~~Développé par Apache, Hadoop est la principale plateforme du Big Data.~~ ~~Il est utilisé pour le stockage et le traitement d’immenses volumes de données.~~ ~~C’est un framework logiciel open source.~~ Ses différents composants sont utilisés par de nombreuses entreprises pour leurs projets Big Data.

~~Hadoop repose sur des modèles de programmation simples pour assurer le traitement des données et les rendre disponibles sur des machines locales.~~

~~Hadoop est un Framework Java libre permettant de développer des applications distribués et échelonables (scalables). Il est utilisé pour les bases de données et regroupe de nombreuses librairies.~~

~~Hadoop est basé sur le HDFS (Hadoop Distributed File System) qui est un modèle de stockage et sur l’algorithme MapReduce qui permet d’effectuer des calculs parallèles et d’avoir des données fiables.~~

### **~~Qu'est-ce qu'Hadoop ?~~**

~~Hadoop est un framework logiciel développé par Apache Software Foundation pour le stockage et le traitement distribués d'énormes quantités d'ensembles de données.~~ **~~[Hadoop](https://data--flair-training.translate.goog/blogs/hadoop-tutorial/?_x_tr_sl=auto&_x_tr_tl=fr&_x_tr_hl=fr)~~** ~~se compose de 3 composants principaux :~~

#### ~~1. HDFS (High Distributed File System = système de fichiers hautement distribué)~~

~~C'est la~~ **~~couche de stockage de Hadoop~~**~~. Les fichiers dans HDFS sont divisés en blocs de la taille d'un bloc. HDFS se compose de deux types de nœuds, à savoir NameNode et DataNodes.~~

* ~~NameNode stocke des métadonnées sur l'emplacement des blocs.~~
* ~~DataNodes stocke le bloc et envoie des rapports de bloc à NameNode dans un intervalle de temps défini.~~

#### ~~2. MapReduce~~

[~~MapReduce~~](https://data--flair-training.translate.goog/blogs/hadoop-mapreduce-tutorial/?_x_tr_sl=auto&_x_tr_tl=fr&_x_tr_hl=fr) ~~est la~~ **~~couche de traitement dans Hadoop~~**~~. Il s'agit d'un cadre logiciel pour écrire une application qui effectue un traitement distribué.~~

#### ~~3. YARN~~

~~C'est la~~ **~~couche de gestion des ressources~~**~~.~~ **~~YARN~~** ~~est responsable de~~ ~~l'allocation des ressources et de la planification des travaux.~~

~~Résumé Hadoop :~~

~~En bref, on peut dire que Hadoop est un framework open-source. Hadoop est surtout connu pour sa tolérance aux pannes et sa fonctionnalité de haute disponibilité. Les clusters Hadoop sont évolutifs. Le framework Hadoop est facile à utiliser.~~

~~Il assure un traitement rapide des données grâce au traitement distribué. Hadoop est rentable. La fonctionnalité de localisation des données Hadoop réduit l'utilisation de la bande passante du système.~~

**Hadoop**

**Qu'est-ce qu'Hadoop ?**

~~Hadoop est une pile logicielle open source qui~~ s'exécute sur un cluster de machines. Hadoop fournit un stockage distribué et un traitement distribué pour de très grands ensembles de données.

**~~Quelle est la licence de Hadoop ?~~**

~~Hadoop est un logiciel open source. Il s'agit d'un projet Apache publié sous Apache Open Source License v2.0. Cette licence est très commerciale.~~

**~~Qui contribue à Hadoop ?~~**

~~À l'origine, Hadoop a été développé et open source par Yahoo. Maintenant, Hadoop est développé en tant que projet Apache Software Foundation et compte de nombreux contributeurs de Cloudera, Horton Works, Facebook, etc.~~

**A qui est destiné Hadoop ?**

Hadoop est conçu pour traiter le Big Data. Donc, si vous n'avez pas de "problème de Big Data", alors Hadoop n'est probablement pas la meilleure solution pour votre entreprise. Mais avant d'arrêter de lire ici, s'il vous plaît lisez la suite :-)

La quantité de données considérées comme Big Data diffère d'une entreprise à l'autre. Pour certaines entreprises, 10 To de données seraient considérées comme du Big Data ; pour d'autres, 1 PB correspondrait au "Big Data". Donc, vous seul pouvez déterminer combien coûte le Big Data.

De plus, si vous n'avez pas de "problème de Big Data" maintenant, est-ce parce que vous ne capturez pas certaines données ? Dans certains scénarios, les entreprises ont choisi de renoncer à la capture de données, car il n'existait aucun moyen pratique de les stocker et de les traiter. Maintenant que Hadoop peut aider avec le Big Data, il peut être possible de commencer à capturer des données qui n'ont pas été capturées auparavant.

