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

Q.1 Explain PySpark in brief?s

Spark is a platform for cluster computing that uses RDD ( Resilient Distributed Dataset).

The Python API for spark is known as the **PySpark**. The main advantage of PySpark is to do data exploration and Machine learning at scale . Spark is an engine that helps is using huge data sets (BIG Data) using batch and parallel systems. Using its Python API (PySpark) we can bring those capabilities and handle/ process the data effectively using Python.

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

Main characteristics of PySpark are:

* Powerful caching of datasets
* Real Time Computation
* Deployable with Hadoop and spark clusters
* Highly scalable
* Can work easily with RDDs
* Robust and Efficient

Q.3 What do you mean by PySpark SparkContext?

The point of entry to any spark functionality is called SparkContext. When any Spark Application is run, it triggers a driver program to start – this has the main function and from the runtime the spark context is initiated.

Q.4 What is pep 8?

PEPs are Python Enhancement proposals, and they describe and document the way python language evolves. They also provide a reference point (and a standard) for the pythonic way to write code. This is just the style guide for Python Code. It was designed to help python developers write more readable code.

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

|  |  |
| --- | --- |
| List | Tuple |
| * Lists are commonly enclosed with the square bracket. | * Tuples are enclosed with parenthesis (). |
| * They are Mutable | * They are immutable |
| * Iteration operation are slower in lists | * Iterations operations are faster while using tuples |
| * They have lots of functional support (methods/ operation) | * They have less functional support (methods/operation) |

**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 maximum (value1, value2):

// checking if value1 is greater than value2

if value1 > value2:

return value1

// checking if value2 is greater than value1

if value2 > value1:

return value2

// Since both are equal returning any of them (value1)

return value1

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\_list(values):

//creating a list for storing the unique values

unique = []

for value in values:

// if the value is not in unique list then the value is unique

if value not in unique:

// adding the unique value to the unique list

unique.append(value)

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

Code: myrdd.take(10)

Explanation:

Use the sparkContext to make the connection between the driver and the executor, followed by creating the RDD (using HDFS file). Finally execute the code myrdd.take(10). This will give the first 10 records from RDD.

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.

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.

Code**:** Sum ([Profit]) / Sum ([Sales])

Explanation:

Click “Create Calculated Field” and then give the name as “Profit ratio”. Finally execute the code (Sum ( [Profit] ) / Sum ( [Sales] ) to get the profit ratio.