**APACHE SPARK || HADOOP || PYSPARK**

- What is Spark - Apache Spark™ is a multi-language engine for executing data engineering, data science, and machine learning on single-node machines or clusters.

- What is pyspark - Pyspark is library. Spark with python is called pyspark

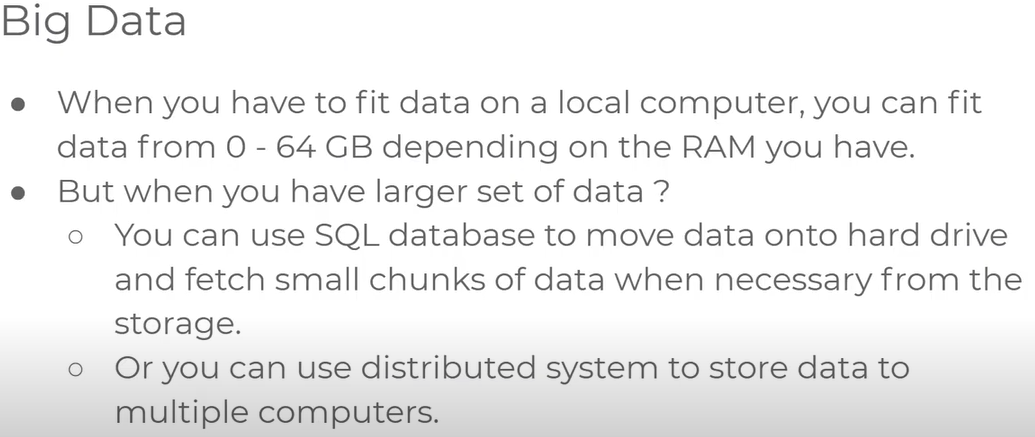
- How we will implement spark using machine learning with spark API called pyspark

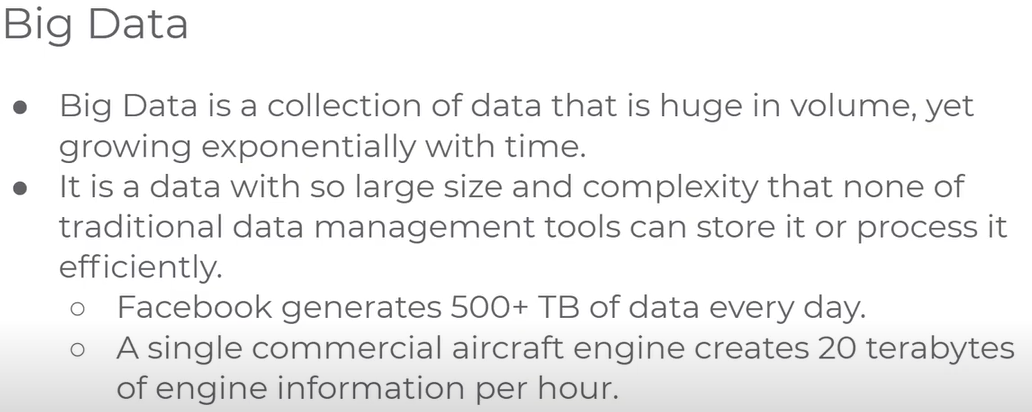
- How spark is used in big data

- What is bigdata

- Setting up Py-Spark on your system

- How to use python to access pyspark

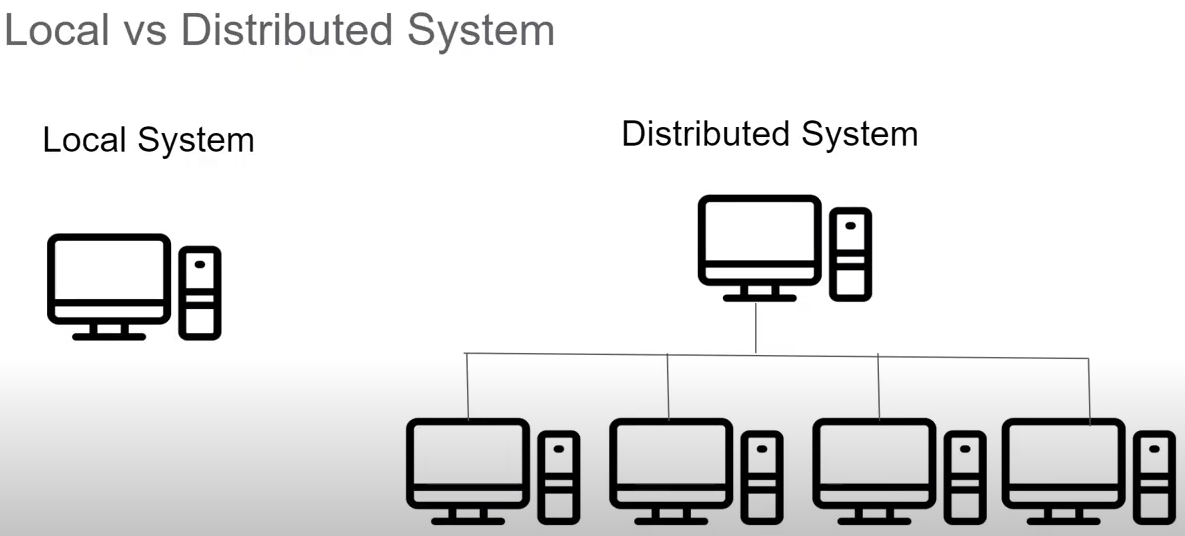




- Single system cannot handle big data that’s why we need to use for distributed system.

- Single system we called as local system

- Local system vs Distributed System



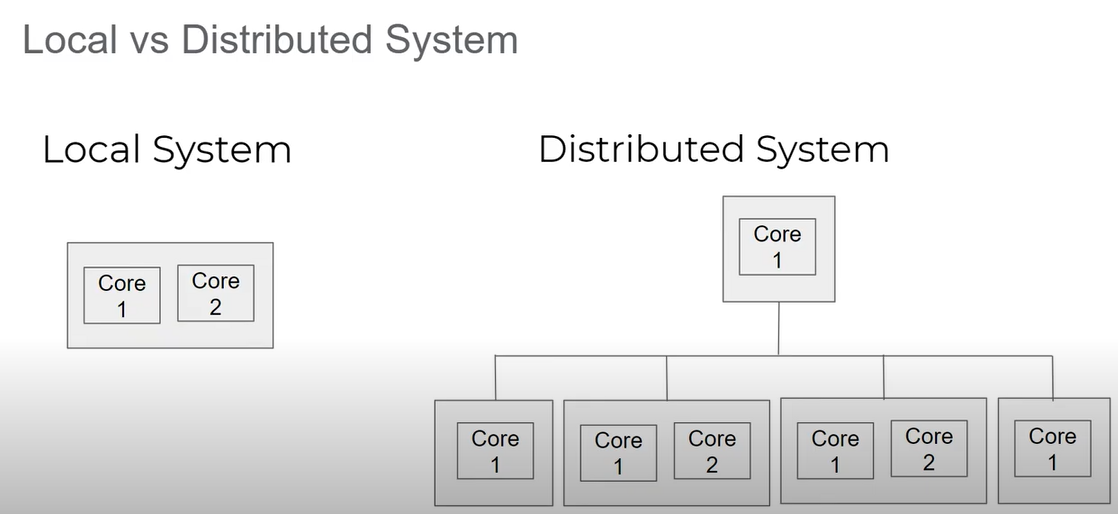
-Local system is a computer system which has only access to his own resources. It cannot access to any other pc’s

