

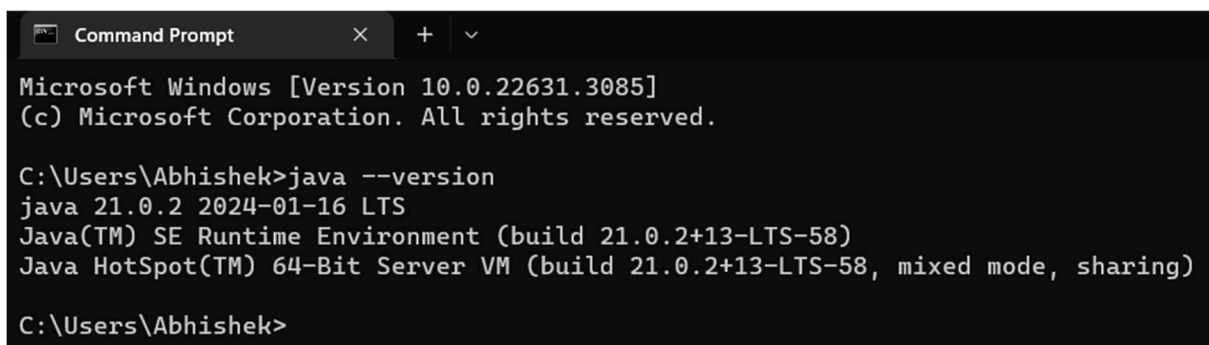
Name: Abhishek Kanoujia

DATA ENGINEERING BATCH 1

DAY 12 ASSIGNMENT

Installing Required Softwares : -

Java : As java was already installed

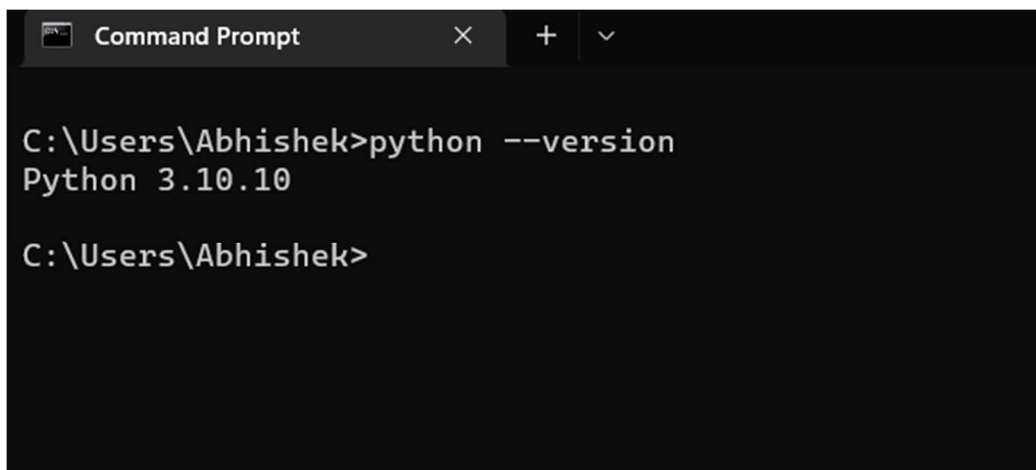


```
Command Prompt
Microsoft Windows [Version 10.0.22631.3085]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Abhishek>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)

C:\Users\Abhishek>
```

- Python : Python was also installed



```
Command Prompt

C:\Users\Abhishek>python --version
Python 3.10.10

C:\Users\Abhishek>
```

