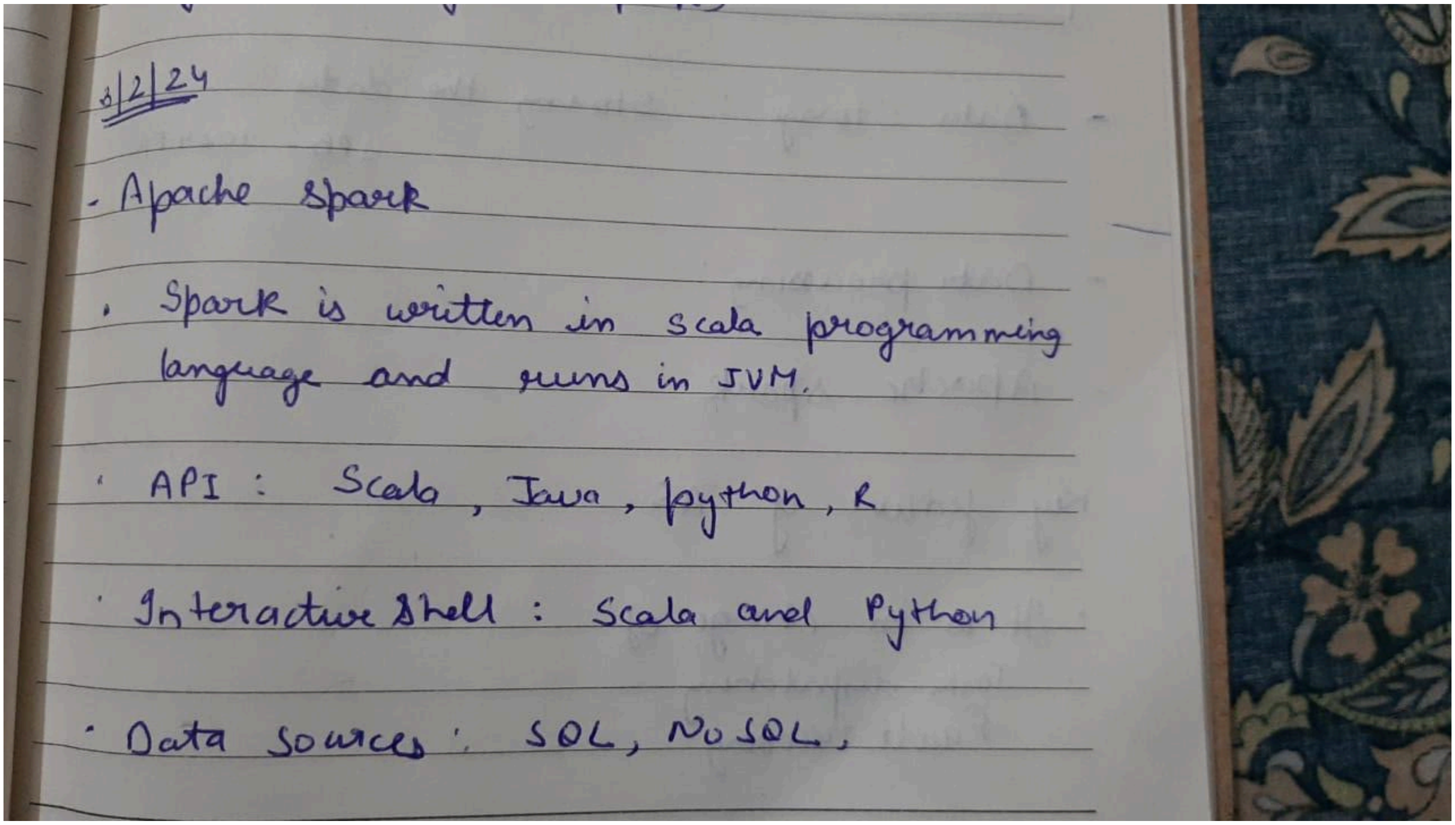
Asmita Porwal
Batch-1
Day-11
3/2/2024
Data engineering

Assignment-11

Hand written notes during the session :