**Quel est le coût d’Hadoop ?**

Hadoop est open source. Le logiciel est gratuit. Cependant, l'exécution de Hadoop a d'autres composants de coût.

* Coût du matériel : Hadoop tourne sur un cluster de machines. La taille du cluster peut aller de 10 nœuds à des milliers de nœuds. Pour un grand cluster, les coûts matériels seront importants.
* Le coût de l'IT / OPS pour mettre en place un grand cluster Hadoop et le prendre en charge devra être pris en compte.
* Étant donné que Hadoop est une technologie plus récente, trouver des personnes pour travailler sur cet écosystème n'est pas facile.

## De quelles compétences ai-je besoin pour apprendre Hadoop ?

Un développeur ou un administrateur expérimenté peut apprendre Hadoop. La liste suivante est un début - sans ordre particulier

* Hadoop est écrit en Java. Donc connaître Java aide
* Hadoop fonctionne sous Linux, vous devez donc connaître les compétences de base de la navigation en ligne de commande Linux
* Certaines compétences en script Linux iront loin

## Je ne suis pas programmeur, puis-je quand même utiliser Hadoop ?

Oui, vous n'avez pas besoin d'écrire du code Java Map Reduce pour extraire des données de Hadoop. Vous pouvez utiliser Pig et Hive. Pig et Hive proposent tous deux une réduction de carte de « haut niveau ». Par exemple, vous pouvez interroger Hadoop en utilisant SQL dans Hive.

## Quels types d'outils de développement sont disponibles pour Hadoop ?

Les outils de développement Hadoop évoluent encore. Voici quelques-uns :

* Karmasphere IDE : optimisé pour le développement pour Hadoop
* Eclipse et autres IDE Java : lors de l'écriture de code Java
* Éditeur de ligne de commande comme VIM : Peu importe l'éditeur que vous utilisez, vous éditerez beaucoup de fichiers/scripts. Il est donc essentiel de se familiariser avec les éditeurs CLI.

## Comment Hadoop résout le problème du Big Data ?

### **Hadoop est conçu pour fonctionner sur un cluster de machines**

Commençons par un exemple. Disons que nous devons stocker beaucoup de photos. Nous allons commencer avec un seul disque. Lorsque nous dépassons un seul disque, nous pouvons utiliser quelques disques empilés sur une machine. Lorsque nous maximisons tous les disques sur une seule machine, nous devons obtenir un groupe de machines, chacune avec un groupe de disques.

C'est exactement ainsi que Hadoop est construit. Hadoop est conçu pour fonctionner sur un cluster de machines dès le départ.

### **Les clusters Hadoop évoluent horizontalement**

Plus de puissance de stockage et de calcul peut être obtenue en ajoutant plus de nœuds à un cluster Hadoop. Cela élimine le besoin d'acheter du matériel de plus en plus puissant et coûteux.

### **Hadoop peut gérer des données non structurées / semi-structurées**

Hadoop n'applique pas de "schéma" aux données qu'il stocke. Il peut gérer du texte arbitraire et des données binaires. Ainsi, Hadoop peut « digérer » facilement toutes les données non structurées.

### **Les clusters Hadoop fournissent le stockage et l'informatique**

Nous avons vu qu'avoir des clusters de stockage et de traitement séparés n'est pas la meilleure solution pour le Big Data. Les clusters Hadoop fournissent le stockage et l'informatique distribuée tout en un.

## Analyse de rentabilisation pour Hadoop

### **Hadoop fournit du stockage pour le Big Data à un coût raisonnable**

Le stockage de Big Data à l'aide d'un stockage traditionnel peut être coûteux. Hadoop est construit autour de matériel de base. Par conséquent, il peut fournir un stockage assez important pour un coût raisonnable. Hadoop a été utilisé sur le terrain à l'échelle Peta octet.

Une étude de Cloudera a suggéré que les entreprises dépensent généralement entre 25 000 et 50 000 dollars par téraoctet et par an. Avec Hadoop, ce coût tombe à quelques milliers de dollars par téraoctet par an. Et le matériel devient de moins en moins cher, ce coût continue de baisser.

### **Hadoop permet de capturer de nouvelles ou plus de données**

Parfois, les organisations ne capturent pas un type de données, car le coût de leur stockage était trop élevé. Étant donné que Hadoop fournit un stockage à un coût raisonnable, ce type de données peut être capturé et stocké.

Un exemple serait les journaux de clics sur le site Web. Étant donné que le volume de ces journaux peut être très élevé, peu d'organisations les ont capturés. Désormais, avec Hadoop, il est possible de capturer et de stocker les journaux

### **Avec Hadoop, vous pouvez stocker des données plus longtemps**

Pour gérer le volume de données stockées, les entreprises purgent périodiquement les anciennes données. Par exemple, seuls les journaux des 3 derniers mois ont pu être stockés et les journaux plus anciens ont été supprimés. Avec Hadoop, il est possible de stocker les données historiques plus longtemps. Cela permet d'effectuer de nouvelles analyses sur des données historiques plus anciennes.