-Processing power of local system is limited & can accessible only for 1 computer

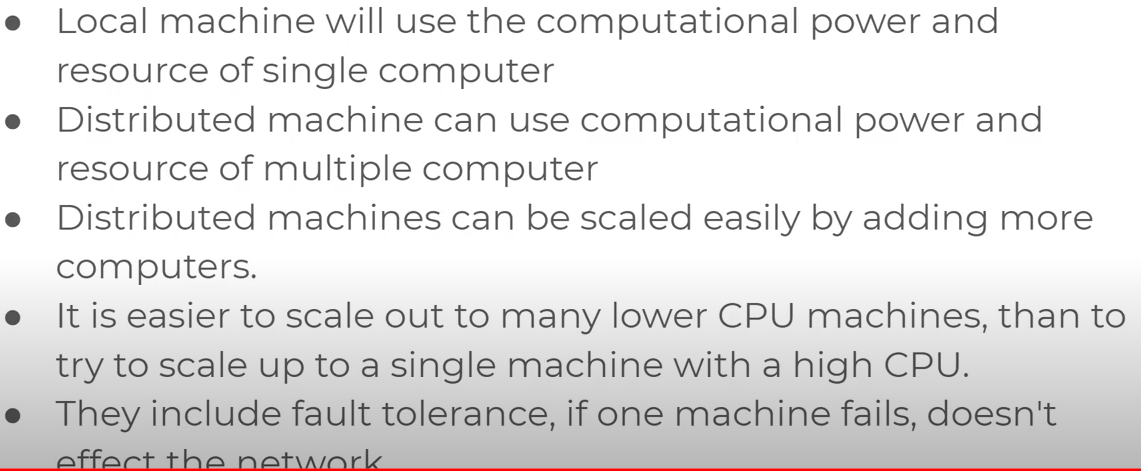
-Distributed system is many computers are connected to single master computer.

The master computer can access the resources for other computer to use for storage and computation.

-Distribution system we can make use of processing power of multiple computer which are connected to distribution system.

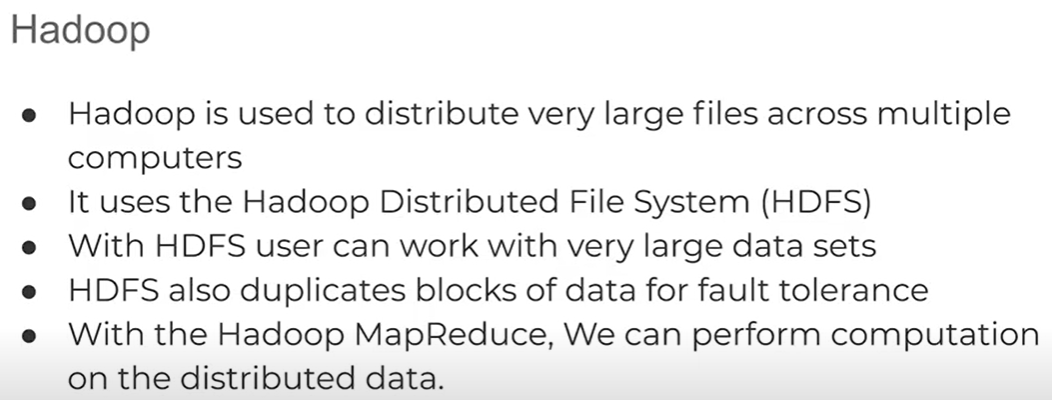


-Instead of buying very expensive single local system. With high computational power we can make use of single cheaper system with low computational power. connect with together and get high computational power.

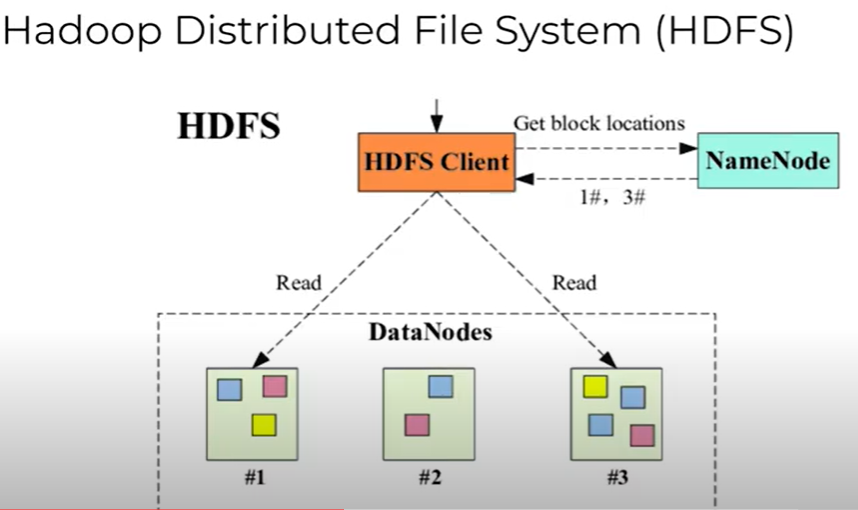


HADOOP:-

-For distributed system we can called as Hadoop



-HDFS will use Name node & data node architecture to implemented distributed file system that provides high speed access to the data.

