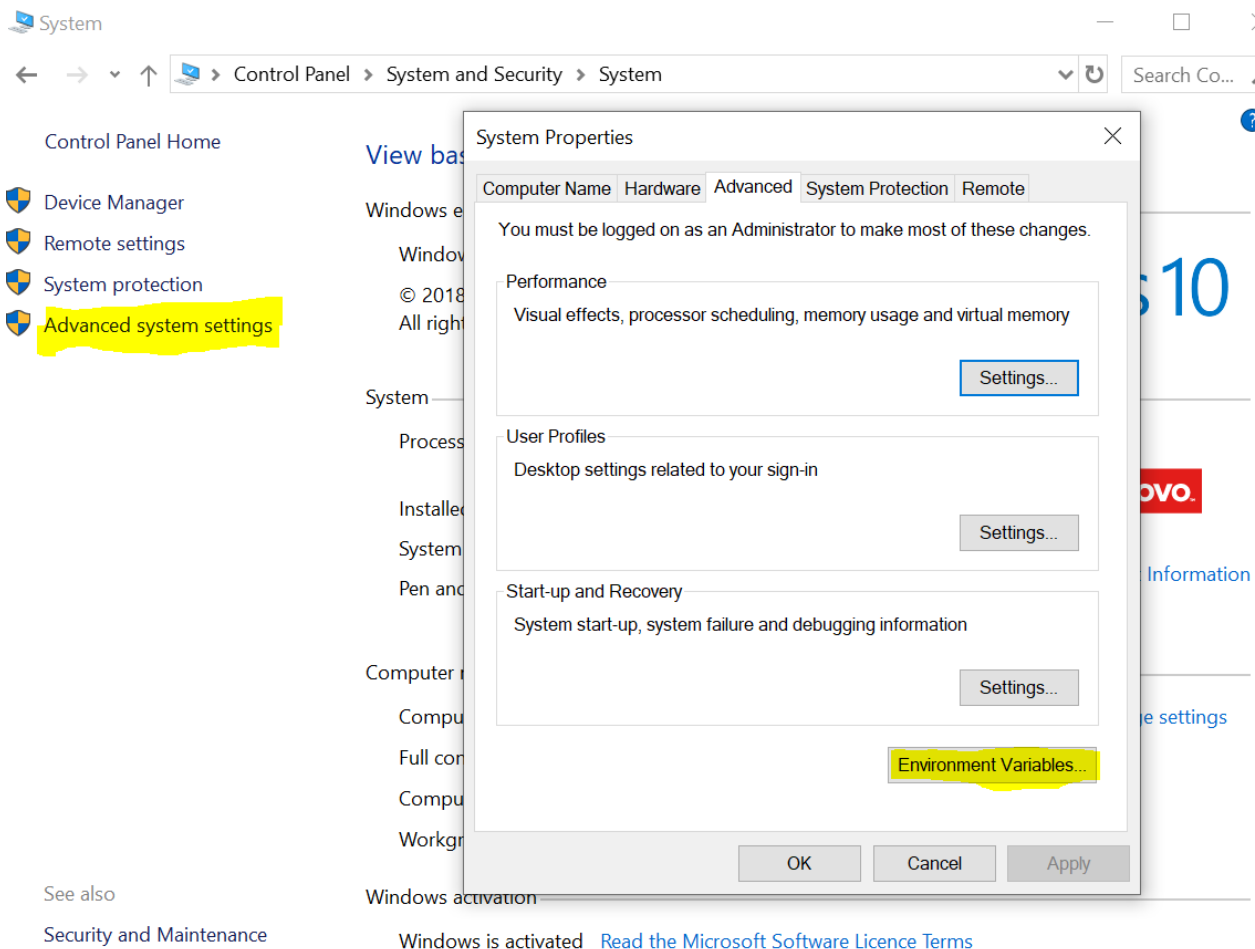
Install Java 8 if it's not been installed. You can use the following link to install Java SE kit 8