3/2/24

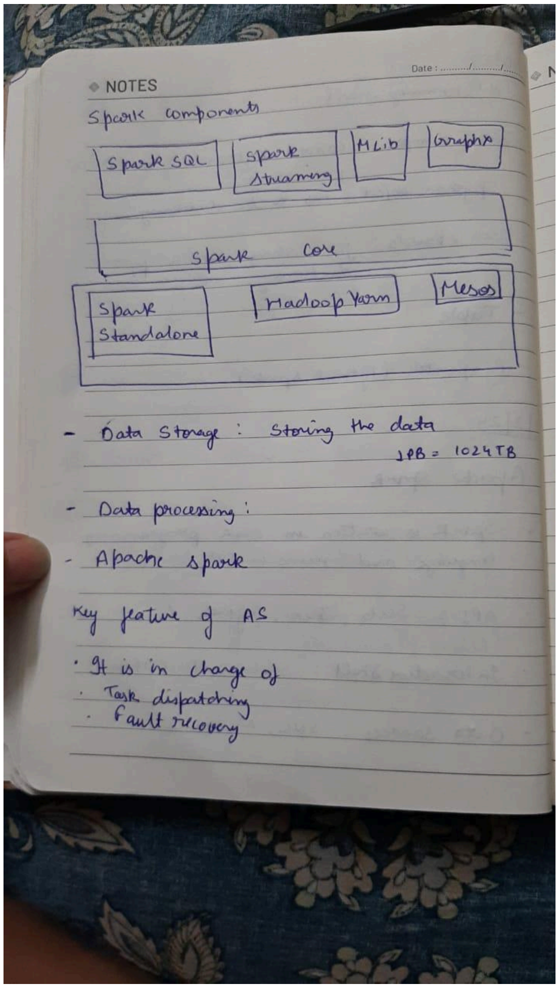
- Apache spark

- Spark is written in scala programming language and runs in JVM.

- API : Scala, Java, python, R

- Interactive Shell : Scala and Python

- Data sources : SQL, NoSQL,



NOTES

Date : ..../......../........

Spark components

Spark SQL

Spark Streaming

MLib

Graphx

Spark core

Spark Standalone

Hadoop Yarn

Mesos

- Data Storage : Storing the data

1PB = 1024TB

- Data processing :

- Apache spark

Key feature of AS

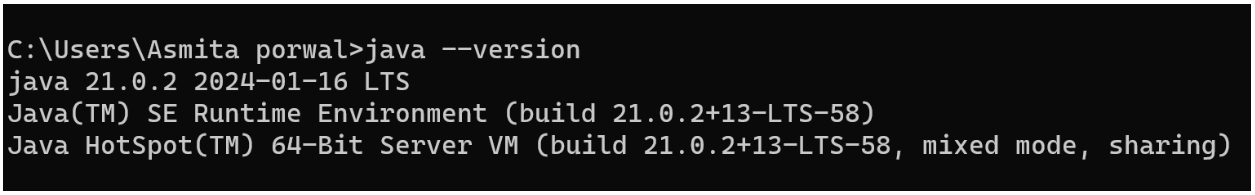
· It is in charge of

· Task dispatching

· Fault recovery

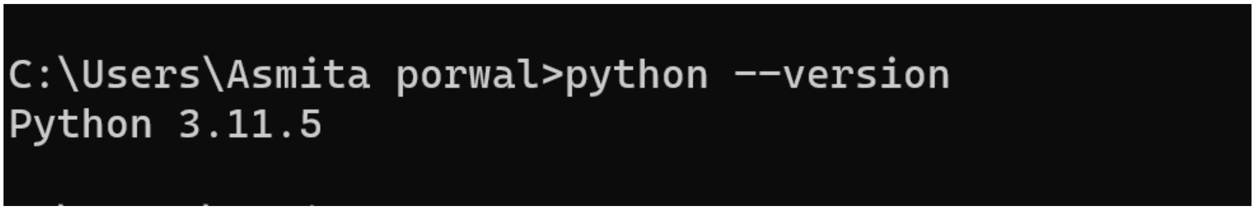
Installing Required Softwares :

Java : As java was already installed on my computer, I checked if it is working fine or not.



```
C:\Users\Asmita porwal>java --version
java 21.0.2 2024-01-16 LTS
Java(TM) SE Runtime Environment (build 21.0.2+13-LTS-58)
Java HotSpot(TM) 64-Bit Server VM (build 21.0.2+13-LTS-58, mixed mode, sharing)
```

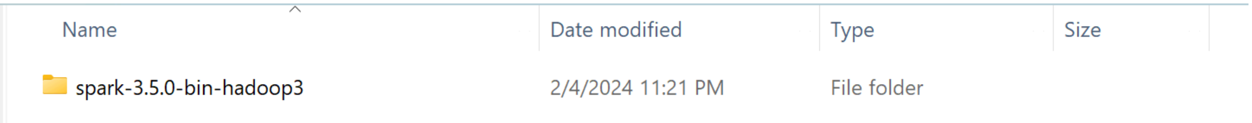
Python : Python was also installed in the system and checked version for it.



```
C:\Users\Asmita porwal>python --version
Python 3.11.5
```