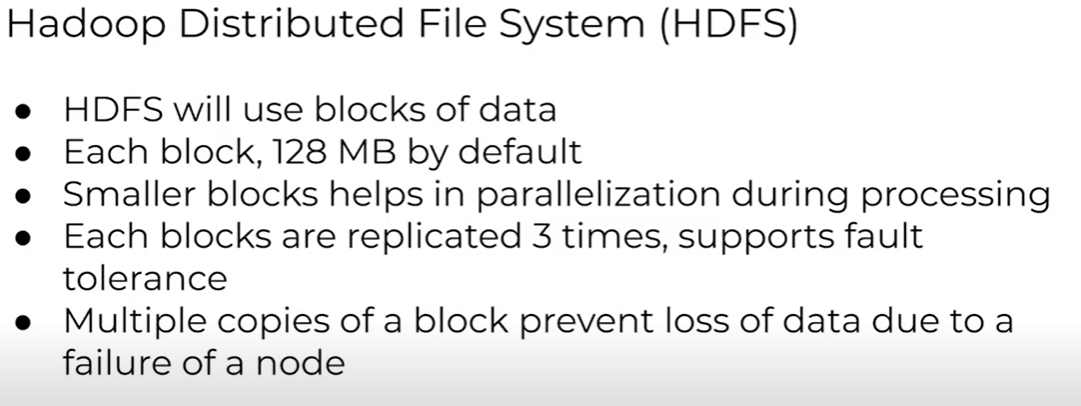


-HDFS has 2 nodes –

Name node – stores the location of data node

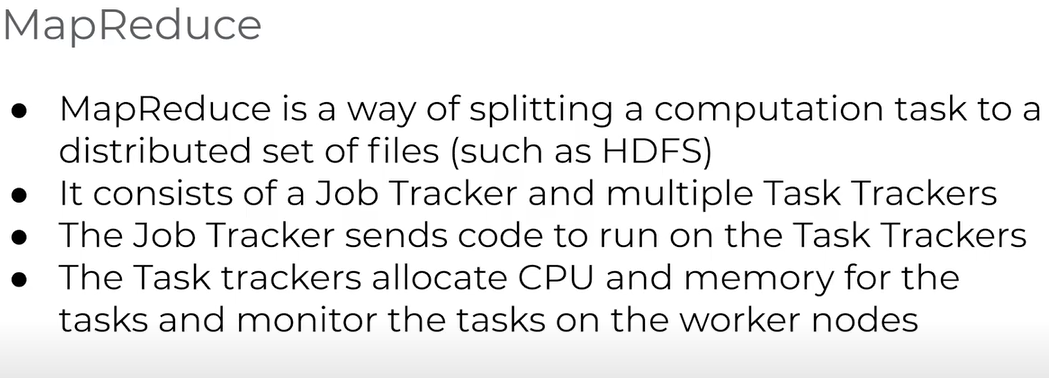
Data – Contains block of the data

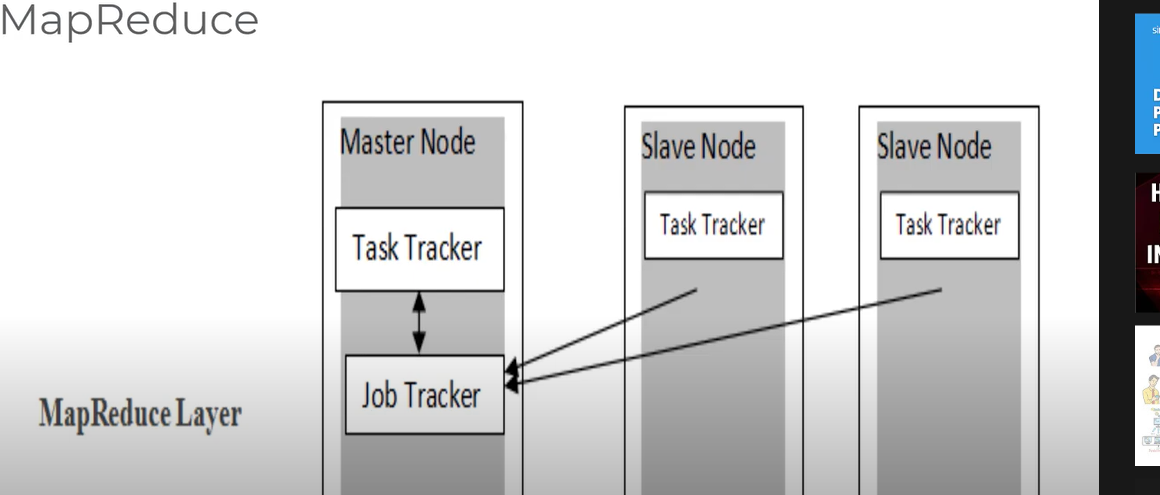
-more import part of the block data –



-With the help of Hadoop, we can distribute large set of data among multiple computers

-To distributed the computational task on distributed file system we must use for MapReduce





- Let’s understand from basic like - How do you read dataset, how do we connect to the data source

-How do we work with Data Frame. We try to understand the Mlib.

-Apache spark MLLIB which is help machine learning regression, classification & clustering.

-Also, we will understand how same operation will happen in the cloud as well.

-We will see some of the huge dataset panda’s data frame won’t work on that case we will go for pyspark data frame

-Advantage of the spark is run workload of 100 time faster

[**https://spark.apache.org/**](https://spark.apache.org/)

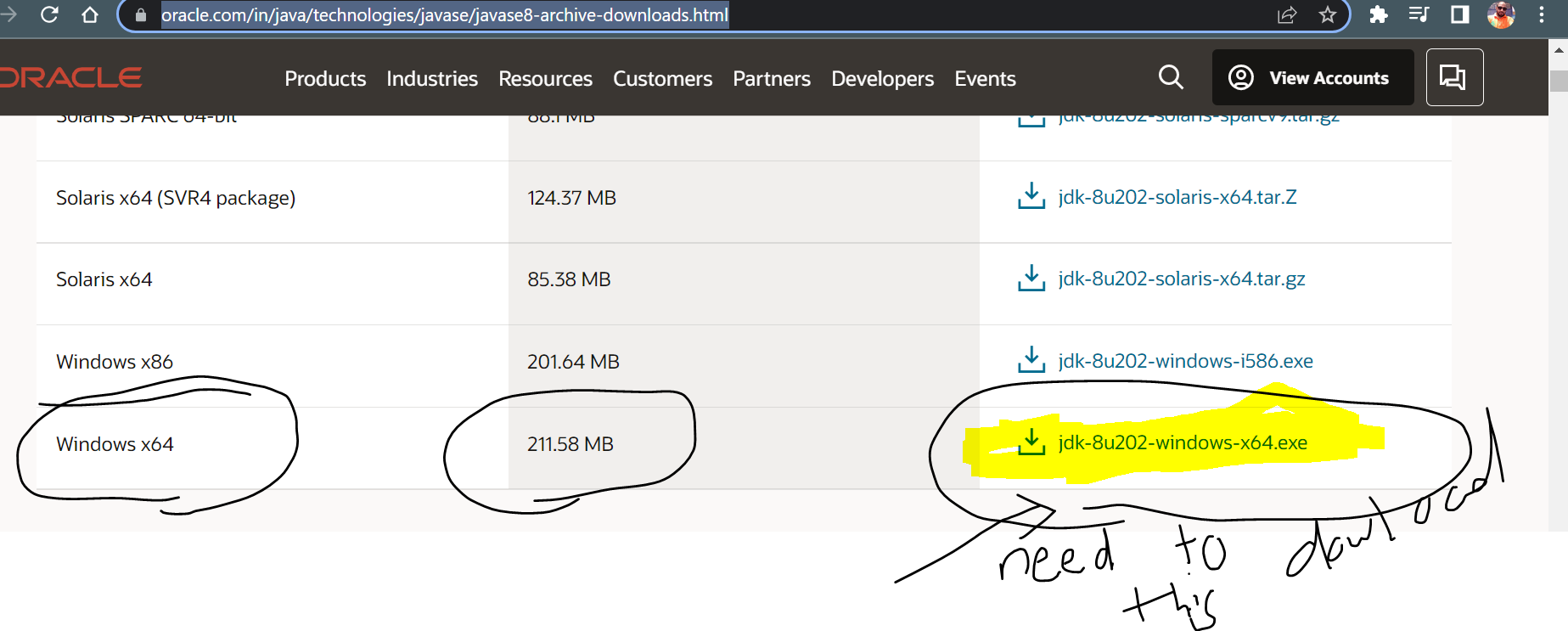
**==================================================================================**

How we can install pyspark on windows and use it with Jupiter notebook. pyspark is used for Data Science( Data Analytics ,Big data, Machine Learning etc)

1: Download & Install Anaconda from given link 🡪 <https://www.anaconda.com/>

2: Download & Install Java 🡪 You need to install java jdk8 download (Another version won’t work)

Without java Apache spark won’t work 🡪 <https://www.oracle.com/in/java/technologies/javase/javase8-archive-downloads.html>

