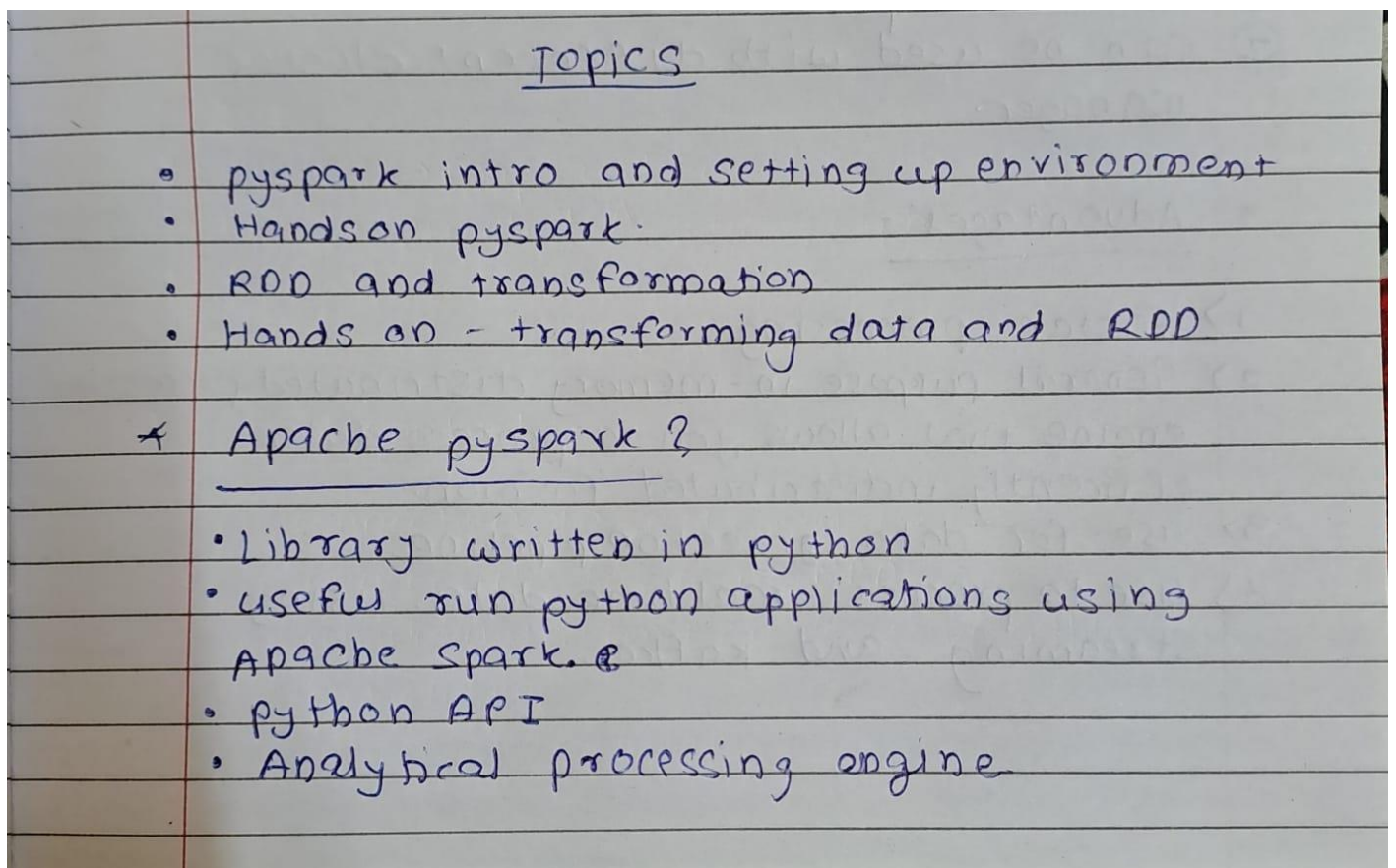


# PySpark Assignment 2

Pratik Wani

- Hand Written Notes:



## \* Apache Spark?

- Unified Analytical engine used for large dataset
- faster
- can run on single node or multiple nodes
- In memory processing
  - Numpy + TensorFlow

## \* Features:-

- ① In memory computation
- ② Distributed processing
- ③ immutable
- ④ Lazy evaluation
- ⑤ supports ANSI SQL
- ⑥ Cache and persistence
- ⑦ can be used with different cluster manager

## \* Advantages:-

- 1) Faster processing
- 2) General purpose in-memory distributed processing engine that allows you to process data efficiently in distributed fashion.
- 3) use for data ingestion pipeline
- 4) use to process real time data using Streaming and Kafka.

## \* Apache kafka:-

- Event streaming platform
- Distributed event streaming platform.

## \* versions

py → 3.8 and above.

Java → Java 8, 11, 13, 17 and above.

Scala → 2.12 and 2.13

R → 3.5

## \* python modules and package:-

- Pyspark RDD
- pyspark dataframe.
- Pyspark Streaming
- pyspark MLlib
- pyspark GraphFrames
- Pyspark Resource (NEW in 3.0)

## \* commands:-

Read data from file

→ create txt file

→ val filepath = "\_\_\_path\_\_"

→ val textRDD = sc.textFile(filepath)

→ textRDD.collect()



\* Rdp creation :-

2 ways:

1st

using `paralize()` :-

- create session
- make dataset
- then passing it in `paralize()`

2nd

using `txtfile`.

`txtfile ("textfilename")`

- Codes:
- Jupyter Notebook:

- Read CSV

```
In [1]: import pyspark
```

```
In [2]: from pyspark.sql import SparkSession
```

```
In [3]: spark = SparkSession.builder.appName("practice").getOrCreate()
```

```
In [4]: spark
```

```
Out[4]: SparkSession - in-memory  
SparkContext
```

[Spark UI](#)

**Version**

v3.5.0

**Master**

local[\*]

**AppName**

practice

```
In [10]: df = spark.read.csv('student_salary_info.csv')  
df
```

```
Out[10]: DataFrame[_c0: string, _c1: string, _c2: string]
```

```
In [11]: df.show()
```

```
+-----+-----+  
|_c0|_c1|_c2|  
+-----+-----+  
| Name|Age|Salary|  
|Pratik| 21| 50000|  
|Vikas| 23|100000|  
|Rushi| 22|120000|  
+-----+-----+
```

