Question 10 Max. score: 2.00

A company wants to make use of Azure Databricks and Azure Data Lake Storage Gen2. You have to ensure that the data in the Data Lake Storage is accessed by using a service principal from Azure Databricks. Which of the following would you implement for this requirement?

- Use the shared access signature in Data lake storage
- Use access keys in Data lake storage
- Create an application registration in Azure AD
- Use a secret from Azure Key vault

Question 9 Max. score: 2.00

You are working on Kubernetes using *minikube*. Which of the following code snippet is the correct way to interact with your Kubernetes cluster and deploy an application to minikube?

Code

1.

```
kubectl create deploy -minikube1 -image=k8s.gcr.io/echoserver:1.4
kubectl expose deploy -minikube1 -type=LoadBalancer -port=8080
```

2.

```
kubectl create deployment -minikube1 -image=k8s.gcr.io/echoserver:1.4
kubectl expose deployment -minikube1 -type=LoadBalancer -port=8080
```

3.

```
kubectl create deploy hello-minikube1 -image=k8s.gcr.io/echoserver:1.4
kubectl expose deploy hello-minikube1 -type=LoadBalancer --port=8080
```

3.

```
kubectl create deploy hello-minikube1 -image=k8s.gcr.io/echoserver:1.4
kubectl expose deploy hello-minikube1 -type=LoadBalancer --port=8080
```

4.

```
kubectl create deployment hello-minikube1 --image=k8s.gcr.io/echoserver:1.4
kubectl expose deployment hello-minikube1 --type=LoadBalancer --port=8080
```

Question 8 Max. score: 4.00

Alice is working on the following dataset. Bob has asked her to pick 3 and 15 from the given dataset and asked her to implement two clusters from the dataset. If she is required to implement K-Means clustering algorithm to perform this action, then determine those two clusters in this scenario:

Dataset

{1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 15, 20, 28, 30}

Clusters

1. {1, 2, 3, 4, 5, 6, 8, 9} and {10, 11, 15, 20, 28, 30}
2. {1, 2, 3, 4, 5, 6, 8} and {9, 10, 11, 15, 20, 28, 30}
3. {1, 2, 3, 4, 5, 6, 8, 9, 10, 11} and {15, 20, 28, 30}
4. {1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 15} and {20, 28, 30}

1

Question 7 Max. score: 6.00

You have a collection 'employees' in a database 'company' as given below.
Collection: employees

```
{ "_id" : ObjectId("60d1bdd023275c561c3d8e7f"), "emp_id" : 3209, "emp_name" : "Lisa Davis", "salary" : 50000 }
{ "_id" : ObjectId("60d1bdd023275c561c3d8e80"), "emp_id" : 9012, "emp_name" : "Tom Cook", "salary" : 20000 }
{ "_id" : ObjectId("60d1bdd023275c561c3d8e81"), "emp_id" : 1035, "emp_name" : "Shital Aggarwal", "salary" : 60000 }
{ "_id" : ObjectId("60d1bdd023275c561c3d8e82"), "emp_id" : 5642, "emp_name" : "Kristen Stokes", "salary" : 17000 }
{ "_id" : ObjectId("60d1bdd023275c561c3d8e83"), "emp_id" : 2465, "emp_name" : "Mike Johnson", "salary" : 40000 }
```

You need to perform the given operation on the database. Find the 'emp_type' for every employee based on the following condition. If the salary of an employee is greater than or equal to 40000, then display 'emp_type' as 'full-time'. Otherwise, display it as 'part_time'. Return the fields 'emp_id' and 'emp_type' for all documents as the result set.
How will you write a query to get the required results using the given condition?

1.

```
db.employees.find([
{ $aggregate: { $addField: { emp_id,  emp_type: { $if: { $salary:{ $gte:[ 40000 ]} ,
"fulltime"}, else: "part-time" } } } }
])
```

2.

```
db.employees.addField([
{ $aggregate:{ emp_id,
 emp_type:{ $cond: { $if: { $gte: ["salary", 40000 ] }, $then: "full-time", $else:"part-time" } } }}
])
```

3.

```
db.employees.aggregate([
{ $project: {emp_id: 1, emp_type:{$cond:{ if:{$gte:[ "$salary", 40000]}, then:"full-time", else: "part-time" }}} }
])
```

4.

```
db.employees.find([
{ $select: {emp_id: 1, emp_type:{ $cond:{{if:{ $gte:[ "$salary", 40000 ], true["full-time"] }, else: "part-time"} } } }
])
```

Question 6

Consider the table *Students* that contains information about Students *id*, their *age*, and the *subject* they opted.

