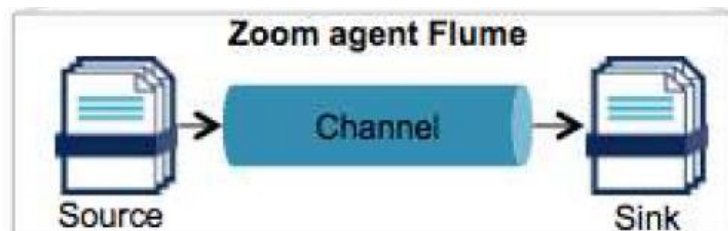
 Se former autrement	EXAMEN	
	Semestre : 1 <input checked="" type="checkbox"/> 2 <input type="checkbox"/>	Session : Principale <input checked="" type="checkbox"/> Rattrapage <input type="checkbox"/>
Module: Big Data Analytics Enseignant(s): Adel Jebali, Ines Channoufi, Ines Slimene Classe(s) :4 ERP-BI Documents autorisés : OUI <input type="checkbox"/> NON <input checked="" type="checkbox"/> Calculatrice autorisée : OUI <input type="checkbox"/> NON <input checked="" type="checkbox"/> Date : 13/01/2020 Heure 09h Nombre de pages : 03 Internet autorisée : OUI <input type="checkbox"/> NON <input checked="" type="checkbox"/> Durée :1h30		

Exercice 1 : (10 points)

1. Présenter l'architecture de *Hadoop2* et définir le rôle de chaque composant. (2 pts)
2. Un cluster *ElasticSearch* est constitué de plusieurs nœuds. Représenter par un schéma, l'architecture d'un cluster *ElasticSearch*. (2 pts)
3. Dans *Kibana*, quel est l'intérêt de définir des Index Pattern ? (1 pt)
4. La figure ci-dessous représente un agent *Flume*. (1,5 pts)



Définir et donner des exemples de chaque composant.

5. Lors de l'importation des données avec *Flume*, on distingue que certains fichiers importés sont stockés avec l'extension *.tmp*. Expliquer. (1 pt)
6. On souhaite importer toutes les données de la base de données **REGIONS** de *MySQL* dans *Hive* avec *Sqoop*. Expliquer la différence entre les deux instructions ci-dessous (1,5 pts)

```
sqoop import-all-tables --connect jdbc:mysql://localhost:3306/regions --username=root
--warehouse-dir=/user/hive/warehouse/regions.db --m=1
```

et

```
sqoop import-all-tables -m 1 \  
--connect jdbc:mysql://localhost:3306/regions \  
--username=root \  
--compression-codec=snappy \  
--as-avrodatafile \  
--warehouse-dir=/user/hive/warehouse
```

7. Donnez les requêtes qui permettent de vérifier que les fichiers ont été bien importés dans *Hive*. (1 pt)

Exercice 2 : (5 points)

On considère le script suivant permettant de créer deux RDD :




```
from pyspark import SparkContext  
sc = SparkContext.getOrCreate()  
rdd1 = sc.parallelize(["Saturday", "Sunday", "Monday", "Tuesday"])  
rdd2 = sc.parallelize(["Wednesday", "Thursday", "Friday", "Saturday"])
```





Ecrire les commandes *Spark* permettant de :

1. Afficher les deux premiers éléments du rdd1.
2. Afficher le nombre d'élément du rdd1.
3. Afficher le nombre de partitions du rdd2.
4. Afficher tous les éléments du rdd2.
5. Afficher les éléments résultants de l'union de rdd1 et rdd2.

Exercice 3 : QCM (5 points)

Cocher la bonne réponse :

1. Que signifie l'action dans *Spark* RDD :
 - a. Le moyen d'envoyer le résultat des exécuteurs vers le driver
 - b. Prend RDD en entrée et produit un ou plusieurs RDD en sortie 
 - c. Crée un ou plusieurs nouveaux RDD
 - d. Tout ce qui précède
2. Les lacunes de *Hadoop MapReduce* ont été surmontées par *Spark* RDD à travers :
 - a. Lazy-evaluation (Évaluation paresseuse) 
 - b. DAG 
 - c. In-memory processing (Traitement en mémoire)
 - d. Tout ce qui précède

3. Lequel des éléments suivants est le point d'entrée de l'application *Spark* ?
- SparkSession
 - SparkContext
 - Aucun des deux
4. *Apache Spark* supporte :
- Batch processing
 - Stream processing
 - Graph processing
 - Tout ce qui précède
5. Lequel des énoncés suivants est vrai pour un RDD ?
- RDD est un paradigme de programmation
 - RDD dans Apache Spark est une collection immuable d'objets 
 - C'est une base de données
 - Aucune de ces réponses
6. Lequel des énoncés suivants n'est pas une transformation ?
- Flatmap 
 - Map 
 - Reduce 
 - Filter
7. Lors de la création d'un nouveau RDD :
- Les données seront chargées directement avec la création de la structure
 - Les données seront chargées lorsque la structure est évaluée
8. Quelle est la différence entre un RDD et un dataframe ?
- Le même concept
 - En plus des données, les dataframes sont accompagnés d'un schéma.
 - Un dataframe consiste en un ensemble de n-uplets bruts.
9. Quelle est la différence entre Apprentissage supervisé et non supervisé ?
- En apprentissage supervisé les données sont étiquetées cependant en apprentissage non supervisé, les données ne sont pas étiquetées.
 - En apprentissage supervisé les données ne sont pas étiquetées cependant en apprentissage non supervisé, les données sont étiquetées.
10. En apprentissage supervisé, quelle est la différence entre les familles Classification et Régression ?
- En Classification, la cible est discrète. En régression, la cible est continue.
 - En Classification, la cible est continue. En régression, la cible est discrète.