Les distributions Hadoop

Cours 2016-2017

Qu’est ce qu’une distribution Hadoop ?

- Une distribution Hadoop est un ensemble de technologies de Hadoop, packagées par un éditeur
- Un package Hadoop contient des outils parfois développés par l’entreprise qui simplifient l’installation et l’utilisation de Hadoop.

Pourquoi une distribution ?

- Simplifie l'installation et la gestion des composants (managers)
  - Déploiement de Hadoop via une interface web
  - Monitoring des services et relance des services
- Package un ensemble de composants de Hadoop avec des outils facilitant leur utilisation (exemple : HUE)
- Support commercial
- VM pour tester

Déployer avec Ambari ou cloudera manager

- Permet d'ajouter des machines à partie de l'adress ip
- Permet de répartir les services sur les noeuds
- Affiche la santé des noeuds

Les principales distributions

- Fournissent une version packagée de Hadoop
- Développent des outils autour de Hadoop
- Proposent du conseil en entreprise
- Proposent un support technique
- Proposent des formations (et délivrent des certificats)

Cloudera

cloudera

- Premier à proposer des VMs tests
- Employeur de Doug Cutting, le “père” de Hadoop

- Interface d’administration : cloudera manager
- Interface d’utilisation : HUE
- Alternative à MapReduce : Impala

Hortonworks

- Technologie 100% open source
- Distribution la plus proche du Hadoop Open Source (nombreux contributeurs Hadoop chez Hortonworks)
- Organise les Hadoop Summit

- Interface d'administration : Ambari
- Interface d'utilisation : Ambari view (très jeune)
- Alternative à MapReduce : Tez

MapR

- Hadoop modifié selon le besoin des entreprises
- Utilise mapR-FS à la place de HDFS (un noeud est à la fois master et slave)
- (d'après les retours d'expérience), couche de sécurité plus facile à mettre en place et à utiliser (Kerberos)

- Interface d'administration : MapR control System

Hortonworks vs Cloudera

	Forrester's Weighting	Cloudera	Hortonworks	IBM	MapR Technologies	Pivotal Software
CURRENT OFFERING	50%	4.53	3.82	4.32	4.34	3.14
Solution configuration	5%	5.00	5.00	5.00	5.00	4.00
Architecture	20%	4.20	3.40	4.00	4.80	2.40
Administration	15%	5.00	4.75	3.75	4.25	3.75
Security	10%	5.00	3.00	4.32	4.34	3.00
Data	15%	4.25	3.50	3.50	4.75	3.00
Data governance	10%	5.00	3.00	5.00	3.00	3.00
Workload flexibility	10%	3.00	3.00	5.00	5.00	3.00
Development	10%	5.00	5.00	5.00	3.00	3.00
Platform integrations	5%	5.00	5.00	5.00	5.00	5.00
STRATEGY	50%	4.63	4.75	4.50	4.50	3.56
Acquisition and pricing	25%	4.50	5.00	3.00	5.00	2.25
Solution road map	25%	5.00	5.00	5.00	4.00	3.00
Ability to execute	25%	5.00	5.00	5.00	5.00	5.00
Implementation support	25%	4.00	4.00	5.00	4.00	4.00
MARKET PRESENCE	0%	4.56	4.45	3.33	3.78	2.21
Evaluated product revenue	33%	4.00	4.00	3.00	3.00	2.00
Customer base	33%	4.67	4.34	4.00	4.67	3.00
Partnerships	34%	5.00	5.00	3.00	3.66	1.66

All scores are based on a scale of 0 (weak) to 5 (strong).

Source : https://www.cloudera.com/content/dam/www/static/documents/analyst-reports/forrester-wave-big-data-hadoop-distributions.pdf

Services entreprise

Solution technique (package hadoop + outils adaptés à l'utilisation de Hadoop en entreprise)

Solution technique + support technique + conseil en entreprise

Solution technique, support, conseil + appliance + maintenance + sla

cloudera

Hortonworks

MAPR

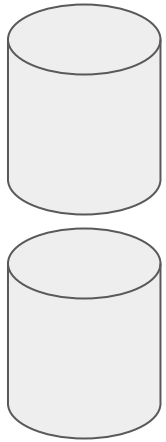
ORACLE + cloudera

TERADATA + Hortonworks

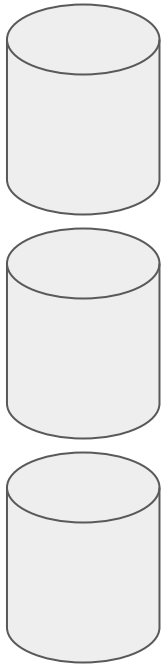
IBM

Bonnes pratiques préconisées par les distributions

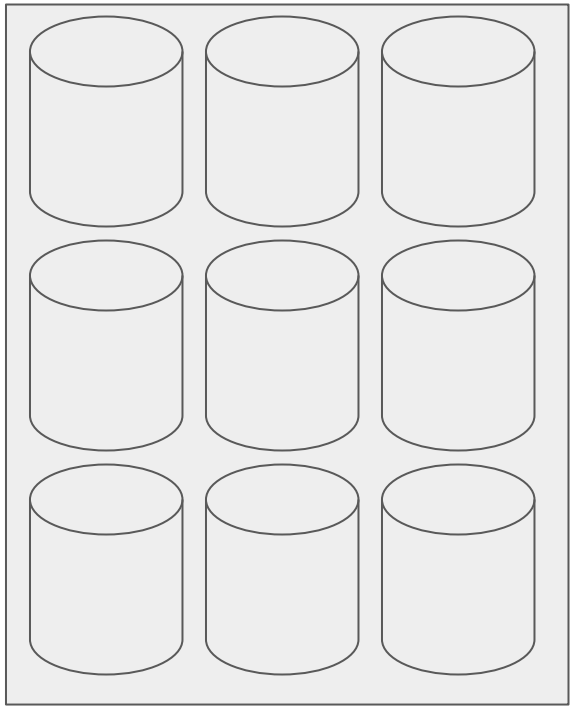
- Choisir le hardware :
  - http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.3.6/bk_cluster-planning-guide/content/conclusion.html
- Architecturer son cluster



clients hdfs, client hive, client pig, etc.



services masters (hdfs namenode, yarn rm, hive server, etc.)



services slaves (hdfs datanode, yarn node manager, etc.)