
Data Mining Big Data & ML

Enseignant: W. Moise CONVOLBO, PhD

Programme

Jour	Cours	TD
19 - 01 - 2022	Intro: Big Data	

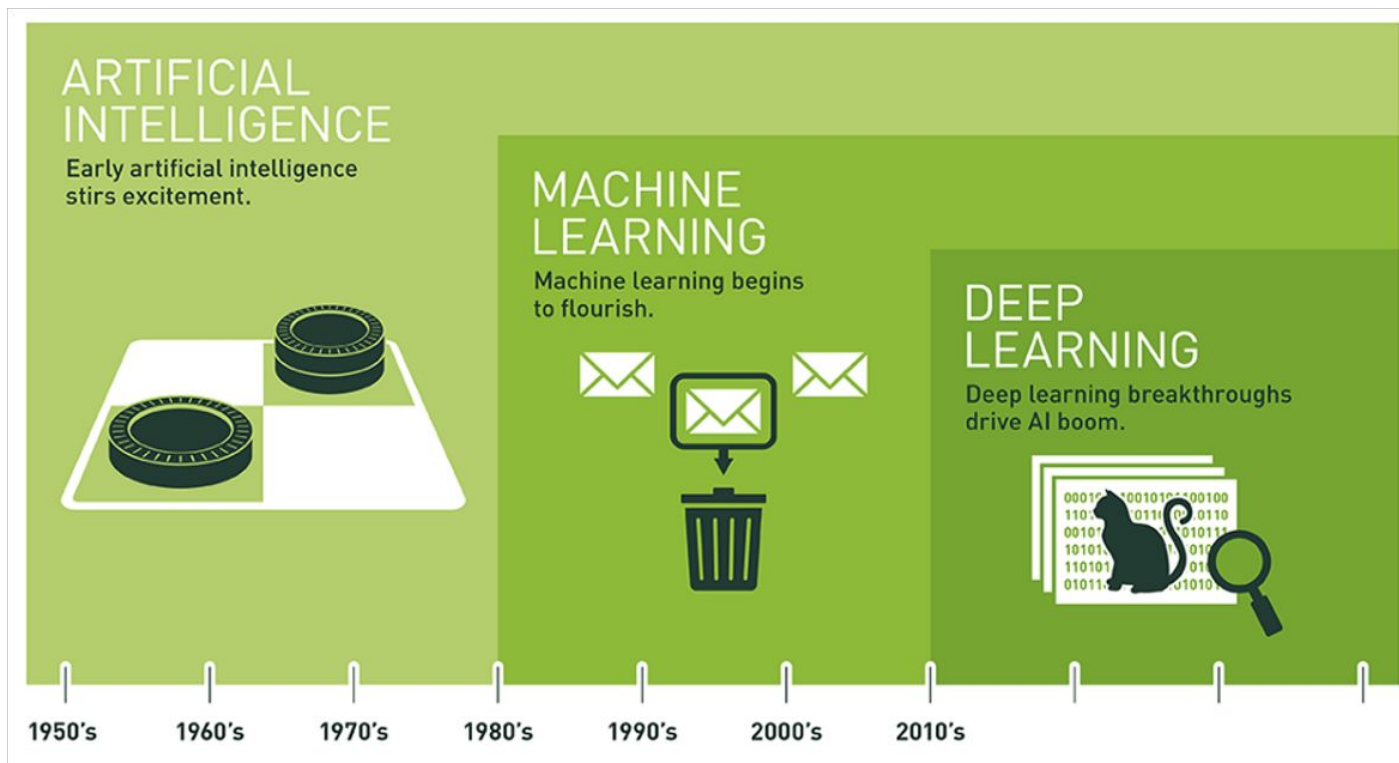


Wendkuuni Moise CONVOLBO, PhD

Inventor, Expert on Cloud,
Geo-Distributed Datacenters & AI

@Convolbo_w, www.convolbo.org, linkedin.com/in/convolbo

Big Data - Cloud - Intelligence Artificielle



Big Data

Les temps ont changé

Homme sur la lune avec 32 Ko (1969); mon ordinateur portable a 32 Go de RAM (2021)

En 2007, Google collectait 270 PB de données en un mois

En 2008, Google collectait 20 000 PB par jour

Début 2020, le nombre d'octets dans l'univers numérique était 40 fois plus grand que le nombre d'étoiles dans l'univers observable.

Avalanche des données



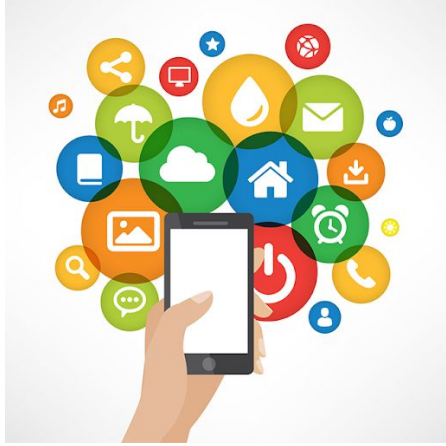
La Bourse de New York génère environ 4 à 5 téra de données par jour.



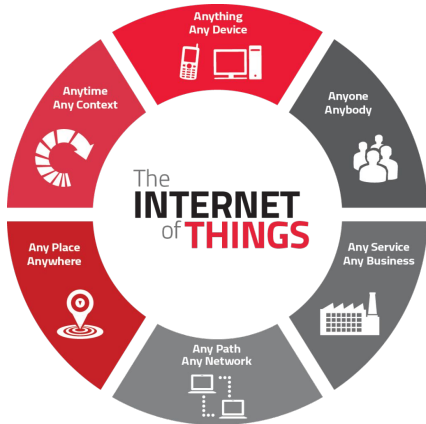
Facebook héberge plus de 240 milliards de photos, avec une croissance de 7 pétaoctets par mois.

Ancestry.com, le site de généalogie, stocke environ 10 pétaoctets de données.

L'Internet Archive stocke environ 18,5 pétaoctets de données.

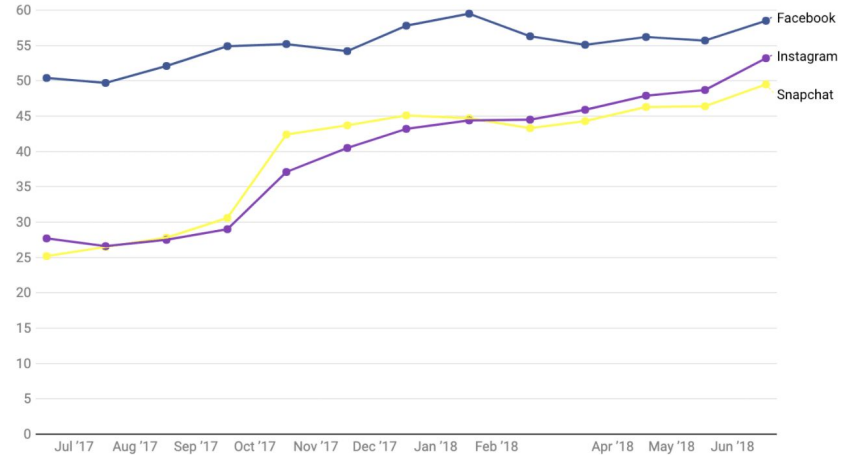


Plus de temps en ligne et sur mobile



Minutes spent per day on social apps

By U.S. Android users



Challenge: Stockage & Utilisation

Vitesse d'accès

- **Le problème est simple** : bien que les capacités de stockage des disques durs aient considérablement augmenté au fil des ans, les vitesses d'accès (le taux auquel les données peuvent être lues à partir des disques) n'ont pas suivi.
- Les disques de 1 TB sont la norme, mais la vitesse de transfert est d'environ 100 Mo/s, il faut donc plus de 2.5 heures pour lire toutes les données du disque.
- C'est long pour lire toutes les données sur un seul lecteur, et l'écriture est encore plus lente

Vitesse de transfert

Vitesse de R/W

Challenge: Stockage & Utilisation

- Pouvoir lire et écrire des données en parallèle???
 - Défaillance matérielle : Plusieurs composants matériels = risque élevé de pannes.
 - Disparité des données dans la localité
-

Requête sur toutes vos données

- Large sites de E-Commerce
 - Réseaux sociaux
 - Moteurs de recherche
 - Industrie Pharmaceutique
 - Le Service de Renseignement
 -
-

Mais avant, un exemple ...

Big Data en Telemedicine

Plan



Bref historique de l'IA et
du ML dans la santé



Pourquoi maintenant?



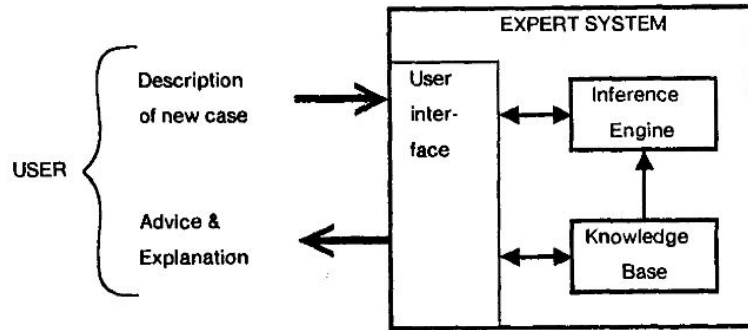
En quoi le ML dans la
santé est unique?

