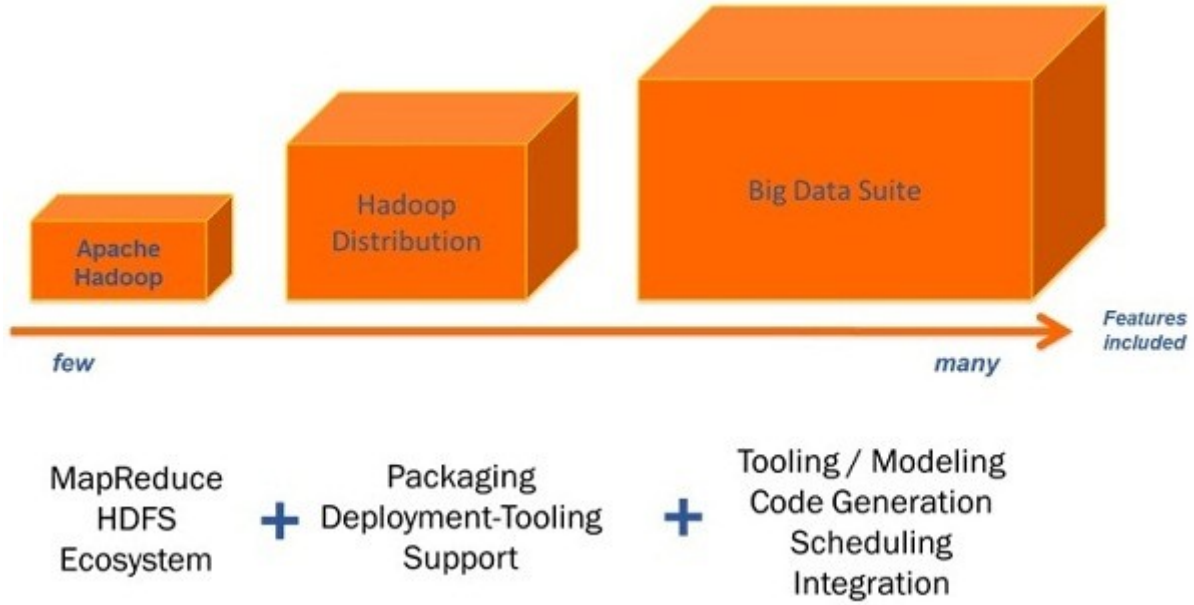
Alternative de plates-formes Hadoop



Cette illustration montre, d'une manière générale les différentes alternatives de plates-formes Hadoop. On peut soit prendre simplement la version Open Source proposer par Apache, soit choisir l'une des différentes distribution proposées par les différents fournisseurs, ou encore décider d'utiliser la package BigData d'un éditeur. Il faut comprendre que chacune de ces alternatives contient Apache Hadoop, et que presque chaque package BigData contient ou utilise une distribution.

Apache Hadoop Open Source :
cette version inclut différents modules :
- Hadoop common, les utilitaires communs pour supporter les autres modules d'Hadoop
- HDFS : un système de fichiers distribués qui fournit un accès haut-débit aux données de l'application
- Hadoop Yarn : framework pour la planification des tâches et la gestion des ressources du cluster
- Hadoop MapReduce : un système basé sur YARN pour le traitement en parallèle des gros volumes de données.

Défauts :
- une installation en mode pseudo-distribué vous aide à simuler une installation multi-noeuds avec un nœud unique. Même dans ce mode, il est nécessaire de faire plusieurs configuration. Si on souhaite faire un cluster sur plusieurs nœuds, c'est plus complexe car il faut gérer les droits d'utilisateurs, les droits d'accès et d'autres problème de ce type.
- tous les projets comme Pig, Hive ou Hbase ne sont pas installés par défaut. Il faut donc les installer manuellement.
- pas de support commercial : en effet vu qu'il s'agit d'un projet Open Source, aucun support commercial n'est disponible, de plus les entreprises qui proposent de telles support le réserve pour leur produits.

Utilisation ?
Permet de réaliser des premiers essai grâce à ses quelques dizaines de minutes d'installation en mode autonome. Bon choix si on ne souhaite pas utiliser de distribution. Cependant, pour de vrai projet hadoop prévilégier plutôt une distribution.

Prendre une distribution Hadoop

Choisir une distribution Hadoop permet de régler un certain nombre des problèmes vu précédemment. Une distribution contient généralement différent projet Hadoop, mais en plus les fournisseurs de distribution proposent des outils graphiques pour le déploiement, l'administration et le monitoring des clusters.
Il y a donc 3 grandes distribution Hadoop qui se distinguent : HortonWorks, Cloudera et MapR.

