esprit Se former autrement	**EXAMEN** Semestre : 1 ☐ 2 ■ Session : Principale ■ Rattrapage ☐
Unité d'enseignement : Valorisation des données massives Module (s) : Big data analytics & Deep learning Classe(s) : 4DS Nombre des questions : ……. Nombre de pages : …… Date : 01/07/2020 Heure 16h15 Durée : 1h00.	

1. Laquelle des commandes suivantes est correcte si vous souhaitez vérifier les contenus d'un fichier texte stocké sur HDFS sans premièrement le copier manuellement dans le stockage local ?
   a. C'est impossible de le faire
   b. Hdfs dfs -vi /user/Hadoop/myFile.txt
   c. Hdfs dfs -nano /user/Hadoop/myFile.txt
   d. Hdfs dfs -cat /user/Hadoop/myFile.txt
2. Quel type de contrainte peut-on ajouter à une table Hive ?
   a. Clé primaire
   b. Clé étrangère
   c. Clé unique
   d. Aucune des clés précédentes
3. Qu'est ce qui se passe à la suite de la suppression d'une base de données Hive ?
   a. Les tables seront aussi supprimées
   b. Le répertoire dédié à la base sera supprimé s'il n'y a pas de table
   c. Les blocks HDFS seront formatés
   d. Aucune des réponses précédentes
4. Etant donné la tâche suivante : vous souhaitez importer une table depuis une base de données MySQL appelée « retail_db » dans Hive. Qu'est ce qu'il manque dans la commande d'importation Sqoop ci-dessous à la ligne 5 ?

```
1  sqoop import \
2  --connect jdbc:mysql://localhost:3306/retail_db \
3  --username=root \
4  --hive-import \
5  ****
6  --m=1
```

   a. Une requête de type formulaire libre en utilisant une clause WHERE
   b. Le nom de la table que vous souhaitez importer

c. Des informations de partitionnement

d. L'emplacement du nœud HDFS.

5. Quel est l'effet de la ligne 7 dans la commande d'importation Sqoop suivante ?

```
1  sqoop import-all-tables \
2  --connect jdbc:mysql://localhost:3306/retail_db \
3  --username=root \
4  --compression-codec=snappy \
5  --as-avrodatafile \
6  --warehouse-dir=/user/hive/warehouse \
7  -m 1
```

a. Elle spécifie le nombre de partitions à importer

b. Elle spécifie le nombre de fichiers résultants sur HDFS

c. Elle spécifie le nombre de tâches de mappage à utiliser pour l'importation en parallèle

d. Elle est spécifique uniquement à l'exportation Sqoop et ignorée pour l'importation Sqoop

6. Laquelle des réponses suivantes n'est pas un type valide de sink pour Flume ?

a. HDFS sink

b. HBASE sink

c. HTTP sink

d. JDBC sink

7. Etant donné le fichier de configuration Flume suivant, quel est le but de l'agent Flume résultant ?

```
# Name the source, channel and sink
flume_importer.sources = avro-source
flume_importer.channels = jdbc-channel
flume_importer.sinks = file-sink

# Source configuration
flume_importer.sources.avro-source.type = avro
flume_importer.sources.avro-source.port = 11112
flume_importer.sources.avro-source.bind = localhost

# Describe the sink
flume_importer.sinks.file-sink.type = hdfs
flume_importer.sinks.file-sink.hdfs.path = /user/hadoop/sink
flume_importer.sinks.file-sink.hdfs.fileType = DataStream
flume_importer.sinks.file-sink.hdfs.fileSuffix = .avro
flume_importer.sinks.file-sink.serializer = avro_event
flume_importer.sinks.file-sink.serializer.compressionCodec=snappy

# Describe the type of channel
flume_importer.channels.jdbc-channel.type = jdbc

# Bind the source and sink to the channel
flume_importer.sources.avro-source.channels = jdbc-channel
flume_importer.sinks.file-sink.channel = jdbc-channel
```

a. Il diffuse les données encodées avro depuis une source avro vers HDFS

b. Il diffuse les données encodées avro depuis HDFS vers une source avro
c. Il diffuse les données encodées avro depuis kafka vers HDFS
d. Il diffuse les données encodées snappy depuis une source avro vers HDFS

8. Laquelle des réponses suivantes est considérée comme le principal avantage de spark par rapport à Mapreduce ?
   a. Spark supporte scala
   b. Spark peut exécuter des calculs en mémoire
   c. Spark supporte les RDD
   d. MapReduce supporte le calcul distribué
9. Lequel des langages suivants n'est pas supporté par Apache spark
   a. Java
   b. Scala
   c. Python
   d. Ruby
10. Lequel des composants suivants ne fait pas partie de l'écosystème de Spark ?
   a. Spark core
   b. Spark Name-Node
   c. Spark GraphX
   d. Spark MLlib
11. Laquelle des réponses suivantes est l'une des responsabilités de Spark Driver ?
   a. Il est un nœud travailleur (worker node) responsable des processus individuels d'un job de spark.
   b. Il stocke les métadonnées sur le job en cours mais il ne contient pas les données réelles.
   c. Il est responsable de l'allocation des ressources aux job lancés.
   d. Il stocke les résultats des calculs en mémoire, cache ou sur le disque.
12. Laquelle des réponses suivantes n'est pas une caractéristique valide d'un RDD ?
   a. L'évaluation paresseuse
   b. La tolérance à la panne
   c. L'immuabilité
   d. L'évolutivité
13. Etant donné la commande suivante exécutée avec pyspark, quel est le type de données résultant ?

```
sqlContext = SQLContext(sc)
data = sqlContext.read.json("data_cours/sparksql/employee.json")
```

   a. Un objet FileInputStream
   b. Un dataframe ou chaque élément est une ligne depuis le fichier
   c. Un pair RDD ou la clé est le chemin de fichier et la valeur du contenu du fichier

d. Un RDD ou chaque élément est une ligne depuis le fichier.

14. Laquelle des réponses ci-dessous n'est pas une opération valide supportée par un RDD ?
    a. Regroup()
    b. Count()
    c. parallelize()
    d. filter()
15. Etant le fragment de code suivant, quel est le résultat du programme ?

```
from pyspark import SparkContext
sc =SparkContext.getOrCreate()
RDD = sc.parallelize([0,1,2,3,4,5,6,7,8,9,10])
RDD1 = RDD.map(lambda x : x+2)
```

    a. Le code ne sera pas compilé
    b. [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]
    c. Le code affichera les numéros dans un ordre non déterminé
    d. Le code sera compilé mais il n'affichera rien