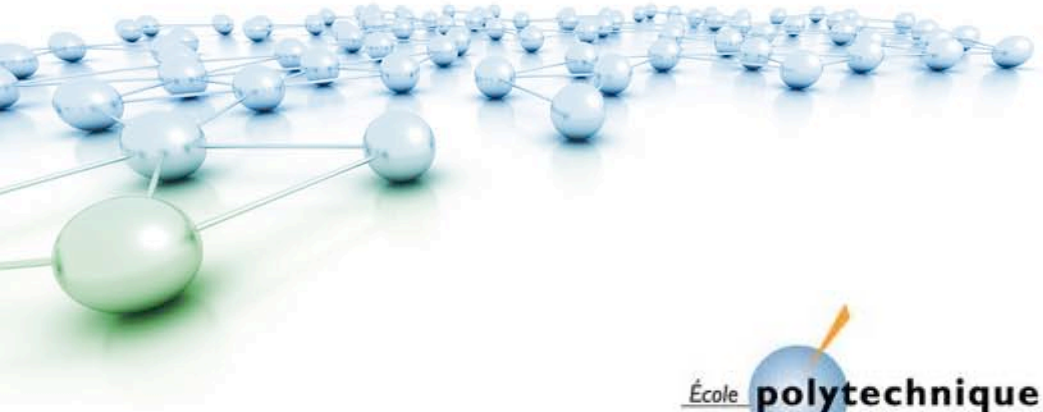


INGENSI
BIG DATA BY CYRES

École polytechnique de l'université de Tours

LI
Laboratoire d'Informatique
EA 2101

UNIVERSITÉ FRANÇOIS - RABELAIS TOURS

Tables Rondes

Le « Big Data »

INGENSI
BIG DATA BY CYRES

Plan

Introduction

1 - Présentation Ingensi
2 - Le Big Data c'est quoi ?
3 - L'histoire
4 - Le monde du libre : Hadoop
5 - Le système HDFS
6 - Les algorithmes distribués : Map Reduce
7 - Hbase : La base NoSQL

Ingensi, team du groupe Cyrès

Ingensi est une division du groupe Cyrès spécialisée dans le traitement et l'analyse temps réel de **gros volumes de données**.
Cette division est animée par Christophe Cerqueira (directeur de projets) et Guillaume Polaert (Responsable R&D).

Les 3 pôles complémentaires du Groupe CYRES :

HOSTEAM Cloud by CYRÈS — Architectures à la demande et services hébergés dans le Cloud. Leader Français de la messagerie Collaborative Exchange.

INGENSI Big data by CYRÈS — Consultants confirmés **et experts de la data** : DBA Oracle et PostgreSQL, ETL-EAI, ODI, Business Intelligence

CYRÈS interactive — Agence de communication digitale spécialisé dans les solutions mobiles et 2.0

Quelques expertises clés

- **Spécialiste** en intégration de données avec des outils EAI/ETL/ELT (ODI, Talend, etc.)
- **Infogérance** de milliers de comptes Exchange/ Sharepoint
- Mise en place d'outils alternatifs de reporting et de « **datavizualtion** » comme Tableau Software
- 3 datacenters, **disponibilité garantie 99,9%**
- Sécurisation des données, **certification PCI-DSS Level 1** Visa & Mastercard pour NORDPAY

Partenariat **DELL - Ingensi**

- **Complémentarité** : offre globale de services et de support autour de la solution Cloudera (Hadoop)
- Acteurs Majeur de la fourniture d'infrastructures de machines Hautes Performances pour les solutions Big Data

1- Ingensi, notre vision du Big Data

Lancement du projet Ingensi en septembre 2009,
Création d'une équipe R&D (Lauréat JCEF-37 en 2011)

Constat

- De plus en plus de difficultés à traiter les données et canaliser les flux entrants
- Sauvegarde et restitution de l'information de plus en plus complexes
- Coût des solutions éditeurs croissant (Oracle, Microsoft, etc.)
- Maturité des « pure-players » de l'Internet (Google, Yahoo, Facebook, etc.)

