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

Q.1 Explain PySpark in brief?

**Answer:** PySpark is a tool which was built for the collaboration of Spark and Python as Spark is written in Scala language. PySpark is useful for establishing the interface between Resilient Distributed data (RDD) in spark and python programming.

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

**Answer:** Here are the following characteristics of PySpark:

1. It is much faster than other data processing frameworks like Hadoop.

2. Network is abstracted, that means there is only implicit communication possible.

3. PySpart is real-time because computation in the PySpark SQL takes place in the memory.

4. PySpark can deploy through the Hadoop or own cluster manager.

Q.3 What do you mean by PySpark SparkContext?

**Answer:** An entry point to a spark functionality is known as SparkContext. SparkContext uses Py4J to launch a JVM object to create a JavaSparkContext. PySpark has a default SparkContext available as ‘sc’.

Q.4 What is pep 8?

**Answer:** The full form of PEP is Python Enhancement Proposal. It provides the detailed guidlines and standard practices for writing a python code. This helps the developers for readability of the code.

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

**Answer:** Here are differences between list and tuple:

1. Lists are mutable. Tuples are immutable

2. Lists consume more memory. Tuples consumer less memory compared to lists.

3. Lists have built-in methods. Tuples don’t have built-in methods.

**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)

**Answer:** The definition of max\_num is given here, which calculates the maximum of two number a,b:

*def max\_num(a,b):*

*if a>b:*

*return a*

*else:*

*return 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.

**Answer:** The following is the function definition:

*def get\_unique(input\_list):*

*unique\_list = []*

*for i in range(len(input\_list)):*

*if input\_list[i] in unique\_list:*

*continue*

*else:*

*unique\_list.append(input\_list[i])*

*return unique\_list*

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

**Answer:** Here are the steps to get the first 10 record from RDD:

Step1. First we need to establised a link between the driver and executor using SparkContext,

Step2. Making a RDD by loading hdfs file

Step3. First we’ll use .takeOrdered to sort the entries (as per user’s requirement) and then fetch the top 10 outputs using .take(10)

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.

**Answer:** Here is the Tableau Case Statement:

CASE ([Ship Mode])

WHEN ("Same Day") THEN 0

WHEN ("First Class") THEN 1

WHEN ("Second Class") THEN 3

WHEN ("Standard Class") THEN 6

END

Steps to follow: We need to choose “Create Calculated Field” , which can be find on data tab. Then we need to provide the name in the first field. Then we need to formulate the column using the values in the case statement. Then, we execute it and now column will be available in the left pane. After that we can create a Tableau calculated field to calculate profit ratio.

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

**Answer:** Statement is as follows:

Sum ( [Profit] ) / Sum ( [Sales] )

Steps to follow: As mentioned in the previous case, We need to choose “Create Calculated Field” , which can be find on data tab. Then we need to provide the name in the first field. After that, we need to enter the statement “Sum ( [Profit] ) / Sum ( [Sales] )” to calculate the profit ratio. Then, we execute it and now column will be available in the left pane