

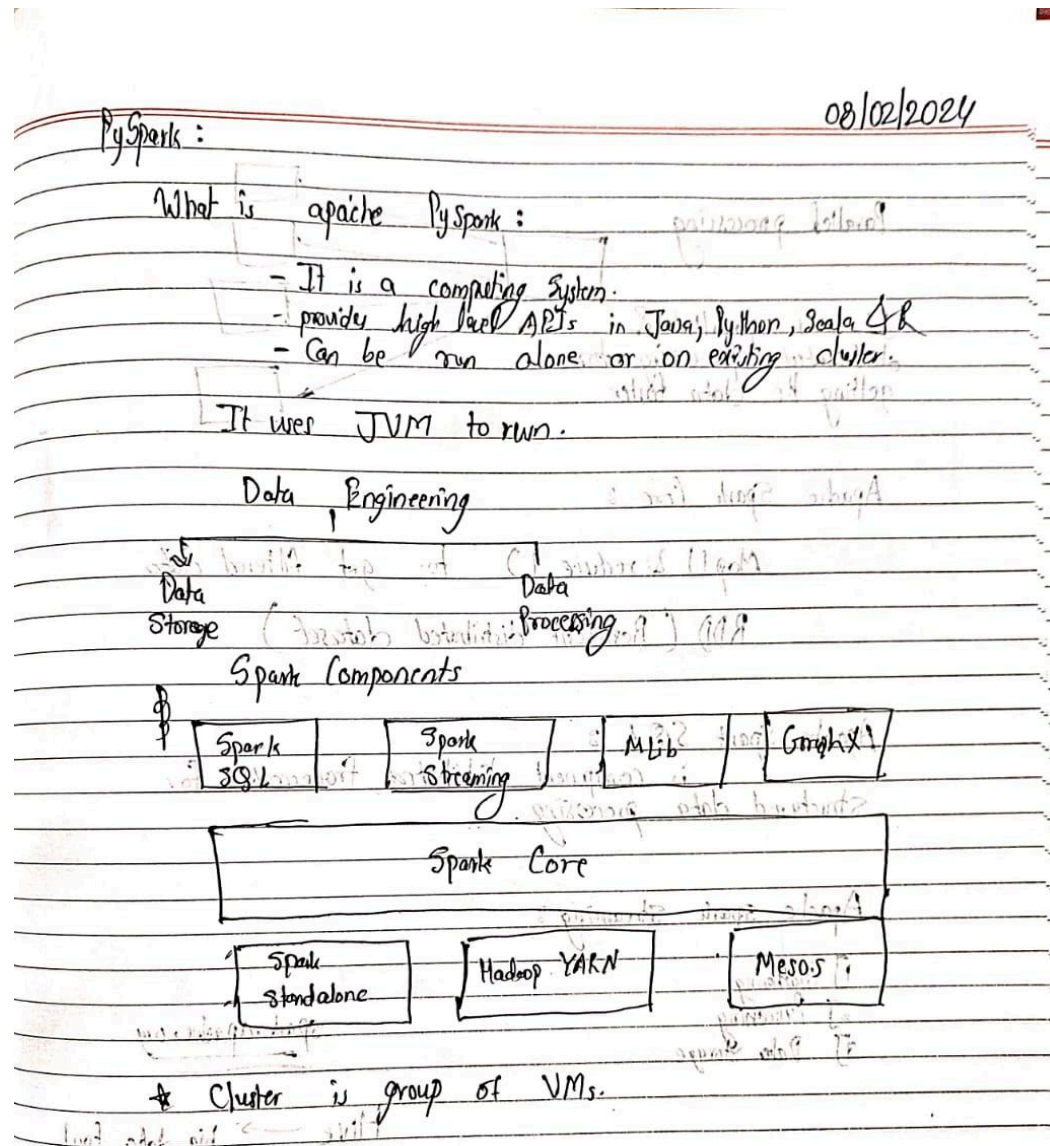
Name : Parth Nandedkar

Date : 02 Feb 2024

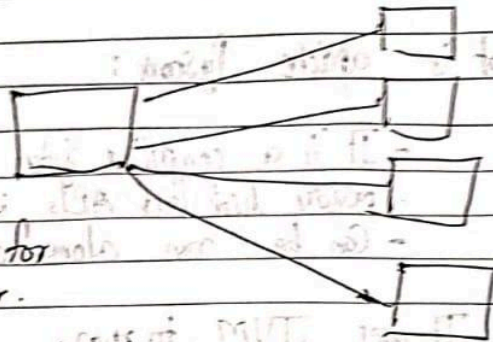
Topics : Introduction to Pyspark

Batch : Data Engineering Batch-1

Hand written notes during the session :



Parallel processing



It uses parallel & distributed processing for getting the data faster.