Problèmes

- Le coût des dépenses en soins de santé aux États-Unis est supérieur à 1500 milliards de dollars et en hausse
- Des meilleurs cliniciens au monde, mais les maladies chroniques sont:
 - Souvent diagnostiquées tardivement
 - Souvent mal gérées
- Les erreurs médicales sont omniprésentes

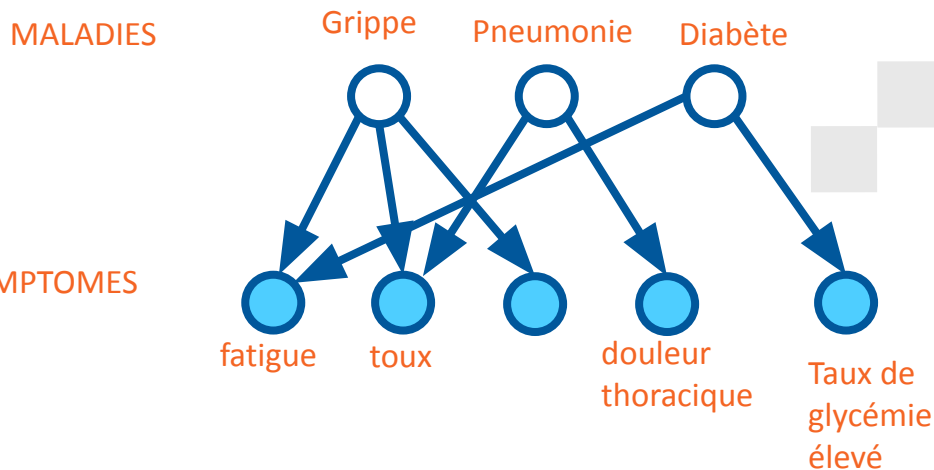
Années 1970: MYCIN Système Experts

- 70 (Stanford): Système Expert de MYCIN pour l'identification des bactéries causant plusieurs infections
- Proposer une thérapie ~69% des cas mieux que les experts en maladies infectieuses

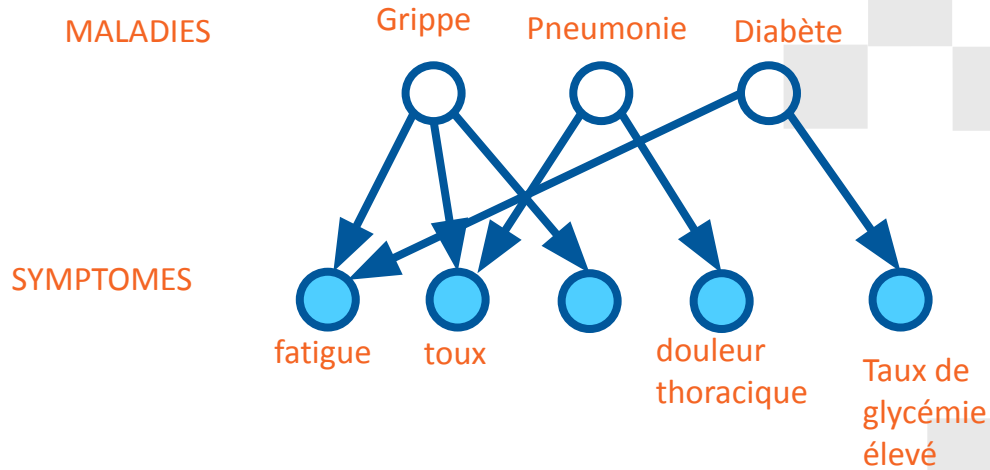


Années 1980: Modèle INTERNIST-1/QMR

- 80 (Pittsburgh): INTERNIST-1/ Quick Medical Reference
- Diagnostic pour la médecine interne



Années 1980: Modèle INTERNIST-1/QMR



Problèmes:

- ❑ Saisie manuelle
- ❑ difficile à entretenir
- ❑ S'étend exponentiellement



- ❑ 570 variables binaires de la maladie
- ❑ 4075 variables binaires de symptômes
- ❑ 45 470 liens entre les maladies et les symptômes

Années 1980: automatisation de la découverte

RX PROJECT: AUTOMATED KNOWLEDGE ACQUISITION

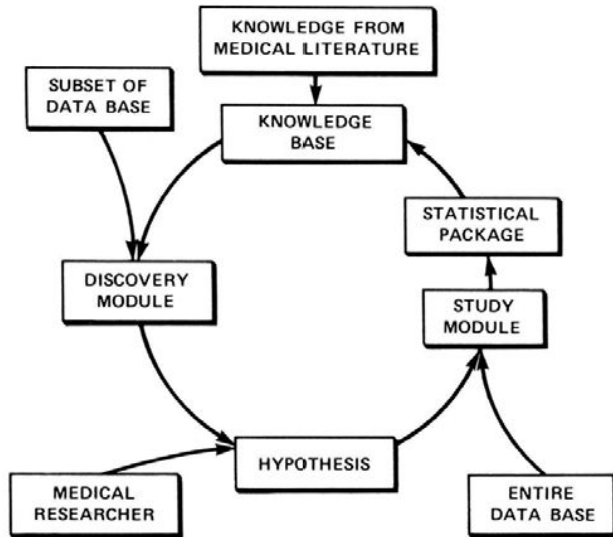
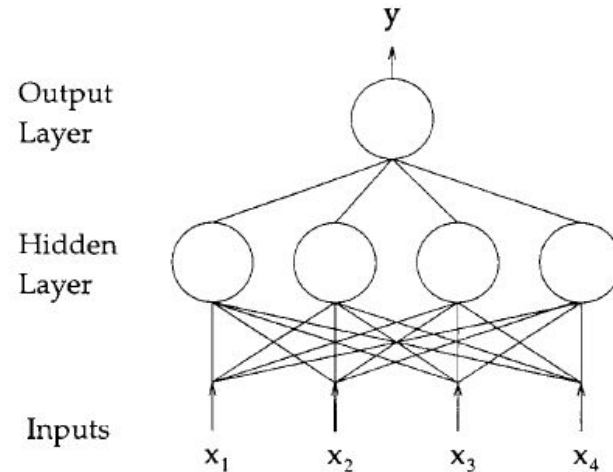


FIG. 2. Discovery and confirmation in RX.

Années 1990: Réseaux de neurones en médecine

- Les réseaux de neurones avec des données cliniques ont décollé en 1990, avec 88 nouvelles études cette année-là
- Petit nombre de fonctionnalités
- Données souvent collectées et présentées | tableaux

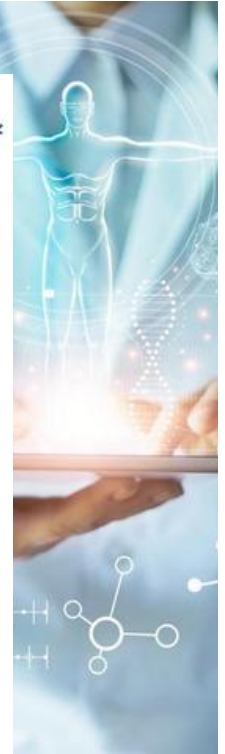
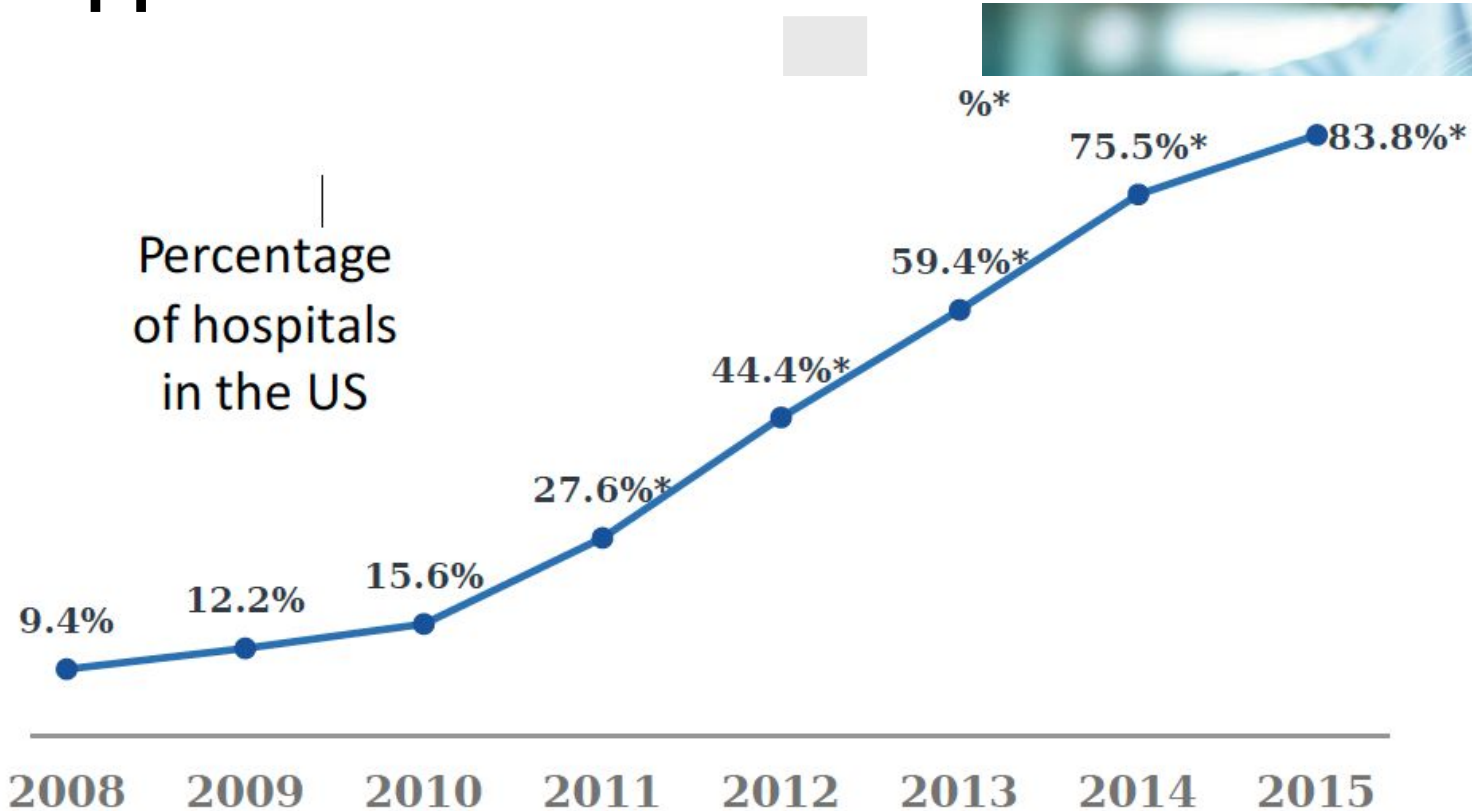
- Ne s'intègre pas bien dans le flux de travail clinique
- Difficile d'obtenir suffisamment de données d'apprentissage
- Mauvaise généralisation à de nouveaux endroits





**Pourquoi
maintenant?**

Opportunités



Courtesy of Health and Human Services. Image is in the public domain.

