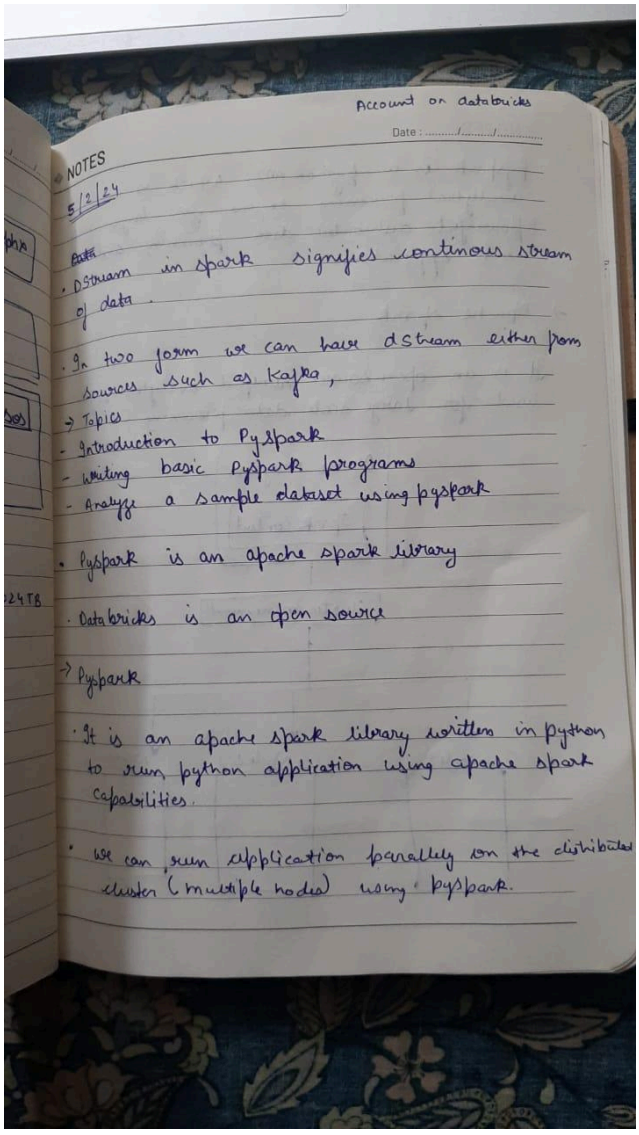
Asmita Porwal
Batch-1
Day-12
5/2/2024
Data engineering

Assignment-12

Hand written notes during the session :



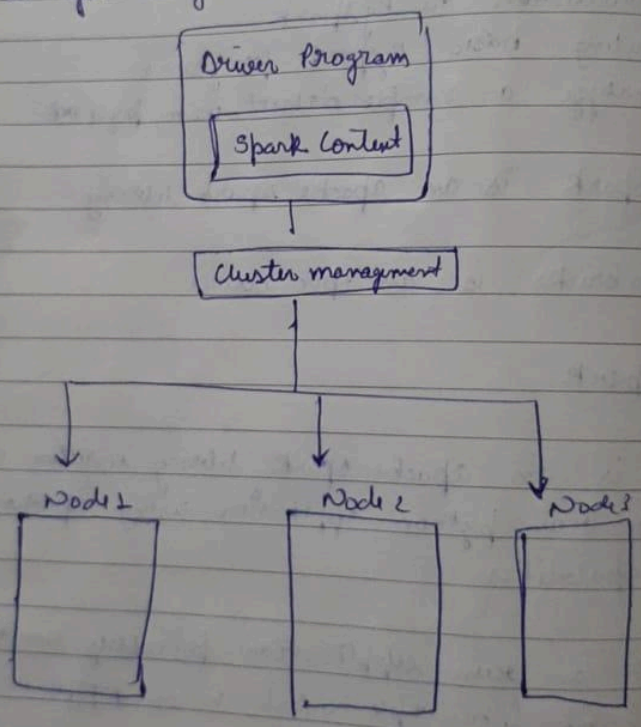
Date : ...../...../.....

