**Note : Kindly Answer the below 5 Questions (Explanation Format)**

Q.1 Explain PySpark in brief?

PySpark is basically python + spark. It is built upon the Aparche spark. Aparche Spark is open source cluster computing framework which is basically used for managing, processing and analysing large datasets. Spark has written using scala programming language. Spark handles all operation in the memory and thus it is faster than MapReduce.

Q.2 What are the main characteristics of (Py)Spark?

As Sparke uses RDD concept so doing any operation is very intuitive, Operations are very fast as it reduces the write-read while doing operations, As it uses RDD concept so if any node gets corrupted then we can trace back to the parent RDD and generate the corrupted data thus we don’t need to store back-up as in Mapreduce, It can handle streaming data.

Q.3 What do you mean by PySpark SparkContext?

When shell is launched it initializes the spark Context(sc). Sc represent the connection between spark and its clusters. Using sc we can load data into memory from a specified location. Sc is used to defined RDD.

Q.4 What is pep 8?

PEP8 is a documentation for writing a clean and understandable python code. The main aim of PEP-8 is to increase reusability and readability of a code. PEP stands for python enhancement proposal.

Q.5 What is the difference between list and tuples in Python?

List is mutable where tuple aren’t, tuple can a made hashable where list can’t thus tuple can be faster than list, we use list when position of elements are not important whereas in tuple position of elements are important. We generally store homogeneous elements in a list whereas in a tuple we store relatable heterogeneous elements.

**Note : Kindly Answer the below 5 Questions (Code Explanation with Pseudo Code Format)**

Q. 1 Write a function that returns the maximum of two numbers. (Python Code)

def getmax(a, b):

if a > b:

return a

else:

return b

getmax(a, b)

Q.2 Write a program (function!) that takes a list and returns a new list that contains all the elements of the first list minus all the duplicates.

def uni\_arr(arr):

mem = {}

for key in arr:

if key not in mem:

mem[key] = key

out = []

for key in mem:

out.append(mem[key])

return out

uni\_arr(arr)

# time complexity is O(n) and space complexity is O(n + n + n)

Q. 3 Write a pyspark program to get the first 10 record from RDD. (Give Complete Explanation with Steps.)

1. First start the ipython shell and thus it will initialize the spark context.

2. Make a RDD either loading the dataset or using sc.parallelize(arr, n)(arr : is your dataset array, n is the number of partition)

3. After making a RDD we can take an action on it to get first 10 records. We can use .take(10) but as RDD stores data in chunks thus .take() and .first() usually return different answer on each call. So to get deterministic output all the time we can first sort it(asc or desc) and call the first 10 records. For that we can use .takeOrder(10) by default it will arrange the record in asc order or if you don’t want to use .takeOrder() we can store data in the memory using .perssitance() and call .take(10) to get deterministic records every time.

Q.4 Write a Tableau Case statement Name: Days to Ship Scheduled If Ship Mode is Same Day, First Class, Second Class, and Standard Class then respective ship days will be 0,1,3,6 Days.

1. Open worksheet

2. click on analysis > create calculated field

3. name new field as “Days”

4. write the below formula in formula space and click on OK

CASE [SHIP MODE]

When “Same Day” Then 0

When “First Class” Then 1

When “Second Class” Then 3

When “Standard Class” Then 6

ELSE -1

End

Q.5 Create a Tableau Calculated Field to calculate Profit Ratio. Where your column names are Profit and Sales.

1. Open WorkSheet

2. Click on analysis > create calculated field

3. name new field as “Profit Ratio”

4. Write the below formula in the formula space and click on “OK”

SUM([Profit])/SUM([Sales])