Apache PySpark :

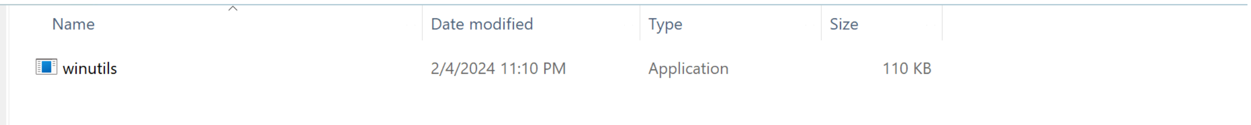
From URL :
https://www.apache.org/dyn/closer.lua/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz
Downloaded the zip file :



Name	Date modified	Type	Size
spark-3.5.0-bin-hadoop3	2/4/2024 11:21 PM	File folder	

For hadoop :

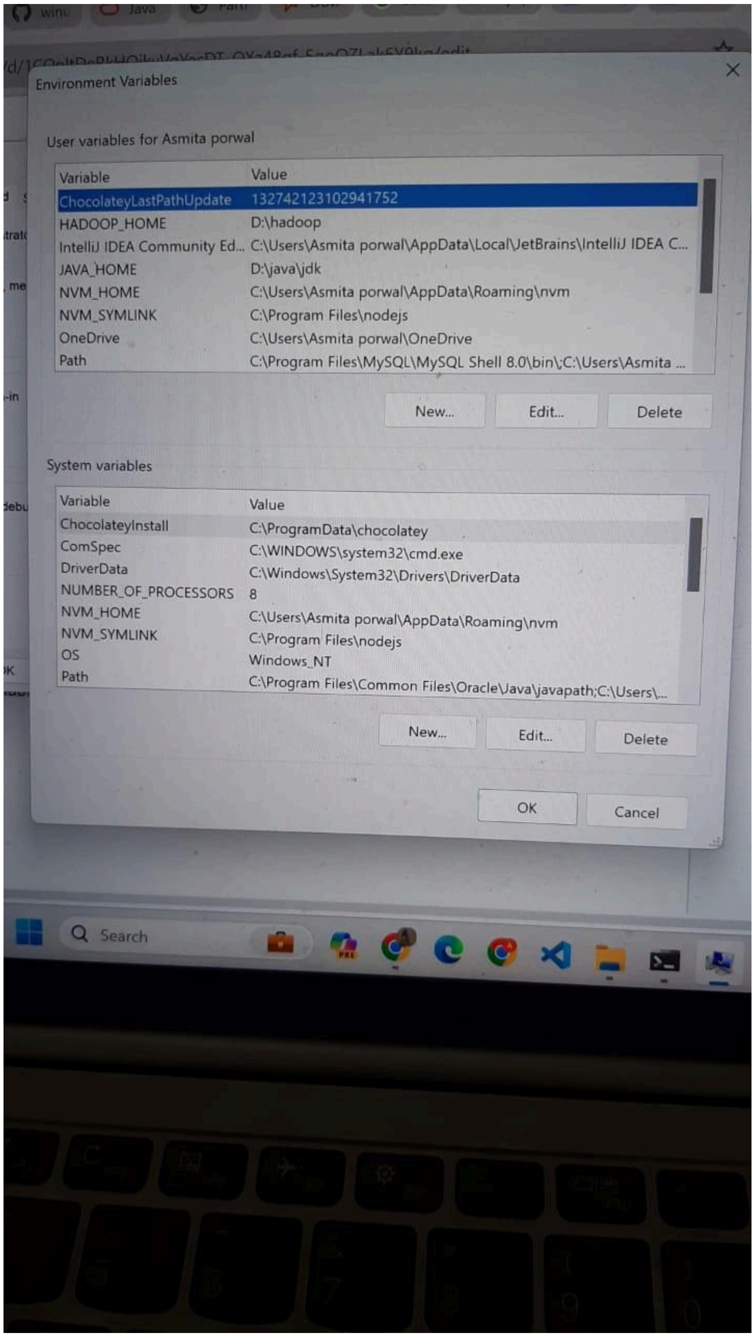
Downloaded the winutils.exe file :



Name	Date modified	Type	Size
winutils	2/4/2024 11:10 PM	Application	110 KB

Setting environment variables :

Added environment variables in system variables :



