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

Q.1 Explain PySpark in brief?

Solu-

The base is the Apache Spark. Spark support different languages such as Java, Scala, Python. When we are using python language then it is called Pyspark. Using the package py4j, we can also use RDDs (Resilient Distributed Datasets) on different nodes connect as a cluster.

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

Solu-

Py Spark is one of the api of Apaches Spark which supports Python language. Here cluster nodes are connected in such a fashion that only implicit communication is possible. It is based on Map reduce method where many data nodes are mapped to reducer. Which working on this one needs to define the Map and Reduce function for it to work. Also we cannot address the individual nodes.

Q.3 What do you mean by PySpark SparkContext?

Solu-

It is called as an entry point to any spark functionality.The core of the Spark is JVM, so in order to lainch jvm, Sparkcontext uses Py4J. The driver program then runs the operations inside the executors on worker nodes.PySpark has SparkContext available as ‘sc’, by default

Q.4 What is pep 8?

Solu-

PEP stnds for Python enhancement proposals. It is a kind of document for standardizing the best code practises of python. This helps in improving the consistency, readability of the code. It provides the guidelines to a python practitioner.

Example- function name should start from small letter.

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

Solu-

* List cam be edited like, we can replace the value in the list but we cannot replace the values in tuples
* List is represented as [] where as tuples are represented by ()
* Lists are mutable and tuples are immutable.
* List is of variable sze, tuples are of fixed size
* List has several built in methods while tuples don’t have

**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)

Solu-

Def max\_of\_two(x,y):

if x>y:

return x

else:

return y

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 unique\_elements(list):

Unique\_list=[]

For I in range(len(list)):

If list[i] not in Unique\_list:

Unique\_list.append(list[i])

Return Unique\_list

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

Solu-

* Firstly, we need to macke driver and executor connection using Spark Context (sc)
* Then making an RDDs by loading an hdf file or using sc.parallelize command.
* Take(10) will give the 10 records from the RDD.
* If we want to have the deterministic output then using takeOrdered to first order the entries in asc or desc order and then fetch the top 10 outputs
* By default it is asc order

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.

Solu-

CASE (Ship Mode)

When (‘Same Day) Then 0

When (‘First Class’) Then 1

When (‘Second Class’) Then 3

When (‘Standard Class’) Then 6

END

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

Solu-

SUM(Profit)/SUM(Sales)