- Using Paralyze()

```
In [1]: from pyspark.sql import SparkSession

In [2]: spark = SparkSession.builder \
        .master("local[1]") \
        .appName("SparkByExamples.com") \
        .getOrCreate()

In [3]: dataList = [{"Shiva", 20000}, {"Pratik", 100000}, {"vikas", 3000}]

In [4]: rdd=spark.sparkContext.parallelize(dataList)

In [5]: rdd2 = spark.sparkContext.textFile("/path/test.txt")

In [7]: result = rdd.collect()

In [8]: result
Out[8]: [('Shiva', 20000), ('Pratik', 100000), ('vikas', 3000)]

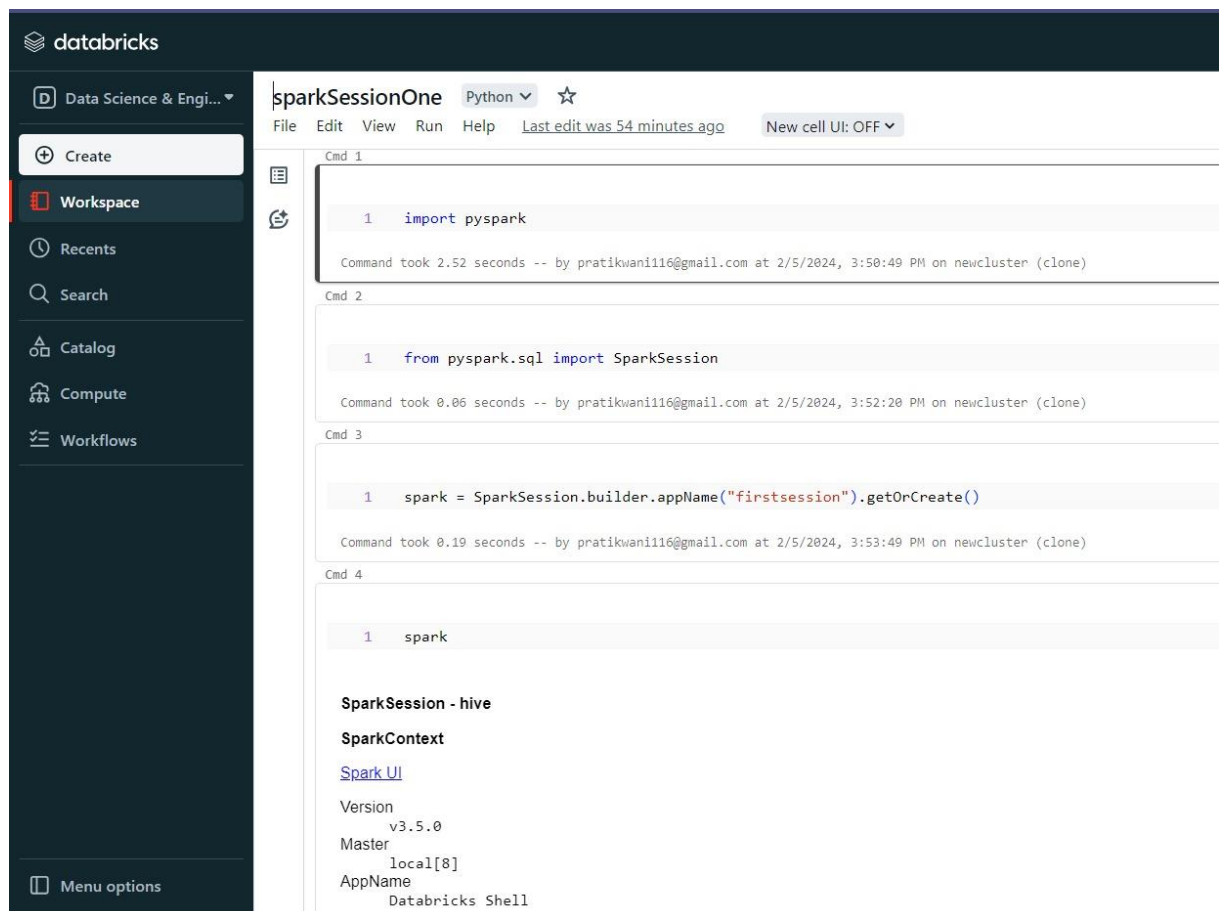
In [9]: result
Out[9]: ('RDD Contents:', [('Shiva', 20000), ('Pratik', 100000), ('vikas', 3000)])

In [10]: rdd2=spark.sparkContext.textFile("myfile.txt")

In [11]: rdd2.collect()
Out[11]: ['Hello!!', 'My name is pratik arun wani', 'Pyspark', 'Databricks']
```

- Databricks:

- Read CSV



**databricks**

Data Science & Engi...

Create

Workspace

Recents

Search

Catalog

Compute

Workflows

Menu options

**sparkSessionOne** Python ☆

File Edit View Run Help Last edit was 54 minutes ago New cell UI: OFF

```
1 df = spark.read.csv('/FileStore/tables/student_salary_info-2.csv')
2 df
```

▶ (1) Spark Jobs

df: pyspark.sql.dataframe.DataFrame = [\_c0: string, \_c1: string ... 1 more field]

DataFrame[\_c0: string, \_c1: string, \_c2: string]

Command took 1.49 seconds -- by pratikwani116@gmail.com at 2/5/2024, 4:03:30 PM on newcluster (clone)

Cmd 6

```
1 df.show()
```

▶ (1) Spark Jobs

_c0	_c1	_c2
Name	Age	Salary
Pratik	21	50000
Vikas	23	100000
Rushi	22	120000

Command took 0.58 seconds -- by pratikwani116@gmail.com at 2/5/2024, 4:03:43 PM on newcluster (clone)

[Shift+Enter] to run  
[Shift+Ctrl+Enter] to run selected text

## ○ Using Paralyze()

**databricks**

Data Science & Engi...