Par exemple, prenez les journaux de clics d'un site Web. Il y a quelques années, ces journaux étaient stockés pendant une brève période pour calculer des statistiques telles que des pages populaires, etc. Maintenant, avec Hadoop, il est viable de stocker ces journaux de clics plus longtemps.

### **Hadoop fournit des analyses évolutives**

Il ne sert à rien de stocker toutes les données, si nous ne pouvons pas les analyser. Hadoop fournit non seulement un stockage distribué, mais également un traitement distribué. Cela signifie que nous pouvons traiter un grand volume de données en parallèle. Le framework de calcul de Hadoop s'appelle Map Reduce. Map Reduce a été prouvé à l'échelle des péta octets.

### **Hadoop fournit des analyses riches**

Native Map Reduce prend en charge Java comme langage de programmation principal. D'autres langages comme Ruby, Python et R peuvent également être utilisés.

Bien sûr, écrire du code Map Reduce personnalisé n'est pas le seul moyen d'analyser les données dans Hadoop. Map Reduce de niveau supérieur est disponible. Par exemple, un outil nommé Pig prend l'anglais comme langage de flux de données et les traduit en Map Reduce. Un autre outil, Hive, prend les requêtes SQL et les exécute à l'aide de Map Reduce.

Les outils de Business Intelligence (BI) peuvent fournir un niveau d'analyse encore plus élevé. De nombreux outils de BI peuvent fonctionner avec Hadoop et analyser les données stockées dans Hadoop.

**HDFS**

### Problème : Les données sont trop volumineuses sur un seul ordinateur

Les mégadonnées d'aujourd'hui sont « trop volumineuses » pour être stockées dans UN SEUL ordinateur, quelle que soit sa puissance et la quantité de stockage dont il dispose. Cela élimine de nombreux systèmes de stockage et bases de données conçus pour des machines uniques. Nous allons donc construire le système pour qu'il fonctionne sur plusieurs ordinateurs en réseau. Le système de fichiers ressemblera à un système de fichiers unique unifié pour le monde "extérieur"

#### Solution Hadoop : Les données sont stockées sur plusieurs ordinateurs

### Problème : Les machines très haut de gamme coûtent cher

Maintenant que nous avons décidé que nous avons besoin d'un groupe d'ordinateurs, de quel type de machines s'agit-il ? Les machines de stockage traditionnelles sont chères avec des composants haut de gamme, parfois avec des composants «exotiques» (par exemple, fibre channel pour les baies de disques, etc.). Évidemment, ces ordinateurs coûtent un joli centime.

Nous voulons que notre système soit rentable, nous n'allons donc pas utiliser ces machines « chères ». Au lieu de cela, nous choisirons d'utiliser du matériel de base. Nous n'entendons pas par là des machines de bureau bon marché. Nous utiliserons des machines de classe serveur performantes - mais il s'agira de serveurs de base que vous pourrez commander auprès de l'un des fournisseurs (Dell, HP, etc.)

Alors, à quoi ressemblent ces machines serveurs ? Consultez le [chapitre 14, *Matériel et logiciel pour le*](https://hadoopilluminated-com.translate.goog/hadoop_illuminated/Hardware_Software.html?_x_tr_sl=auto&_x_tr_tl=fr&_x_tr_hl=fr&_x_tr_sch=http) guide Hadoop.

#### Solution Hadoop : Exécuter sur du matériel de base

### Problème : le matériel de base échouera

À l'époque de l'informatique distribuée, les pannes étaient une exception et les erreurs matérielles n'étaient pas bien tolérées. Ainsi, les entreprises fournissant du matériel pour l'informatique distribuée ont veillé à ce que leur matériel tombe rarement en panne. Ceci est réalisé en utilisant des composants de haute qualité et en ayant des systèmes de sauvegarde (dans certains cas, une sauvegarde sur des systèmes de sauvegarde !). Ainsi, les machines sont conçues pour résister aux défaillances des composants, tout en continuant à fonctionner. Cette ligne de pensée a créé un matériel impressionnant, mais CHER !

D'un autre côté, nous allons avec du matériel de base. Ceux-ci n'ont pas de composants haut de gamme comme les cadres principaux mentionnés ci-dessus. Ils vont donc échouer - et échouer souvent. Nous devons nous préparer à cela. Comment?

L'approche que nous adopterons consiste à intégrer « l'intelligence » dans le logiciel. Ainsi, le logiciel du cluster sera suffisamment intelligent pour gérer les pannes matérielles. Le logiciel détecte les pannes matérielles et prend automatiquement des mesures correctives, sans intervention humaine. Notre logiciel sera plus intelligent !

