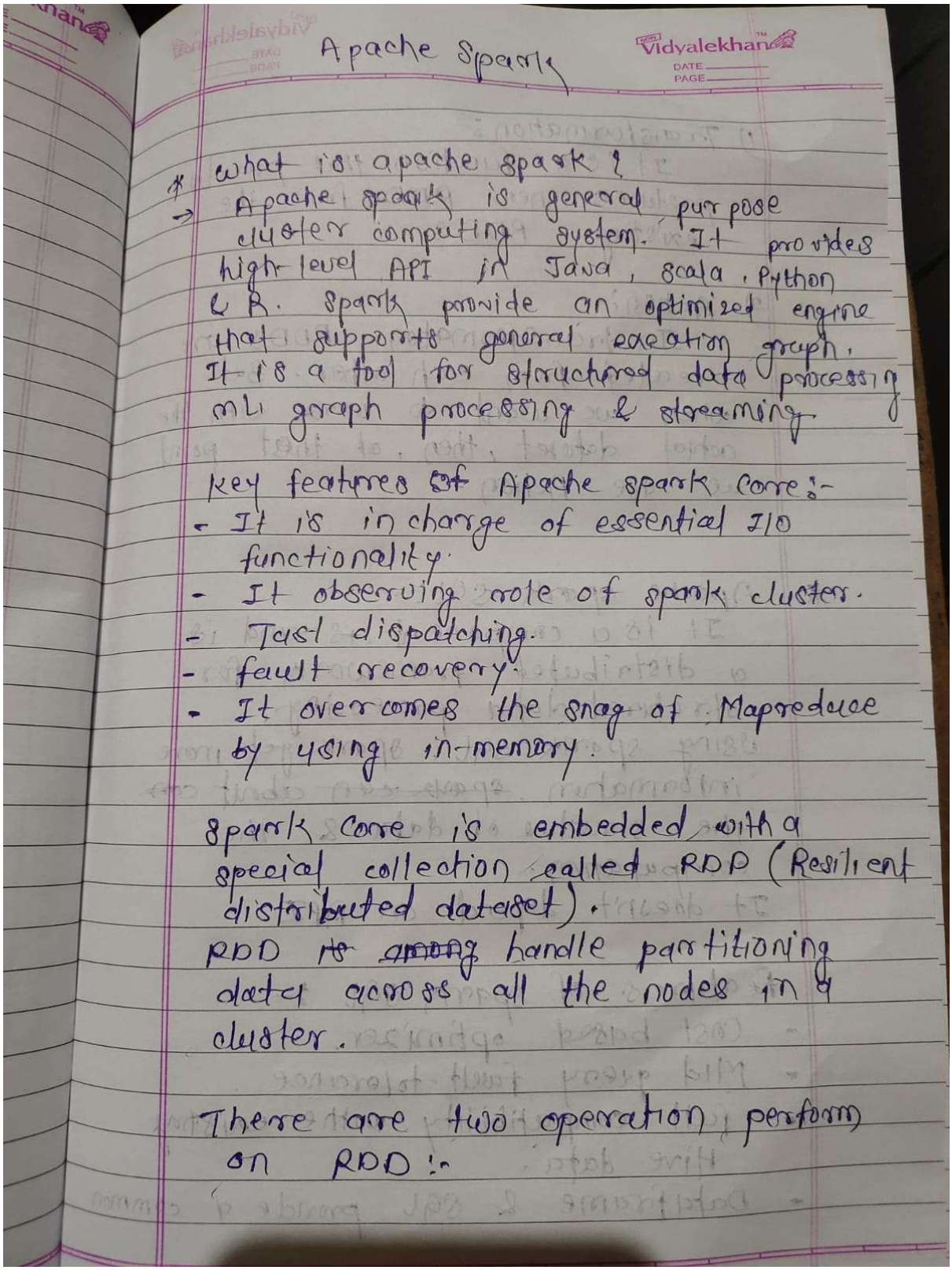
Day 11 Assessment

Name: Krushnakumar Patle

Email: krishnapatle128@gmail.com

Batch: Data Engineering Batch-1



Apache Spark

* what is apache spark ?
→ Apache spark is general purpose cluster computing system. It provides high level API in Java, scala, Python & R. Spark provide an optimized engine that supports general execution graph. It is a tool for structured data processing ML, graph processing & streaming.

key features of Apache spark core :-
- It is in charge of essential I/O functionality.
- It observing role of spark cluster.
- Task dispatching.
- fault recovery.
- It overcomes the snag of Mapreduce by using in-memory.

Spark Core is embedded with a special collection called RDD (Resilient distributed dataset).
RDD ~~is among~~ handle partitioning data across all the nodes in a cluster.

