HBase, BigTable sans Google

Formateur : Franck MARIN, société Radioactiv'IT

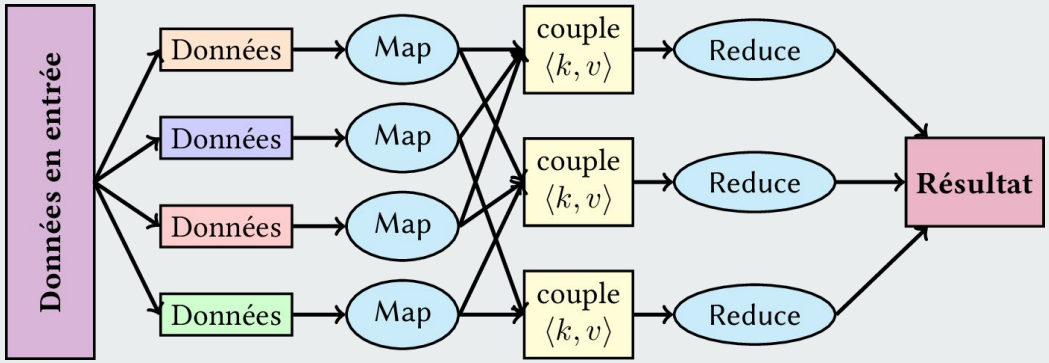
Historique de la solution

De Google à Hadoop en repassant par Google jusqu'à la fondation Apache

Historique de la solution

2004 - Au début Google

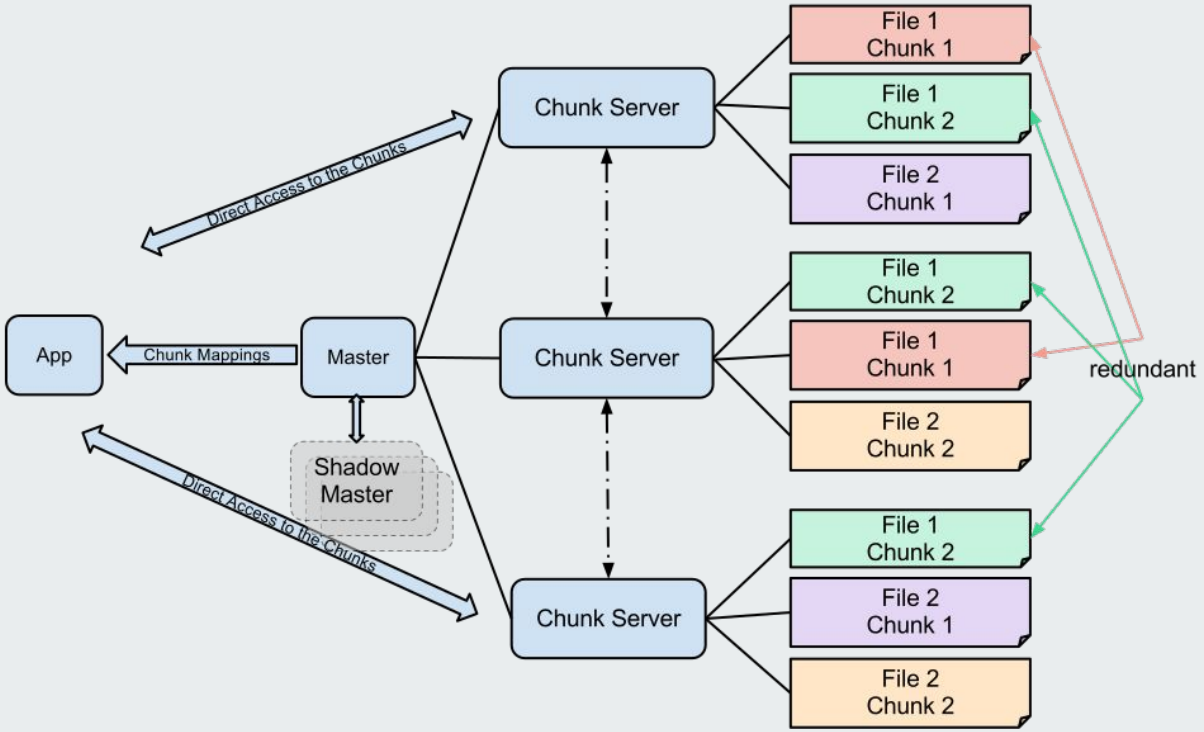
Un article de recherche est publié par Google, il présente deux technologies clefs pour le big data.
Pour commencer : MapReduce qui permet de manipuler de grandes quantités de données en les distribuant sur un cluster de machines