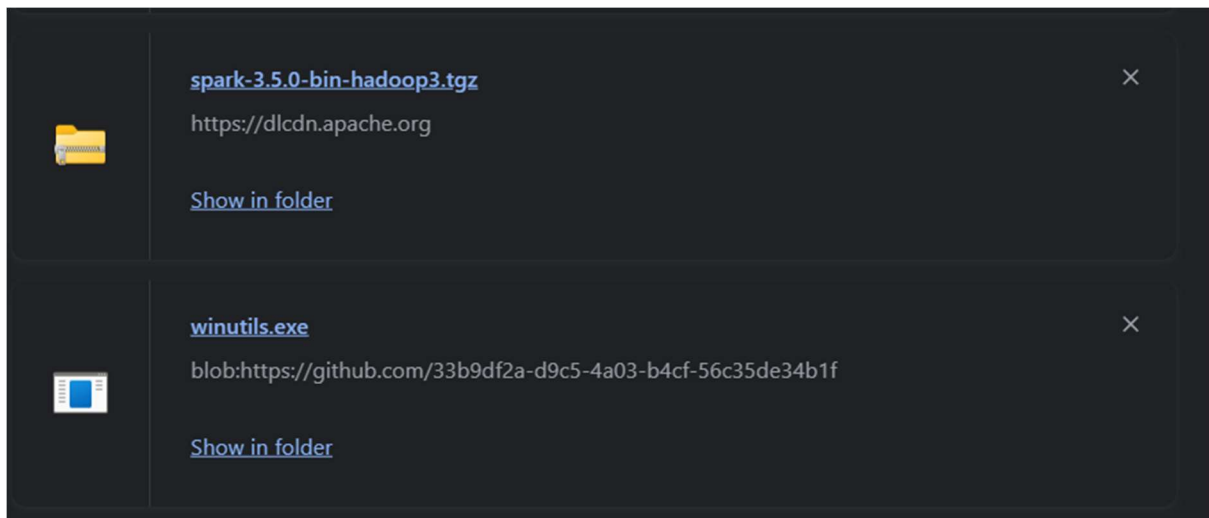
Apache PySpark :

From URL : <https://www.apache.org/dyn/closer.lua/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz>

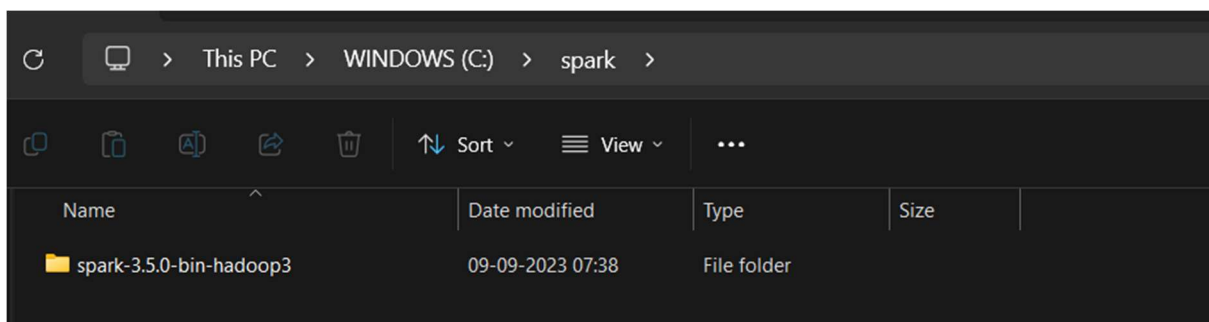
Downloaded the zip file :

For hadoop :

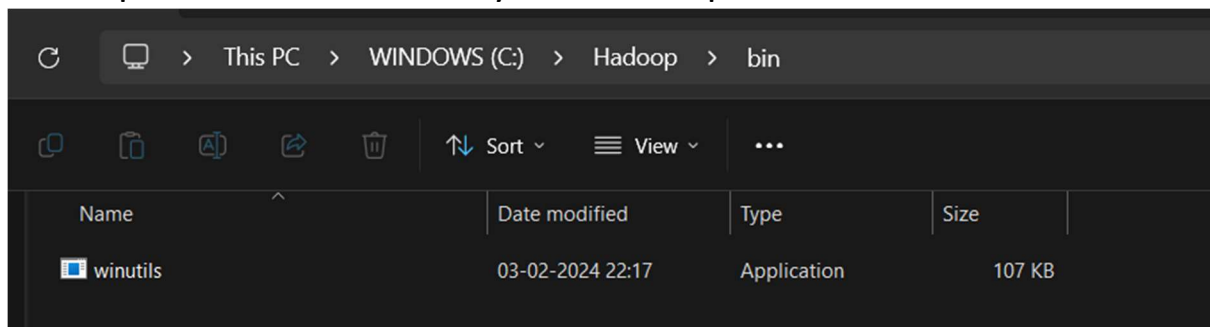
Downloaded the winutils.exe file :