Apache Spark Core :

Map() & reduce() to get Altered data

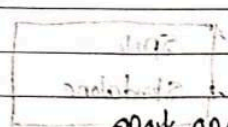
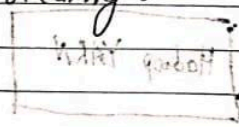
RDD (Resilient distributed dataset)

Apache Spark SQL :

is component distributed framework for structured data processing.

Apache Spark Streaming :

- 1] Gathering
- 2] Processing
- 3] Data Storage



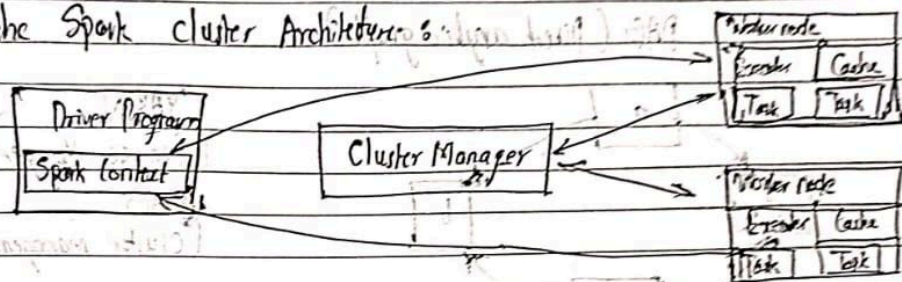
spark.apache.org

Hive → big data tool

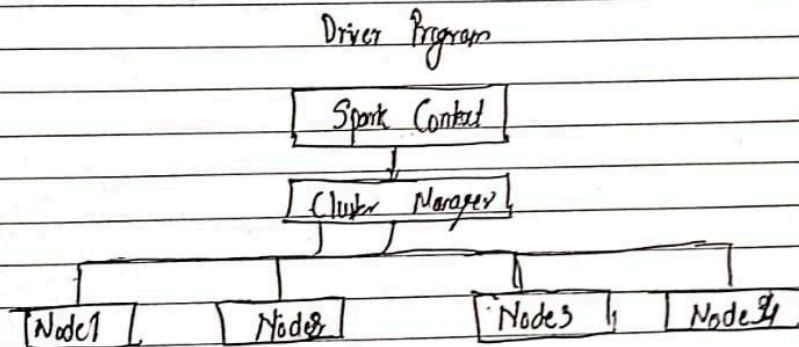
integrating hadoop & pyspark :