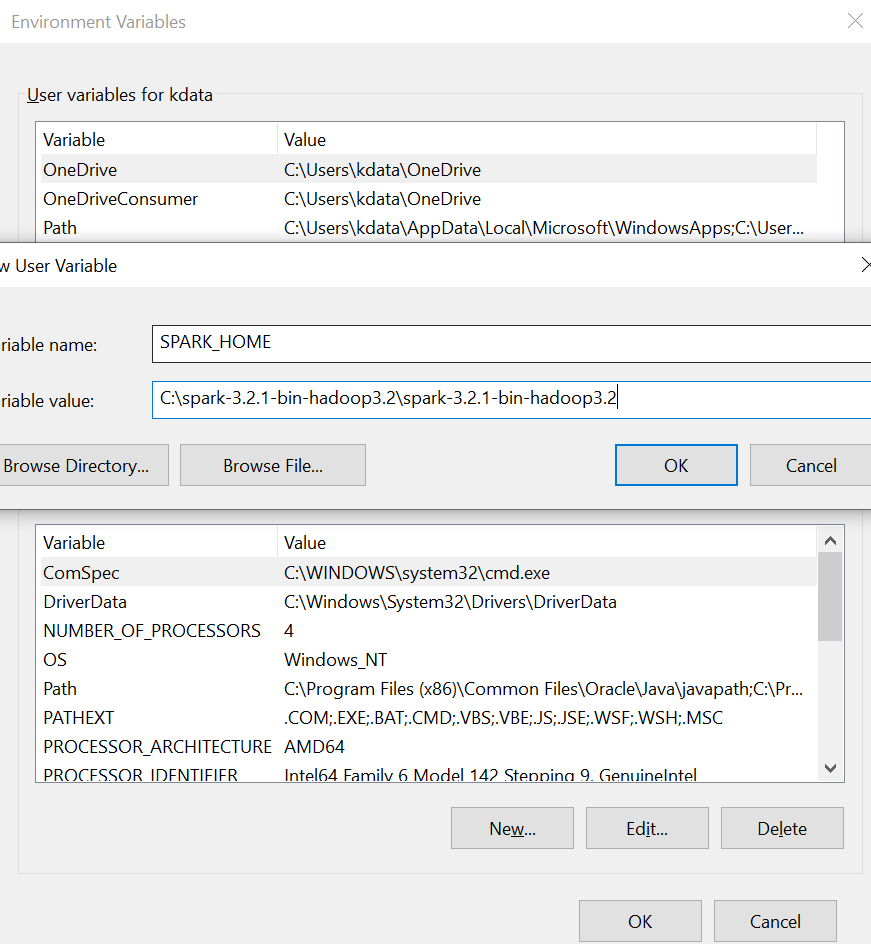


3: Download & Install Apache Spark 🡪

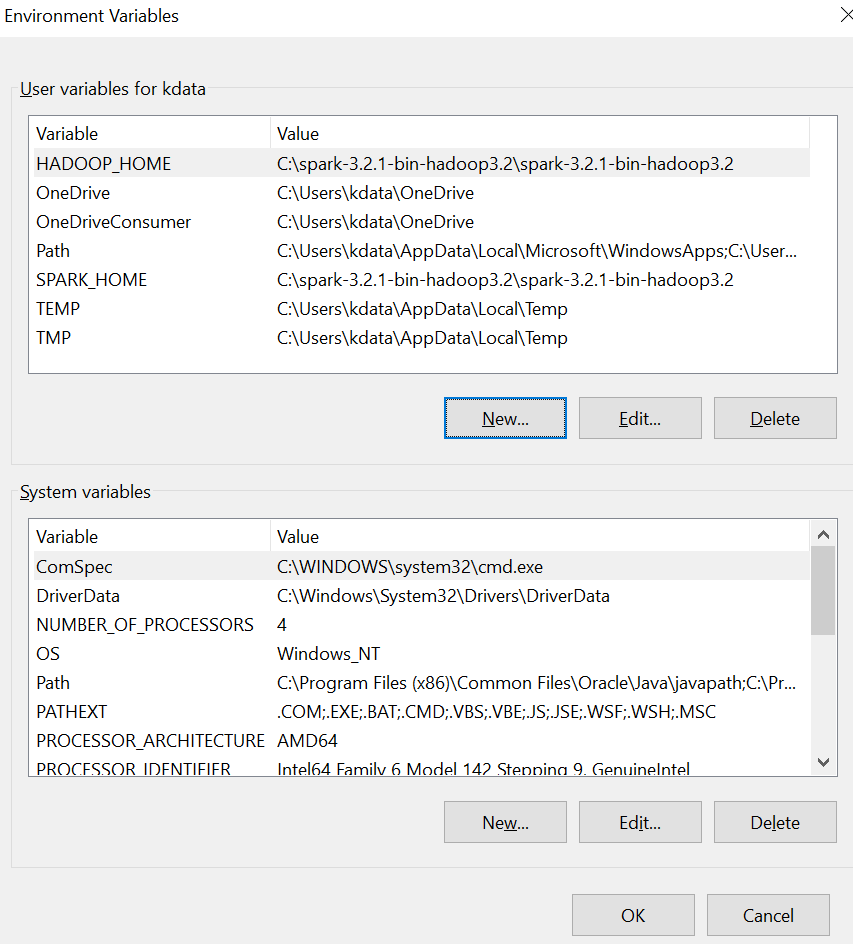
<https://spark.apache.org/downloads.html> --> click on download spark: [spark-3.2.1-bin-hadoop3.2.tgz](https://www.apache.org/dyn/closer.lua/spark/spark-3.2.1/spark-3.2.1-bin-hadoop3.2.tgz) 🡪 click on HTTP link : [**https://dlcdn.apache.org/spark/spark-3.2.1/spark-3.2.1-bin-hadoop3.2.tgz**](https://dlcdn.apache.org/spark/spark-3.2.1/spark-3.2.1-bin-hadoop3.2.tgz) --> After download the file need to extract the file and move to C-drive

4: Download winutils 🡪 winutils is available in GitHub (It is required so that Hadoop can work in windows) 🡪 <https://github.com/cdarlint/winutils/tree/master/hadoop-2.7.7/bin> --> download winutils.exe from Github 🡪 after download just copy the winutils.exe file 🡪 go to C-drive 🡪 open Apache spark folder 🡪 open bin folder 🡪 paste the winutils.exe file

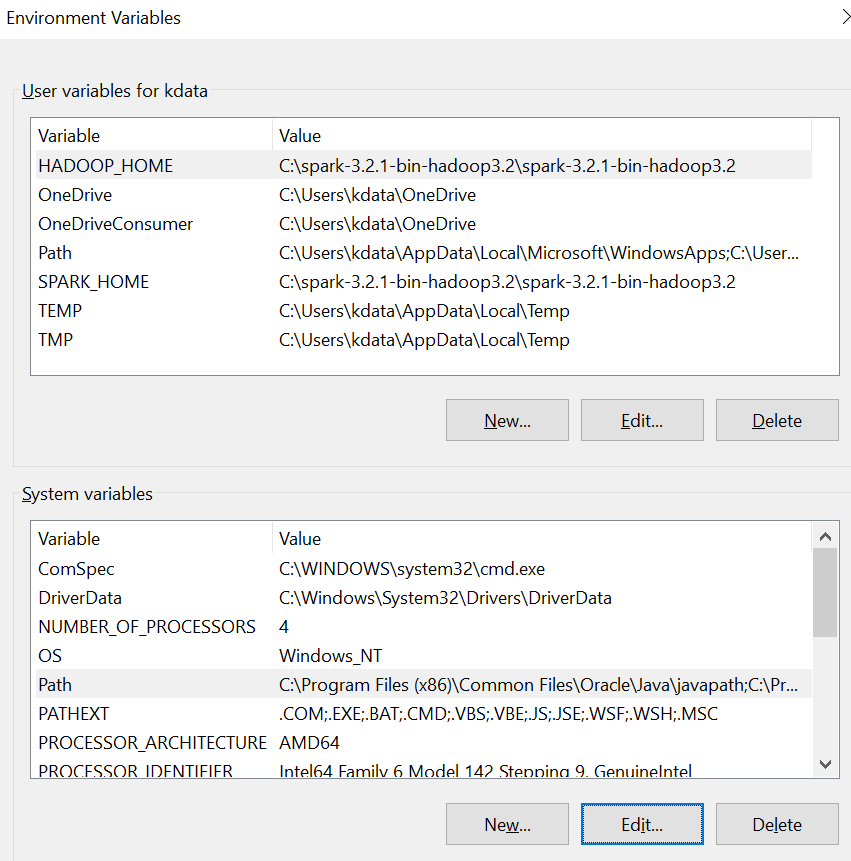
5: Set Environment variables and path 🡪 this is the most important step 🡪 right click on windows taskbar – right click on windows icon - select system – in setting window under related settings click Advanced system settings – Advanced – Environment variable – user variable – variable name SPARK\_HOME 🡪 paste the file path where you saved the SPARK folder