Objectifs du projet

- Structurer une offre de services et de conseils autour des Big Data et de l'écosystème Hadoop
- Mise à disposition d'une offre SaaS pour le traitement ponctuel et/ou économique de grands volumes de données
- Création d'un pôle de compétences et de formations autour Hadoop

Offre de services Ingensi

Sensibilisation des acteurs informatiques aux problématiques « Big Data » et aux solutions adaptées

Réalisation de « **Proof Of Concept** » et de maquette

Définition des architectures techniques/ fonctionnelles

Formations et **support** (Cloudera)

Administration et optimisation d'un cluster Hadoop

Solutions : Mahout (Analyse),Hive (BI),HBase (NoSQL)

La R&D, au cœur de l'offre

- Partenariat avec l'université de Tours
- Mise en place d'une thèse
- Volonté de créer un pôle « français » contributeur des solutions Hadoop
- Projet Européen pour construire un Cluster de 1600 Cores

2 - Big Data, solution au data déluge ?

Quelques chiffres

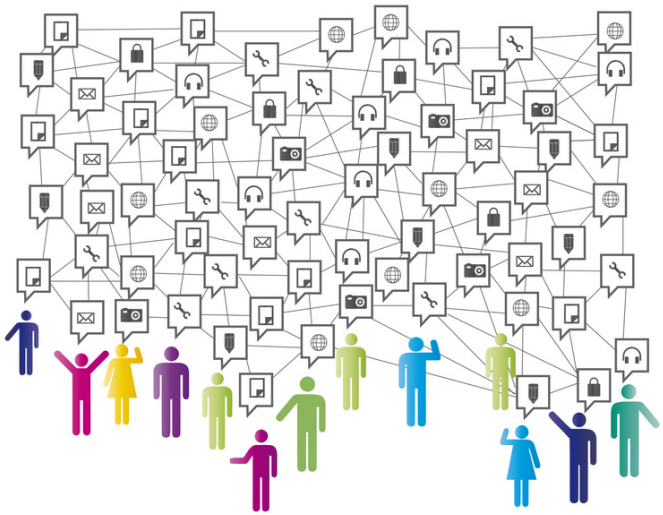
1,8 zettaoctets en 2011
soit une pile de blu-ray qui ferait 7 fois le tour de la Terre [1]

60% de croissance/an des volumes d'informations
5% pour les budgets informatiques [2]

Un Boeing produit 20 To/heure
de données [2]

250 milliards d'emails
envoyés par jour (80 % de spam) [3]

72h de vidéos déposées par minute
sur Youtube



Les systèmes actuels sont incapables de gérer de telles quantités :

80% de l'information est « non-structurée » [4]

95% de l'information est non-exploitée [4]

…Début d'une nouvelle ère

(1) IDC, 2011 - (2) Gartner, 2011 - (3) Radicati Group, 2009 - (4) Forrester, 2011

2 - Big Data, préparer et anticiper

Les solutions actuelles répondent mal (pas) aux problématiques liées, avec un TCO élevé (Exadata d'Oracle, Netezza d'IBM, etc.)

Les applications doivent **changer**

- Dimensionnées à l'échelle de la planète
- Flux de données complexes, multiples et en temps réel
- Agilité à tous les niveaux : analyse, stockage, restitution

