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

Q.1 Explain PySpark in brief?

Ans: PySpark is a Python API build to access Apache-Spark, which is written in Scala. This therefore helps in accessing all the features that Spark by use of pythonic codes. As the name suggests it brings with it the power of Python alongwith Spark. Now python is an industry standard for building ML models and Spark is known to be the best method to handle big-data. Therefore PySpark is one of the ideal tools to build pipelines that can process and feed large amounts of data into the ML models.

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

Ans: The main characteristics of PySpark are:

1. Node Abstraction: It is impossible to address a single node
2. Network Abstraction: Only implicit communication is possible with spark and its nodes.
3. Map Reduce functionality: The developer can define custom map and reduce functions which are easily implemented by the pythonic interface.

Another important aspect is the REPL feature added due to the introduction of Python over Spark. Therefore, the codes can be tested with immediate effect.

Q.3 What do you mean by PySpark SparkContext?

Ans: Spark Context (SC) represents the connection of Pyspark to the Spark cluster installed on the machine. SC provides the interface to launch and manage the JVM. For simplicity in interfacing, it comes as an inbuilt “sc” class within PySpark. It uses the Py4J library, thus creating a Java Spark Context to connect to the JVM. Thus the PySpark SC is an api to access the available Java SC.

Q.4 What is pep 8?

Ans: PEP stands for Python Enhancement Proposal. This is a document that provides guidelines and best practices on how to write Python code with its first version coming out in 2000. PEP-8 is the 8th iteration of the document and was released in 2001. The primary focus of PEP 8 is to improve the readability and consistency of the programs written in python. The most important change in this documentation was that the line length was bound to 79 characters to enhance readability on top of the existing norms previously.

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

|  |  |
| --- | --- |
| Lists | Tuples |
| 1. Lists are mutable, that is we can change the contents of the list without changing the identity of the list. 2. The list is better for performing operations, such as insertion and deletion. 3. Iterating the contents of a list is slower 4. Lists consume more memory as each entry can be stored at different locations in the memory. 5. Lists have several built-in methods 6. The unexpected changes and errors are more likely to occur | Tuple are immutable  Tuple data type is appropriate for accessing the elements  Tuples can be iterated over comparatively faster as data is stored in single blocks of memory.  Tuple consume less memory as compared to the list as data is stored as blocks.  Tuple does not have built-in methods.  In tuples, unexpected changes and errors are hard to take place. |

**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)

Code:

def max\_2(a,b):

if a >b:

return a

else:

return b

Explanation: If a>b then the function return a to be greater than b. Else, if b>a then the second command is executed. Now if a=b then the else part is executed and value of b is returned which is perfectly fine as max(a,b) = a = 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.

Code:

Def rem\_duplicates( list\_):

unique = []

for i in range(len(list\_)):

if list\_[i] not in unique:

unique.append(list\_[i])

return unique

# An easier one line code is:

List2 = list(set(list\_))

Explanation: the first function iterates over the list and checks if a value is there or not in the unique. Now since unique was initialised as null the first entry goes in by default, Rest entries are then compared by the “not in” check. For the second code snippet, the set function is called which removes the duplicates in the list.

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

Code:

sc

rdd\_var =sc.textFile(<path>)

rdd\_var.take (10)

Explanation: We first execute sc to connect to the spark cluster. Then we make a RDD by loading hdfs file. (The second can be executed with any other file. Basically we want to create a RDD from the data.) Once the RDD is made we collect the first 10 rows by the take(10) command.

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.

Code:

CASE ([Ship\_mode])

WHEN ("Same Day") THEN 0

WHEN ("First Class") THEN 1

WHEN ("Second Class") THEN 3

ELSE 6

END

Explanation:

The case statement switches between the entries of the variable “Ship\_mode”. When compares the values and then compares associates it to the modified shipping days.

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: To execute this code, first choose “Create Calculated Field” and fill the first field as the desired name of the column( Say “Profit\_ratio”). Type in “the code written above. This calculates the profit ratio, i.e. profit per unit sales.