Assignment-3

The due date for submitting this assignment has passed. As per our records you have not submitted this assignment. Due on 2019-03-20, 23:59 IST.

1) In spark, a ______________________ is a read-only collection of objects partitioned across a set of machines that can be rebuilt if a partition is lost.

- Spark Streaming
- Resilient Distributed Dataset (RDD)
- FlatMap
- Driver

No, the answer is incorrect.

Score: 0

Accepted Answers:
Resilient Distributed Dataset (RDD)

2) Given the following definition about the join transformation in Apache Spark:

```scala
def join[W](other: RDD[(K, W)]: RDD[(K, (V, W))]
```

Where join operation is used for joining two datasets. When it is called on datasets of type (K, V) and (K, W), it returns a dataset of (K, (V, W)) pairs with all pairs of elements for each key.

Output the result of joinrdd, when the following code is run.

```scala
val rdd1 = sc.parallelize(Seq(("m",55),("m",56),("e",57),("e",58),("s",59),("s",54)))
val rdd2 = sc.parallelize(Seq(("m",60),("m",65),("s",61),("s",62),("h",63),("h",64)))
val joinrdd = rdd1.join(rdd2)
joinrdd.collect
```

Output:

```
Array[(String, (Int, Int))]: Array((m,(55,60)), (m,(55,65)), (m,(56,65)), (m,(56,65)), (s,(59,61)), (s,(59,62)), (h,(63,64)), (s,(54,61)), (s,(54,62)))
```

© 2014 NPTEL - Privacy & Terms - Honor Code - FAQs -
Array[(String, (Int, Int))] = Array((m,(55,60)), (m,(55,65)), (m,(56,60)), (m,(56,65)), (s,(59,61)), (s,(59,62)), (s,(54,61)), (s,(54,62)))

3) Consider the following statements in the context of Spark:

Statement 1: Spark also gives you control over how you can partition your Resilient Distributed Datasets (RDDs)

Statement 2: Spark allows you to choose whether you want to persist Resilient Distributed Dataset (RDD) onto disk or not.

- Only statement 1 is true
- Only statement 2 is true
- Both statements are true
- Both statements are false

No, the answer is incorrect.
Score: 0

Accepted Answers:
Both statements are true

4) ___________ leverages Spark Core fast scheduling capability to perform streaming analytics.

- MLlib
- Spark Streaming
- GraphX
- RDDs

No, the answer is incorrect.
Score: 0

Accepted Answers:
Spark Streaming

5) _______________ is a distributed graph processing framework on top of Spark.

- MLlib
- Spark streaming
- GraphX
- All of the mentioned

No, the answer is incorrect.
Score: 0

Accepted Answers:
GraphX

6) Consider the following statements:

Statement 1: Scale out means grow your cluster capacity by replacing with more powerful machines

Statement 2: Scale up means incrementally grow your cluster capacity by adding more COTS machines (Components Off the Shelf)

- Only statement 1 is true
- Only statement 2 is true
- Both statements are true
Both statements are false
No, the answer is incorrect.
Score: 0
Accepted Answers:
Both statements are false

7) Which of the following is not a NoSQL database?

- HBase
- SQL Server
- Cassandra
- None of the mentioned

No, the answer is incorrect.
Score: 0
Accepted Answers:

8) Which of the following are the simplest NoSQL databases?

- Key-value
- Wide-column
- Document
- All of the mentioned

No, the answer is incorrect.
Score: 0
Accepted Answers:
Key-value

9) Point out the incorrect statement in the context of Cassandra:

- It is originally designed at Facebook
- It is a centralized key-value store
- It is designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure.
- It uses a ring-based DHT (Distributed Hash Table) but without finger tables or routing

No, the answer is incorrect.
Score: 0
Accepted Answers:
It is a centralized key-value store