Welcome to HealthData.gov

This site is dedicated to making high value health data more accessible to entrepreneurs, researchers, and policy makers in the hopes of better health outcomes for all.



Jeux de données


Order by:

Datasets ordered by Popular

Filter by location
[Clear](#)

218,110 datasets found

Allegheny County Property Sale Transactions  **118 recent views**

Allegheny County / City of Pittsburgh / Western PA Regional Data Center — This dataset contains data on all Real Property parcels that have sold since 2013 in Allegheny County, PA. Before doing any market



Jeux de données

Registration

New User
Change Password
User's Agreement

Project

FAQ
Overview
History
HMD Events

People

Acknowledgements
Research Teams
HMD Publications

Methods

Brief Summary
Full Protocol
Special Methods

Data

What's New
Explanatory Notes
Data Availability
Zipped Data Files
Citation Guidelines

Links

Max Planck Institute
UC Berkeley
UC Berkeley Demography
INED
Human Life Table

The Human Mortality Database

Vladimir Shkolnikov, *Director*

Max Planck Institute for Demographic Research

Magali Barbieri, *Associate Director*

University of California, Berkeley and INED, Paris

John Wilmoth, *Founding Director*

United Nations and formerly University of California, Berkeley

In response to the COVID-19 pandemic, the HMD team decided to establish a new data resource: **Short-term Mortality Fluctuations (STMF) data series**. Objective and internationally comparable data are crucial to determine the effectiveness of different strategies used to address epidemics. Weekly death counts provide the most objective and comparable way of assessing the scale of short-term mortality elevations across countries and time. [Here](#) we provide weekly death counts for 37 countries: Austria, Australia (Doctor certified deaths), Belgium, Bulgaria, Chile, Canada, Croatia, Czech Republic, Denmark, England and Wales, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Israel, Italy, Latvia, Lithuania, Luxembourg, Netherlands, New Zealand, Northern Ireland, Norway, Poland, Portugal, Republic of Korea, Russia, Scotland, Slovenia, Slovakia, Spain, Sweden, Switzerland and the USA. The same data in the pooled CSV file are available for download [here](#). Data formats and methods are described in the [STMFNote](#). We also strongly recommend reading the [metadata text](#). Following the HMD practice, we also publish [original input data in standardized format](#). During the next few weeks data will be frequently updated and new countries will be added. The most recent STMF update is: 2020-11-19.

New: We invite you to explore this data with our online [STMF visualization toolkit](#).

Jeux de données



CORONAVIRUS DISEASE 2019 (COVID-19)

Explore the latest COVID-19 data and sign up to receive updates from NCHS



Navigation arrows (left and right) and a progress indicator (a solid black bar followed by four empty white bars) are located at the bottom of the banner.

Jeux de données



Open-source APIs

[LEARN MORE →](#)

Jeux de données

DATA PLATFORM

INDICATOR CATEGORIES >

CITIES >

POLICIES & PRACTICES >

ABOUT

BACKGROUND

METHODOLOGY

FAQS

DOWNLOAD

OPIOID-RELATED UNINTENTIONAL DRUG OVERDOSE MORTALITY RATE

[Local Policies & Practices](#) [Indicator CSV](#)

Cities

- Boston, MA
- Columbus, OH
- Denver, CO
- Fort Worth, TX
- Houston, TX
- Kansas City, MO

Year

2010

Sex

Both

Race

All

Display Results

10 results

Jeux de données

COVID-19 is an emerging, rapidly evolving situation.



Expand for resources



Español

1-800-4-CANCER

Live Chat

Publications

Dictionary

ABOUT CANCER CANCER TYPES RESEARCH GRANTS & TRAINING NEWS & EVENTS ABOUT NCI

search



NCI is the nation's
leader in cancer
research

Jeux de données




MIMIC

Documents 

Data 

Community 

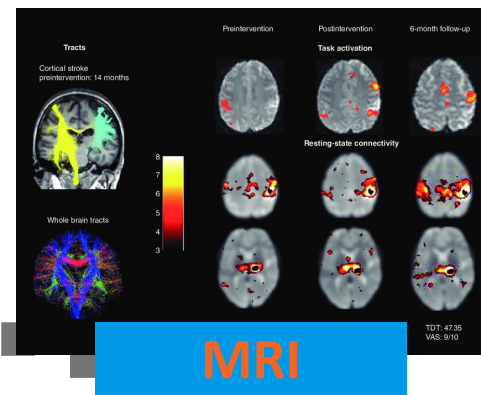
Code (GitHub) 

Jeux de données



**Pourquoi
maintenant?**

Diversité des données de la Santé



Standardisations



ICD-9 codes 290–319: mental disorders

ICD-9 codes 320–359: diseases of the nervous system

ICD-9 codes 360–389: diseases of the sense organs

ICD-9 codes 390–459: diseases of the circulatory system

ICD-9 codes 460–519: diseases of the respiratory system

ICD-9 codes 520–579: diseases of the digestive system

ICD-9 codes 580–629: diseases of the genitourinary system

ICD-9 codes 630–679: complications of pregnancy, childbirth,



Standardisations

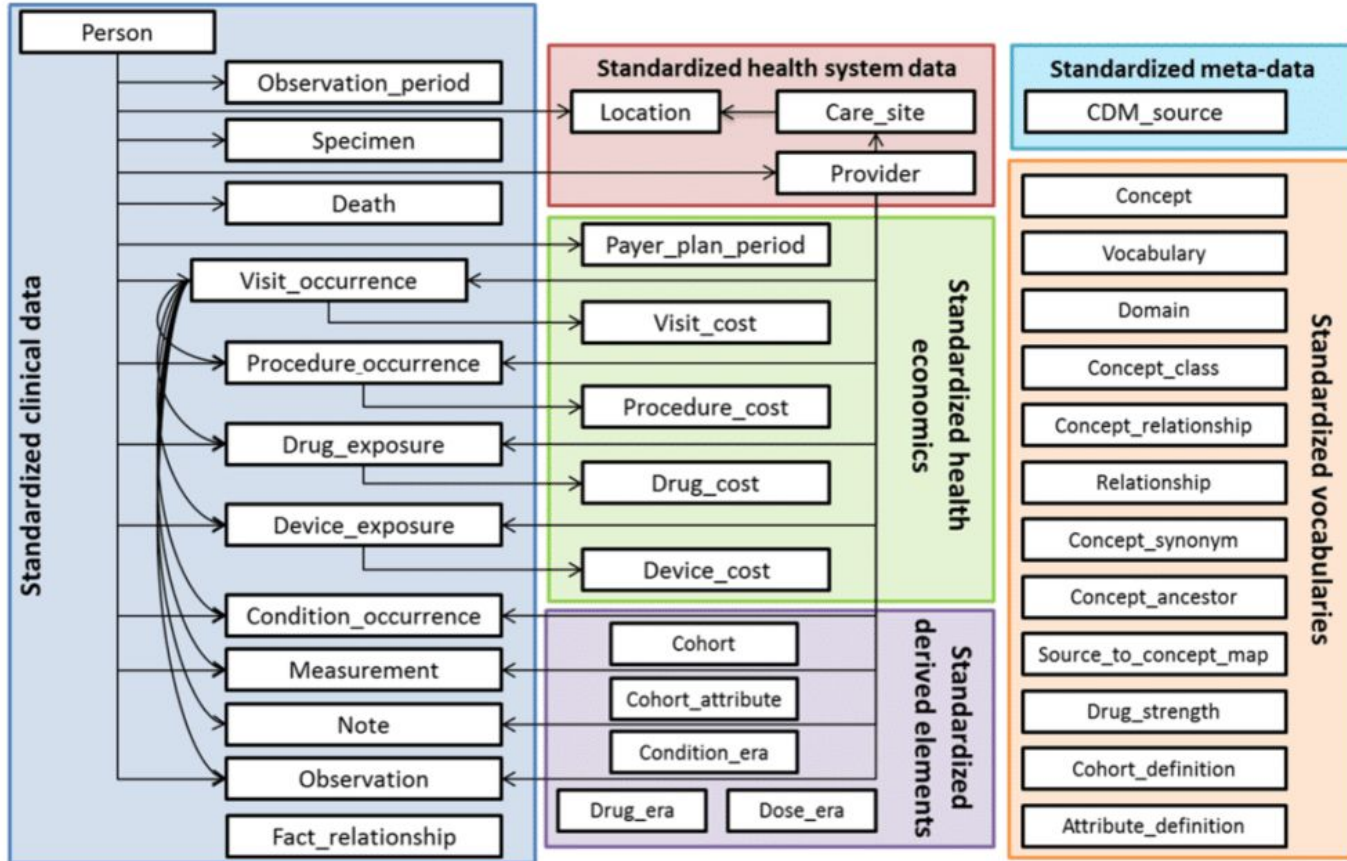
Tests de laboratoire: LOINC

Pharmacie: Code national des médicaments (NDCs)