Table: Students

id	subject	age
1	A	21
1	B	21
2	C	23
3	D	24

Now, to convert the above *Students* table into ***2NF*** you have decomposed it into 2 tables named *Students_1, Students_2*.

Table: Students_1

id	age
1	21
2	23
3	24

Now, which of the following tables will be *Students_2* so that the table *Students* will be converted into *2NF*?

Now, which of the following tables will be *Students_2* so that the table *Students* will be converted into *2NF*?

Tables

1.

id	subject
1	A
1	B
2	C
3	D

2.

id	subject
1	A,B
2	C
3	D

3.

id	subject
1	A
NULL	B
2	C
3	D

4.

id	subject
NULL	A
1	B
2	C
3	D

Question 5 **Max. score: 4.00**

You are calling the *reduceByKey(func, [numTasks])* property on a DStream of *(K, V)* pairs that return a new DStream of *(K, V)* pairs.
It uses the default number of parallel tasks that are available in the Spark framework. Which of the following correctly represents the value if you are calling the property in the cluster mode?

- 0
- 1
- 2
- Determined by the config property spark.default.parallelism

Reset Answer

Question 4 Max. score: 6.00

You are using the *dataFrame.cache()* method that is provided by the Spark SQL module to cache tables by using an in-memory columnar format. To minimize the memory usage and GC pressure, the Spark SQL module scans only the required columns and automatically tunes the compression process. Which of the following will you use to remove the table from the memory?

- spark.catalog.uncacheTable("tableName")
- spark.catalog.cacheTable("tableName")
- spark.sql.inMemoryColumnarStorage.compressed("tableName")
- spark.sql.inMemoryColumnarStorage.catchTable("tableName")

Reset Answer

Question 3 Max. score: 4.00

In Apache HBase, you are working on the HBase shell. You have a table named *MyHackData* that contains the list of skills that are available in HackerEarth's library. You have observed there are more than 1000 skills in the table. Now, if you are required to fetch only 100 records at a time from your entire data, then which of the following commands can be used to perform this action in this scenario:

Options

1.

```
countData 'MyHackData', CACHE => 1000
```

2.

```
countRows 'MyHackData', CACHE = 1000
```

3.

```
count 'MyHackData', CACHE =>1000
```

3.

```
count 'MyHackData', CACHE =>1000
```

4.

```
count 'MyHackData', CACHE ==1000
```

Question 2 **Max. score: 2.00**

In Hadoop, you are working on HDFS architecture. If you are required to perform the following actions, then which of the following elements is used to perform these actions in this scenario:

Actions

1. Manage the file system namespace
2. Regulate the client's access to files
3. Execute file system operations such as renaming, closing, and opening files and directories

- Datanode
- Namenode
- Mapreduce cluster
- None of these

Question 1 **Max. score: 6.00**

You want to create a table in Hive such that the table is clustered by a hash function of userid into 32 buckets. Within each bucket, the data is sorted in increasing order of viewTime. In order to achieve this, you use the code snippet given alongside.

Analyze the given scenario and determine which of the following can be achieved by doing this?
1. Allows the user to do efficient sampling on the clustered column - in this case, userid
2. Allows internal operators to take advantage of the better-known data structure while evaluating queries with greater efficiency.

```
CREATE TABLE page_view(viewTime INT, userid BIGINT,
                page_url STRING, referrer_url STRING,
                ip STRING COMMENT 'IP Address of the User')
COMMENT 'This is the page view table'
PARTITIONED BY(dt STRING, country STRING)
CLUSTERED BY(userid) SORTED BY(viewTime) INTO 32 BUCKETS
ROW FORMAT DELIMITED
        FIELDS TERMINATED BY '1'
        COLLECTION ITEMS TERMINATED BY '2'
        MAP KEYS TERMINATED BY '3'
STORED AS SEQUENCEFILE;
```