ASSIGNMENT------SOLUTION SUBMISSION
ON
AZURE ANALYTICS
BY

NAME : SAI KIRAN ANCHE	**ROLL NO:** DXC262AB12021
BATCH:DXC-262-ANALYTICS-B12-AZURE	**COMPANY –** DXC TECHNOLOGY
TRAINING UNDER : MANIPAL PRO LEARN	**TRAINER NAME** – MR. AJAY KUMAR
DATE OF SUBMISSION : 06-06-2022	**NO OF QUESTIONS** :10
EMPLOYEE DOMAIN - AZURE ANALYTICS	

QUESTIONS:

Assignement - 6th June 2022:
----

1. Explain what is in-Memory computation in details?

2. Explain advantages of Spark framework ?

3. Explain components of Spark with block diagram ?

4. Explain benifits of in-Memory computation ?

5. Explain major difference between Hadoop & Spark ?

6. Explain features of Spark?

7. Write a Py-Spark program to create Dataframe from RDD & explain with screenshots & steps ?

8. Explain what is RDD & why it is needed ?

9. Write a Py-Spark program to make the column in Upper case & explain with screenshots & steps ?

Please create a word / pdf document, and send it to : avyuktitraining1@gmail.com

INTRODUCTION

This Assignment is given by manipal pro learn team on the basis of the training done in the forenoon session of this morning. The main objective behind this assignment is to master the theory and enhance knowledge over spark , py-spark etc...

There are 9 questions and they are of easy to moderately difficult level. All the questions have been focused on what the trainer taught in the earlier sessions.Some questions have been answered partially due to unavailability of access.

This assignment gave me immense confidence in mastering the domain that has been assigned to me.

ANSWERS

1. Explain what is in-memory computation in detail?

A: *In memory computing is a computing technique in which all the computer calculations are done in the Computer memory i.e, in the Computer RAM storage. The entire computer calculations are done in the RAM and there will be the elimination of all slow running process in the background and thus it runs faster. The in-memory computation is extensively applied to solve complex problems in the RAM of the Computer Pools server.*

2. Explain advantages of spark frame work.

A: *Some of the advantages of Spark frame work include :*

- *It is an Open-source frame work.*
- *It is very fast in processing data.*
- *It is very easy to use.*
- *Supports various libraries.*
- *Light weight.*
- *Supports real-time streaming.*

3. Explain components of spark with block diagram.

A: *The fundamental components of spark include:*

SPARK SQL	*SPARK* *STREAMING* *REAL-TIME*	*MILB MACHINE* *LEARNING*	*GRAPH X* *GRAPH PROCESSING*

SPARK CORE

SPARK CORE:

It is the heart if the spark frame work and it looks after the core functionality. It holds various components required for performing various actions.

SPARK SQL:

The Spark SQL is build on the spark core and it also provides support to the structured data

SPARK STREAMING:

Spark Streaming is a Spark component that supports scalable and fault-tolerant processing of streaming data.

MILB MACHINE LEARNING:

It is a Machine Learning Library that has various machine learning algorithms.

GRAPH X GRAPH PROCESSING:

It is a library that is used to manipulate the graphs and perform graph- parallel operations.

4. Explain benifits if in-memory computation.

A:

Some of the benefits of the in-memory computation include:

- *Better and faster decision making.*
- *Economic*
- *Profitable.*
- *Less risky*
- *Highly efficient.*
- *Identification of competitive opportunities.*

5. Explain the differences between Hadoop and spark.

A:

Hadoop	**Spark**
Hadoop is an open source framework which uses a MapReduce algorithm.	Spark is lightning fast cluster computing technology, which extends the MapReduce model to efficiently use with more type of computations.
Hadoop's MapReduce model reads and writes from a disk, thus slow down the processing speed	Spark reduces the number of read/write cycles to disk and store Intermediate data in-memory, hencefaster-processing speed.
Hadoop is a high latency computing framework, which does not have an interactive	Spark is a low latency computing and mode can process data interactively.
With Hadoop MapReduce, a developer can only process data in batch mode only	Spark can process real-time data, from real time events like twitter, facebook
Hadoop is a cheaper option available while comparing it in terms of cost	requires a lot of RAM to run in memory, thus increasing the cluster and hence cost.