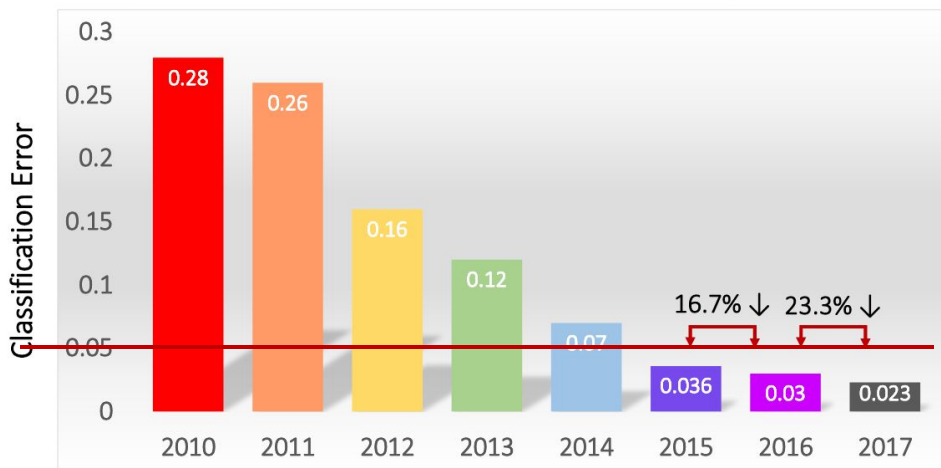
Système de langue médicale unifié (UMLS): des millions de concepts médicaux



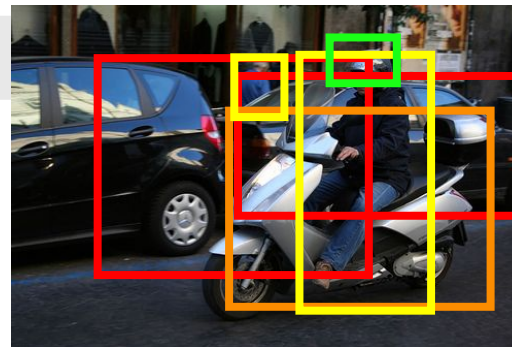
Standardisations



Avancée dans l'apprentissage automatique



Precision humaine(5%)



Personne
voiture
Moto
Casque

Pourquoi?

Big Data

Avancée dans les

Algorithmes

Communautés Open-Source

Avancée dans l'apprentissage automatique



PYTORCH

theano

gensim

Caffe



SM

将軍
sho gun

- **Des avancées majeures en ML et en IA**
 - Apprentissage avec des caractéristiques de grande dimension (par exemple, régularisation l1)
 - Apprentissage semi-supervisé et non supervisé
 - Techniques modernes d'apprentissage en profondeur (par exemple, convnets, variantes de SGD)
- Démocratisation de l'apprentissage automatique
 - Logiciels open source de haute qualité, tels que scikit-learn de Python, TensorFlow, Torch, Theano



**qu'est-ce qui
différencie la santé?**

Avancée dans l'apprentissage automatique

- **Décisions de vie ou de mort**
 - Besoin d'algorithmes robustes
 - Contrôles et équilibres intégrés au déploiement ML
 - (Se pose également dans d'autres applications de l'IA comme la conduite autonome)
 - Besoin d'algorithmes justes et responsables
- De nombreuses questions portent sur l'apprentissage non supervisé
 - Découvrir des sous-types de maladies ou répondre à des questions telles que «caractériser les types de personnes qui sont très susceptibles d'être réadmis à l'hôpital»?
- **Bon nombre des questions auxquelles nous voulons répondre sont causales**
 - L'utilisation naïve de l'apprentissage automatique supervisé est insuffisante

Avancée dans l'apprentissage automatique

- **Très peu de données étiquetées**
 - Motive les algorithmes d'apprentissage semi-supervisé
- **Parfois un petit nombre d'échantillons (par exemple, une maladie rare)**
 - Apprenez autant que possible d'autres données (par exemple, des patients en bonne santé)
 - Modélisez le problème avec soin
- **Beaucoup de données manquantes, intervalles de temps variables, étiquettes censurées**



Avancée dans l'apprentissage automatique

- **Difficulté à désidentifier les données (*anonymiser*)**
 - Besoin d'accords de partage de données et sensibilité
- **Difficulté à déployer le ML**
 - Le logiciel commercial de dossier de santé électronique est difficile à modifier
 - Les données sont souvent en silos; tout le monde reconnaît le besoin d'interopérabilité, mais la lenteur des progrès
 - Des tests et une itération minutieux sont nécessaires



Q & A?



Data Mining

Big Data & ML: Hadoop

Enseignant: W. Moise CONVOLBO, PhD

Programme

Jour	Cours	TD
19 - 01 - 2022	Intro: Big Data	

Plan

1. Decouvrir Hadoop
 2. MapReduce
 3. Hadoop Distributed File System
HDFS
 4. YARN
 5. Hadoop I/O
 6. Comment Hadoop Execute MR
-

1-Découvrir Hadoop

Les temps ont changé

Homme sur la lune avec 32 Ko (1969); mon ordinateur portable a 32 Go de RAM (2021)

En 2007, Google collectait 270 PB de données en un mois En 2008, Google collectait 20 000 PB par jour

Début 2020, le nombre d'octets dans l'univers numérique était 40 fois plus grand que le nombre d'étoiles dans l'univers observable.

Avalanche des données



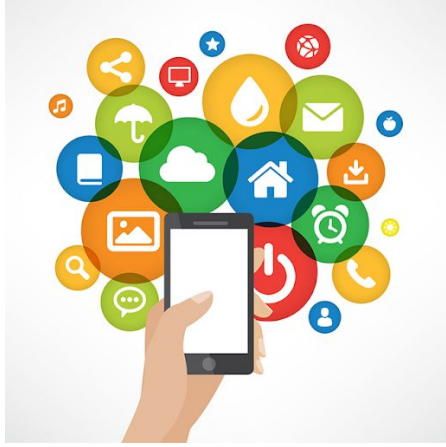
La Bourse de New York génère environ 4 à 5 téra de données par jour.



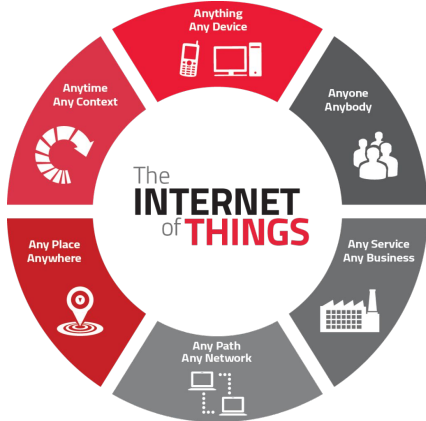
Facebook héberge plus de 240 milliards de photos, avec une croissance de 7 pétaoctets par mois.

Ancestry.com, le site de généalogie, stocke environ 10 pétaoctets de données.

L'Internet Archive stocke environ 18,5 pétaoctets de données.

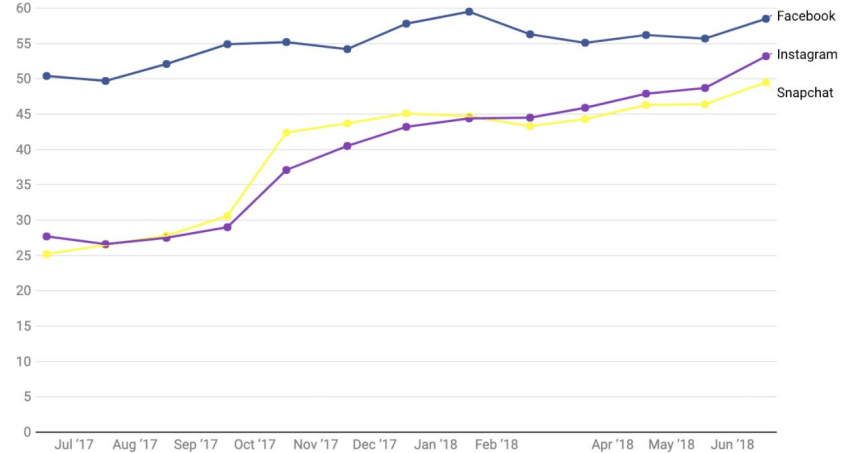


Plus de temps en ligne et sur mobile



Minutes spent per day on social apps

By U.S. Android users



Challenge: Stockage & Utilisation

Vitesse d'accès

- **Le problème est simple** : bien que les capacités de stockage des disques durs aient considérablement augmenté au fil des ans, les vitesses d'accès (le taux auquel les données peuvent être lues à partir des disques) n'ont pas suivi.
- Les disques de 1 TB sont la norme, mais la vitesse de transfert est d'environ 100 Mo/s, il faut donc plus de 2.5 heures pour lire toutes les données du disque.
- C'est long pour lire toutes les données sur un seul lecteur, et l'écriture est encore plus lente

Vitesse de transfert

Vitesse de R/W

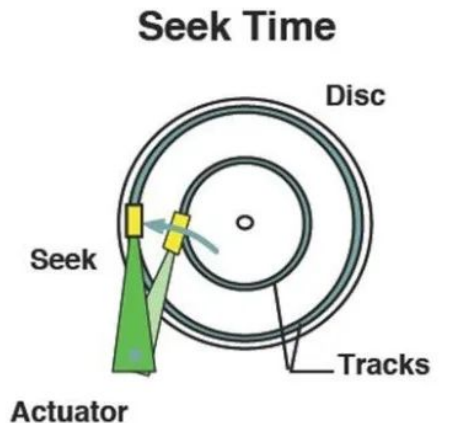
Challenge: Stockage & Utilisation

- Pouvoir lire et écrire des données en parallèle???
 - Défaillance matérielle : Plusieurs composants matériels = risque élevé de pannes.
 - Disparité des données dans la localité
-

