

Name: Rohan Vinayak Chaudhari

Batch: Data Engineering

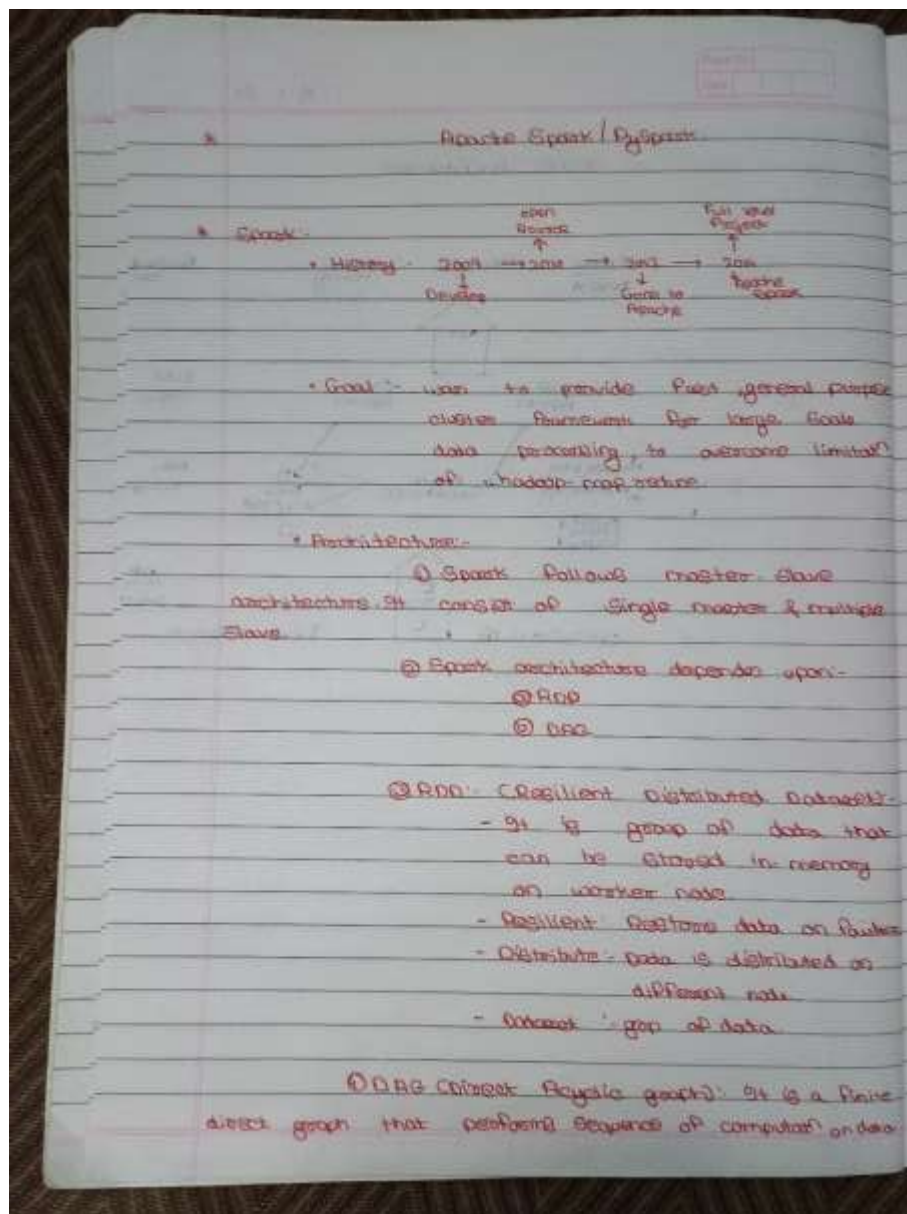
Date: 03/02/2024

Topic: Pyspark

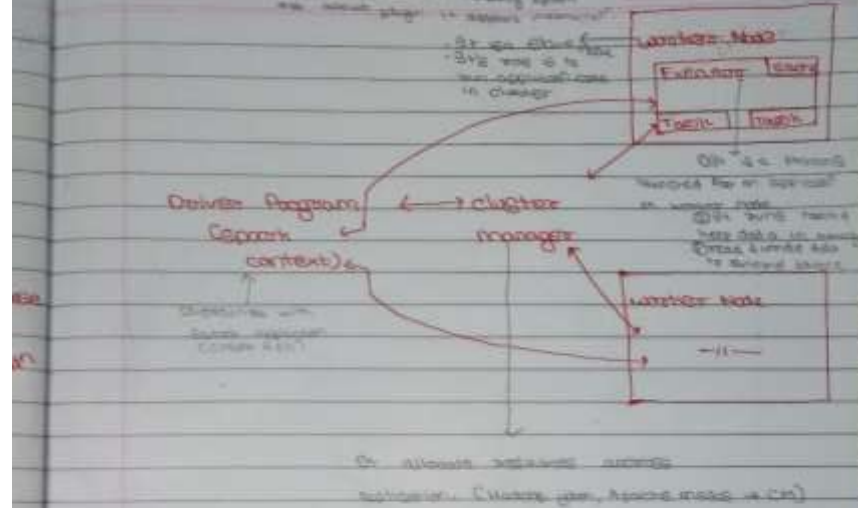
Solution:

## 1. Pyspark:

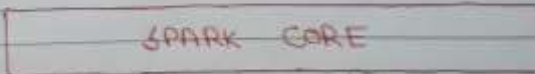
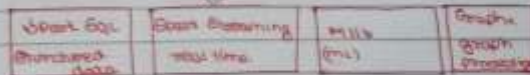
### Introduction to pyspark and Basic Architecture:



- DAG Scheduler: Creates a DAG of stages for each task & sends them to executors.
- Task Scheduler: Schedules tasks to executors, taking into account dependencies.
- Resource Manager: Manages resources in the cluster, allocating them to tasks.



#### \* Spark components:



Master of Spark  
On the server side, it holds the information for task scheduling.  
Fault recovery, including with storage system & memory management.

## Pyspark:

```
C:\Users\chaud>pySpark
Python 3.10.9 (tags/v3.10.9:1dd9be6, Dec 6 2022, 20:01:21) [MSC v.1934 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).
24/02/03 21:40:30 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Welcome to

  ____      _
 / ___|  __| | | |
| |  | |__| | | |
| |  | |__| | | |
| |  | |__| | | |
|_|  |____|_|_|_|

 version 3.3.0

Using Python version 3.10.9 (tags/v3.10.9:1dd9be6, Dec 6 2022 20:01:21)
Spark context Web UI available at http://192.168.1.36:4040
Spark context available as 'sc' (master = local[*], app id = local-1706976632391).
SparkSession available as 'spark'.
>>> 24/02/03 21:40:46 WARN ProfMetricsGetter: Exception when trying to compute pagesize, as a result reporting of ProcessTree metrics is stopped
```

## Spark shell:

```
exit()

C:\Users\chaud>SUCCESS: The process with PID 11396 (child process of PID 8292) has been terminated.
SUCCESS: The process with PID 8292 (child process of PID 9736) has been terminated.
SUCCESS: The process with PID 9736 (child process of PID 7588) has been terminated.

C:\Users\chaud>spark-shell
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
20/02/03 21:43:15 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Spark context Web UI available at http://192.168.1.36:8080
Spark context available as 'sc' (master = local[*], app id = local-1706976796559).
Spark session available as 'spark'.
Welcome to

      ____
     / __ \
    / /_< \
   / ___> \
  /_/_____\

 version 1.1.0

Using Scala version 2.12.15 (Java HotSpot(TM) Client VM, Java 1.8.0_401)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```