

Name: Rohan Vinayak Chaudhari

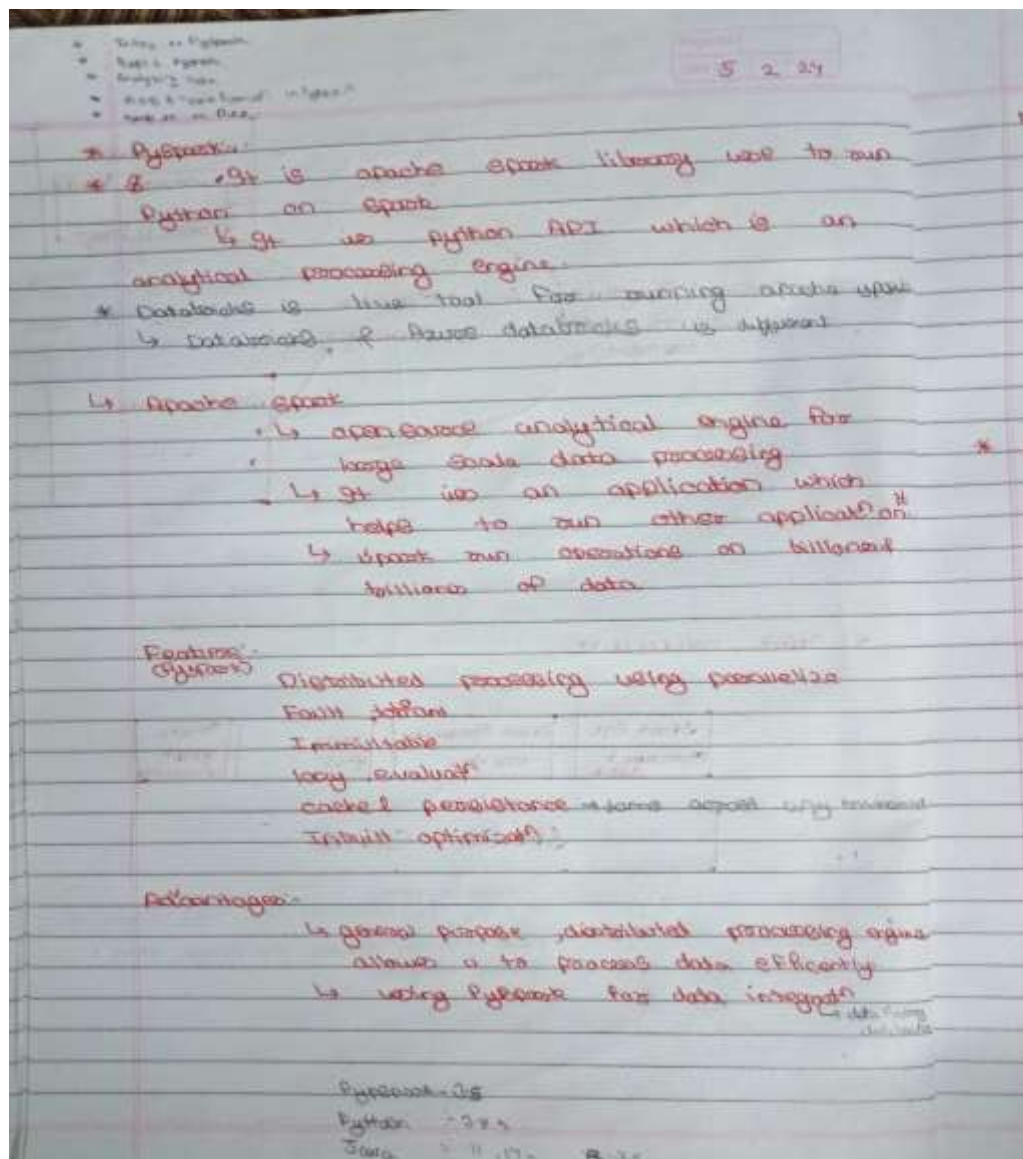
Batch: Data Engineering

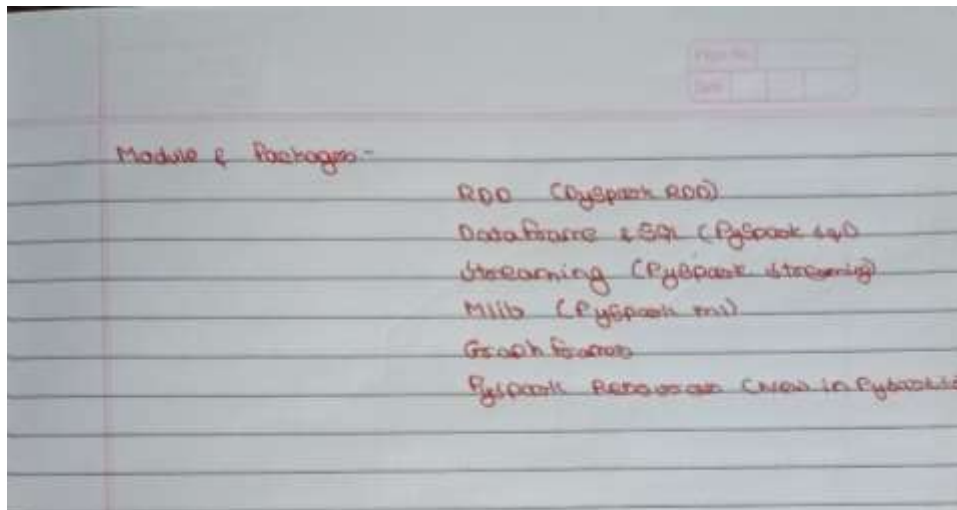
Date: 05/02/2024

Topic: Pyspark

Solution:

1. Pyspark:





```
jupyter PySpark Last Checkpoint: Last Saturday at 4:19 PM (autosaved)
File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (pykernel)
In [5]: import pyspark
In [8]: import findspark
In [10]: from pyspark.sql import SparkSession
In [11]: spark=SparkSession.builder.appname('Dataframe').getOrCreate()
In [12]: spark
Out[12]: SparkSession - in-memory
SparkContext
Spark UI
Version
v3.5.0
Master
local[*]
AppName
Dataframe
```

```
jupyter PySpark Last Checkpoint: Last Saturday at 4:19 PM (autosaved)
File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (pykernel)
In [13]: df_pyspark=spark.read.option('header','true').csv("D:\Hexaware\Data_Engineering\Python\output_file.csv",inferSchema=True)
In [14]: df_pyspark.printSchema()
root
 |-- PassengerId: integer (nullable = true)
 |-- Survived: integer (nullable = true)
 |-- Pclass: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- Sex: string (nullable = true)
 |-- Age: double (nullable = true)
 |-- SibSp: integer (nullable = true)
 |-- Parch: integer (nullable = true)
 |-- Ticket: string (nullable = true)
 |-- Fare: double (nullable = true)
 |-- Cabin: string (nullable = true)
 |-- Embarked: string (nullable = true)
```

```
jupyter Pyspark Last Checkpoint: Last Saturday at 4:19 PM (autosaved)
File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (pykernel)
In [15]: df_pyspark=spark.read.csv('D:\Hexaware\Data_Engineering\Python\output_file.csv',header=True,inferSchema=True)
In [16]: df_pyspark.show()