D stream in spark :
high level abstraction

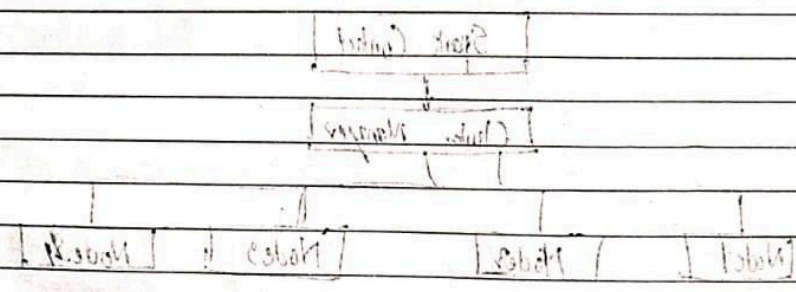
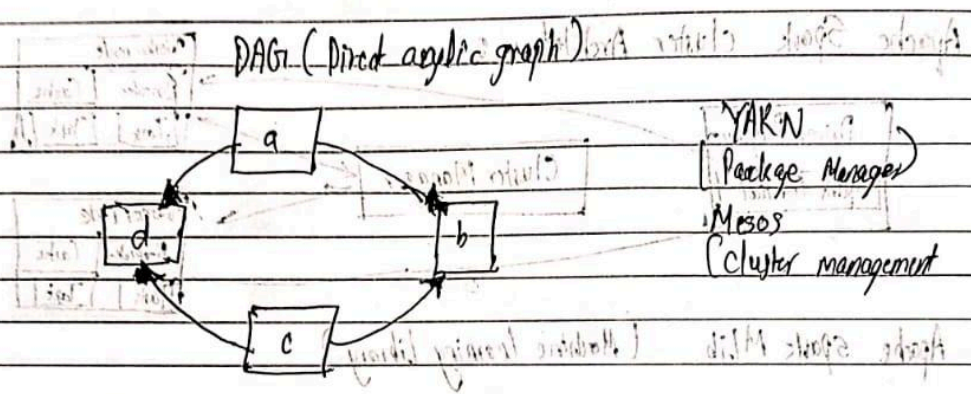
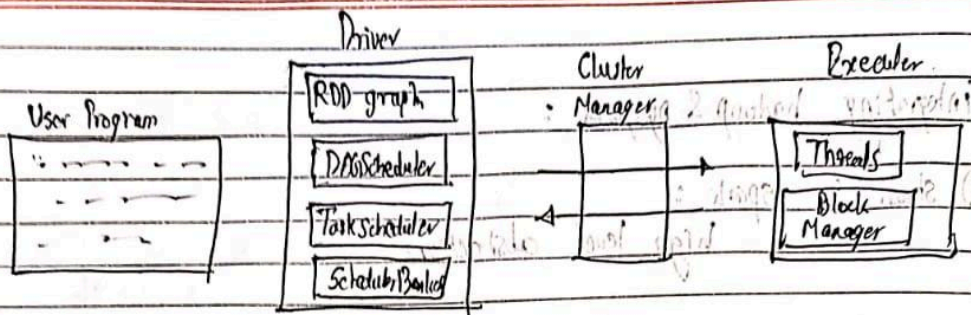
Apache Spark cluster Architecture :



Apache Spark MLlib (Machine Learning Library)



Core components → map & reduce

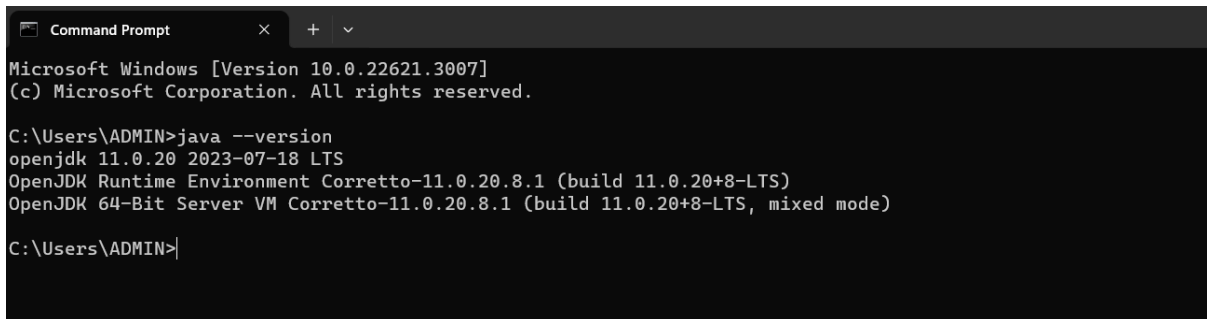


Cluster Manager & Driver

Installing Required Softwares :

- Java :

As java was already installed on my computer checked as if it is working fine or not.



