**Pourquoi choisir Hadoop ?**

<https://data-flair.training/blogs/features-of-hadoop-and-design-principles/>

**1. Hadoop est Open Source**

Hadoop est un projet open source, ce qui signifie que son code source est disponible gratuitement pour inspection, modification et analyse, ce qui permet aux entreprises de modifier le code selon leurs besoins.

**2. Le cluster Hadoop est hautement évolutif**

Le cluster Hadoop est évolutif signifie que nous pouvons ajouter n'importe quel nombre de nœuds ( évolutif horizontal ) ou augmenter la capacité matérielle des nœuds ( évolutif vertical ) pour atteindre une puissance de calcul élevée. Cela fournit une évolutivité horizontale et verticale au framework Hadoop.

**3. Hadoop fournit une tolérance aux pannes**

La tolérance aux pannes est la caractéristique la plus importante de Hadoop. [HDFS](https://translate.google.com/website?sl=auto&tl=fr&hl=fr&u=https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html) dans Hadoop 2 utilise un mécanisme de réplication pour fournir une tolérance aux pannes.

Il crée une réplique de chaque bloc sur les différentes machines en fonction du facteur de réplication (par défaut, c'est 3). Ainsi, si une machine d'un cluster tombe en panne, les données sont accessibles à partir des autres machines contenant une réplique des mêmes données.

Hadoop 3 a remplacé ce mécanisme de réplication par un codage d'effacement . Le codage d'effacement offre le même niveau de tolérance aux pannes avec moins d'espace. Avec le codage Erasure, la surcharge de stockage ne dépasse pas 50 %.

Lisez l'article sur le [codage](https://data--flair-training.translate.goog/blogs/hadoop-hdfs-erasure-coding/?_x_tr_sl=auto&_x_tr_tl=fr&_x_tr_hl=fr) d'effacement pour apprendre l'algorithme de codage d'effacement.

**4. Hadoop offre une haute disponibilité**

Cette fonctionnalité de Hadoop garantit la haute disponibilité des données, même dans des conditions défavorables.

En raison de la fonction de tolérance aux pannes de Hadoop, si l'un des DataNodes tombe en panne, les données sont disponibles pour l'utilisateur à partir de différents DataNodes contenant une copie des mêmes données.

De plus, le cluster Hadoop haute disponibilité se compose de 2 ou plusieurs NameNodes (actifs et passifs) en cours d'exécution dans une configuration de secours automatique. Le nœud actif est le NameNode, qui est actif. Le nœud passif est le nœud de secours qui lit la modification des journaux d'édition du NameNode actif et les applique à son propre espace de noms.

Si un nœud actif tombe en panne, le nœud passif prend le relais du nœud actif. Ainsi, même si le NameNode tombe en panne, les fichiers sont disponibles et accessibles aux utilisateurs.

**5. Hadoop est très rentable**

Étant donné que le cluster Hadoop se compose de nœuds de matériel de base peu coûteux, il fournit ainsi une solution rentable pour le stockage et le traitement de données volumineuses. Étant un produit open source, Hadoop n'a besoin d'aucune licence.

**6. Hadoop est plus rapide dans le traitement des données**

Hadoop stocke les données de manière distribuée, ce qui permet de traiter les données de manière distribuée sur un cluster de nœuds. Ainsi, il fournit une capacité de traitement ultra-rapide au framework Hadoop.

**7. Hadoop est basé sur le concept de localité des données**

Hadoop est populairement connu pour sa fonctionnalité de localisation des données qui consiste à déplacer la logique de calcul vers les données, plutôt que de déplacer les données vers la logique de calcul. Cette fonctionnalité de Hadoop réduit l'utilisation de la bande passante dans un système.

Pour installer et configurer Hadoop suivez ce [guide d'installation](https://data--flair-training.translate.goog/blogs/installation-of-hadoop-3-on-ubuntu/?_x_tr_sl=auto&_x_tr_tl=fr&_x_tr_hl=fr) .

**8. Hadoop fournit la faisabilité**

Contrairement au système traditionnel, Hadoop peut traiter des données non structurées. Ainsi, offrez aux utilisateurs la possibilité d'analyser des données de tous formats et de toutes tailles.

**9. Hadoop est facile à utiliser**

Hadoop est facile à utiliser car les clients n'ont pas à se soucier de la distribution de l'informatique. Le traitement est géré par le framework lui-même.

**10. Hadoop assure la fiabilité des données**

Dans Hadoop, en raison de la réplication des données dans le cluster, les données sont stockées de manière fiable sur les machines du cluster malgré les pannes de machine.

Le framework lui-même fournit un mécanisme pour assurer la fiabilité des données par Block Scanner, Volume Scanner, Disk Checker et Directory Scanner. Si votre machine tombe en panne ou si les données sont corrompues, vos données sont également stockées de manière fiable dans le cluster et sont accessibles depuis l'autre machine contenant une copie des données.