Historique de la solution

2004 - Au début Google

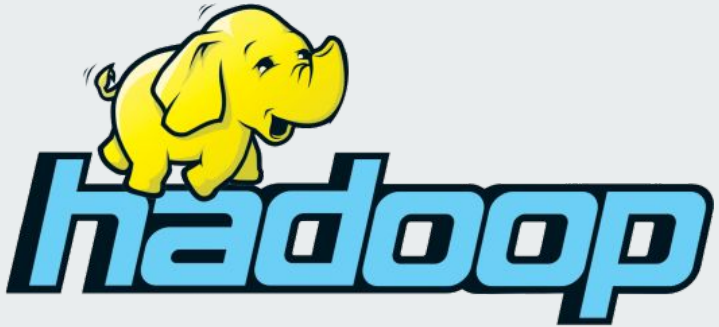
Google File System (GFS)
Un système de fichier distribué dont la théorie a été posé par Google



Historique de la solution

2006 - Doug Cutting et Hadoop

Après Lucene, Doug Cutting adopte la théorie de Google sur le MapReduce et GFS et créé sur la même idée Hadoop et HDFS



Historique de la solution

2006 - Powerset, fondation apache

Le projet HBASE a commencé sur les bases de Hadoop/HDFS par la société Powerset dans le but de fournir une fonctionnalité de recherche en langage naturelle sur de grosses quantités de data. C'est plus tôt cette même année 2006 que Google a publié un article sur BigTable.

Le projet HBASE a ensuite rejoint le giron de la fondation Apache en 2008 pour intégrer la famille des produits gravitant autours de Hadoop. Il est considéré en top projet de la fondation en 2010.

Powerset
NATURAL LANGUAGE SEARCH

THE APACHE® SOFTWARE FOUNDATION

Historique de la solution

De nos jours...

Hbase est depuis février 2017 en version stable 1.2.x , la version 2 date du 13 mars 2015.

L'essentiel des articles et des documentations disponibles en ligne vont dater de 2015, année où la base semble avoir connu son apogée.

Intérêt de la solution

Hbase est une base de données dite orientée colonnes.

Intérêt de la solution

Stockage orienté lignes

id	type	lieu	spec	intérêts
Nicolas	prof	CNAM	BDD, NoSQL	BZH, Star Wars
Régis		OC	Machine Learning, Dev	escalade, nouilles chinoises
Luc	resp formation OC	OC	formation, audiovisuel	
Céline	prof	CentraleSupelec	Ontologie, logique formelle, visualisation	

Stockage orienté colonnes

id	type
Nicolas	prof
Céline	prof
Luc	resp formation OC

id	lieu
Céline	Centrale Supelec
Nicolas	CNAM
Régis	OC
Luc	OC

id	spec
Nicolas	BDD
Nicolas	NoSQL
Régis	Machine Learning
Régis	Dev
Luc	formation
Luc	audiovisuel
Céline	Ontologie
Céline	logique formelle
Céline	visualisation

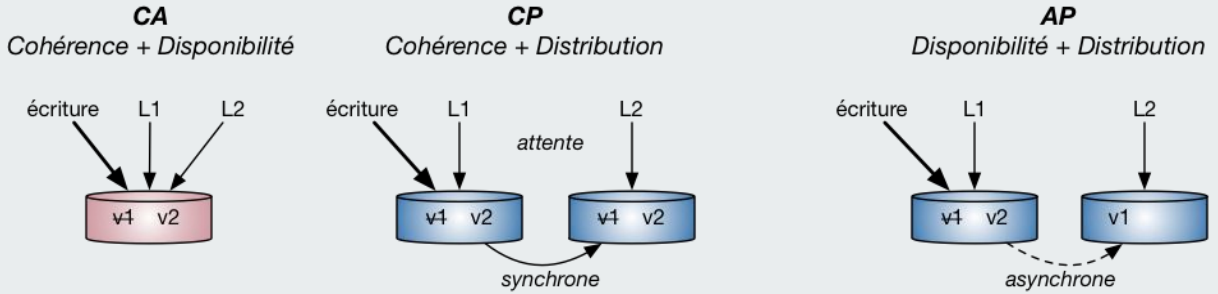
id	intérêts
Nicolas	BZH
Nicolas	Star Wars
Régis	escalade
Régis	nouilles chinoises