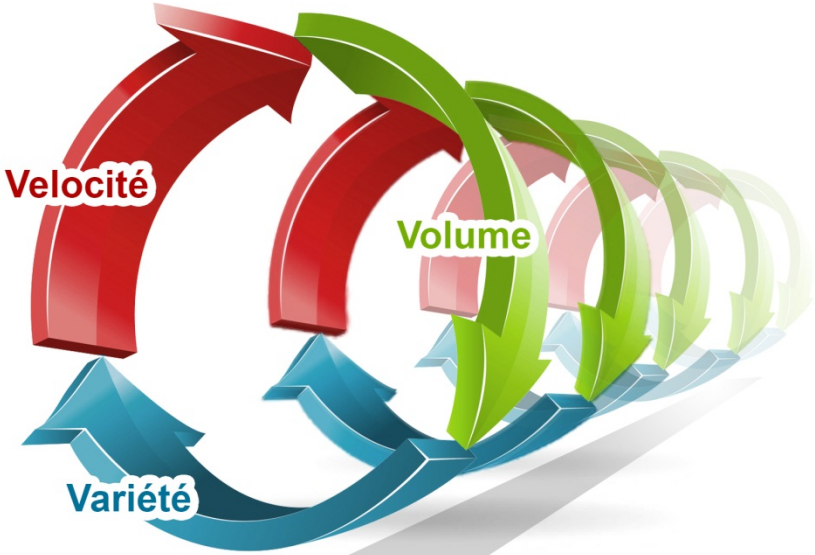
Pour

- tirer profit de ses données mais également de celles qui sont à portée de main,
- répondre à des besoins qui pour le moment n'étaient pas adressables

… et tout ça en **temps réel**

2 - Big Data, trois « V »

Les solutions mises en œuvre doivent répondre aux **3** « **V** » dans leur globalité



Volume

Saturation des systèmes actuels avec toujours plus de données

Vélocité

Quel délai pour prendre une décision à partir de l'information collectée ?

Variété

Intégrer une multitude de formats différents provenant d'une multitude de sources de données

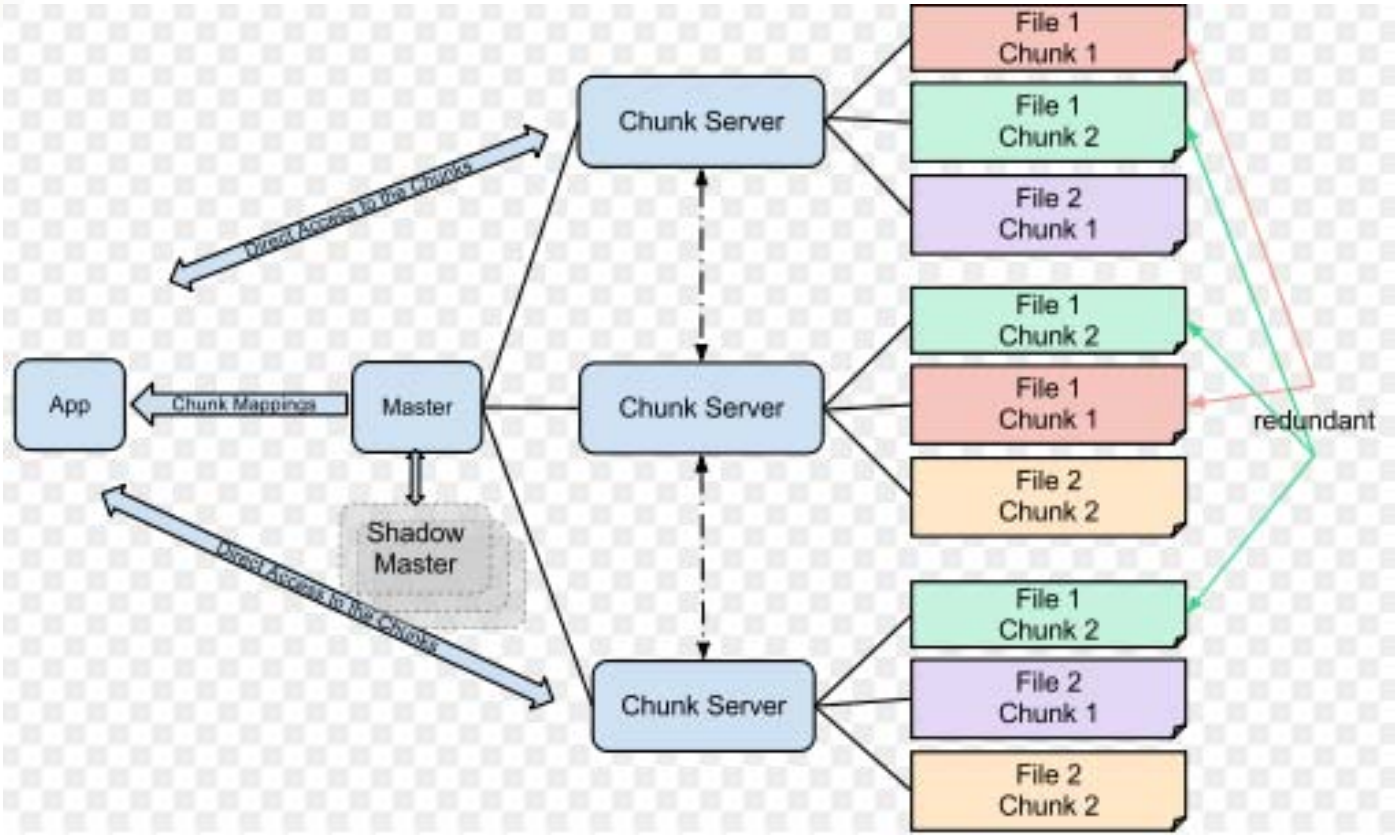
3 - L'histoire : le Big Data, Google : Le système de fichier GFS