Environment Variables

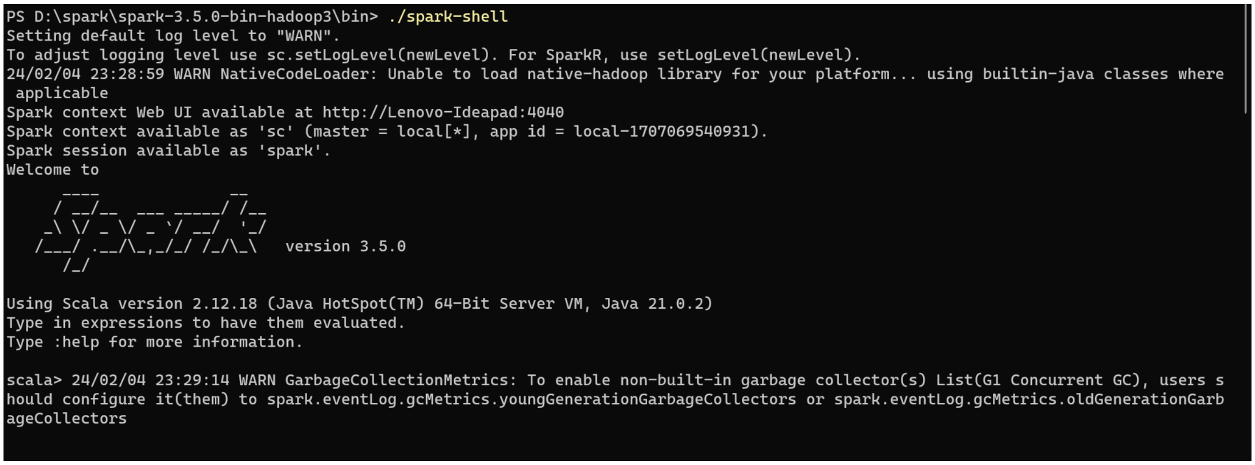
User variables for Asmita porwal

Variable	Value
ChocolateyLastPathUpdate	132742123102941752
HADOOP_HOME	D:\hadoop
IntelliJ IDEA Community Ed...	C:\Users\Asmita porwal\AppData\Local\JetBrains\IntelliJ IDEA C...
JAVA_HOME	D:\java\jdk
NVM_HOME	C:\Users\Asmita porwal\AppData\Roaming\nvm
NVM_SYMLINK	C:\Program Files\nodejs
OneDrive	C:\Users\Asmita porwal\OneDrive
Path	C:\Program Files\MySQL\MySQL Shell 8.0\bin\;C:\Users\Asmita ...

System variables

Variable	Value
ChocolateyInstall	C:\ProgramData\chocolatey
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
NUMBER_OF_PROCESSORS	8
NVM_HOME	C:\Users\Asmita porwal\AppData\Roaming\nvm
NVM_SYMLINK	C:\Program Files\nodejs
OS	Windows_NT
Path	C:\Program Files\Common Files\Oracle\Java\javapath;C:\Users\...

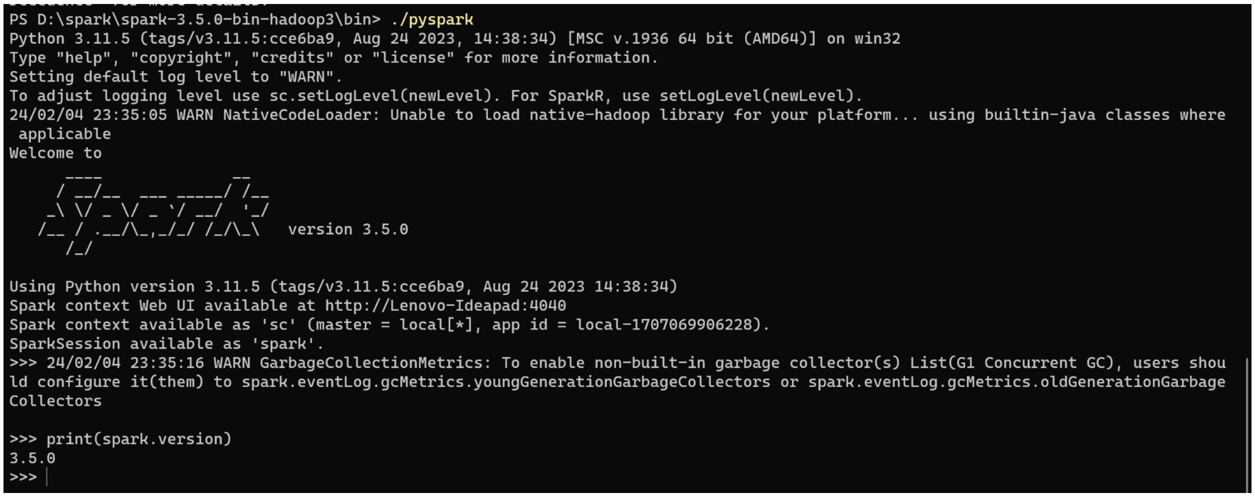
Then on Powershell started shell service :



```
PS D:\spark\spark-3.5.0-bin-hadoop3\bin> ./spark-shell
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/02/04 23:28:59 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
 applicable
Spark context Web UI available at http://Lenovo-Ideapad:4040
Spark context available as 'sc' (master = local[*], app id = local-1707069540931).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.5.0
      /_/

Using Scala version 2.12.18 (Java HotSpot(TM) 64-Bit Server VM, Java 21.0.2)
Type in expressions to have them evaluated.
Type :help for more information.

scala> 24/02/04 23:29:14 WARN GarbageCollectionMetrics: To enable non-built-in garbage collector(s) List(G1 Concurrent GC), users s
hould configure it(them) to spark.eventLog.gcMetrics.youngGenerationGarbageCollectors or spark.eventLog.gcMetrics.oldGenerationGarb
ageCollectors
```