#### Solution Hadoop : le logiciel est suffisamment intelligent pour faire face aux pannes matérielles

### Problème : une panne matérielle peut entraîner une perte de données

Nous avons donc maintenant un réseau de machines servant de couche de stockage. Les données sont réparties sur tous les nœuds. Que se passe-t-il lorsqu'un nœud échoue (et rappelez-vous, nous nous attendons à ce que les nœuds échouent). Toutes les données de ce nœud deviendront indisponibles (ou perdues). Alors, comment pouvons-nous l'empêcher?

Une approche consiste à faire plusieurs copies de ces données et à les stocker sur différentes machines. Ainsi, même si un nœud tombe en panne, les autres nœuds auront les données. C'est ce qu'on appelle la "réplication". La réplication standard est de 3 exemplaires.

#### Solution Hadoop : répliquer (dupliquer) les données

### Problème : comment les nœuds distribués vont-ils se coordonner entre eux

Étant donné que chaque machine fait partie du "stockage", nous aurons un "démon" en cours d'exécution sur chaque machine pour gérer le stockage de cette machine. Ces démons se parleront pour échanger des données.

OK, maintenant que nous avons tous ces nœuds qui stockent des données, comment nous coordonnons-nous entre eux ? Une approche consiste à avoir un MASTER pour être le coordinateur. Bien que la construction de systèmes distribués avec un coordinateur centralisé puisse sembler une idée étrange, ce n'est pas un mauvais choix. Il simplifie l'architecture, la conception et la mise en œuvre du système

Alors maintenant, notre architecture ressemble à ceci. Nous avons un seul nœud maître et plusieurs nœuds de travail.

#### Solution Hadoop : il existe un nœud maître qui coordonne tous les nœuds de travail

## 8.1. Architecture HDFS

Nous sommes maintenant à peu près arrivés à l'architecture de HDFS

Passons en revue quelques principes de HDFS. Examinons d'abord les parallèles entre «notre conception» et la conception HDFS réelle.

### Conception maître / ouvrier

Dans un cluster HDFS, il y a UN nœud maître et de nombreux nœuds de travail. Le nœud maître est appelé le nœud de nom (NN) et les travailleurs sont appelés nœuds de données (DN). Les nœuds de données stockent en fait les données. Ce sont les bêtes de somme.

Name Node est en charge des opérations du système de fichiers (comme la création de fichiers, les autorisations des utilisateurs, etc.). Sans cela, le cluster sera inopérant. Personne ne peut écrire des données ou lire des données.   
C'est ce qu'on appelle un point de défaillance unique. Nous examinerons cela plus tard.

### Fonctionne sur du matériel de base

Comme nous l'avons vu, hadoop n'a pas besoin de matériel sophistiqué et haut de gamme. Il est conçu pour fonctionner sur du matériel standard. La pile Hadoop est conçue pour faire face aux pannes matérielles et le système de fichiers continuera de fonctionner même si les nœuds échouent.

### HDFS est résilient (même en cas de panne de nœud)

Le système de fichiers continuera à fonctionner même si un nœud tombe en panne. Hadoop y parvient en dupliquant les données entre les nœuds.

### Les données sont répliquées

Alors, comment Hadoop assure-t-il la sécurité et la résilience des données en cas de défaillance d'un nœud ? Simple, il conserve plusieurs copies de données autour du cluster.

Pour comprendre le fonctionnement de la réplication, examinons le scénario suivant. Le segment de données #2 est répliqué 3 fois, sur les nœuds de données A, B et D. Supposons que le nœud de données A échoue. Les données sont toujours accessibles depuis les nœuds B et D.

### HDFS est mieux adapté aux fichiers volumineux

Les systèmes de fichiers génériques, comme les systèmes de fichiers Linux EXT, stockent des fichiers de taille variable, de quelques octets à quelques gigaoctets. HDFS, cependant, est conçu pour stocker des fichiers volumineux. Grand comme dans quelques centaines de mégaoctets à quelques gigaoctets.

Pourquoi est-ce?

HDFS a été conçu pour fonctionner avec des lecteurs de disque mécaniques, dont la capacité a augmenté ces dernières années. Cependant, les temps de recherche ne se sont pas beaucoup améliorés. Hadoop essaie donc de minimiser les recherches de disque.

### Les fichiers sont à écriture unique (non modifiables)

HDFS prend en charge l'écriture unique des fichiers (ils ne peuvent pas être mis à jour). Il s'agit d'une différence frappante entre HDFS et un système de fichiers générique (comme un système de fichiers Linux). Les systèmes de fichiers génériques permettent de modifier les fichiers.

Cependant, l'ajout à un fichier est pris en charge. L'ajout est pris en charge pour activer des applications telles que HBase.