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|PassengerId|Survived|Pclass|Name|Sex|Age|SibSp|Parch|Ticket|Fare|Cabin|Embarked|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|1|0|3|Braund, Mr. Owen ...|male|22.0|1|0|A/5 21171|7.25|NULL|S|
|2|1|1|Cumings, Mrs. Joh...|female|38.0|1|0|PC 17599|71.2833|C85|C|
|3|1|3|Heikkinen, Miss. ...|female|26.0|0|0|STON/O2. 3101282|7.925|NULL|S|
|4|1|1|Futrelle, Mrs. Ja...|female|35.0|1|0|113803|53.1|C323|S|
|5|0|3|Allen, Mr. Willia...|male|35.0|0|0|373450|8.05|NULL|S|
|6|0|3|Moran, Mr. James|male|NULL|0|0|330877|8.4583|NULL|Q|
|7|0|1|McCarthy, Mr. Tim...|male|54.0|0|0|17463|51.8625|E46|S|
|8|0|3|Palsson, Master. ...|male|2.0|3|1|349909|21.075|NULL|S|
|9|1|1|Johnson, Mrs. Osc...|female|27.0|0|2|347742|11.1333|NULL|S|
|10|1|2|Nasser, Mrs. Nich...|female|14.0|1|0|237736|30.0708|NULL|C|
|11|1|3|Sandstrom, Miss. ...|female|4.0|1|1|PP 9549|16.7|G6|S|
|12|1|1|Bonnell, Miss. El...|female|58.0|0|0|113783|26.55|C103|S|
|13|0|3|Saunderscock, Mr. ...|male|20.0|0|0|A/S. 2151|8.05|NULL|S|
|14|0|3|Andersson, Mr. An...|male|39.0|1|5|347082|31.275|NULL|S|
|15|0|3|Vestrom, Miss. Hu...|female|14.0|0|0|350406|7.8542|NULL|S|
|16|1|2|Hewlett, Mrs. (Ma...|female|55.0|0|0|248706|16.0|NULL|S|
|17|0|3|Rice, Master. Eugene|male|2.0|4|1|382652|29.125|NULL|Q|
|18|1|2|Williams, Mr. Cha...|male|NULL|0|0|244373|13.0|NULL|S|
|19|0|3|Vander Planke, Mr...|female|31.0|1|0|345763|18.0|NULL|S|
|20|1|3|Masselmani, Mrs. ...|female|NULL|0|0|2649|7.225|NULL|C|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 20 rows
```

```
jupyter Pyspark Last Checkpoint: Last Saturday at 4:19 PM (autosaved)
File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (pykernel)
In [17]: type(df_pyspark)
Out[17]: pyspark.sql.dataframe.DataFrame
In [7]: exit()
```