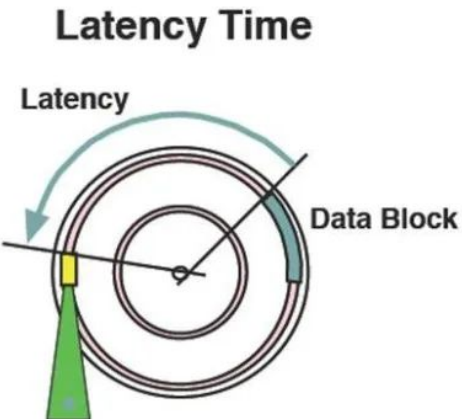
Requête sur toutes vos données

- Large sites de E-Commerce
 - Réseaux sociaux
 - Moteurs de recherche
 - Industrie Pharmaceutique
 - Le Service de Renseignement
 -
-

Pourquoi pas les RDBMS



- Pourquoi ne pouvons-nous pas utiliser des bases de données avec beaucoup de disques pour effectuer des analyses à grande échelle ? Pourquoi Hadoop est-il nécessaire ?
- La réponse à ces questions vient d'une autre tendance des lecteurs de disque : le temps de recherche s'améliore plus lentement que le taux de transfert.



Comparaison

	Traditional RDBMS	Autre possibilité
Data size	Gigabytes	Petabytes
Access	Interactive and batch	Batch
Updates	Read and write many times	Write once, read many times
Transactions	ACID	None
Structure	Schema-on-write	Schema-on-read
Integrity	High	Low
Scaling	Nonlinear	Linear

Question:

Que
represente
ACID en
base de
données?

Grid Computing

- Le HPC et Grid computing faisaient du traitement de données à grande échelle depuis des années
 - API
 - Passage de message (MPI)

Coordination

Gestion des
Ressources

Tolérance à la
panne

Histoire de Hadoop



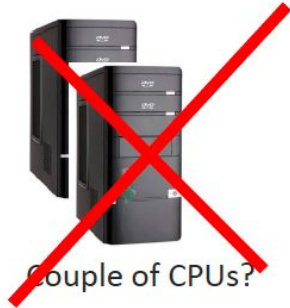
Doug Cutting



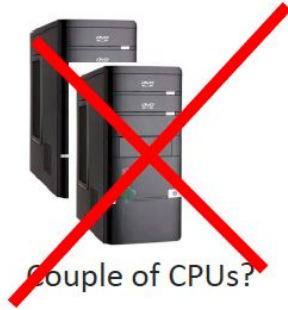
Histoire de Hadoop

- 2003 Google File System ([GFS](#));
 - 2004 MapReduce Paper ([lien](#));
 - Février 2006 Lucene → Hadoop;
 - Février 2008 Cluster Hadoop de 10 000 cœurs;
 - En avril 2008, Hadoop a battu un record du monde en trie (1T de données en entier);
-

De quelle échelle parlons-nous ?



De quelle échelle parlons-nous ?



De quelle échelle parlons-nous ?



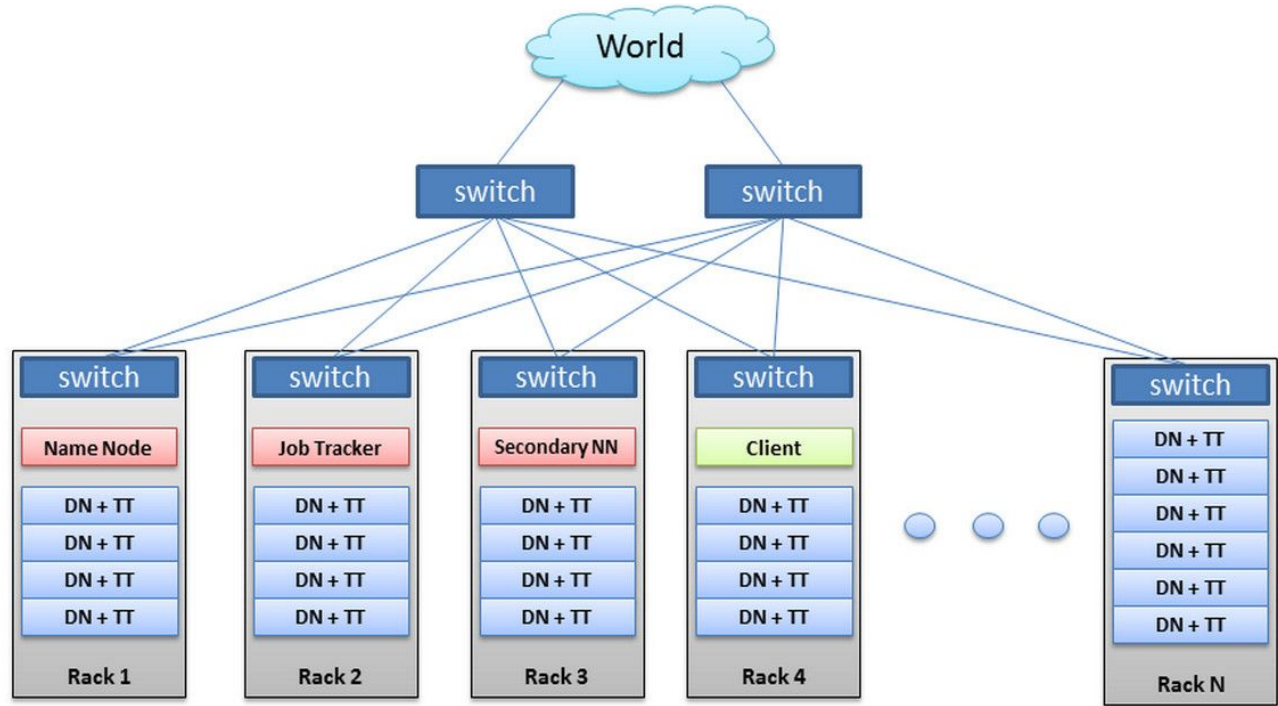
De quelle échelle parlons-nous ?



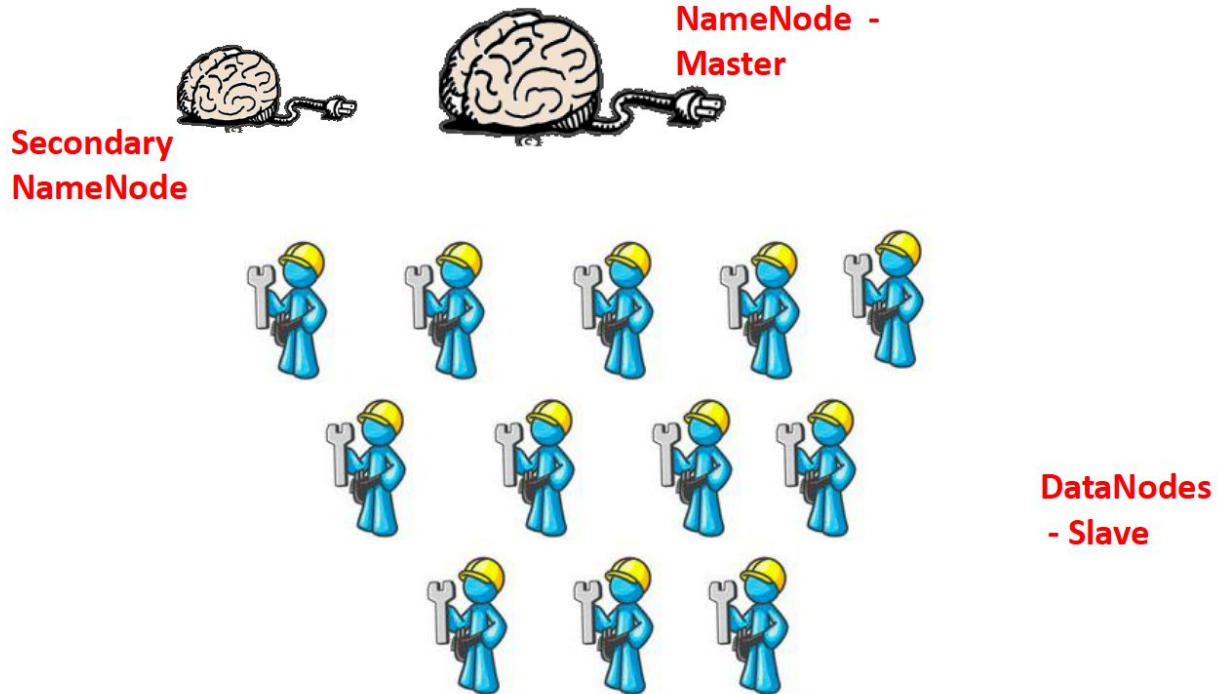
Problematique

- **Panne de matériel**
 - Un ordinateur = tombe en panne une fois tous les 1000 jours
 - 1000 ordinateurs = 1 par jour
 - **Pétaoctets de données à traiter en parallèle**
 - 1 disque dur = 100 Mo/sec
 - 1000 HDD = 100 Go/sec
 - **Évolutivité facile**
 - Augmentation/diminution relative des performances en fonction de l'augmentation/diminution des nœuds
-

Hadoop cluster



Comment ca marche?



