Chapitre 10
Mettre en œuvre un cluster Hadoop

1. Objectif du chapitre

L'objectif de ce chapitre est de comparer différentes options possibles pour mettre en œuvre un cluster Hadoop, ainsi que les différents coûts à prendre en compte. En effet, considérer que Hadoop, parce qu'il est un projet Open Source, ne coûte rien serait aller un peu vite en besogne : mettre en place un cluster, serait-ce en version totalement open source, entraîne au minimum des coûts matériels et humains qui ne sont pas toujours négligeables.

2. Cluster dédié ou cluster dans le Cloud ?

Le premier choix qui se présente à un directeur informatique qui se lance dans Hadoop est de décider s'il souhaite opter :

– Pour un cluster dédié, qu'il soit physiquement installé dans les locaux de l'entreprise ou chez un hébergeur.

– Pour un cluster dans le Cloud, dont il partagera éventuellement les ressources avec d'autres clients du prestataire de services.

Le tableau ci-après essaye de résumer les principaux avantages et inconvénients de ces trois grandes options, que l'organisme concerné soit une entreprise ou un autre organisme.

Option	Avantages	Inconvénients
1. Cluster dédié installé dans les locaux de l'organisme	– Totale maîtrise du cluster (matériels, logiciels, réseaux). – Grande disponibilité des ingénieurs système, qui sont présents sur place. – Totale liberté au niveau des moyens et du planning pour faire fonctionner et évoluer le cluster. – Seule option qui garantisse à 100% la confidentialité des données (si pas de réseau vers l'extérieur).	– Investissement initial en euros et en temps peut être plus important que pour les autres options. – Risque de tout vouloir faire en interne (Hadoop est un outil disposant de très nombreuses possibilités de paramétrage, pas toujours faciles à maîtriser).

Option	Avantages	Inconvénients
2. Cluster dédié installé chez un hébergeur	– Tarifs souvent compétitifs (mutualisation des ressources). – Tarifs plus progressifs que dans l'option 1 (effets de seuils réduits). – Équipes gérant l'hébergement normalement très professionnelles. – Fiabilité et disponibilité élevées. – Dans certaines offres, les principaux logiciels de l'écosystème de Hadoop sont préinstallés. – Permet de concentrer les efforts sur les tâches directement productives.	– Configurations matérielle, logicielle et réseau pas forcément connues de manière précise. – Limitations techniques (tout n'est pas autorisé). – Possibilités d'optimisation du cluster plus limitées. – Moins grande disponibilité des ingénieurs système (au moins pour la partie du système gérée par l'hébergeur). – Pas de véritable assurance quant à la confidentialité des données. – Moins de liberté de choix au niveau des logiciels de l'écosystème de Hadoop dans certains cas.

Option	Avantages	Inconvénients
3. Cluster partagé dans le Cloud	– Tarifs souvent compétitifs (pas d'investissement + mutualisation des ressources). – Tarifs très progressifs (peu ou pas d'effets de seuil). – Aspects techniques, en particulier concernant les réseaux, largement masqués pour le client. – Équipes gérant le Cloud normalement très professionnelles. – Fiabilité et disponibilité du cluster élevées. – Si les données Hadoop sont déjà stockées dans le Cloud, il est logique du point de vue de Hadoop de les traiter au même endroit (cf. proximité des données).	– Irréversibilité du processus (surtout si les volumes de données en jeu sont importants). – Configurations matérielle, logicielle et réseau pas forcément connues de manière précise. – Possibilités d'optimisation du cluster réduites. – Les aspects juridiques sont plus complexes car la localisation du cluster n'est pas forcément connue, ou qu'elle se situe à l'étranger. – Pas de véritable assurance quant à la confidentialité des données. – Moins de liberté de choix au niveau des logiciels de l'écosystème de Hadoop dans certains cas.

Comme souvent, chaque option a ses avantages et ses inconvénients. Le choix entre l'une ou l'autre se fait donc au cas par cas. On peut en particulier distinguer différentes phases dans un projet Hadoop, l'option retenue pour une phase pouvant être différente de celle choisie pour une autre phase.

En première approche on peut considérer que :

- Opter pour un cluster partagé dans le Cloud :
  - Est une option intéressante dans la phase de démonstration du concept (*Proof of Concept*) car elle est très souple et facile à mettre en œuvre, pour un coût forcément réduit (un démonstrateur requiert une fraction de la puissance du cluster opérationnel).
  - Peut être plus problématique dans la phase d'exploitation du fait que ses performances risquent d'être pénalisées par la virtualisation, et une architecture réseau qui n'est pas forcément optimisée.
- Opter pour un serveur hébergé :
  - Est une option peut-être un peu contraignante à mettre en œuvre dans la phase de démonstration.
  - Est une option envisageable dans la phase d'exploitation, sous réserve.
    - De s'assurer que les connexions réseau sont adaptées aux types de traitement qu'envisage l'organisme.
    - De disposer d'une équipe d'informaticiens de haut niveau, car la résolution de problèmes affectant des systèmes distants peut s'avérer plus complexe que dans le cas d'une solution purement interne.
- Opter pour un cluster dédié installé dans les locaux de l'entreprise :
  - Est une option envisageable dans la phase de démonstration (le plus souvent les coûts matériels sont limités du fait du faible nombre de nœuds nécessaires).
  - Est une option qui est aussi envisageable en exploitation, même si elle peut sembler, parfois à juste titre, plus coûteuse que les deux autres options.

En résumé, on peut dire :

- Que n'importe laquelle des trois options peut être mise en œuvre dans la phase de démonstration, en prêtant une attention particulière aux performances réseau si l'option 2 ou l'option 3 est retenue.
- Que le choix de l'option dans la phase d'exploitation doit être précédée d'une analyse de type forces-faiblesses - opportunités-menaces qui tienne compte des besoins et des moyens (matériels, humains, financiers, etc.) de l'organisme. Une évaluation du TCO (*Total Cost of Ownership*), ou coût du cycle de vie, pourra être faite.

3. Les coûts

3.1 Cluster dédié installé dans les locaux de l'organisme

3.1.1 Les coûts de formation

La première chose à faire avant de se lancer dans l'aventure que constitue la mise en œuvre d'un cluster Hadoop, quelle que soit l'option retenue, est de se former un minimum.

Dans tous les cas, les décideurs pourront suivre une formation courte présentant de manière non technique Hadoop et son écosystème. À l'issue de cette formation, ils auront une idée précise de ce qu'est Hadoop, ainsi que de la manière dont le Big Data peut offrir de nouvelles perspectives à leur organisme.

Si l'organisme souhaite maîtriser l'installation et l'exploitation de son cluster Hadoop, il est indispensable :

– Qu'au moins un ingénieur système suive une formation d'administrateur Hadoop.
– Qu'au moins un développeur suive une formation de programmeur Hadoop incluant différents langages (Java, Python, HiveQL, Pig Latin, etc.).

Le cas échéant, un analyste de données pourra aussi suivre une ou deux formations spécifiques.

Cloudera propose, directement ou par le biais de la société Global Knowledge, une série de formations assez complète (cf. tableau ci-après). Ces formations peuvent se donner en présentiel (en inter ou en intra) ou à distance via Internet.

La plupart des autres grands fournisseurs de logiciels et de services Hadoop proposent des formations similaires à celles de Cloudera, à des prix très proches.