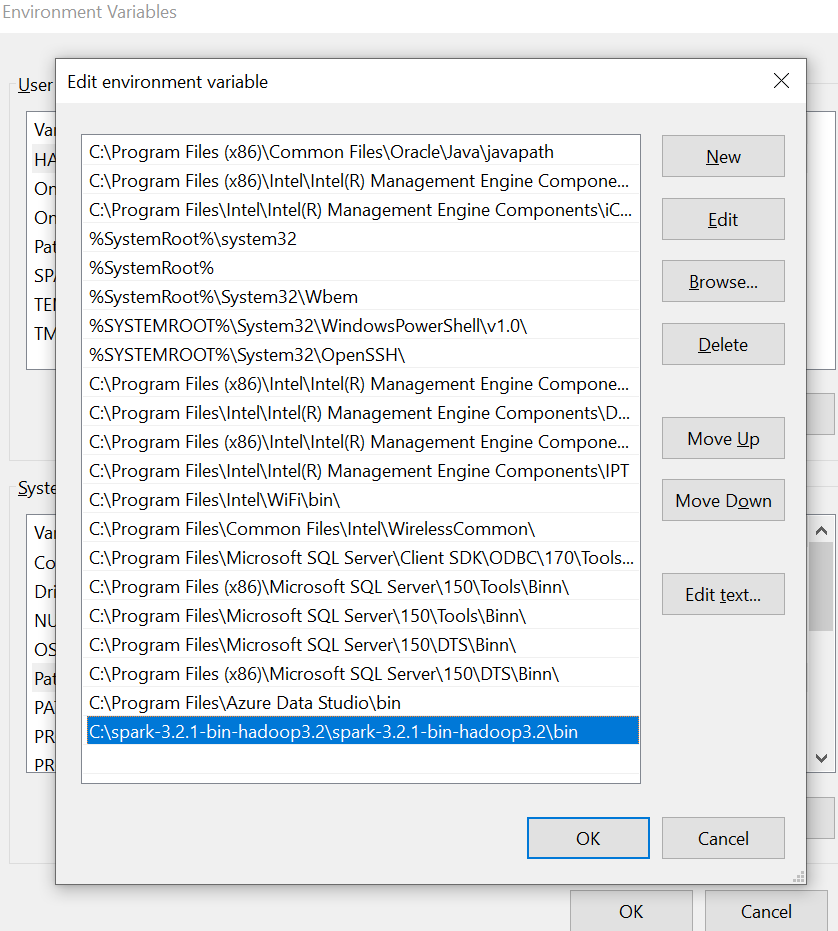


Same way creates another environment variable 🡪 crate user name as HADOOP\_HOME 🡪 paste the same path because we already add Hadoop in spark folder

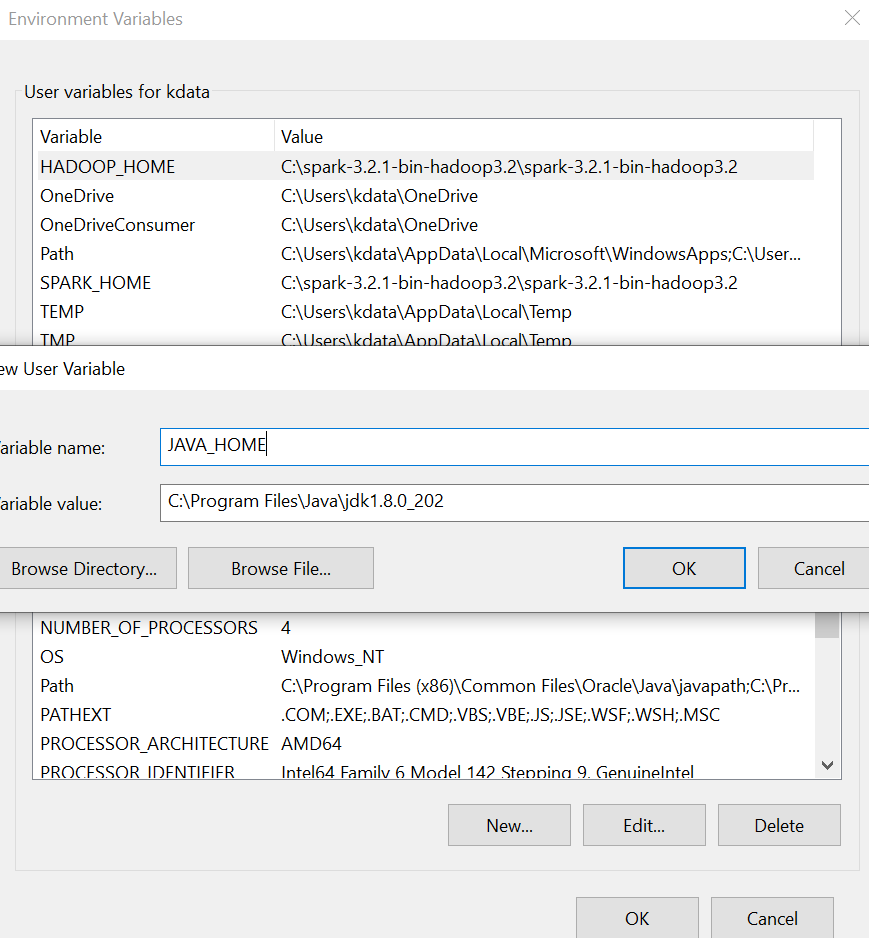


Next step -- system variable – Select the path option -- Add a new path (C: drive and select bin folder’s path) -





You need to add 1 more variable – JAVA\_ HOME – Need to give java jdk path which you can get it file from (C-Drive: Programme file: java folder: jdk path)



6: Install Find spark package

Open Anaconda navigator cmd – install findspark package from link – (<https://anaconda.org/conda-forge/findspark>) (Need to try with all packages)

7: Using Jupyter notebook: import pyspark 🡪 if you get any error 🡪 next step would be type below codes (import findspark // findspark.init() // findspark.find() 🡪 then again import pyspark 🡪 this should work and congratulation successfully pyspark package is installed successfully in your machine.

