Cloud computing et Big Data

Examen – session 2

Daniel Hagimont

Durée: 1h30, documents autorisés

Questions Cloud (10 points)

Quel est le mécanisme essentiel permettant de mettre en œuvre la consolidation des machines virtuelles dans
un datacenter ? (1 point)
Quel est l'avantage de la virtualisation assistée par le matériel par rapport aux autres techniques de
virtualisation (1 point)
Dans AWS, à quoi servent les key-pairs ? (1 point)
Quels sont les 2 mécanismes de AWS permettant de mettre en œuvre la scalabilité horizontale (1 point)
Un Utilisateur de cloud veut utiliser une application. Un fournisseur lui propose une solution SaaS et une solution IaaS. Quel différence cela fait il pour l'utilisateur ? (1 point)

Vous disposez d'un datacenter composé de 15 machines physiques, chacune d'elle ayant 2 processeurs physiques. Ce datacenter héberge 28 VMs ayant chacune un processeur virtuel (vCPU). Chaque propriétaire de VM a réservé la capacité d'un processeur physique pour sa VM. Le tableau suivant présente le niveau d'utilisation de chaque vCPU pendant 1 heure.

Nombre de VM à ce niveau de charge CPU	Niveau de charge CPU (%)
6	100 % chacune
15	75 % chacune
3	50 % chacune
4	25 % chacune

Dans les questions suivantes, on ne s'intéresse qu'à la ressource CPU. On dispose de deux consolidateurs. Le premier (consolidateur par défaut) se base sur les quantités CPU réservées par les machines virtuelles (même si celles-ci ne sont pas utilisées). Le second consolidateur dispose d'un outil de mesure de l'utilisation du CPU dans chaque machine virtuelle et se base sur cette utilisation réelle du CPU.

- 1. Combien de machines physiques le consolidateur par défaut pourra t-il atteindre ? (1 point)
- 2. Qu'en sera-t-il avec le second consolidateur ? (2 points)
- 3. Sachant que la charge d'une VM peut varier brusquement, quel est l'inconvénient du second consolidateur ? (1 point)
- 4. Que proposez vous pour y remédier ? (1 point)

Questions Big Data (10 points)

On considère un grand fichier contenant les notes des tous les étudiants de France au Baccalauréat dans les différentes matières (une ligne donne pour un étudiant sa note dans une matière). Vous disposez dans un programme Spark du RDD suivant (qui a été initialisé avec un fichier disponible dans HDFS) :

JavaRDD <string> data = sc.textFile(inputFile);</string>
Une ligne de ce fichier a le format suivant : 12345 français 14
Le premier champs est le numéro (INE) de l'étudiant, le second est la matière et le troisième est la note obtenue. On suppose que chaque étudiant a 10 notes et que le fichier est cohérent (il ne manque aucune note).
Donnez l'implantation Java d'une fonction (PairFunction) permettant d'extraire d'une ligne de ce fichier le numéro de l'étudiant et la note (1 points)
Donnez le programme Spark qui calcule la moyenne de chaque étudiant (3 points)

Donnez l'implantation Java de la fonction (PairFunction) permettant d'extraire d'une ligne du fichier les
données utiles et expliquez ce que cette fonction retourne (2 points)
Donnez le programme Spark qui calcule la moyenne nationale par matière (4 points)
Donnez le programme spark qui calcule la moyenne nationale par matiere (4 points)

On suppose maintenant que les étudiants ont un nombre variable de notes. Cela implique que le nombre de notes par matière est lui aussi variable. On veut calculer la moyenne au niveau national dans chaque matière.