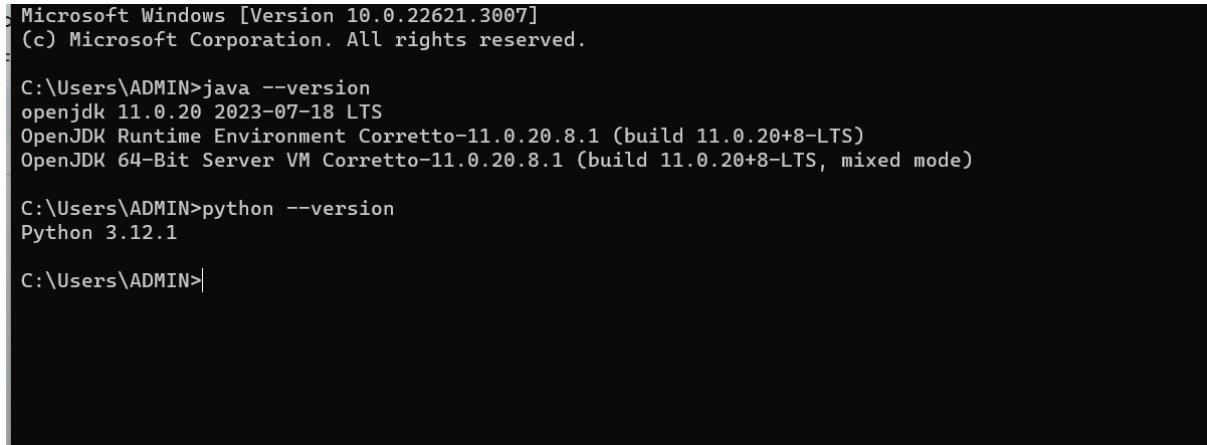
```
Command Prompt
Microsoft Windows [Version 10.0.22621.3007]
(c) Microsoft Corporation. All rights reserved.

C:\Users\ADMIN>java --version
openjdk 11.0.20 2023-07-18 LTS
OpenJDK Runtime Environment Corretto-11.0.20.8.1 (build 11.0.20+8-LTS)
OpenJDK 64-Bit Server VM Corretto-11.0.20.8.1 (build 11.0.20+8-LTS, mixed mode)

C:\Users\ADMIN>
```

- Python :

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



```
Microsoft Windows [Version 10.0.22621.3007]
(c) Microsoft Corporation. All rights reserved.

C:\Users\ADMIN>java --version
openjdk 11.0.20 2023-07-18 LTS
OpenJDK Runtime Environment Corretto-11.0.20.8.1 (build 11.0.20+8-LTS)
OpenJDK 64-Bit Server VM Corretto-11.0.20.8.1 (build 11.0.20+8-LTS, mixed mode)

C:\Users\ADMIN>python --version
Python 3.12.1

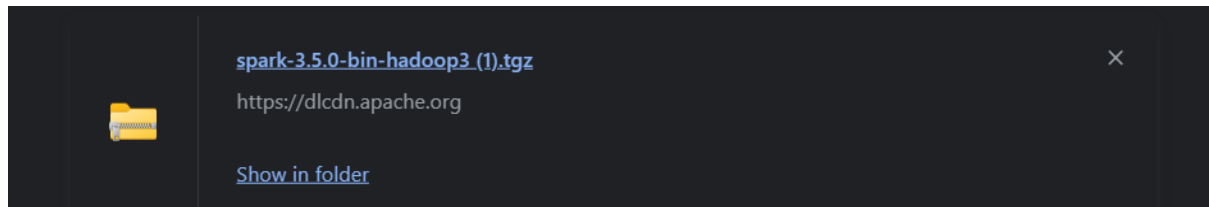
C:\Users\ADMIN>
```

Apache PySpark :

From URL :

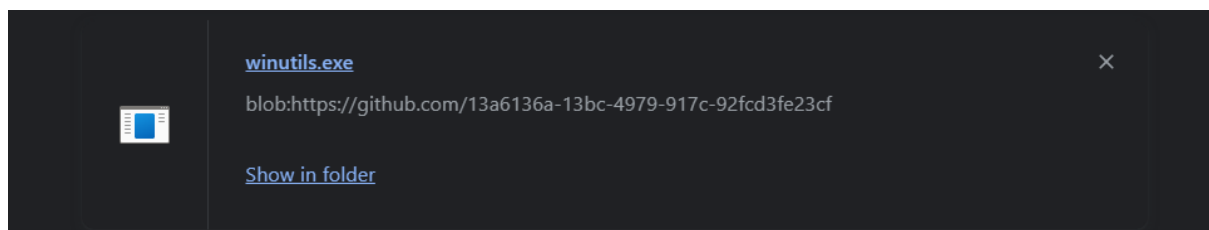
<https://www.apache.org/dyn/closer.lua/spark/spark-3.5.0/spark-3.5.0-bin-hadoop3.tgz>

Downloaded the zip file :