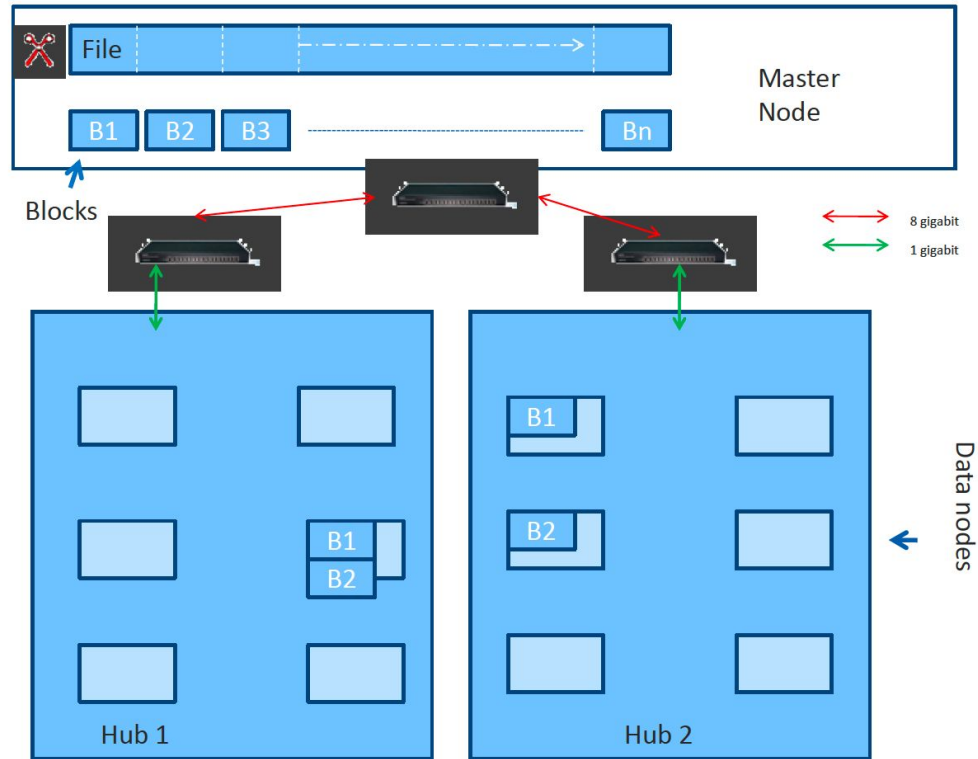
Stockage de fichier sur HDFS

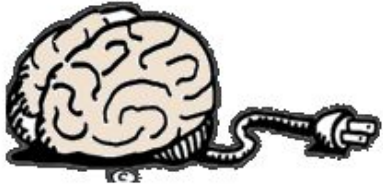
Motivation

- Fiabilité,
- Disponibilité,
- Bande passante du réseau
- ❑ Le fichier d'entrée (disons 1 To) est divisé en petits morceaux / blocs de 64 Mo (ou multiples de 64 Mo)
- ❑ Les morceaux sont stockés sur plusieurs nœuds en tant que fichiers indépendants sur les nœuds esclaves
- ❑ Pour s'assurer que les données ne sont pas perdues, les répliques sont stockées de la manière suivante :
 - Un sur **le nœud local**
 - Un sur **le rack distant** (en cas de panne du rack local)
 - Un sur **le rack local** (en cas de défaillance du nœud local)
 - Autre placé au **hasard**
 - Le facteur de réplication par défaut est 3



Data node





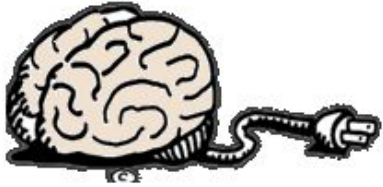
NameNode-master

The master node: NameNode

Les fonctions:

- Gère les fichiers de mappage du système de fichiers vers les blocs et les blocs vers les nœuds de données
- Maintien de l'état des nœuds de données
 - Battement de coeur
 - Datanode envoie des battements cardiaques (heartbeat) à intervalles réguliers
 - Si le heartbeat n'est pas reçu, le datanode est déclaré mort
 - Bloquer le rapport
 - DataNode envoie la liste des blocs dessus
 - Utilisé pour vérifier la santé de HDFS





NameNode-master

The master node: NameNode

- **Réplication**
 - En cas d'échec du Datanode
 - En cas d'échec du disque
 - Sur la corruption de bloc
 - **Intégrité des données**
 - Somme de contrôle pour chaque bloc
 - Stocké dans un fichier caché
 - **Outil de rééquilibrage “équilibreur”**
 - Ajout de nouveaux nœuds
 - Démantèlement
 - Suppression de certains fichiers
-

Robustesse HDFS

- Mode sans échec
 - **Au démarrage** : Aucune réplication possible
 - Reçoit Heartbeats et Blockreports de Datanodes
 - Seul un pourcentage de blocs est vérifié pour le facteur de réplication défini

Tout va bien → Quitter le mode sans échec

- Répliquer les blocs si nécessaire
-

Résumé HDFS

- Tolérant aux pannes
 - Évolutif
 - Fiable
 - Les fichiers sont distribués en gros blocs pour
 - Lectures efficaces
 - Accès parallèle
-

Questions?

2- MapReduce

Qu'est ce que MapReduce?

- MapReduce est un modèle de programmation que Google a utilisé avec succès pour traiter ses ensembles de « grandes données » (environ 20 000 péta-octets par jour)
 - Les utilisateurs spécifient le calcul en termes d'une carte et d'une fonction de réduction,
 - Le système d'exécution sous-jacent parallélise automatiquement le calcul sur des clusters de machines à grande échelle, et
 - Le système sous-jacent gère également les pannes de machine, les communications efficaces et les problèmes de performances.

--Référence : Dean, J. et Ghemawat, S. 2008. MapReduce : traitement simplifié des données sur les grands clusters. Communication de l'ACM 51, 1 (janv. 2008), 107-113.

De la fondation CS a MapReduce?

- Considérons une grande collecte de données :

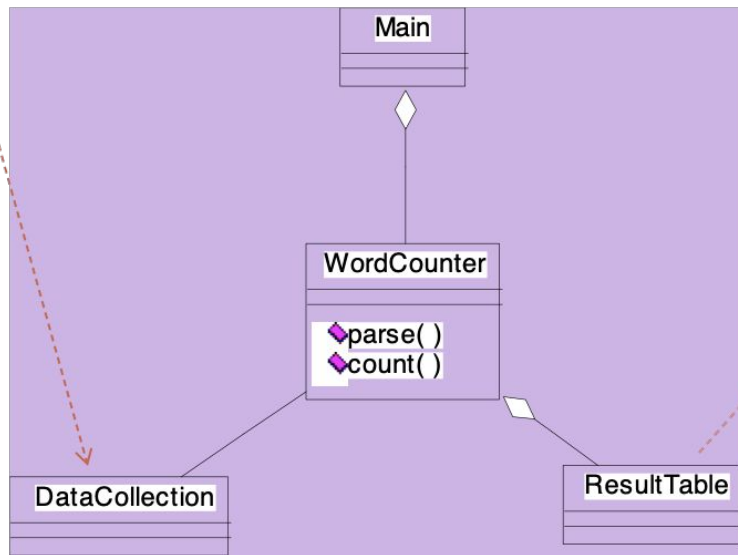
{web, weed, green, sun, moon, land, part, web, green,...}

Problème : Comptez les occurrences des différents mots de la collection.

- Concevons une solution à ce problème ;
-


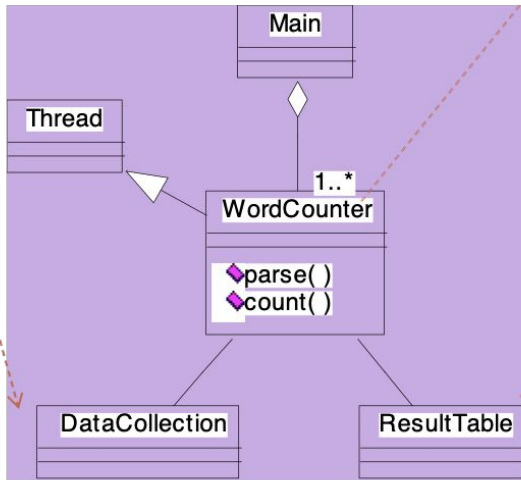
De la fondation CS a MapReduce?

{web, weed, green, sun, moon, land, part,
web, green,...}



web	2
weed	1
green	2
sun	1
moon	1
land	1
part	1

De la fondation CS a MapReduce?