**==================================================================================**

How to install Hadoop 3.2.2 spark 3.1.2 in windows

**https://medium.com/@pedro.a.hdez.a/hadoop-3-2-2-installation-guide-for-windows-10-454f5b5c22d3**

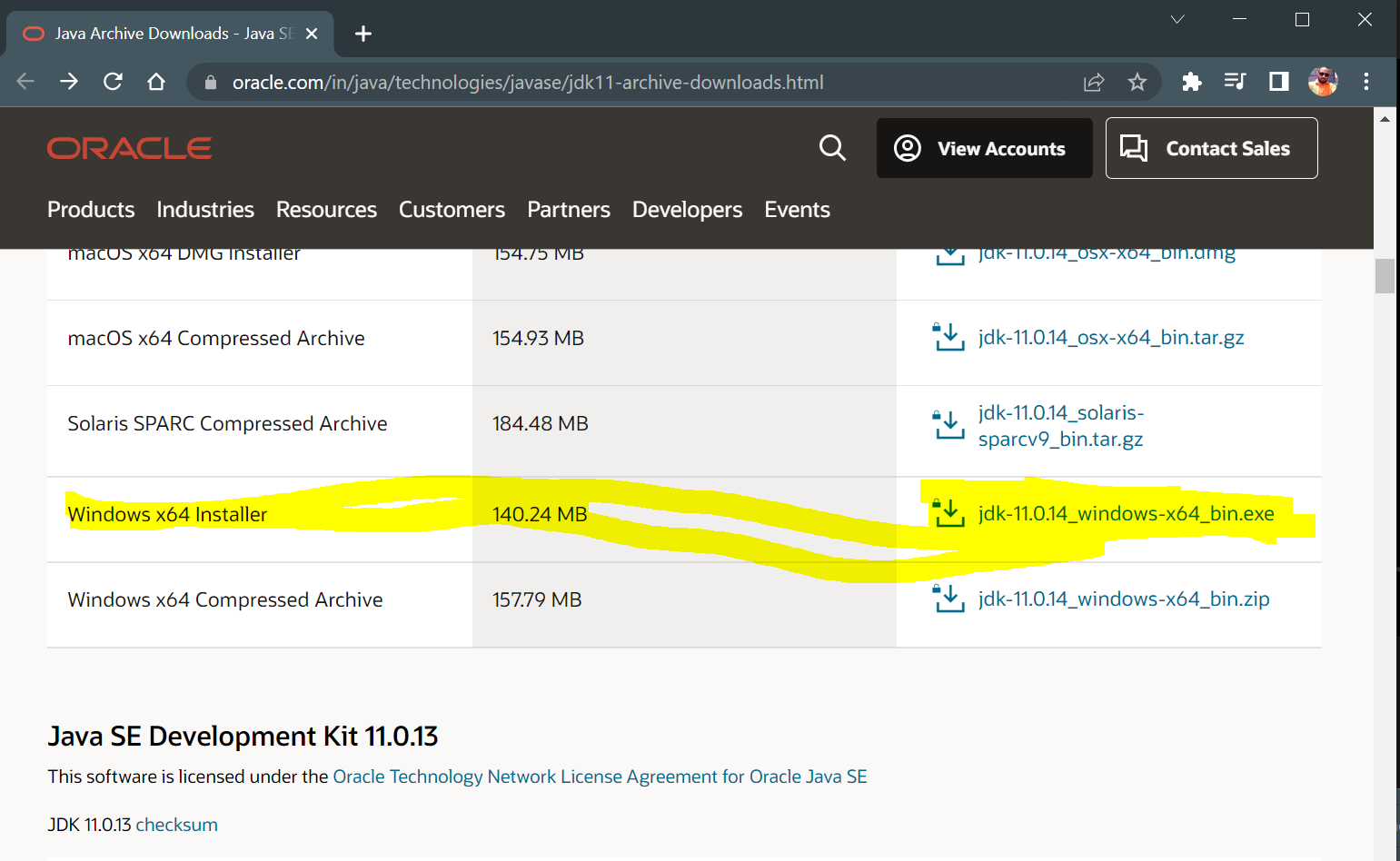
Below 4 files are mandatory to download the software

1. Spark file version (3.2.1)
2. Hadoop file
3. [Java](https://www.oracle.com/in/java/technologies/javase/jdk11-archive-downloads.html) 11 Version
4. Down load the file from GitHub

I will show you how to install Hadoop, spark, Java 11 in windows.

Simple way. Nowadays Hadoop 3.2.2 and spark 3.2.1 are only highly stable.

I recommend install these versions only. Sometimes you will get compatible issues if u use different Hadoop & spark versions. Its recommended: Java 11 spark 3.2.1, hadoop 3.2.2, and python 3.8 (if Pyspark) or Java 1.8.202, Spark 2.4.8 and hadoop 2.7.2 and python 3.6 (if Pyspark)

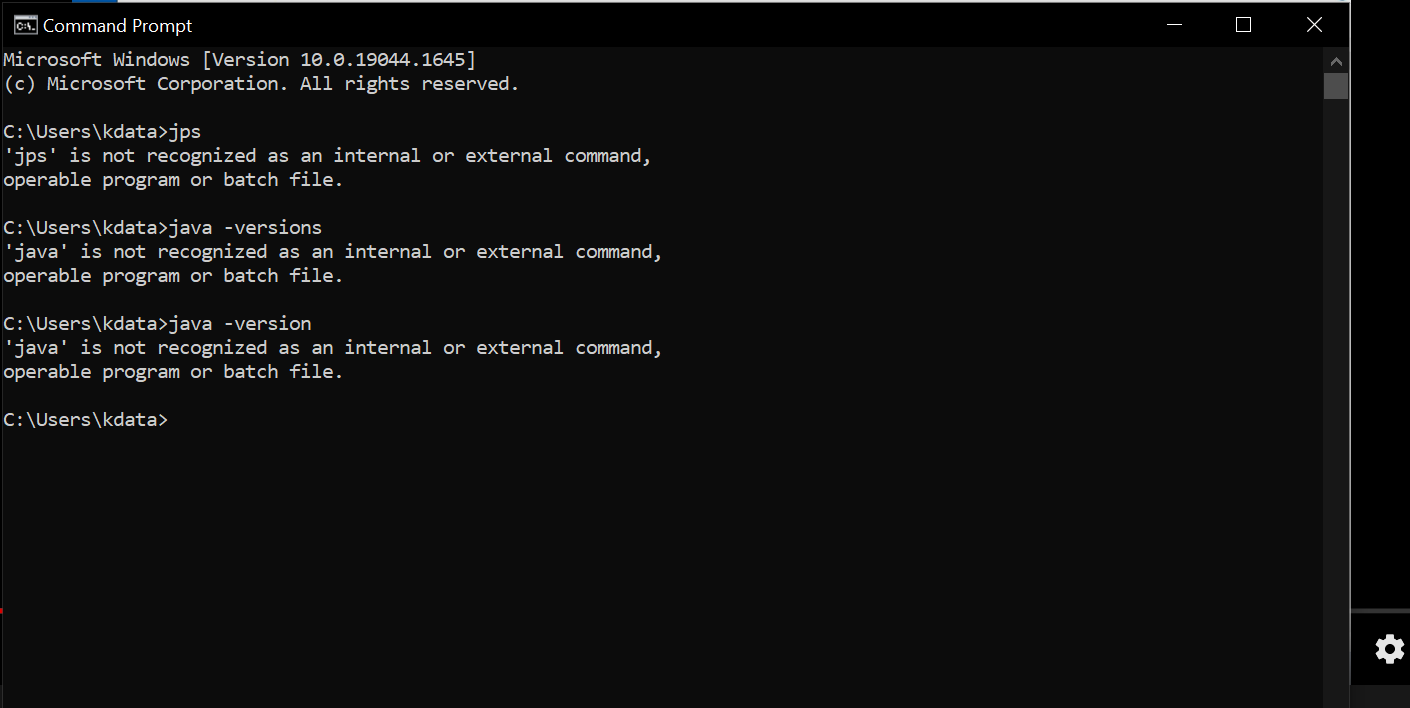


STEP – 1:

- First check java is installed or not

- Go to the terminal – type (jps) means java is not installed

- Only Java 11 is recommended to download these files



-Create the folder data engineer in C-Drive

-Install the Java 11 to C-drive\Data Engineering so that java is installed in Data engineering folder

-After install go to java – jdk – copy the path (e.g- C:\Data Engineering\Java\jdk-11.0.15)

-Create the environment variable – make the variable called JAVA\_HOME & paste the path

-Also change the path to data engineering (path\bin) and save the path (Refer to attached documentation)

- Java installation is completed and please check the java is installed or not

- cmd 🡪 type commans – (java) – (java -version)

STEP-2:

-Install the spark 3.2.1 directly stored in c-drive

-Install the winutils for same version 3.2.1

-Same set the environment variable for both spark & Hadoop

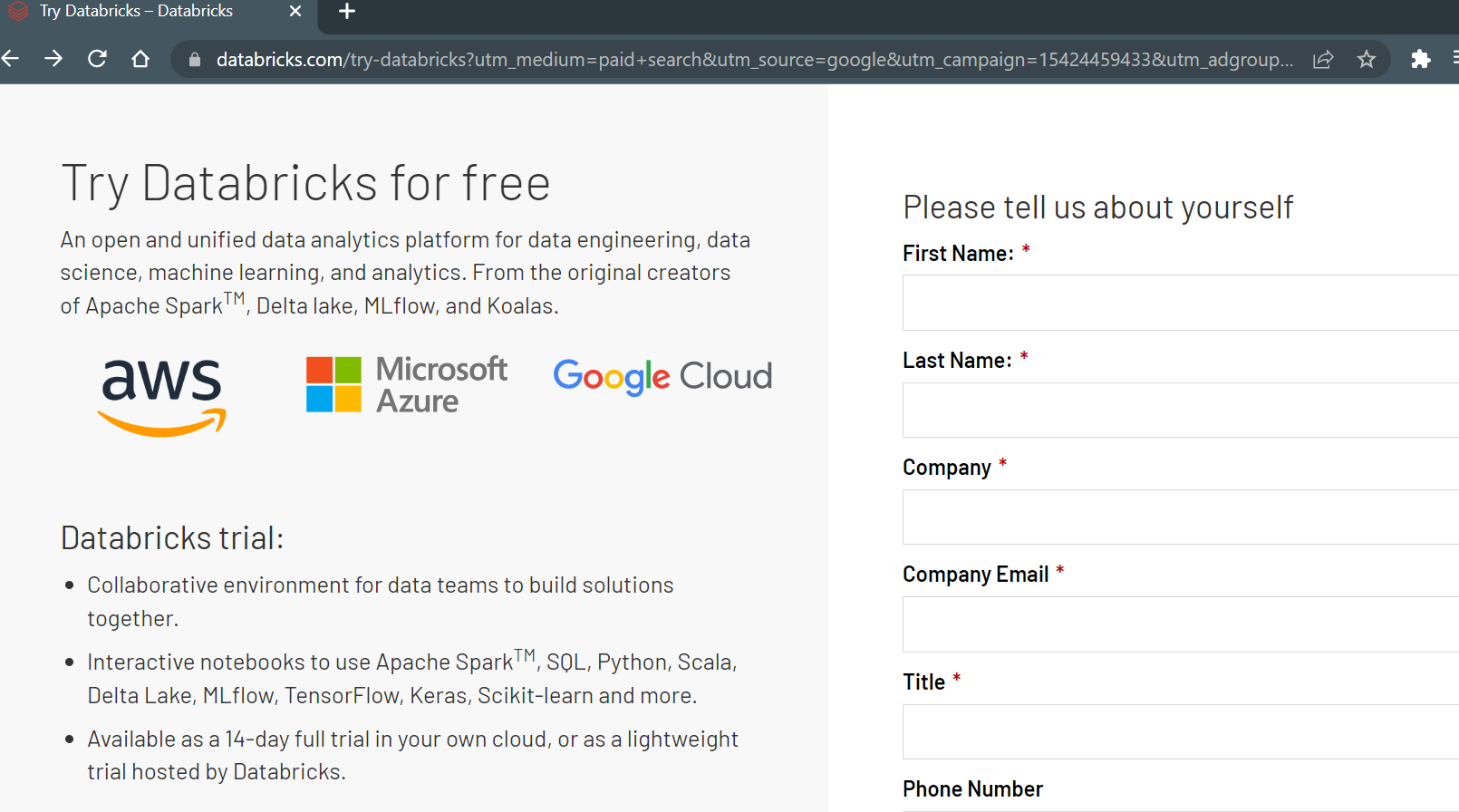
-Install the findspark package from conda

-Open command prompt try to execute spark-shell

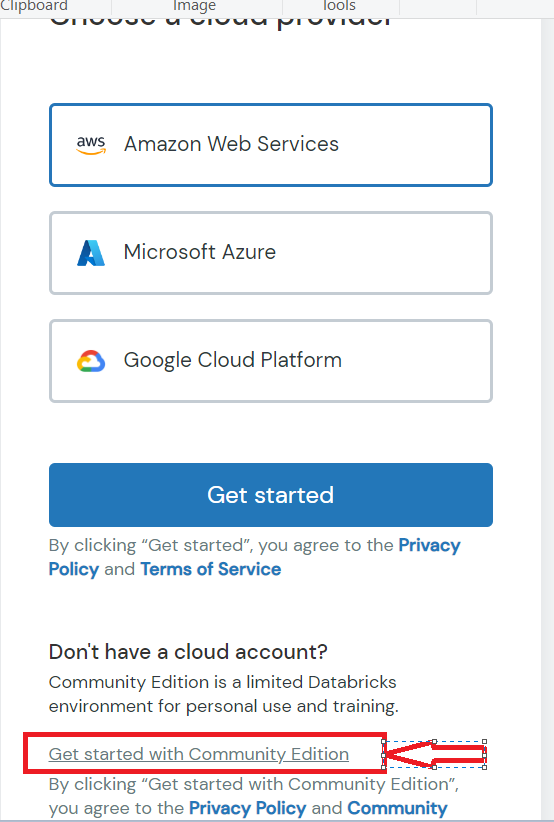
==================================================================================