Create

Workspace

Recents

Search

Catalog

Compute

Workflows

Menu options

**SparkSession2** Python ☆

File Edit View Run Help Last edit was 47 minutes ago New cell UI: OFF

Cmd 1

```
1 from pyspark.sql import SparkSession
```

Command took 0.06 seconds -- by pratikwani116@gmail.com at 2/5/2024, 4:04:59 PM on newcluster (clone)

Cmd 2

```
1 spark = SparkSession.builder \
2     .master("local[1]") \
3     .appName("SparkByExamples.com") \
4     .getOrCreate()
```

Command took 0.08 seconds -- by pratikwani116@gmail.com at 2/5/2024, 4:05:14 PM on newcluster (clone)

Cmd 3

```
1 dataList = [("Shiva", 20000), ("Pratik", 100000), ("vikas", 3000)]
```

Command took 0.13 seconds -- by pratikwani116@gmail.com at 2/5/2024, 4:05:23 PM on newcluster (clone)

Cmd 4

```
1 rdd=spark.sparkContext.parallelize(dataList)
```

Command took 0.16 seconds -- by pratikwani116@gmail.com at 2/5/2024, 4:05:36 PM on newcluster (clone)

databricks

Data Science & Engi...

Create

Workspace

Recents

Search

Catalog

Compute

Workflows

SparkSession2

Python

File

Edit

View

Run

Help

Last edit was 48 minutes ago

New cell UI: OFF

Command took 0.16 seconds -- by pratikwani116@gmail.com at 2/5/2024, 4:05:36 PM on newcluster (clone)

Cmd 5

1

rdd.collect()

▶ (1) Spark Jobs

[('Shiva', 20000), ('Pratik', 100000), ('vikas', 3000)]

Command took 0.30 seconds -- by pratikwani116@gmail.com at 2/5/2024, 4:05:56 PM on newcluster (clone)

Cmd 6

1

rdd2=spark.sparkContext.textFile("/FileStore/tables/test.txt")

Command took 0.20 seconds -- by pratikwani116@gmail.com at 2/5/2024, 4:08:56 PM on newcluster (clone)

Cmd 7

1

rdd2.collect()

▶ (1) Spark Jobs

['Hello!!', 'My name is', 'Pratik', 'Arun', 'Wani']

Command took 0.99 seconds -- by pratikwani116@gmail.com at 2/5/2024, 4:09:04 PM on newcluster (clone)

[Shift+Enter] to run

[Shift+Ctrl+Enter] to run selected text

- Local Host 4040/jobs

Spark Jobs (2)

User: dell

Total Uptime: 2.0 h

Scheduling Mode: FIFO

Completed Jobs: 2

▼ Event Timeline

☐ Enable zooming

Executors

Added

Removed

Jobs

Succeeded

Failed

Running

14:56

14:57

14:58

14:59

15:00

15:01

15:02

15:03

15:04

15:05

15:06

15:07

15:08

15:09

15:10

15:11

Mon 5 February

Executor driver added

▼ Completed Jobs (2)

Page: 1

1 Pages. Jump to 1. Show 100 items in a page. Go

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
1	showString at NativeMethodAccessorImpl.java:0 showString at NativeMethodAccessorImpl.java:0	2024/02/05 15:10:13	95 ms	1/1	1/1
0	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2024/02/05 15:09:53	0.5 s	1/1	1/1