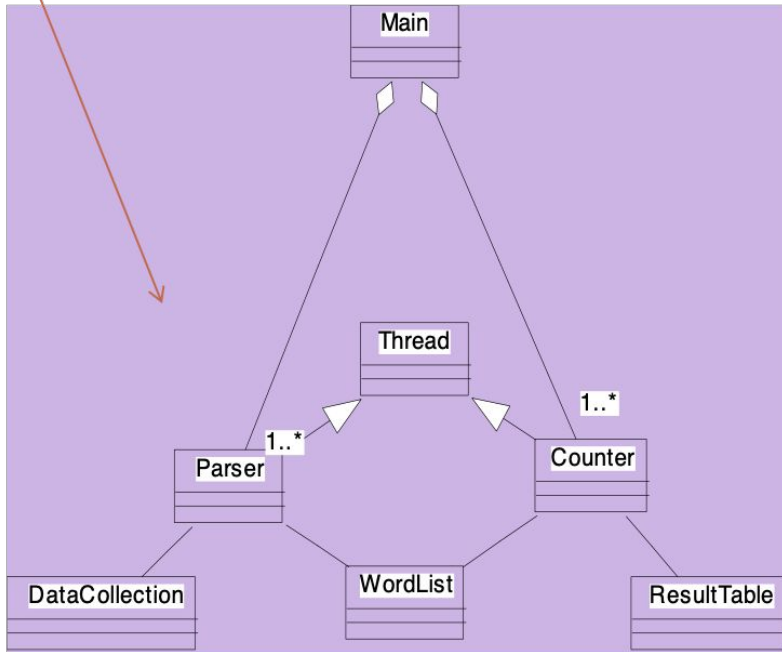
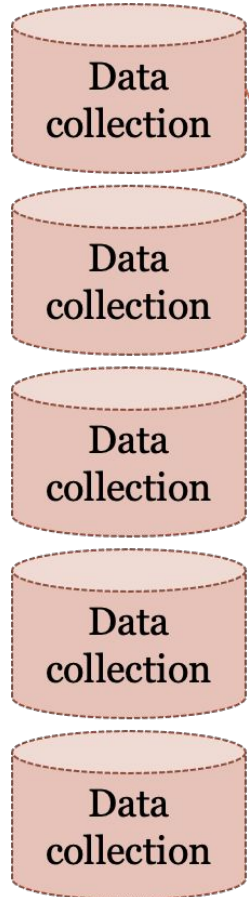


web	2
weed	1
green	2
sun	1
moon	1
land	1
part	1

Observe:
Multi-thread
Lock on shared data

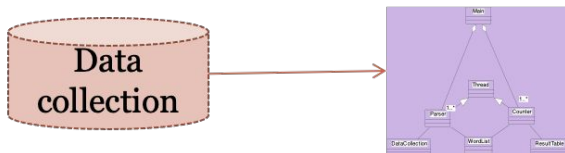
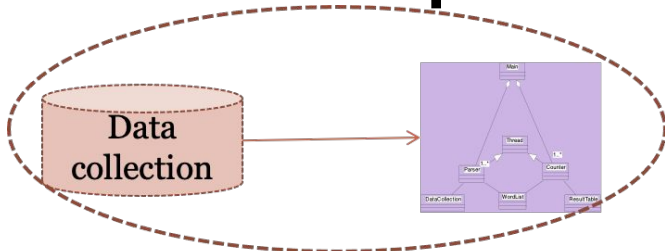
Résoudre le problème d'échelle

- Une seule machine ne peut pas traiter toutes les données : vous avez besoin d'un système (de fichiers) spécial distribué
 - Grand nombre de disques matériels de base : disons, 1000 disques de 1 TB chacun
 - Problème : avec un temps moyen entre les pannes (MTBF) ou un taux de panne de 1/1000, alors au moins 1 des 1000 disques ci-dessus serait en panne à un moment donné.
 - Ainsi, l'échec est la norme et non une exception.
 - Le système de fichiers doit être tolérant aux pannes : réplication, somme de contrôle
 - La bande passante de transfert de données est critique (emplacement des données)
 - Aspects critiques : tolérance aux pannes + réplication + équilibrage de charge, surveillance
 - Exploiter le parallélisme offert par la division de l'analyse et du comptage
 - Provisionner et localiser en faisant le traitement là où il y a les données
-



1. Données avec des caractéristiques WORM :
Traitement parallèle ;
2. Données sans dépendances :
Traitement en offline

Diviser pour régner: Localité des données

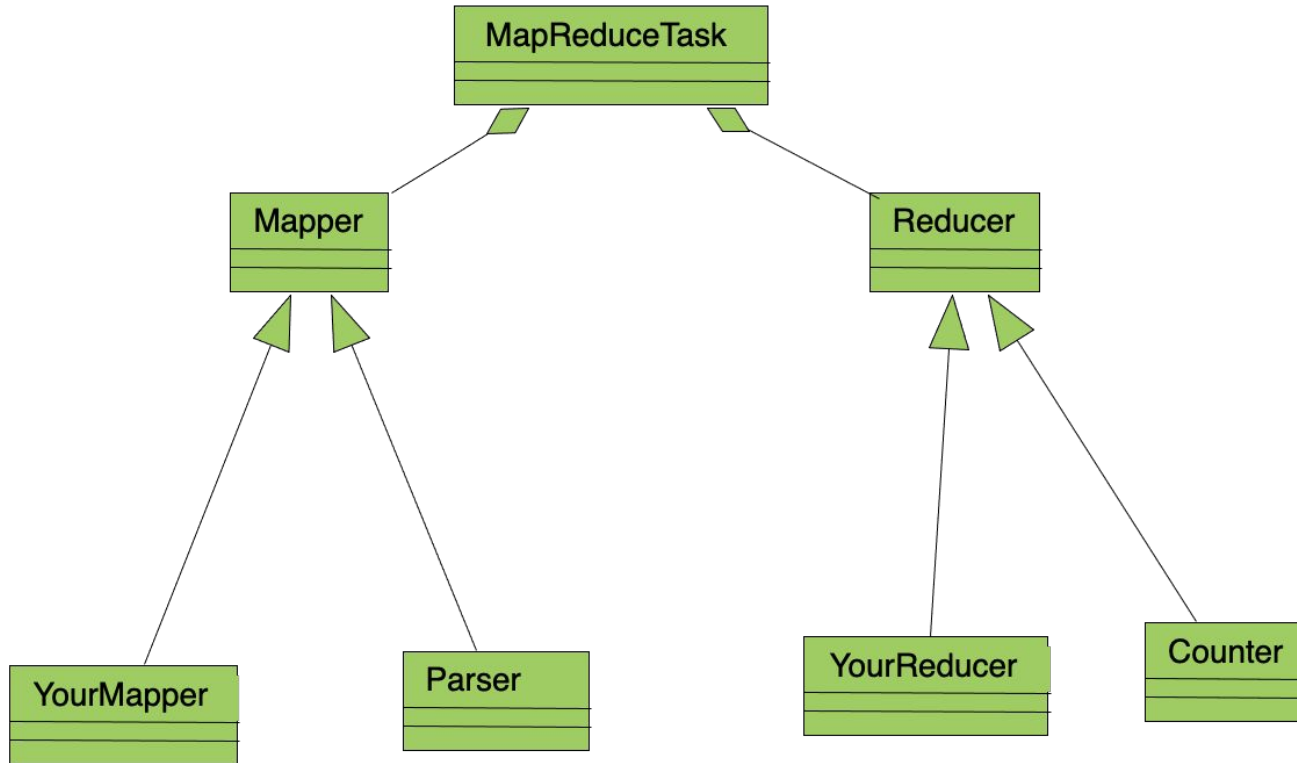


Exemple,

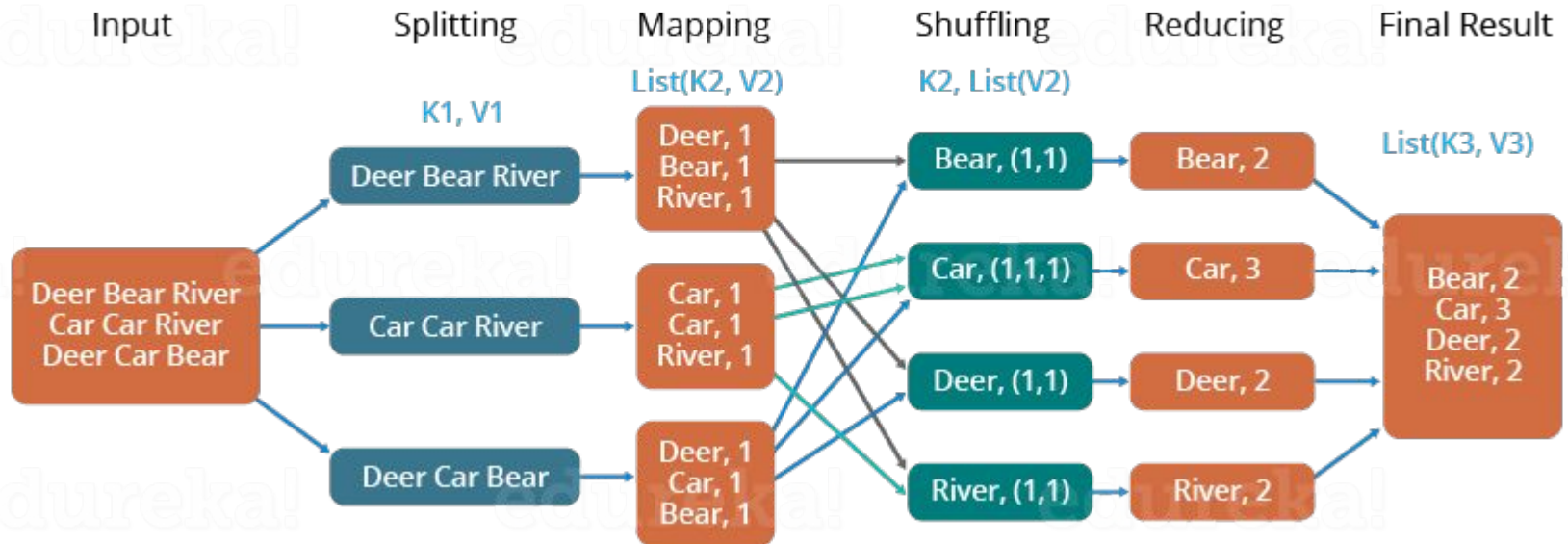
#1 : Planifier des tâches d'analyse parallèle

#2 : Planifiez des tâches de comptage parallèles

Mapper et Reducer



MapReduce en Action



Mapreduce walkthrough: Wordcount

```
/**
 * Counts the words in each line.
 * For each line of input, break the line into words and emit them as
 * (<b>word</b>, <b>1</b>).
 */
public static class MapClass extends MapReduceBase
    implements Mapper<LongWritable, Text, Text, IntWritable> {

    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(LongWritable key, Text value,
        OutputCollector<Text, IntWritable> output,
        Reporter reporter) throws IOException {
        String line = value.toString();
        StringTokenizer itr = new StringTokenizer(line);
        while (itr.hasMoreTokens()) {
            word.set(itr.nextToken());
            output.collect(word, one);
        }
    }
}
```