Intérêt de la solution

Rappel du théorème de brewer

Une base de données ne peut répondre qu'à deux des trois propriétés dites CAP :

- Consistency (Cohérence)
- Availability (Disponibilité)
- Partition tolerance (Distribution)

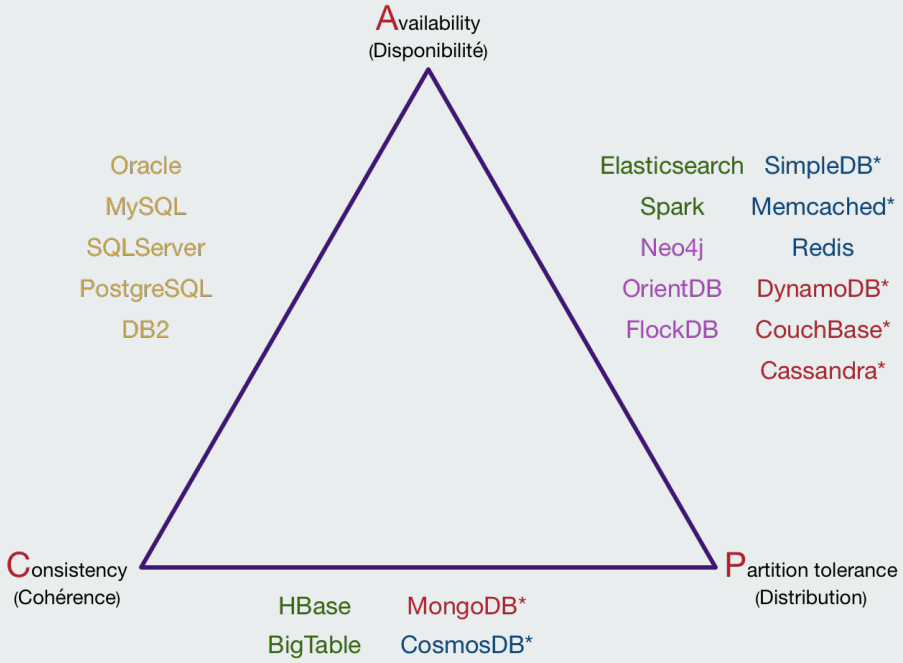


Intérêt de la solution

Modèle

Relationnel
Oriented Clé-Valeur
Oriented Colonne
Oriented Document
Oriented Graphe

* Possibilité de changer
le mode de cohérence
Cohérence <-> Disponibilité



Intérêt de la solution

Pour quels besoins ?

Hbase est un choix qui s'impose pour de la manipulation analytique de base de données de plus de 1 TéraOctet. Sans une quantité de data massive, la technologie n'a pas de réelle utilité.

Intérêt de la solution / Pour quels besoins ?

Ce que sait faire HBase

- Scalability
- Sharding
- Stockage distribué
- Consistance en lecture et écriture
- Failover automatique
- Support API : Java API et thrift
- Support du MapReduce
- Cache et filtre de bloom pour des opérations en temps réel

Intérêt de la solution

Quels sont les équivalents/alternatives ?

Bien entendu BIGTABLE

Intérêt de la solution / Quels sont les équivalents ?

Stable depuis 2016, Hypertable.
Utilisé en production par Baidu
Monté aussi en surcouche de Hadoop

Intérêt de la solution

Quand l'utiliser et quand préférer autre chose ?

Intérêt de la solution / Quand l'utiliser et quand préférer autre chose ?

HBase n'est pas une base de données relationnelle !!!

- Il ne comprend pas le SQL
- Il ne supporte pas de transactions sur plusieurs enregistrements ou de jointure

HBase sera adapté pour :

- Traiter des données non relationnelles en énorme quantité (de l'ordre du To)
- C'est souvent un bon intermédiaire de stockage de données comme on va le voir sur le cas client

Qui l'utilise ?