There are two operation perform on RDD :-

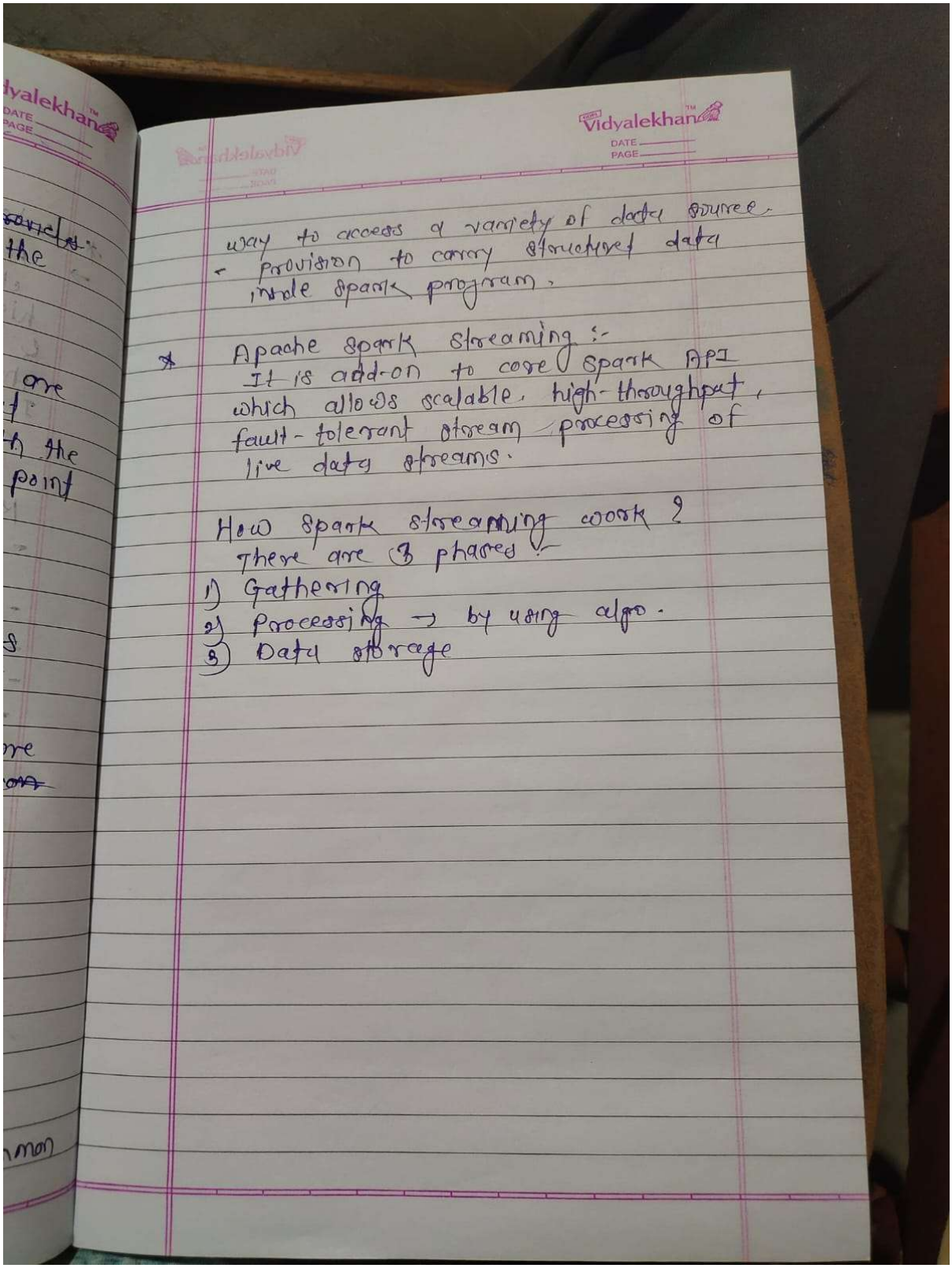
1) Transformation :-
It is a function that produces new RDD from the existing RDDs.

2) Action :-
In transformation, RDDs are created from each other. But when we want to work with the actual dataset, then, at that point we use Action.

* Apache spark SQL :-
It is a component and is a distributed framework for structured data processing.
Using spark SQL, spark get more information about the structure of data & the computation.
It doesn't depend on API

features of spark SQL :-
- Cost based optimizer.
- Mid query fault-tolerance
- full compatibility with the existing Hive data.
- Dataframe & SQL provide a common

Day 11 Assessment



way to access a variety of data source.
- provision to carry structured data inside Spark program.

* Apache Spark Streaming :-
It is add-on to core Spark API which allows scalable, high-throughput, fault-tolerant stream processing of live data streams.

