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

Q.1 Explain PySpark in brief?

Apache Spark is a general-purpose data analysis and processing engine. It was originally written in Scala. So, in order to add support Python with Spark, Spark developer community released a tool, called “PySpark”. PySpark allows us to work with RDDs (main programming abstraction in Spark). It is possible due to its library name Py4j.

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

Some of the main characteristics of PySpark:

* It is an API for spark build to include support for python programming language.
* Node abstraction is present, which means that direct interaction with a single data node is not possible
* Network abstraction is also present, which means that only abstract communication is possible. These features mean that PySpark and RDDs take care of many of the complexities under the hood without the need of the user.

Q.3 What do you mean by PySpark SparkContext?

“Spark Context” is the basic entry point for accessing any of the spark functionality. Spark Context represents a connection to the Spark Cluster and is crucial for the creation of RDDs. When we talk about PySpark, by default it already has a SparkContext available as ‘sc’. In PySpark Spark Context uses Py4J(library) in order to launch a JVM. In this way, it creates a “JavaSparkContext”.

Q.4 What is pep 8?

PEP 8 stands for “Python Enhancement Proposal”. It was launched in 2001 and is a document containing guidelines and a set of best practices for writing python code. The aim of the document is to improve the general readability and consistency of the python code.

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

List:

* Lists are mutable
* Lists are more prone to accidental or unexpected errors and changes
* Lists are better suited for insertion, deletion and update operations.
* Lists consume more memory when compared with tuples
* Lists have some in-built functions

Tuples:

* Tuples are immutable
* Tuples reduce the possibility of accidental or unexpected errors or changes.
* Tuples are more suited for accessing the elements operations
* Tuples in general consume less memory.
* Tuples do not have in-built functions

**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)

# Function Definition

def maximum\_number(x,y):

‘’’ Function returns the maximum of 2 values’’’

if( x > y ):

return x

else:

return y

# Driver Code:

x = 5

y = 3

print(maximum\_number(x, 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.

# function definition

def unique\_list( z ):

‘’’ Function takes a list as input and returns a new list without the duplicate values’’’

y = []

for element in z:

if element not in y:

y.append(element)

return y

# driver code

z = [1, 2, 2, 3, 3, 3, 4, 4, 4, 4]

print(unique\_list( z ))

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

# PySpark steps

1. Firstly, establish a connection to the spark cluster using Spark Context (sc)
2. Secondly, read the data either using the sc.textfile() method or the sc.parallize method.

dataRDD = sc.textfile(<filepath>) (file can either be in local system or HDFS) or

dataRDD = sc.parallize(List(1,2,3,4))

1. Thirdly we can use the .take() method to extract the first 10 rows (unordered) from the RDD

dataRDD.take(10)

1. .take(10) will give first 10 unordered values from the RDD, We can also use .takeOrdered function to select 10 ordered values as well

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.

* In the tableau, select the “data” tab on the left side. Then select “Create Calculated Field” from the drop-down menu.
* Enter “Days to Ship Scheduled” in the first field. (name of the field)
* Type in the following code-

CASE ([Ship Mode])

WHEN ("Same Day") THEN 0

WHEN ("First Class") THEN 1

WHEN ("Second Class") THEN 3

ELSE 6

END

* Click APPLY or OK to execute. This calculated column will not be ready for use

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

* In the tableau, select the “data” tab on the left side. Then select “Create Calculated Field” from the drop-down menu.
* Enter “Days to Ship Scheduled” in the first field. (name of the field)
* Type in the following code-

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

* Click APPLY or OK to execute. This calculated column will not be ready for use