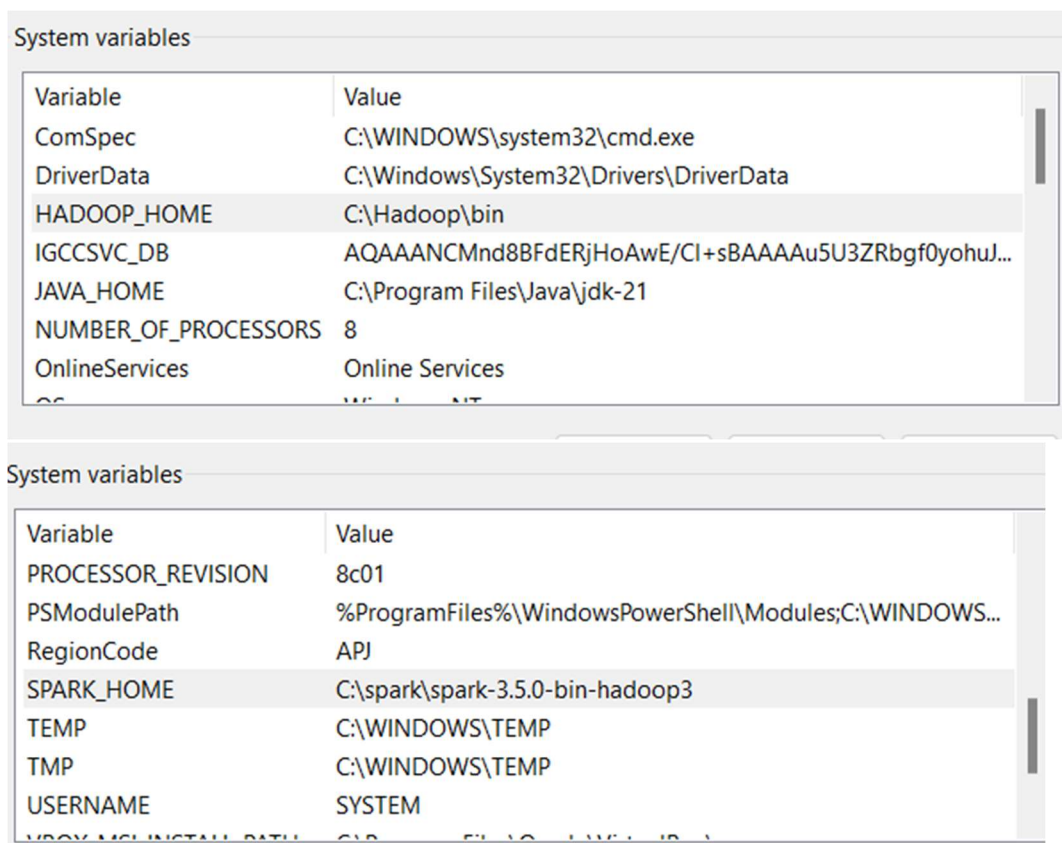
Then extracted the tbz file in my C:/Spark/



hadoop file created directory : C:/Hadoop/bin :



Setting environment variables :



Then on Powershell started shell service :

```
Windows PowerShell
PS C:\spark\spark-3.5.0-bin-hadoop3\bin> ./spark-shell
24/02/03 22:28:18 WARN Shell: Did not find winutils.exe: java.io.FileNotFoundException: Hadoop bin directory does not exist: C:\Hadoop\bin\bin -see https://
wiki.apache.org/hadoop/WindowsProblems
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/02/03 22:28:22 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Spark context Web UI available at http://LAPTOP-42FDQ6JC:4040
Spark context available as 'sc' (master = local[*], app id = local-17069795037744).
Spark session available as 'spark'.
Welcome to

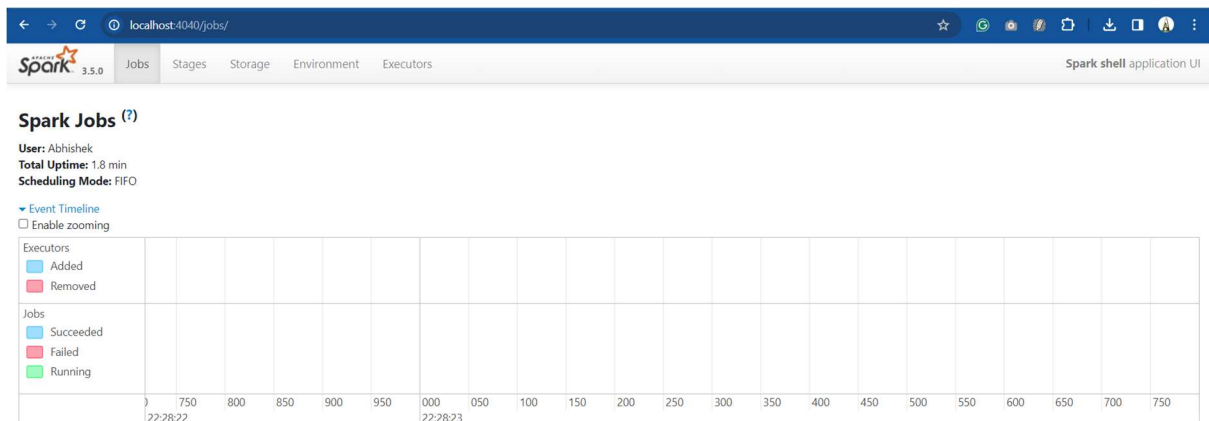
  ____      _
 / ___|    / \
| |  | |  / _ \
| |  | | / ___ \
| |  | |/_/   \_\
|_|  |_____/___\_\

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> 24/02/03 22:28:39 WARN GarbageCollectionMetrics: To enable non-built-in garbage collector(s) List(G1 Concurrent GC), users should configure it(them)
to spark.eventLog.gcMetrics.youngGenerationGarbageCollectors or spark.eventLog.gcMetrics.oldGenerationGarbageCollectors
```

