PUBLICIS SAPIENT ASSESSMENT

SUBJECTIVE ANSWERS

Answer 1)

PySpark is an API that is built to support Apache Spark in python. In order to handle Big Data Analysis, we use Apache Spark since it is a distributed framework. Apache Spark is written in Scala and its integration with languages such as Python, Java & SQL is relatively easy which makes it more developer friendly. Spark is basically a computational engine that work with huge sets of data by processing them in batches parallelly.

Answer 2)

Characteristics of PySpark:

1) Easy Integration with other languages – PySpark framework supports a variety of languages such as Python, Java & SQL
2) Resilient Distributed Datasets – PySpark helps developers to easily work with RDD
3) Coaching & Disk Persistence – This has a powerful caching and disk persistence mechanism for datasets that makes it incredibly faster than the others
4) Network – The nodes are abstracted which is why it is not possible to address an individual node. Also, the network is abstracted so no implicit communication is possible
5) Map-Reduce – PySpark is based on Map-Reduce so the developer provides a map and a reduce function.

Answer 3)

The entry point to any Spark functionality is known as SparkContext. On running any Spark application, a driver programs having a main function starts which leads to initiation of SparkContext. The driver program then runs the operations inside the executors on worker nodes.
SparkContext uses Py4j to launch a JVM and creates a Java SparkContext. By default, PySpark has SparkContext available as ‘sc’, so creating a new SparkContext won’t work.

Answer 4)

PEP-8 is a document that provide guidelines and best practices on how to write a code in Python. PEP-8 convention was written in 2001 with the primary focus to improve the readability and consistency of Python codes.
PEP stands for Python Enhancement Proposal and there as several of them. PEP-8 is just one of its examples. A PEP is a document that describes new features proposed for Python and documents aspects of Python, like design and style of the community.

Answer 5)

List	Tuple
Lists are mutable	Tuples are immutable
Lists have several inbuilt methods	Tuples don’t have built-in methods
Lists consume more memory	Tuples consume less memory when compared to Lists
Iterations in List is time consuming	Iterations in Tuples are faster than Lists
Lists are better for performing operations such as insertion and deletion	Tuples are generally used for accessing the elements
Unexpected changes and errors are more likely to occur in Lists	Unexpected changes and errors are less likely to occur in Tuples

PSEUDO CODE / CODING ANSWERS

Answer 1)

```
def max_of_two(x , y):

        if x >= y:
                return x
        else:
                return y

num1 = 199
num2 = 299

maximum = max_of_two(num1 , num2)
print(maximum)
```

Answer 2)

```
def remove_duplicates(sample_list):
        temp = []

        for index in sample_list:
                if index not in temp:
                        temp.append(index)

        return temp

test_list = [1,1,2,3,4,4,5,6,7,7]
unique_list = remove_duplicates(test_list)
print(unique_list)
```

Answer 3)

CASE 1 – If we need to sort and then take the top 10

STEPS –

1) Using SparkContext, make a connection between a driver and an executor
2) Using HDFS (Hadoop Distributed File System) or sc.parallelize command, make a RDD
3) After the RDD is make, we need to order the entries in ascending or descending order according to our requirements. Perform an action .takeOrdered to get the deterministic output.
4) Then use .take(10) to get the first 10 records.

CASE 2 – If we do not need to sort and take the top 10 as they are in the database

STEPS –

1) Using SparkContext, make a connection between a driver and an executor
2) Using HDFS (Hadoop Distributed File System) or sc.parallelize command, make a RDD
3) After the RDD is made, we use .take(10) to get the first 10 records.

Answer 4)

Code:

```
CASR ([Ship Mode])
WHEN ("Same Day") THEN 0
WHEN ("First Class") THEN 1
WHEN ("Second Class") THEN 3
WHEN ("Standard Class") THEN 6
END
```

Explanation:

1) In the data tab on the left side, go to the drop-down box beside the filter icon and choose "Create Calculated Field"
2) Type in the name "Days to Ship Scheduled" in the first field
3) Type in "<CODE Any Option>" to formulate the column using the values in the case statement. Based on the class of the shipping that the customer has ordered with the number of days being defined by "THEN" statement for each "WHEN" clause

4) Click “OK” to execute. This column will now be available in the left pane in the data tab.

Answer 5)

Code:

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

Explanations:

1) In the data tab on the left side go to the drop-down box beside the filter icon and choose “Create Calculated Field”.
2) Type in the name “PROFIT RATIO” in the first field.
3) Type in the above code to calculate the profit ratio (profit per unit sales).
4) Click “OK” to execute. This column will now be available in the left pane in the data tab.