Pour stocker son Index Grandissant Quelle solution pour Google ?

- *Utilisation d'un SGBDR ?*
  - ➔ *Problème de distribution des données*
  - ➔ *problème du nombre d'utilisateurs*
  - ➔ *problème de Vitesse du moteur de recherche*

- Invention d'un nouveau système Propriétaire : GFS (Google File Système en 2003)

3 - L'histoire : le Big Data, Google : Le système de fichier GFS



3 - L'histoire : le Big Data, Comment exploiter ce système de fichier ?

La notion de Big Data est intimement lié à la capacité de traitements de gros volumes ➔ Un nouvel Algorithme a été mis au point….

• Le premier Article a été publié en 2004 : Jeffrey Dean and Sanjay Ghemawat

MapReduce : Simplified Data Processing on Large Clusters

• C'est un algorithme inventé par Google, Inc afin de distribuer des traitements sur un ensemble de machines avec le système GFS

• Google possède aujourd'hui plus de 1 000 0000 de serveurs interconnectés dans le monde

3 - L'histoire : le Big Data, Google et les autres

YAHOO!

twitter

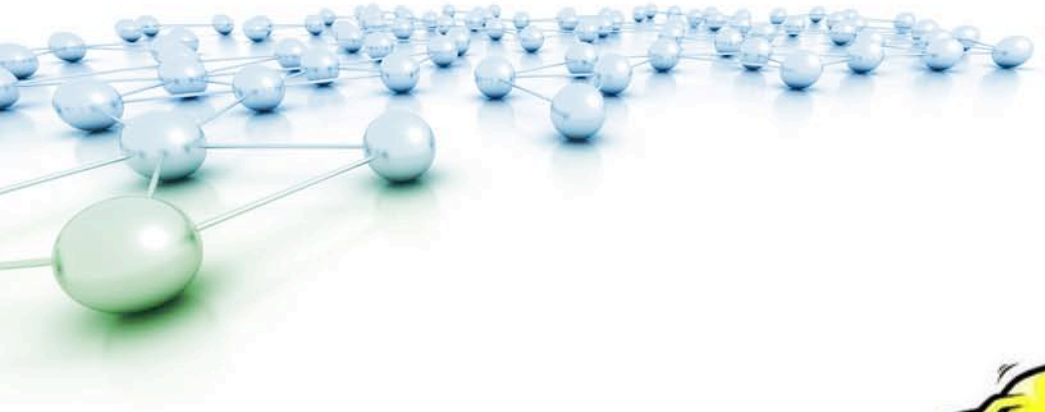
✓ **Contributeur de l'implémentation Libre (Dugg Ketting)**

Linked in

✓ **Les pures players de l'internet ont choisi d'utiliser ces algos distribués. (HDFS et MAPREDUCE)**

facebook

-***Facebook***
-***Twitter***
-***LinkedIn***
-***….***



INGENSI
BIG DATA BY CYRES

Le monde du Libre : Hadoop

Hortonworks

cloudera

MAPR
TECHNOLOGIES

4 - Big Data, solution Apache Hadoop

Hadoop

- **Projet initié** par les pure-players de l'Internet (Yahoo, Facebook, Twitter) **dès 2008**
- Inspiré des travaux de Google
- Libre, **fondation Apache**
- **Enrichit chaque jour** par de nombreuses sociétés dédiées : Cloudera, Hortonworks, etc.