URL : <http://localhost:4040/jobs/> :



(Saturday)

Date _____

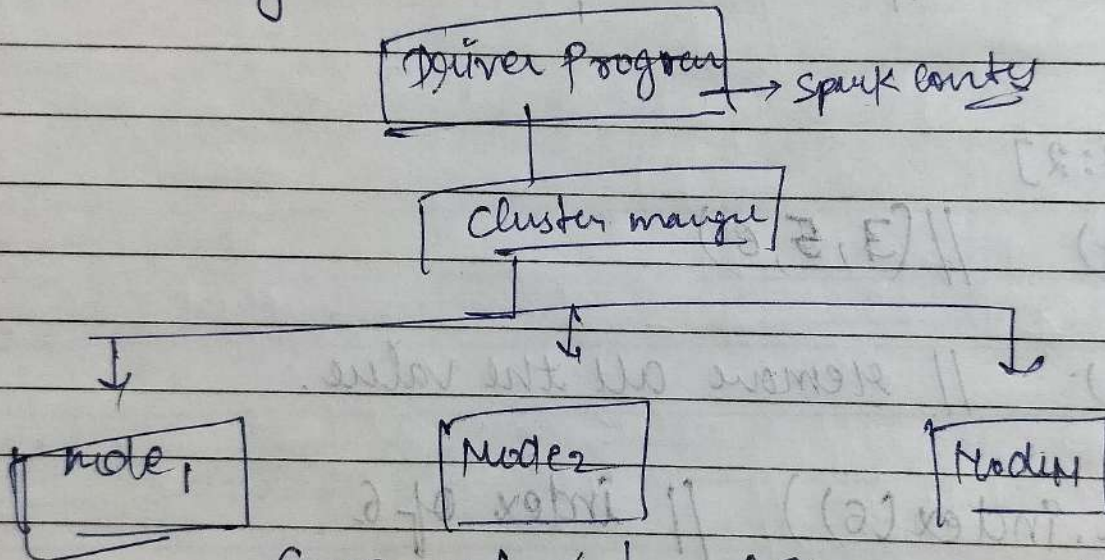
(PySpark)

Q. Introduction of Spark

(a. p

(b. Processing

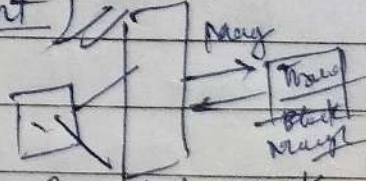
(c. data storage.



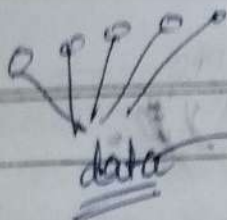
Core Concept

- Job
- Stage
- Task
- DAG
- Executor
- Master
- Slave

(Spark Component)

- (1) Spark driver : 
- (2) Spark Context : Connect to Spark cluster.
- (3) DAG Scheduler : Compute DAG → TS
- (4) Task Scheduler : Send task to cluster.
- (5) Scheduler Backend : backend

(Interface)



cluster manager

spark cluster → different

cluster → group of mach

SQL + Python

1000 GB of data → store in → cluster

(Stream data)
(Banking data)

(Apache spark streaming)

Spark component:

RDD ?

1) spark driver

2) Spark context: Represent connection to Spark Cluster, Create RDD

3) DAG Scheduler: Compute DAG, sent to task scheduler.

4) Task Scheduler: send task to cluster

5) Scheduler Backend: Backend for Scheduler.

What is DAG?

Concept of Airflow and collection task together.

declare: { id =
start date =
task =

1) Spark standalone

2) Yarn.

3) Mesos.

How Spark Work?