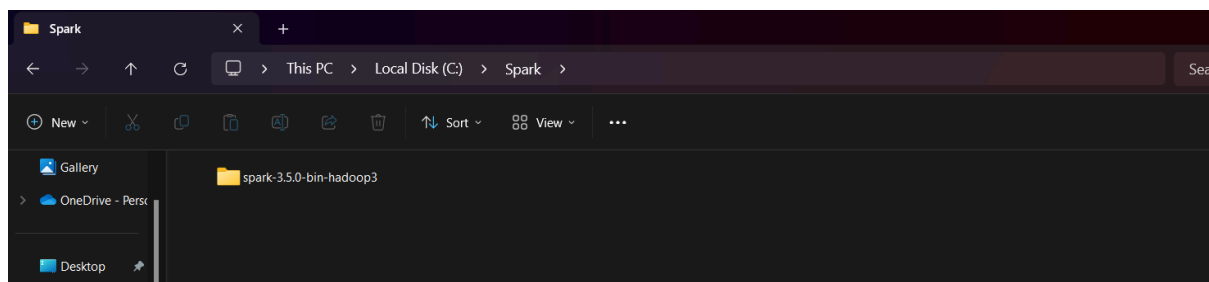


For hadoop :

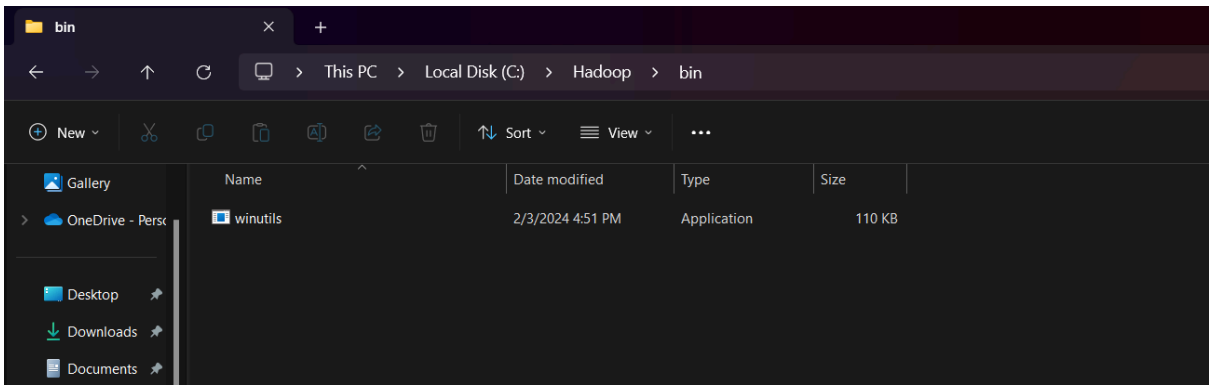
Downloaded the winutils.exe file :