How Spark streaming work ?
There are 3 phases :-
1) Gathering
2) Processing → by using algo.
3) Data storage

Installation Of Apache Spark and Pyspark:

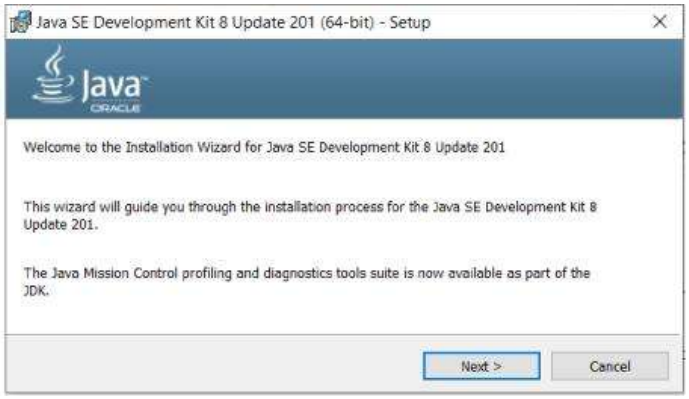
Step 1:

1. Go to the official Java site mentioned below the page.

Accept Licence Agreement for Java SE Development Kit 8u201

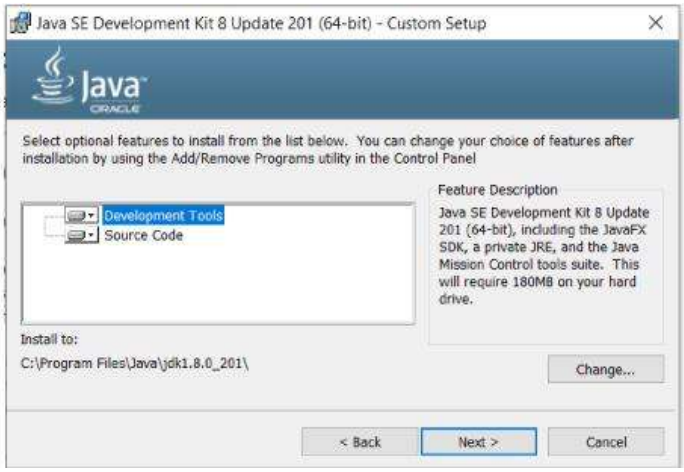
2. Download jdk-8u201-windows-x64.exe file

3. Double Click on the Downloaded .exe file, and you will see the window is shown below.



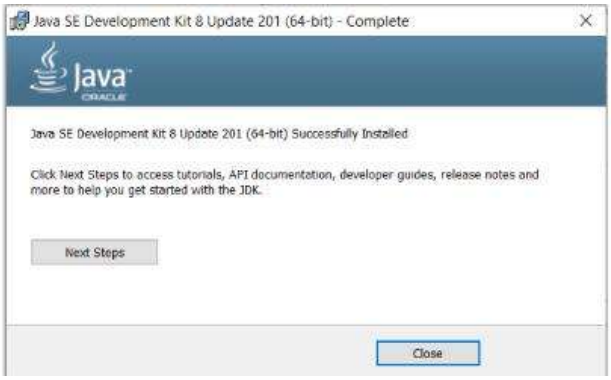
4. Click Next.

5. Then below window will be displayed.



6. Click Next.

7. Below window will be displayed after some process.



8. Click Close.

Step 2: 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.

Step 3:

Go to Apache Spark's official download page and choose the latest release. For the package type, choose ‘Pre-built for Apache Hadoop’.

The page will look like the one below.

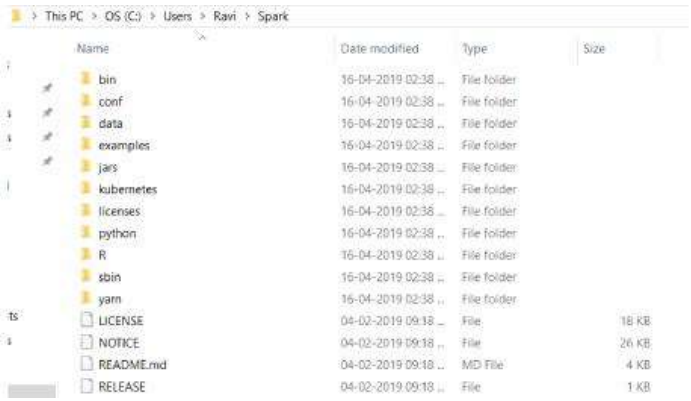
Step 4: Once the download is completed, unzip the file, unzip the file using WinZip or WinRAR, or 7-ZIP.

Step 5: Create a folder called Spark under your user Directory like below and copy and paste the content from the unzipped file.

```
C:\Users\<USER>\Spark
```

Copy Code

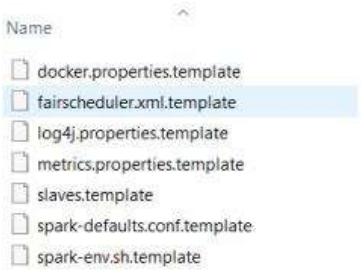
It looks like the below after copy-pasting into the Spark directory.