2 concepts clés

- **HDFS** : Stockage **économique et extensible**, pour de grandes quantités d'information bénéficiant d'**une haute tolérance aux pannes**
- **MapReduce** : Algorithme de traitements parallèles et distribués des données.

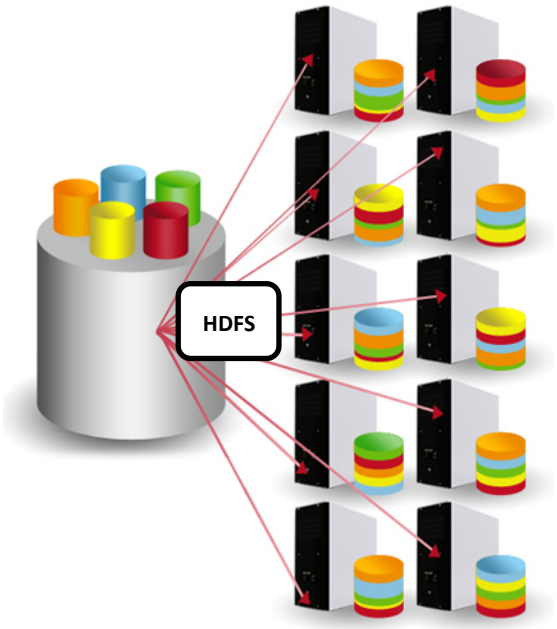
Écosystème riche

- **Mahout** : machine-learning (dataming, clustering)
- **HBase** : base de données temps-réel NoSQL
- **Hive** : traitement batch analytique BI
- **3 éditeurs** (Cloudera, Hortonworks, MapR)

5 - HDFS, système de fichiers distribués

Objectifs et principes

- Stocker de grandes quantités d'information à moindre coût **» utilisation de serveurs courants**
- Idéalement des fichiers volumineux
- Haute tolérance aux pannes **» donnée répliquée 3 fois sur 3 serveurs géographiquement distants**
- Stockage extensible à volonté **» ajout à chaud de serveurs** pour augmenter les **capacités de stockage et de traitement** de l'architecture



Techniquement

- **namenode** : serveur maître. Cartographie des blocs de données sur le cluster. Vital pour la plateforme
- **datanode** : stocker localement les blocs de données. Informe le namenode de son état « via un battement de cœur » toutes les secondes. Possibilité de définir sa position géographique (site de données, rack) pour que le namenode contrôle au mieux les différents emplacements des blocs de données

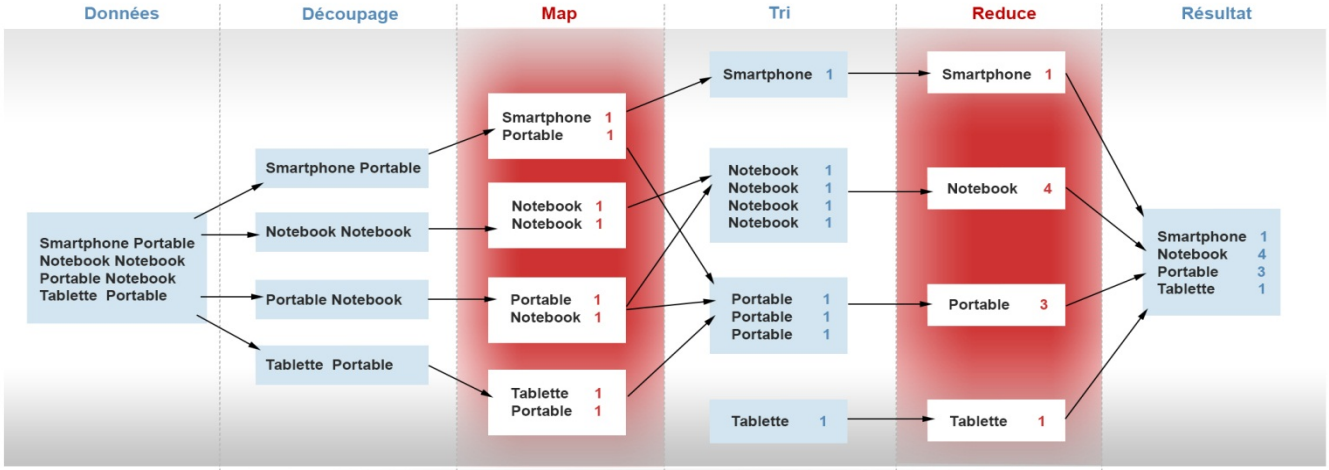
6 - MapReduce, algorithme de traitement des données