- **Cloudera** : la distribution de loin la plus installée avec le plus grand nombre de déploiements référencés. Un outillage puissant pour le déploiement, la gestion et le suivi est disponible. Impala est développée par Cloudera pour offrir des traitements temps réel de big data.
- **Hortonworks** : le seul fournisseur qui utilise 100% du projet open source Apache Hadoop sans ses propres (non-open) modifications. Hortonworks est le premier vendeur à utiliser les fonctionnalités Apache HCatalog pour des services de méta-données. Par ailleurs, leur initiative Stinger optimises massivement le projet Hive. Hortonworks offre un très bon bac à sable, facile à utiliser pour commencer. Hortonworks a développé et committé des améliorations au niveau du coeur qui rend Apache Hadoop exécutable nativement sur les plate-formes Microsoft Windows incluant Windows Server et Windows Azure.
- **MapR** : utilise quelques concepts différents de ses concurrents, en particulier du support pour un système de fichier Unix natif au lieu de HDFS(avec des composants non open source) pour une meilleure performance et une facilité d'utilisation. Les commandes natives Unix peuvent être utilisées à la place des commandes Hadoop. De plus, MapR se différencie de ses concurrents avec des fonctionnalités de haute disponibilité comme les snapshots, la réplication ou encore le basculement avec état ("stateful failover"). L'entreprise est aussi à la tête du projet Apache Drill, un projet open source réinventé à partir de Dremel de Google pour des requêtes de type SQL sur des données Hadoop afin d'offrir des traitements temps réels.

Utilisation ?
Avantages : packaging, outils et support commercial, une distribution Hadoop peut être utilisée dans la plus part des cas. Cependant, même une distribution demande un peu d'effort, il faut quand même écrire beaucoup de code pour les jobs MapReduce, ainsi que pour intégrer toutes les différentes sources de données dans Hadoop.

Le package BigData

Au dessus d'Apache Hadoop ou d'une distribution Hadoop, vous pouvez utiliser un package Big Data. Ce dernier supporte souvent différentes distributions Hadoop sous son capot. Cependant, certains fournisseurs implémentent leur propre solution Hadoop. De toute façon un package Big Data ajoute plusieurs autres caractéristiques aux distributions pour le traitement des données :

- **Outillage** : habituellement, un package Big Data est basée au dessus d'un IDE comme Eclipse. Des plugins additionnels facilitent le développement des applications big data. Vous pouvez créer, construire et déployer des services big data avec un environnement de développement familier.
- **Modélisation** : Apache Hadoop ou une distribution Hadoop offrent une infrastructure pour des clusters Hadoop. Cependant, vous devez toujours écrire beaucoup de code complexe pour construire votre programme MapReduce. Vous pouvez écrire ce code entièrement en Java, ou vous pouvez utiliser des langages optimisés tels que PigLatin ou le langage de requête Hive, qui génère du code MapReduce. Un package Big Data offre un outillage graphique pour modéliser vos services big data. Tout le code requis est généré. Vous avez juste à configurer vos jobs (c-a-d définir tous les paramètres). Réaliser les jobs big data devient plus facile et plus efficace.
- **Génération de code** : tout le code est généré. Vous n'avez pas à écrire, debugger, analyser et optimiser votre code MapReduce.
- **Plannification** : l'exécution des jobs big data doit être programmée et surveillée. Au lieu d'écrire des jobs cron ou d'autres codes pour l'ordonnancement, vous pouvez utiliser un package Big Data pour définir et gérer les plans d'exécutions facilement.
- **Intégration** : Hadoop a besoin d'intégrer des données de toutes sortes de technologies et de produits. En plus des fichiers et des bases de données SQL, vous avez aussi à intégrer des bases de donnéess NoSQL, des médias sociaux tels que Twitter ou Facebook, des middleware de messagerie ou des données de produits B2B tels que Salesforce ou SAP. Un package Big Data aide beaucoup en offrant des connecteurs à toutes ces interfaces pour Hadoop et la partie back. Vous n'avez pas à coder à la main la "colle" qui liera les différents composants, vous avez juste à utiliser les outils graphiques pour intégrer et cartographier toutes ces données. Les capacités d'intégration incluent souvent des fonctionnalités sur la qualité des données tel que le nettoyage de données pour améliorer la qualité des données importées.