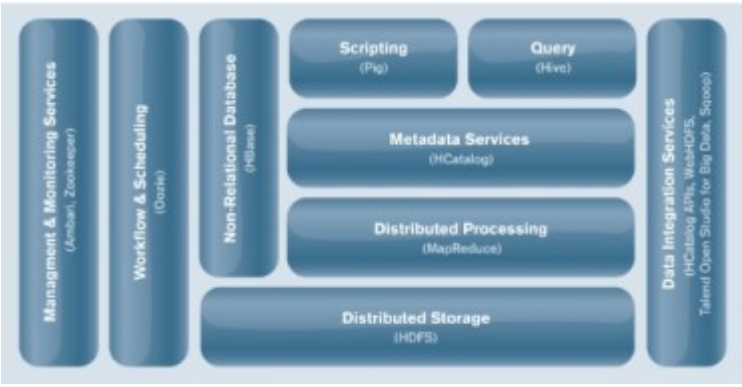
# Hortonworks révèle sa première distribution Hadoop

*Hadoop Summit 2012* : La première distribution Apache Hadoop finalisée d'Hortonworks, **HDP** (Hortonworks Data Platform), assemblage de composants open source complété d'outils d'administration et d'interopérabilité, est destinée à devenir la tête de file de la communauté Hadoop.

Après sept mois de développement, Hortonworks Data Platform v1 est une réponse aux 'nombreux' sceptiques d'Hadoop qui jugent que la plateforme **Big Data** est trop complexe. L'objectif affiché de la spin out de Yahoo!, portée financièrement par Benchmark Capital, est en effet très clairement de simplifier l'adoption, le déploiement et l'administration de Hadoop.

## Distribution Hadoop « production-ready »

À l'image de ses concurrentes Cloudera et MapR, la distribution HDP – désormais considérée comme la première version "production-ready" de Hadoop – est composée d'un ensemble de composants open source issu de la stack Apache Hadoop v1.0. Y figurent **HDFS** (Hadoop Distributed File System) pour tripler la donnée dans un cluster, la base de données **HBase**, la solution d'entrepôt de données (data warehouse) **Hive**, le moteur de script **Pig** langage de data flow, le cœur algorithmique et analytique de **MapReduce**, **Streaming** pour scripter les jobs MapReduce, et un ensemble d'outils **Hadoop Common**. À la différence de ses concurrents, cette distribution est la seule à être entièrement open source.




Hortonworks a complété cette base avec des composants destinés à l'administration d'Hadoop. **Apache Ambari** surveille et gère le cycle de vie d'une instance Hadoop sur plusieurs serveurs, **Oozie** planifie les flux de jobs MapReduce en ligne et **Zookeeper** suit et nomme les conventions dans un cluster. Les tableaux de bord peuvent s'afficher sur OpenView de HP, Viewpoint de Teradata ou vCenter de VMware via une API (Application Programming Interface) spécifique. Ils peuvent afficher bon nombre de métriques, comme l'usage des CPU, de la mémoire et des disques, l'utilisation des réseaux et la latence, le démarrage des tâches et leur nombre, ou encore le nombre et la localisation des blocs de données utilisés par une tâche.

Un catalogue de métadonnées, basé sur **Apache HCatalog**, place des pointeurs sur les tables

Hadoop afin de faciliter les requêtes fournies par des outils orientés bases de données structurées et relationnelles, de quoi simplifier le requêtage en provenance des solutions de Business Intelligence. À noter, l'intégration de la version open source de **Talend Open Studio for Big Data** et de **Sqoop**, l'outil d'Apache pour déplacer des données de Hadoop vers une base structurée, bien connu du monde Eclipse.

# L'après HDP v1

La v1 de HDP présente une limitation qui sera corrigée dans les prochaines versions : l'outil d'administration ne tourne que sur un seul cluster. Par ailleurs, Hortonworks travaille avec VMware pour permettre à vSphere et sa console vCenter de suivre et de protéger les services de base d'Hadoop que sont **NameNode HA**, la table d'allocation des fichiers sur un disque physique, et **JobTracker**, les jobs MapReduce qui concernent des données particulières résidant sur des nœuds d'un cluster, qui s'exécutent dans des instances serveurs au-dessus de l'hyperviseur ESXi, et de rediriger éventuellement les opérations de sauvegarde vers un autre cluster.

Le modèle économique de Hortonworks, dont rappelons-le la solution est intégralement open source, est basé sur le support, avec trois niveaux de prestation. Par exemple, 12 000 dollars par an pour un cluster de 10 nœuds. Nettement moins cher que ses deux concurrents qui, partant du noyau open source d'Hadoop augmenté de composants propriétaires, facturent leur distribution au nœud.

Rappelons enfin que Hortonworks a été retenu par Microsoft pour l'usage de ses services cloud Windows Azure.

---

**Voir aussi**
Dossier Silicon.fr – Microsoft Windows Azure, une plate-forme cloud enfin complète