Principe

- Traitement divisé en tâche lesquelles sont traitées en parallèle : **MAP**
- Synthèse et agrégation des traitements : **REDUCE**
- **Les traitements sont effectués là où la donnée est stockée** (sur chaque serveur)

Techniquement

- Service« **jobtracker** » coordonne l'exécution du traitement. Découpe et affecte chaque tâche aux tasktracker.
- Service « **tasktracker** » responsable de l'exécution de la tâche localement.
- Si une tâche échoue, le **jobtracker** la relance sur un autre serveur.



INGENSI
BIG DATA BY CYRES

7- HBase, base de données NoSQL temps réel

Pourquoi ?

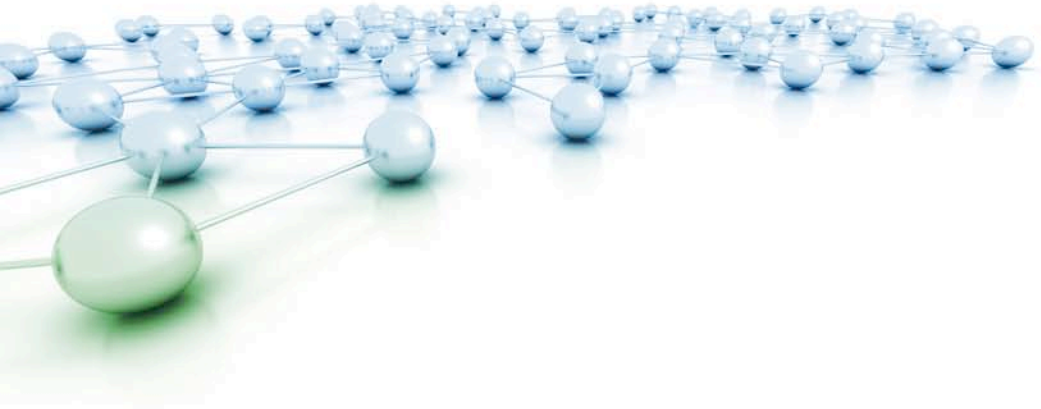
- Nécessité d'un mode « temps réel » pour le système Hadoop.
- Le mode batch ne convient pas à toutes les applications

Concepts clés

- Base de données NoSQL en « mode colonne »
- Gestion des transactions simple
- Déploiement à grande échelle sur un très grand nombre de serveurs
- Partitionnement automatique des tables par l'ajout de serveur (region server)
- Modélisation des données orientée « recherche »

Usages

- Stockage et recherche de n'importe quel type de données (PDF, photos, document word, etc.).
- Données accessibles via de nombreuses API
- Stockage dénormalisé des données
- Insertion et recherche en temps réel via une série de méthode
- Recherche très rapide (concept « in memory »)



INGENSI
BIG DATA BY CYRES

École polytechnique de l'université de Tours

INGENSI

Groupe Cyrès
19 - 21 rue Édouard Vaillant
37000 Tours
Tél : 02 47 68 48 50
contact@ingensi.com

LI
Laboratoire d'informatique
EA 6300

DELL

CYRÈS conseil