◆ NOTES

- pyspark is a python API which is an analytical processing engine for large-scale powerful distributed data processing and ml applications.

→ Apache Spark

It is an open source unified analytics engine used for large scale data processing.



NOTES

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

→ Uses Pyspark

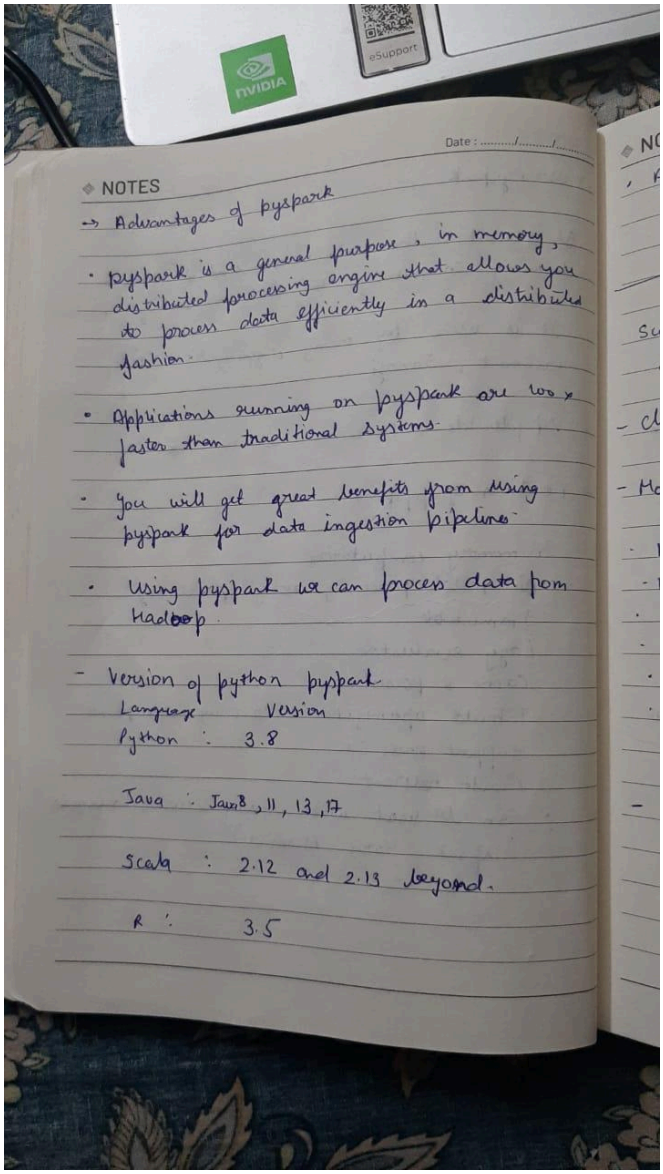
It is used in the Data science and Machine learning community as there are many

. It is used by many organization like walmart, Sanofi

. Spyder Ide, Anaconda we can use

→ Features of pyspark

- In memory computation
- Distributed processing using parallelly
- Immutable
- Lazy evaluation
- Cache & persistence
- Inbuild optimization when using Data frames
- support ANSI SQL.
- Fault tolerant
- Can be used with many cluster managers (Spark, Yarn, Mesos etc)



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

NOTES

→ Advantages of pyspark

- pyspark is a general purpose, in memory, distributed processing engine that allows you to process data efficiently in a distributed fashion.
- Applications running on pyspark are 100 x faster than traditional systems.
- You will get great benefits from using pyspark for data ingestion pipelines.
- Using pyspark we can process data from Hadoop.

- Version of python pyspark

Language	Version
Python	3.8
Java	Java 8, 11, 13, 17
Scala	2.12 and 2.13 beyond.
R	3.5

NOTES

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

- Apache Kafka is an open source distributed event streaming platform used by thousands of companies for high-performance data pipelines, streaming analytics, data integration and mission-critical applications.

