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

Q.1 Explain PySpark in brief?

**Ans**:

Spark is written in Scala and Python is more widely used than Scala. So Spark community built a tool in order to support Python language . It’s called PySpark. We can build RDD in it. The PySpark tool is possible due to Py4j library.

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

**Ans:**

Network is abstracted so only implicit communication is possible.

Nodes are also abstracted means no possible to address each node.

It is based on Map-Reduce means user need to submit Map and Reduce functions.

PySpark is an API for Spark.

Q.3 What do you mean by PySpark SparkContext?

**Ans:**

It is an entry point to any Spark functionality in Spark context. It uses Py4j library to create JVM. It creates Java Spark Context. It is available as “sc” in PySpark.

Q.4 What is pep 8?

**Ans:**

It’s full form is Python Enhancement Proposal. It’s a guideline or set of rules to write python codes matching with industry standards. It’s aim is to improve readability and consistency throughout different users’ code.

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

**Ans:**

List:

* They are mutable.
* They need more memory.
* They have more built in methods.
* They have time consuming iteration.

Tuples:

* They are immutable.
* They have comparatively faster iteration.
* They don’t have built in methods.
* They need less memory.

**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)

**Ans**:

def maxm(a,b):

if a>=b:

return a

else:

return b

We can give 2 inputs in the above function and find the maximum of those numbers.

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.

**Ans**:

Here a is a list input.

def unq(a):

unqlist = []

for i in range(len(a)):

flag = 0

for j in range(i+1,len(a)):

if a[i] == a[j]:

flag = 1

if flag == 0 and a[i] not in unqlist:

unqlist.append(a[i])

return unqlist

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

**Ans**:

1. Use Spark context to make connection between a driver and executor.

 2. Make a RDD either by loading HDFS file or using sc.parallelize() command,

 3. After making a RDD perform an action .take(10) to get 10 records from the RDD. “take” will give to random outputs in each call.

4. We can also use .takeOrdered() to first order the entries in asc or desc order and then fetch the top 10 outputs

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.

**Ans**:

**CODE-**

Case ([Ship Mode])

WHEN(“Same Day”) THEN 0

WHEN(“First Class”) THEN 1

WHEN(“Second Class”) THEN 3

ELSE 6

**Explanation:**

1. On the left side in the data tab, go to drop down beside filter and right click to select “Create Calculated Field”.
2. Type Days to Ship Scheduled in the first field.
3. Then put the above given code in the formula section. It is setting a condition on “Ship Mode” and matching its value with the scenarios given in the question. We are matching it with WHEN and setting the answer with THEN.
4. Click apply or OK to execute it. A column will be created based on the conditions defined in the code.

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

**Ans:**

**CODE** –

SUM([“Profit”]) / SUM([“Sales”])

**Explanation**

1. On the left side in the data tab, right click to select “Create Calculated Field”.
2. In the first field write “Profit Ratio”.
3. Then put the above code in the formula section. It is the formula for the Ratio – Profit per unit sales.
4. Click apply or OK to execute to create a new column of Profit Ratio.