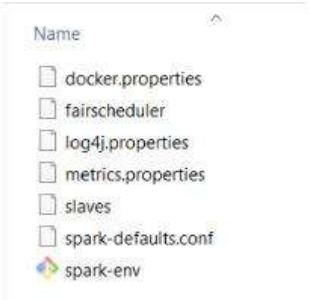
Step 6: Go to the conf folder and open the log file called log4j.properties. template. Change INFO to WARN (It can be an ERROR to reduce the log). This and the next steps are optional.

Remove. template so that Spark can read the file.

Before removing. template all files look like below.



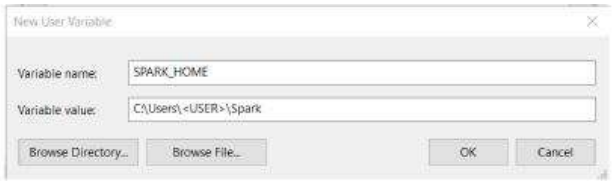
After removing. template extension, files will look like below



Step 7: Now, we need to configure the path.

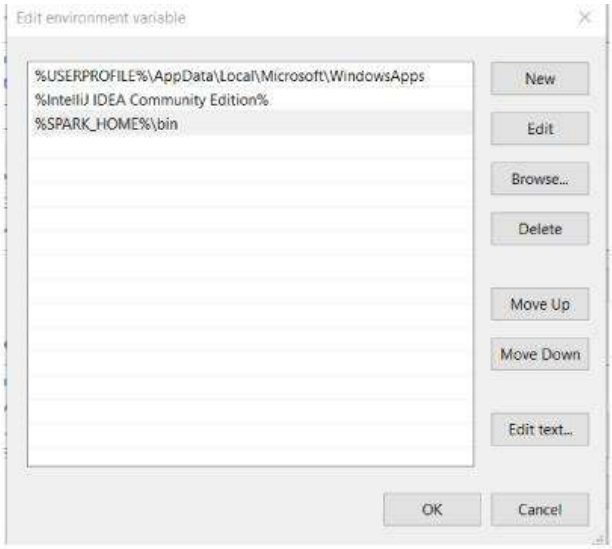
Go to Control Panel -> System and Security -> System -> Advanced Settings -> Environment Variables

Add below new user variable (or System variable) (To add a new user variable, click on the New button under User variable for <USER>)



Click OK.

Add %SPARK_HOME%\bin to the path variable.



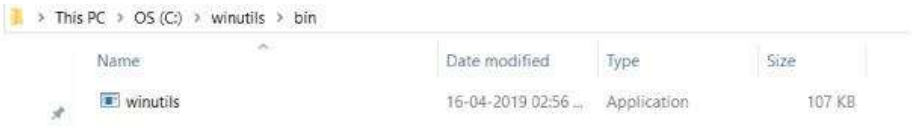
Click OK.

Step 8: Spark needs a piece of Hadoop to run. For Hadoop 2.7, you need to install winutils.exe.

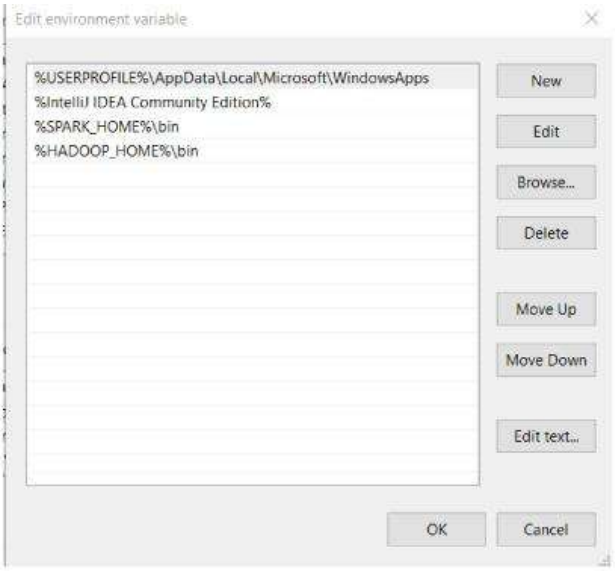
You can find **winutils.exe on this page**. You can download it for your ease.

Step 9: Create a folder called winutils in C drive and create a folder called bin inside. Then, move the downloaded winutils file to the bin folder.

C:\winutils\bin



Add the user (or system) variable %HADOOP_HOME% like SPARK_HOME.



Click OK.

Step 10: After Doing this all we can Run Spark on cmd

Day 11 Assessment