**DATA BRICKS INSTALLATION 🡪**

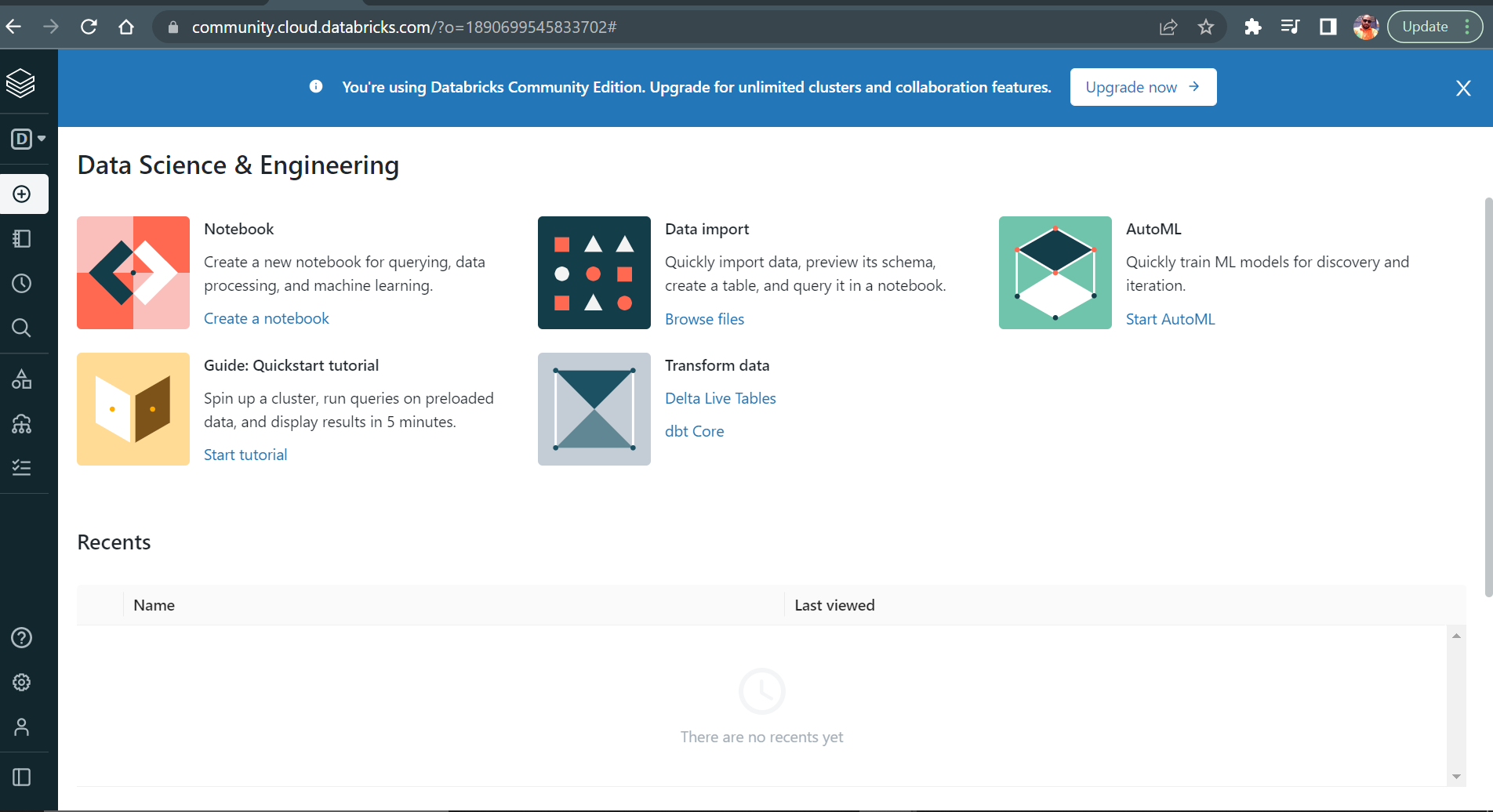
* Databricks is a company that provides clusters that run on top of AWS and adds a convenience of having a Notebook system already set up and the ability to quickly add files
* It has free community version that supports a 6 GB cluster
* It also has its own storage format known as DBFS
* This ‘Table’ format needs to be accessed in a particular way
* Let’s understand how to setup Data Bricks account
* I recommended Databricks for the people who want to quickest online setup



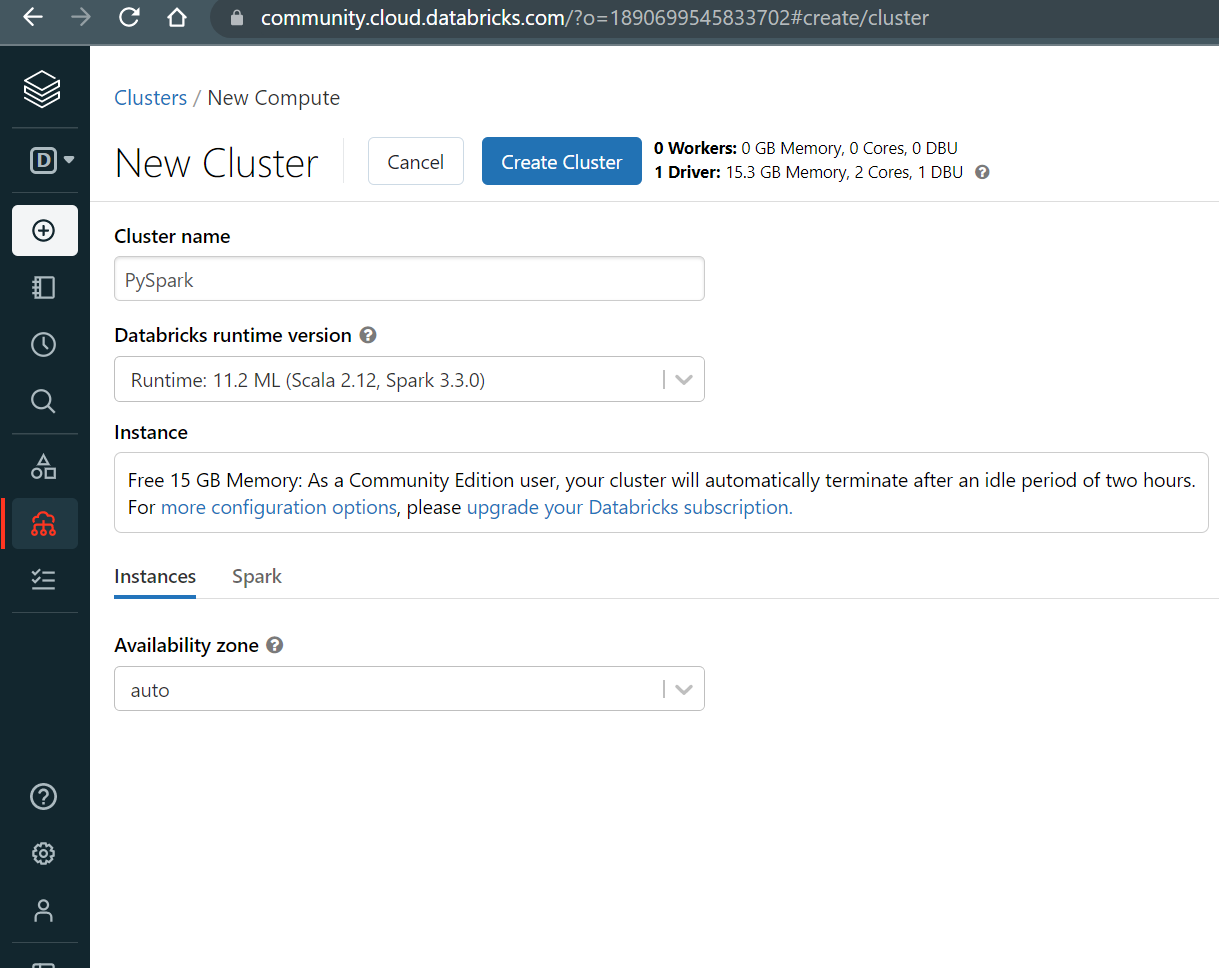
Need to register your details company -personal / no need to enter your phone number – you need to select below option –



Once you click on Community edition – you will get one puzzle game – you need to play well – after you will receive email to registered gmail id – click on the link – you will get below page



Check on left pane – you will get cluster/compute – create cluster – pyspark cluster – you will get below screen shot



Click on data – create table – click on the drop file – it will point towards system folders