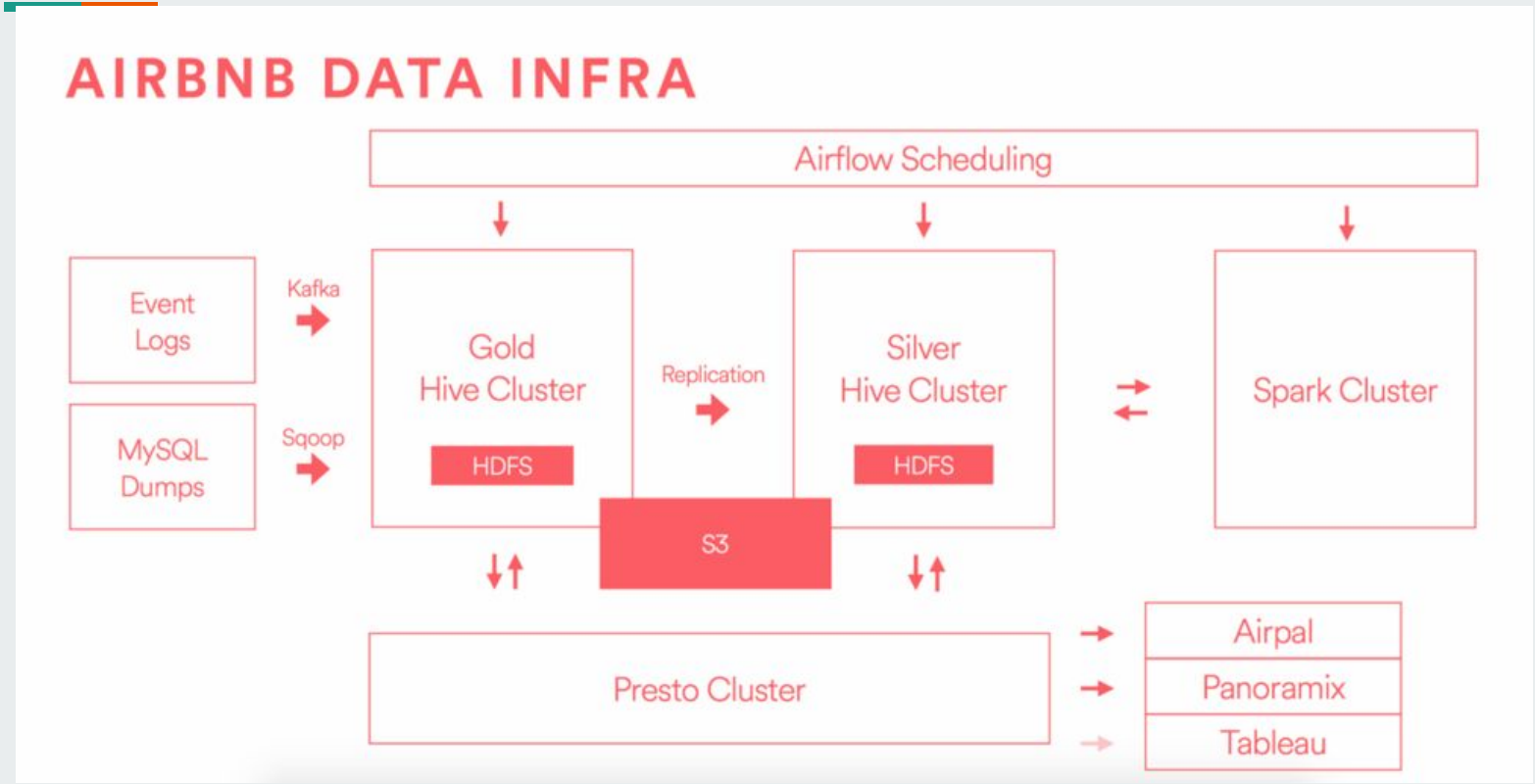
Adobe

AMADEUS

Qui l'utilise ?

Analyse d'un cas client

Analyse d'un cas client



Point sur les offres cloud

HBase dans le cloud ⇔ Google BigTable

Point sur les offres cloud

FONCTIONNALITÉ	PRIX
Nœuds	0,65 $ par nœud/heure
Stockage SSD	0,17 $ (Go/mois)
Stockage HDD	0,026 $ (Go/mois)
Entrées réseau	GRATUIT
Sorties réseau	Les tarifs des sorties Internet et interrégionales s'appliquent.

Bon appétit