jupyter Pyspark Last Checkpoint: Last Saturday at 4:19 PM (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted Python 3 (pykernel)

Run

Code

```
In [1]: import findspark
findspark.init()

In [2]: from pyspark import SparkContext

# Create a SparkContext
sc = SparkContext("local", "RDD Transformation Exercise")

In [3]: # Create an RDD from a list of numbers
data = [1, 2, 3, 4, 5]
rdd = sc.parallelize(data)

In [4]: # Use map transformation to square each element
squared_rdd = rdd.map(lambda x: x ** 2)
print("Squared RDD:", squared_rdd.collect())

Squared RDD: [1, 4, 9, 16, 25]

In [5]: import pyspark
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName('test').getOrCreate()

columns = ["language", "users_count"]
data = [{"Java", "20000"}, {"Python", "100000"}, {"Scala", "3000"}]

df = spark.createDataFrame(data=data, schema=columns)
print(df.printSchema())

root
 |-- language: string (nullable = true)
```

jupyter Pyspark Last Checkpoint: Last Saturday at 4:19 PM (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted Python 3 (pykernel)

Run

Code

```
In [6]: df.show()

+-----+-----+
|language|users_count|
+-----+-----+
|   Java|      20000|
| Python|     100000|
|   Scala|       3000|
+-----+-----+

In [1]: # Import SparkSession
from pyspark.sql import SparkSession

# Create SparkSession
spark = SparkSession.builder \
    .master("local[*]") \
    .appName("SparkByExamples.com") \
    .getOrCreate()

datalist = [{"Java", 20000}, {"Python", 100000}, {"Scala", 3000}]
rdd=spark.sparkContext.parallelize(datalist)
rdd2 = spark.sparkContext.textFile("/path/test.txt")

In [3]: rdd.collect()

Out[3]: [('Java', 20000), ('Python', 100000), ('Scala', 3000)]
```

With Databricks:



The screenshot displays the Databricks web interface. At the top, the header shows the Databricks logo, the notebook name 'First_notebook', the language 'Python', and a star icon. Below this is a menu bar with 'File', 'Edit', 'View', 'Run', and 'Help'. The 'Run' button is highlighted, and a 'New cell (Ctrl+Enter)' button is visible. On the right side of the header, there are buttons for 'Run all', 'Unlink', 'Share', and 'Publish'. The main area of the notebook contains a code editor with the following Python code:

```
1 %use spark
2
3 # Create SparkSession
4 spark = SparkSession.builder \
5     .master("local[*]") \
6     .appName("SparkByExample.com") \
7     .getOrCreate()
8 data1 = [{"id": 1, "name": "John", "age": 30}, {"id": 2, "name": "Jane", "age": 25}]
9 rdd1 = spark.sparkContext.parallelize(data1)
10 rdd2 = spark.sparkContext.textFile("/path/to/text.txt")
11
12 rdd1.collect()
```

Below the code editor, the output of the code is displayed. It shows the Spark version and configuration details, followed by the output of the `rdd1.collect()` command, which is a list of dictionaries: `[{"id": 1, "name": "John", "age": 30}, {"id": 2, "name": "Jane", "age": 25}]`. The output is displayed in a table format with columns for 'id', 'name', and 'age'.