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

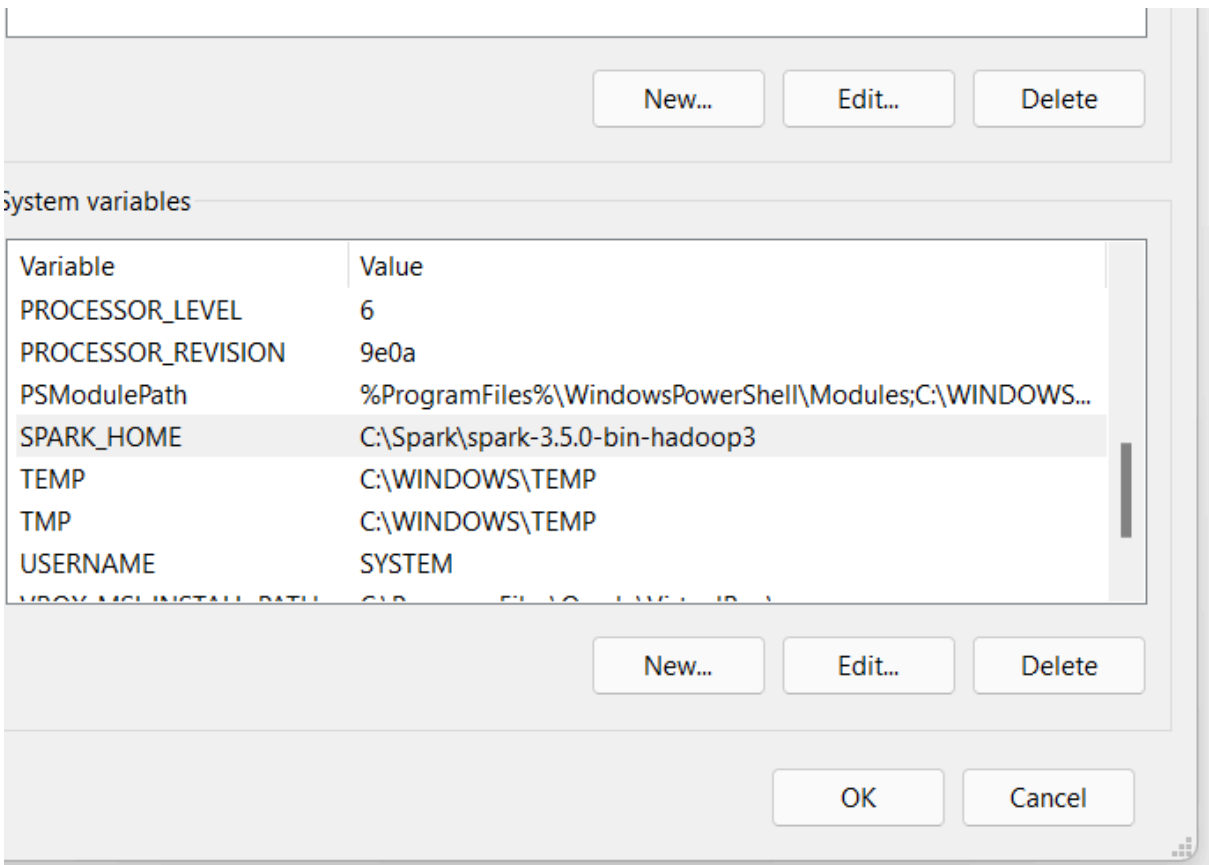


And for hadoop file created directory :
C:/Hadoop/bin :

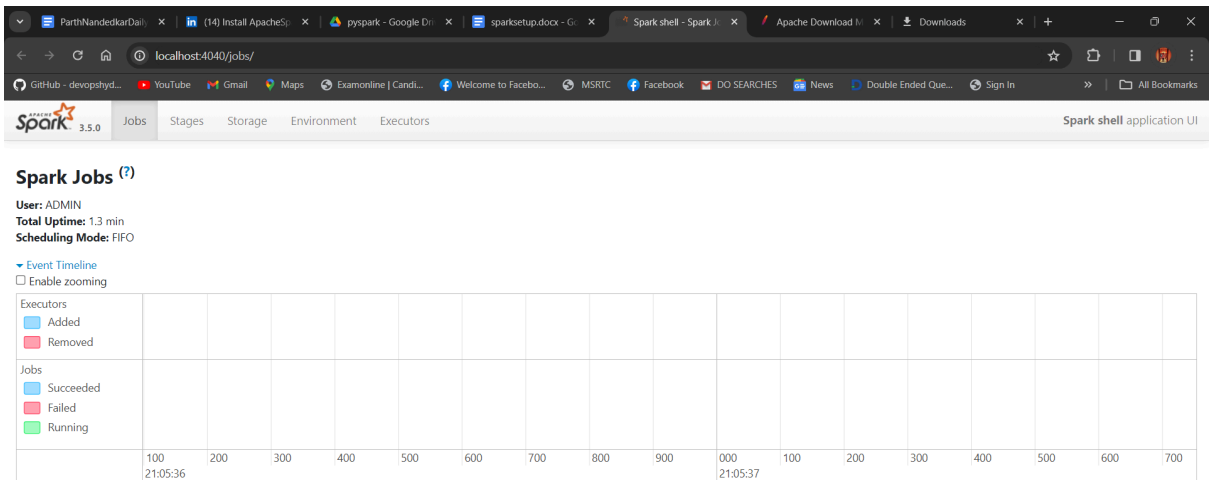


Setting environment variables :

Added environment variables in system variables :



Then on browser by entering URL : <http://localhost:4040/jobs/> :



Got Apache Spark service running.