6. Explain feauters of spark.

A:

Some of the feauters of spark include:

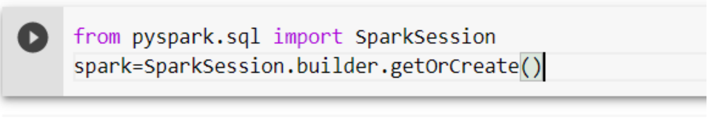
- *Fast/quick .*
- *Less complex and easy to use.*
- *Supports real-time streaming.*
- *Supports various libraries.*

7. Write a py-spark program to create dataframe from RDD and explain with screen shots.

A:

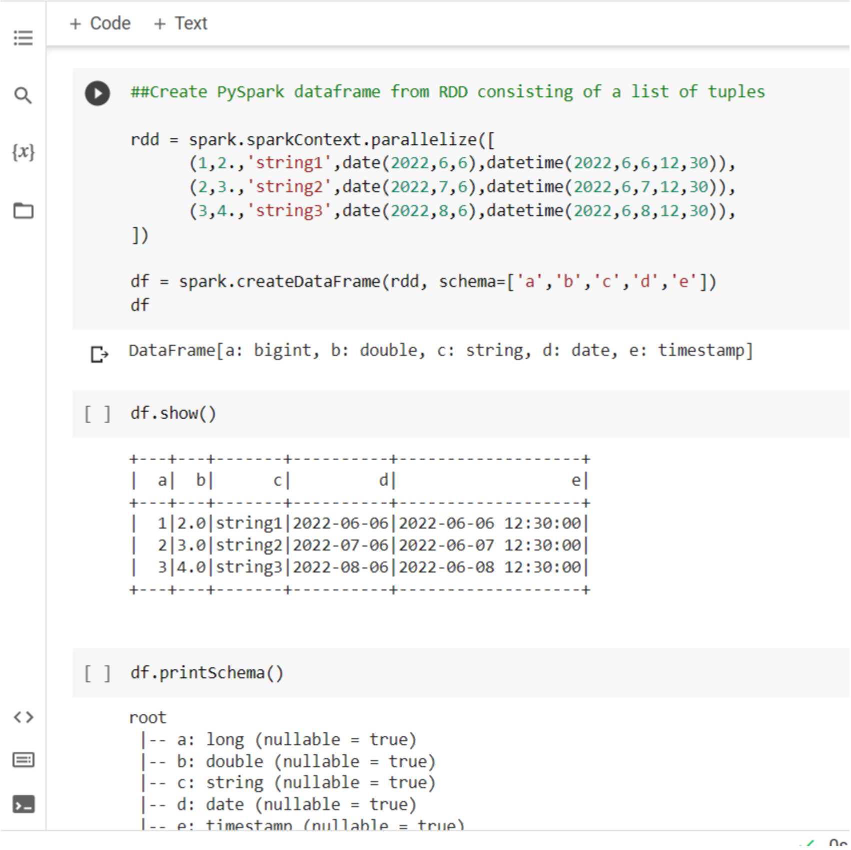
Step1:

Create PySpark RDD



```
from pyspark.sql import SparkSession
spark=SparkSession.builder.getOrCreate()
```

Step2:enter the data into the dataframe



```
##Create PySpark dataframe from RDD consisting of a list of tuples

rdd = spark.sparkContext.parallelize([
      (1,2.,'string1',date(2022,6,6),datetime(2022,6,6,12,30)),
      (2,3.,'string2',date(2022,7,6),datetime(2022,6,7,12,30)),
      (3,4.,'string3',date(2022,8,6),datetime(2022,6,8,12,30)),
])

df = spark.createDataFrame(rdd, schema=['a','b','c','d','e'])
df
```

```
DataFrame[a: bigint, b: double, c: string, d: date, e: timestamp]
```

```
[ ]  df.show()
```

```
+---+---+-------+----------+-------------------+
|  a|  b|      c|         d|                  e|
+---+---+-------+----------+-------------------+
|  1|2.0|string1|2022-06-06|2022-06-06 12:30:00|
|  2|3.0|string2|2022-07-06|2022-06-07 12:30:00|
|  3|4.0|string3|2022-08-06|2022-06-08 12:30:00|
+---+---+-------+----------+-------------------+
```

```
[ ]  df.printSchema()
```

```
root
 |-- a: long (nullable = true)
 |-- b: double (nullable = true)
 |-- c: string (nullable = true)
 |-- d: date (nullable = true)
 |-- e: timestamp (nullable = true)
```

8. Explain what is RDD? And why is RDD needed?

A:

RDD – Resilient Distributed Dataset: It is the fundamental building block of Spark. RDD (Resilient Distributed Dataset) is the core abstraction of Spark.

It's a collection of components that have been partitioned throughout the cluster's nodes so that we may run different concurrent operations on it.

RDDs may be created in two ways:

It provides in-memory processing computation by parallelizing existing data in the driver application and referencing a dataset in an external storage system, such as a shared filesystem, HDFS, HBase, or any data source delivering a Hadoop InputFormat. This implies that the state of memory is stored as an object across all tasks, and the object may be shared across them.