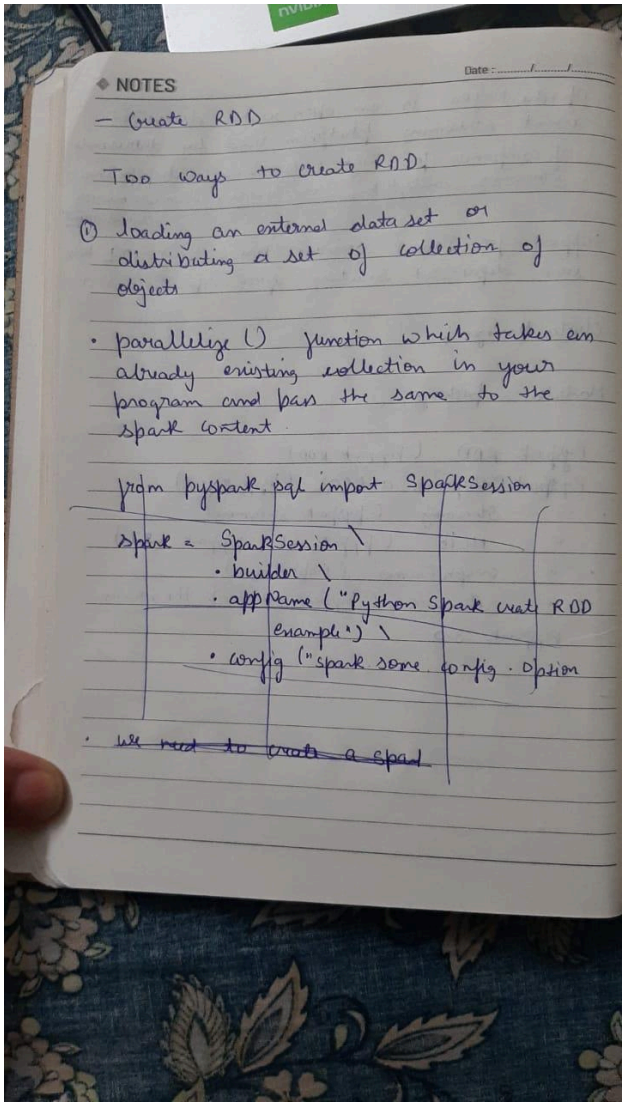
Support for Java 8 versions prior to 8U371 has been deprecated starting from Spark 3.5.0.

- Cluster Manager:

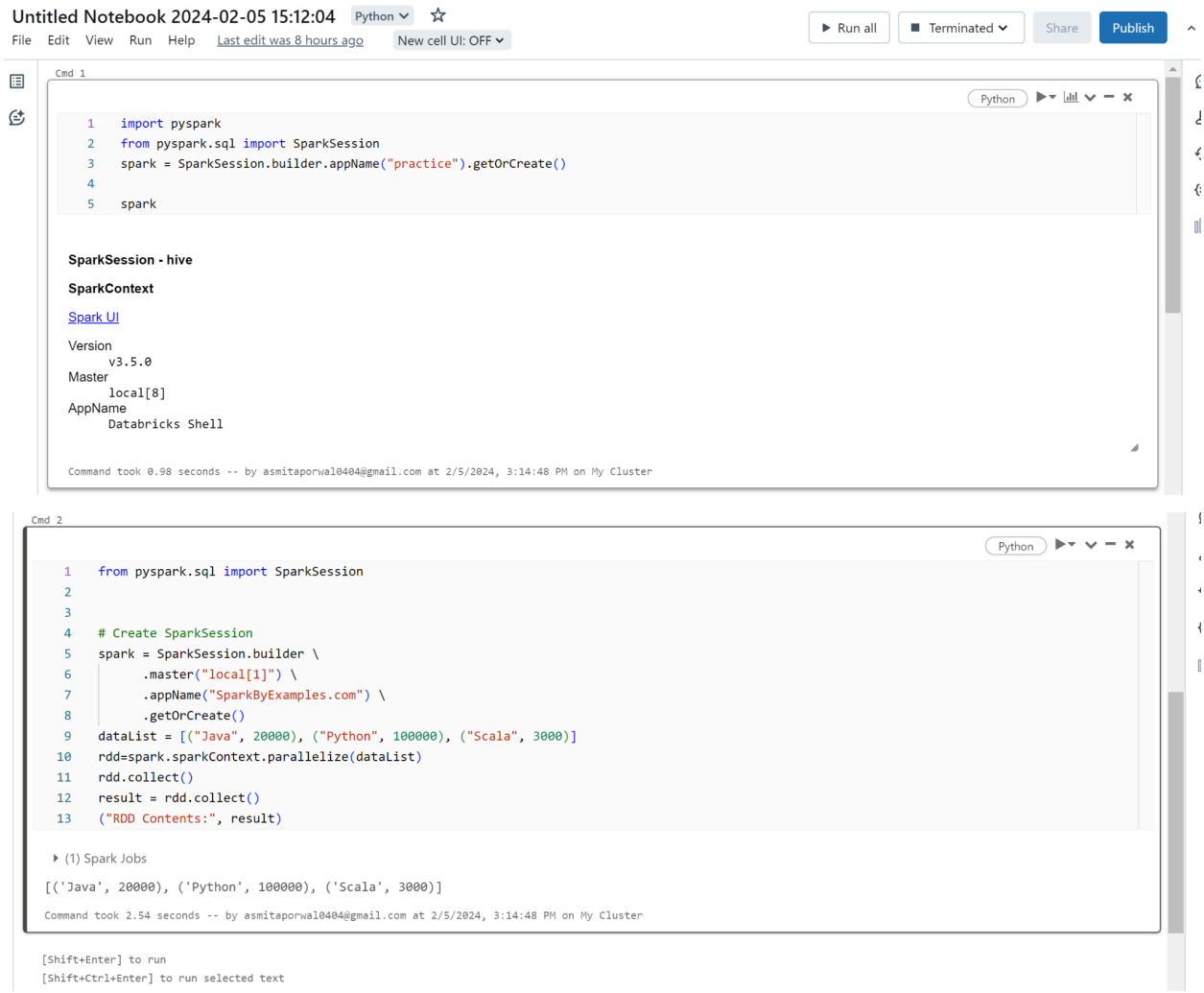
- Modules & packages

- Pyspark RDD (pyspark.RDD)
- Pyspark Dataframe and SQL (pyspark.sql)
- " Streaming (pyspark.streaming)
- " MLib (pyspark.ml, pyspark.mllib)
- " Graph Frames (GraphFrames)
- " Resource (pyspark.resource) It's new in Pyspark 3.0.

-



Create RDD on data bricks



Creating rdd on local system on jupyter notebook

In [5]:
```
df.show()
```

```
+----+--------+--------+---+
| _c0|     _c1|     _c2|_c3|
+----+--------+--------+---+
|Name|M1 Score|M2 Score|age|
|Alex|      62|      80| 20|
|Brad|      45|      56| 19|
|Joey|      85|      98| 21|
|abhi|      54|      79| 20|
+----+--------+--------+---+
```

In [6]:
```
dataList = [("Java", 20000),("Python", 100000),("Scala",3000)]
```

In [7]:
```
rdd = spark.sparkContext.parallelize(dataList)
```

In [8]:
```
result = rdd.collect()
result
```

Out[8]:
```
[('Java', 20000), ('Python', 100000), ('Scala', 3000)]
```

In [10]:
```
rdd = spark.sparkContext.textFile("D:\\pyspark.txt")
```

In [11]:
```
result = rdd.collect()
result
```

Out[11]:
```
['PySpark has been released in order to support the collaboration of Apache Spark and Python, it actually is a Python API for S
park. In addition, PySpark, helps you interface with Resilient Distributed Datasets (RDDs) in Apache Spark and Python programmi
ng language']
```