<https://www.oracle.com/java/technologies/javase/javase-jdk8-downloads.html>

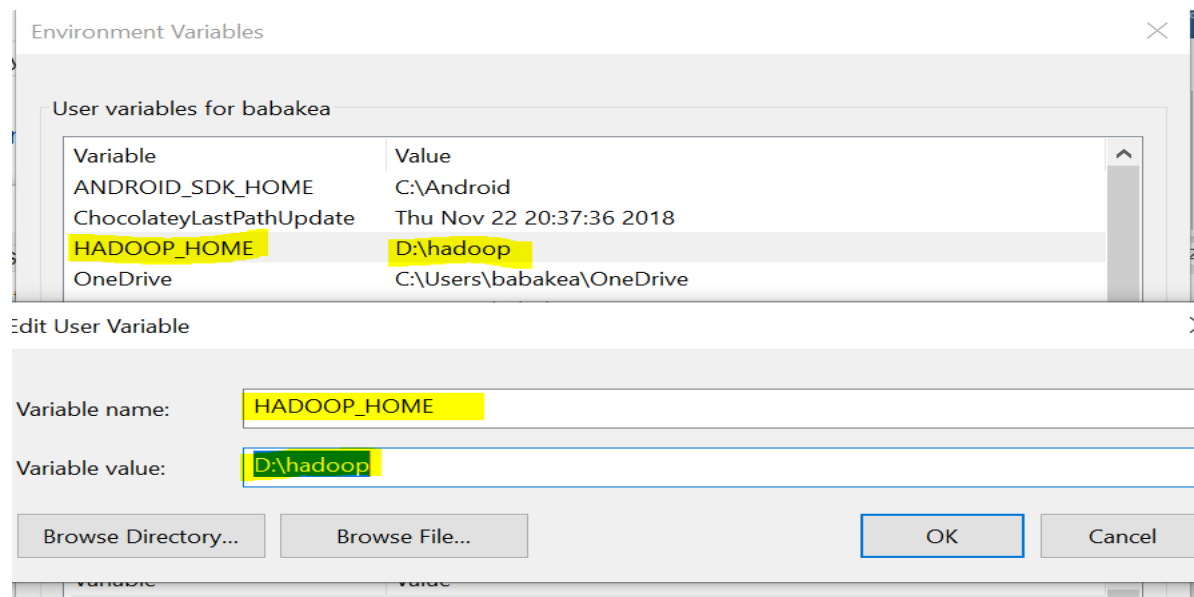
```
C:\Users\babakea>java -version
java version "1.8.0_181"
Java(TM) SE Runtime Environment (build 1.8.0_181-b13)
Java HotSpot(TM) 64-Bit Server VM (build 25.181-b13, mixed mode)
```

Install Apache Spark:

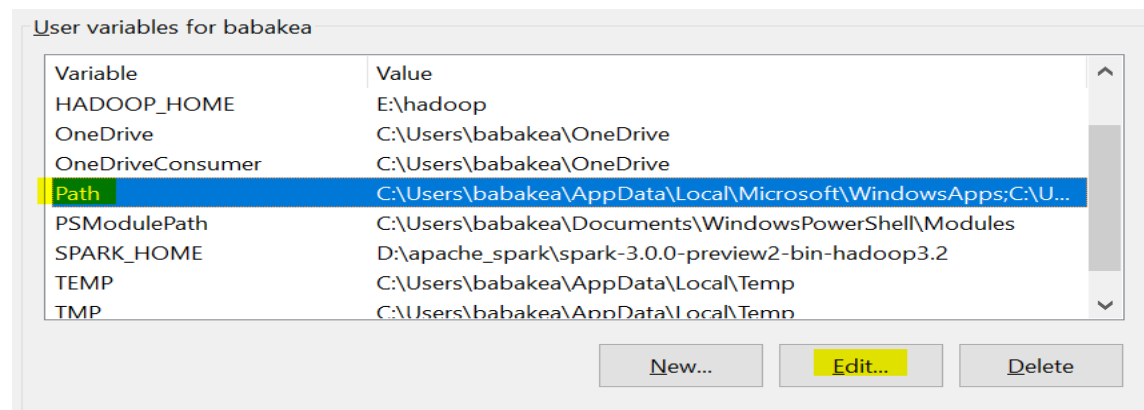
- 1) Download the apache Spark from : <https://spark.apache.org/downloads.html>
 - a. <http://us.mirrors.quenda.co/apache/spark/spark-3.0.0-preview2/spark-3.0.0-preview2-bin-hadoop3.2.tgz>
- 2) Unzip the file to the spark folder in your local computer (D:\apache_spark\spark-3.0.0-preview2-bin-hadoop3.2)
- 3) Install winutils.exe to setup Hadoop using the following link:
 - a. https://www.github.com/BabakEA/localizing_Jupyter_spark_on_Windows/blob/master/winutils.exe
 - b. Create a Hadoop folder
 - c. Create bin folder and copy the winutils.exe in the (D:\Hadoop\bin\)
 - d. Go to Control Panel\System and Security\System and click on advanced system setting



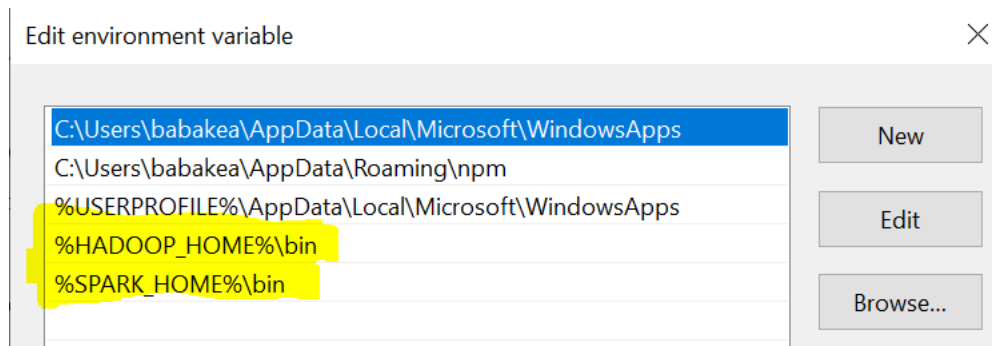
- e. Got to Environment variables, create HADOOP_HOME and SPARK_HOME
 - i. HADOOP:
 1. Variable Name: HADOOP_HOME
 2. Variable value: D:\hadoop
 - ii. SPARK:
 1. Variable Name: SPARK_HOME
 2. Variable value: D:\apache_spark\spark-3.0.0-preview2-bin-hadoop3.2



- f. Add Hadoop and Spark Binary to the path: go to the User Variables for ... and edit the path



- g. Add Hadoop and spark home to the path:
- %HADOOP_HOME%\bin
 - %SPARK_HOME%\bin



- h. Create a new folder name tem. Inside the tem folder create new folder name hive. Change hive permission to accessible for all users by writhing:
 - i. `d:\hadoop\bin\winutils chmod 777 d:\tem\hive`
- i. change the Spark log, from info to error: go to the following address
- j. `D:\apache_spark\spark-3.0.0-preview2-bin-hadoop3.2\conf`
 - i. Make a copy of (log4j.properties.template) and rename it to log4j.properties
 - ii. Go to # Set everything to be logged to the console
 - iii. `log4j.rootCategory=INFO, console` and change it to (ERROR), `log4j.rootCategory=ERROR, console`

Setup Jupyter notebook for Pyspark:

- 1) Install findspark lib using:
 - a. `python -m pip install findspark`
- 2) initialise pyspark using this lib using:
 - a. `import findspark`
 - b. `findspark.init()`