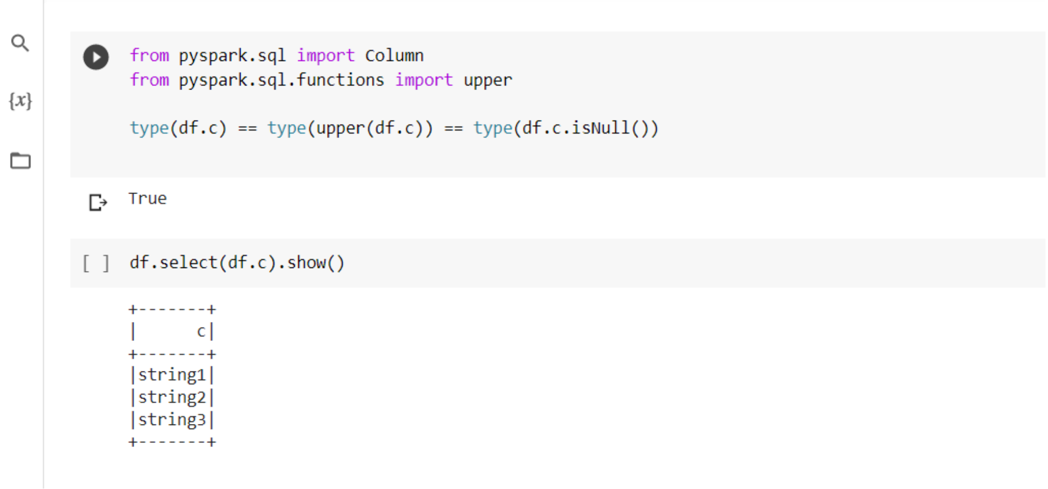
9. Write a Py-spark program to make the column in upper case & explain with screenshots&steps.?

A:

Step 1:

Import pyspark sql functions and run

Type(df,c)== type(upper(df.c))==type(df.c.isNull())



```
from pyspark.sql import Column
from pyspark.sql.functions import upper

type(df.c) == type(upper(df.c)) == type(df.c.isNull())
```

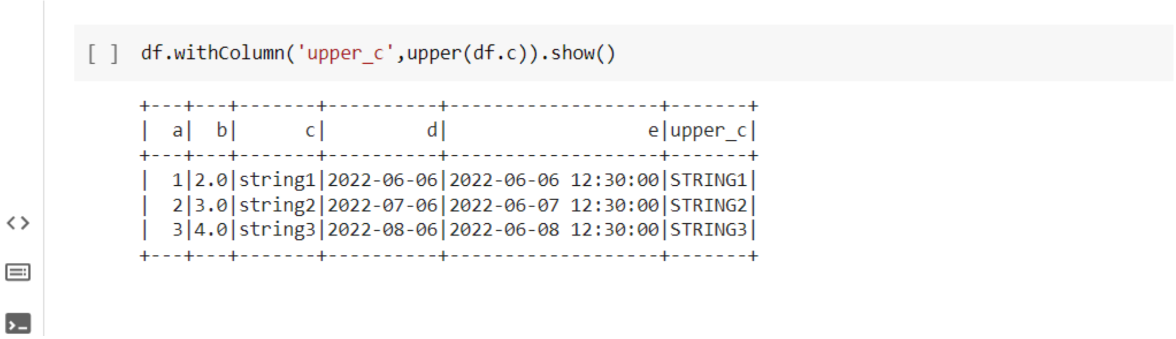
```
True
```

```
df.select(df.c).show()
```

```
+-------+
|      c|
+-------+
|string1|
|string2|
|string3|
+-------+
```

Step2:

Put the command df.withColumn('upper_c',upper(df.c)).show()



```
[ ]  df.withColumn('upper_c',upper(df.c)).show()
```

```
+---+---+-------+----------+-------------------+-------+
|  a|  b|      c|         d|                  e|upper_c|
+---+---+-------+----------+-------------------+-------+
|  1|2.0|string1|2022-06-06|2022-06-06 12:30:00|STRING1|
|  2|3.0|string2|2022-07-06|2022-06-07 12:30:00|STRING2|
|  3|4.0|string3|2022-08-06|2022-06-08 12:30:00|STRING3|
+---+---+-------+----------+-------------------+-------+
```

RESULT

Almost all the test questions have been solved and presented successfully in the present document except few due to lack of data .

CONCLUSIONS

All the questions have been solved successfully with all the concepts that have been covered in the training session. It’s really a great experience of learning while solving the cases. This assignment gave me immense confidence regarding my ability to upskill in new technologies.