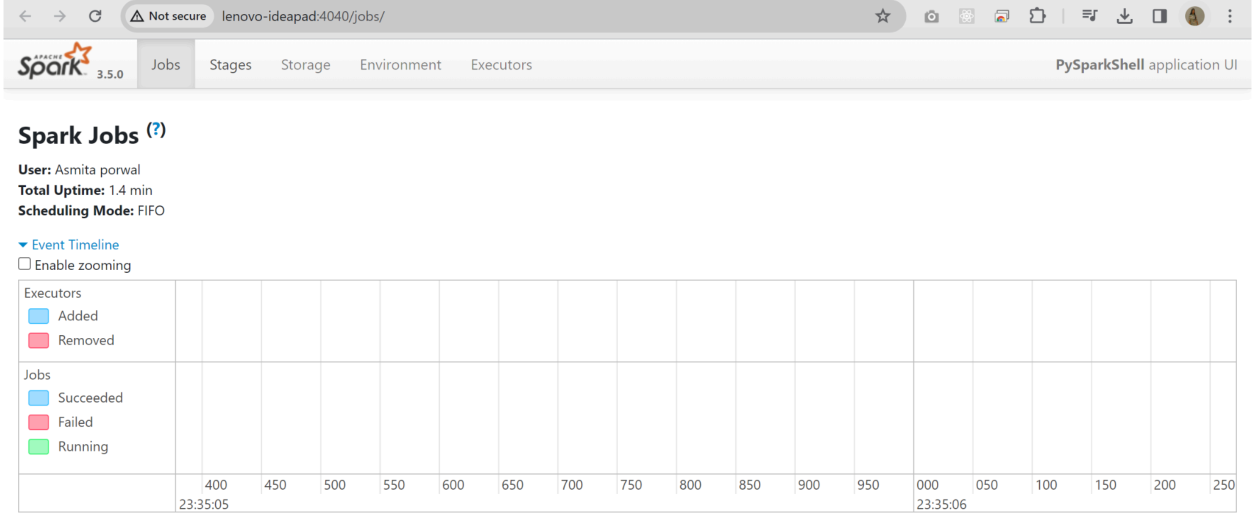


```
PS D:\spark\spark-3.5.0-bin-hadoop3\bin> ./pyspark
Python 3.11.5 (tags/v3.11.5:cce6ba9, Aug 24 2023, 14:38:34) [MSC v.1936 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/02/04 23:35:05 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
 applicable
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.5.0
      /_/

Using Python version 3.11.5 (tags/v3.11.5:cce6ba9, Aug 24 2023 14:38:34)
Spark context Web UI available at http://Lenovo-Ideapad:4040
Spark context available as 'sc' (master = local[*], app id = local-1707069906228).
SparkSession available as 'spark'.
>>> 24/02/04 23:35:16 WARN GarbageCollectionMetrics: To enable non-built-in garbage collector(s) List(G1 Concurrent GC), users shou
ld configure it(them) to spark.eventLog.gcMetrics.youngGenerationGarbageCollectors or spark.eventLog.gcMetrics.oldGenerationGarbage
Collectors

>>> print(spark.version)
3.5.0
>>>
```

Then on browser by entering URL : http://lenovo-ideapad:4040/jobs/



Spark Jobs (?)

User: Asmita porwal
Total Uptime: 1.4 min
Scheduling Mode: FIFO

▾ Event Timeline

□ Enable zooming