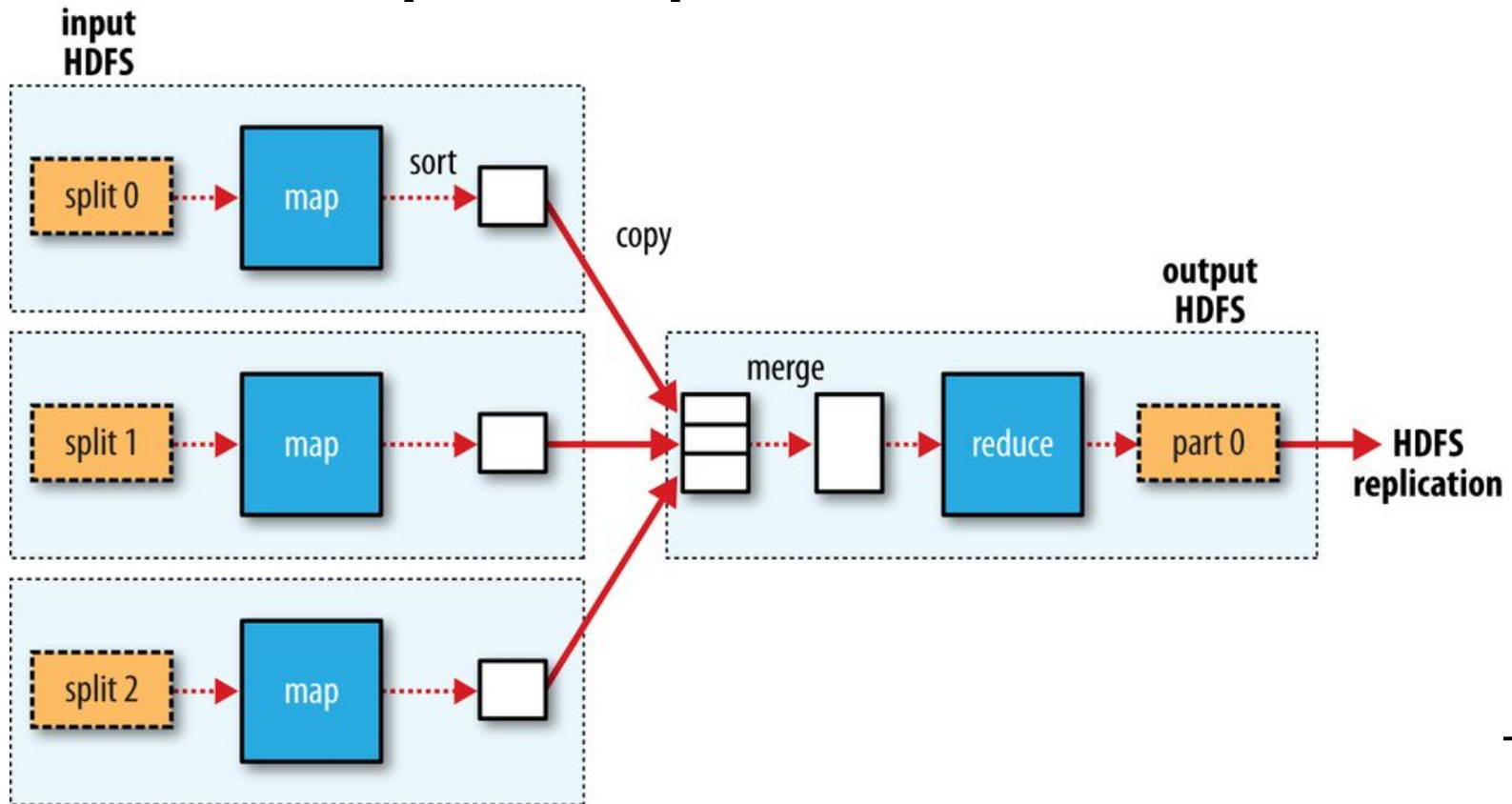
Mapreduce walkthrough: [Wordcount](#)

```
/**
 * A reducer class that just emits the sum of the input values.
 */
public static class Reduce extends MapReduceBase
    implements Reducer<Text, IntWritable, Text, IntWritable> {

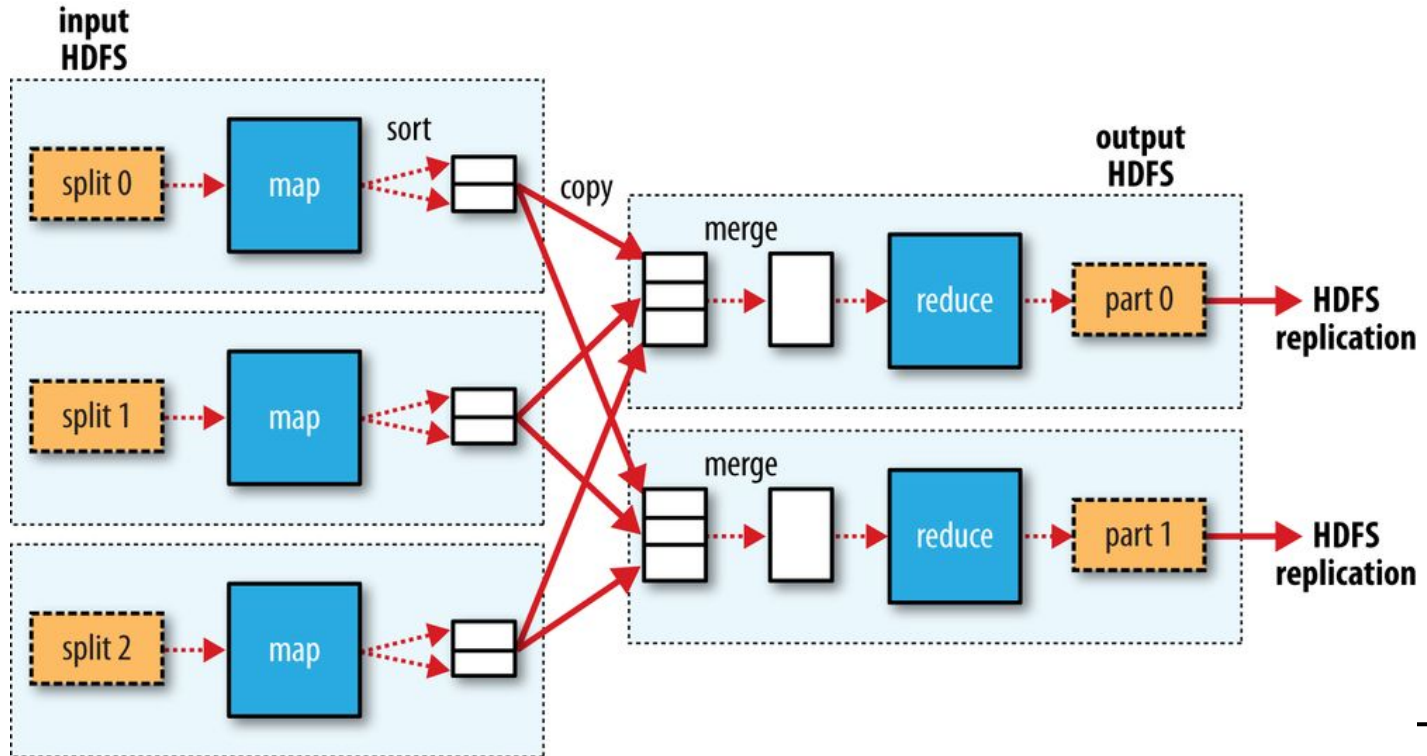
    public void reduce(Text key, Iterator<IntWritable> values,
        OutputCollector<Text, IntWritable> output,
        Reporter reporter) throws IOException {

        int sum = 0;
        while (values.hasNext()) {
            sum += values.next().get();
        }
        output.collect(key, new IntWritable(sum));
    }
}
```

Input/output dans HDFS



Input/output dans HDFS



3- HDFS

Le Design de HDFS

Fichiers très volumineux

Accès aux données en streaming

Matériel de commodité

Accès aux données à faible latence

Beaucoup de petits fichiers

Plusieurs rédacteurs, modifications de fichiers arbitraires

Conceptes de HDFS

Blocks: 128 Mo par défaut

Namenodes et Datanodes: Master et Workers

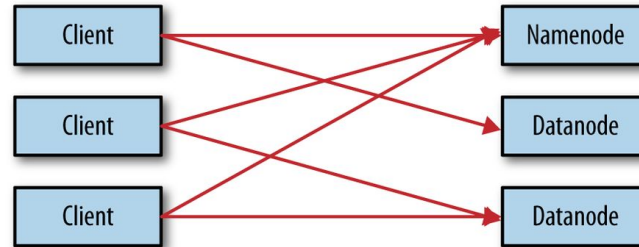
Mise en cache des Blocks: *off-heap block cache*

Haute disponibilité HDFS: Replicas (3 par défaut)

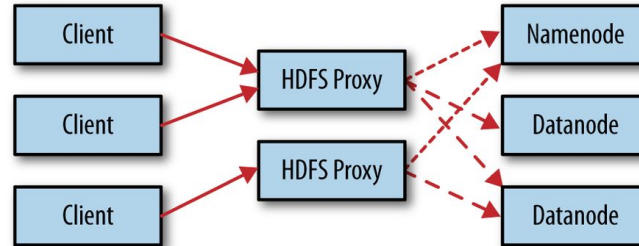
Failover and fencing: ZooKeeper

Interfaces d'accès

i) Direct access

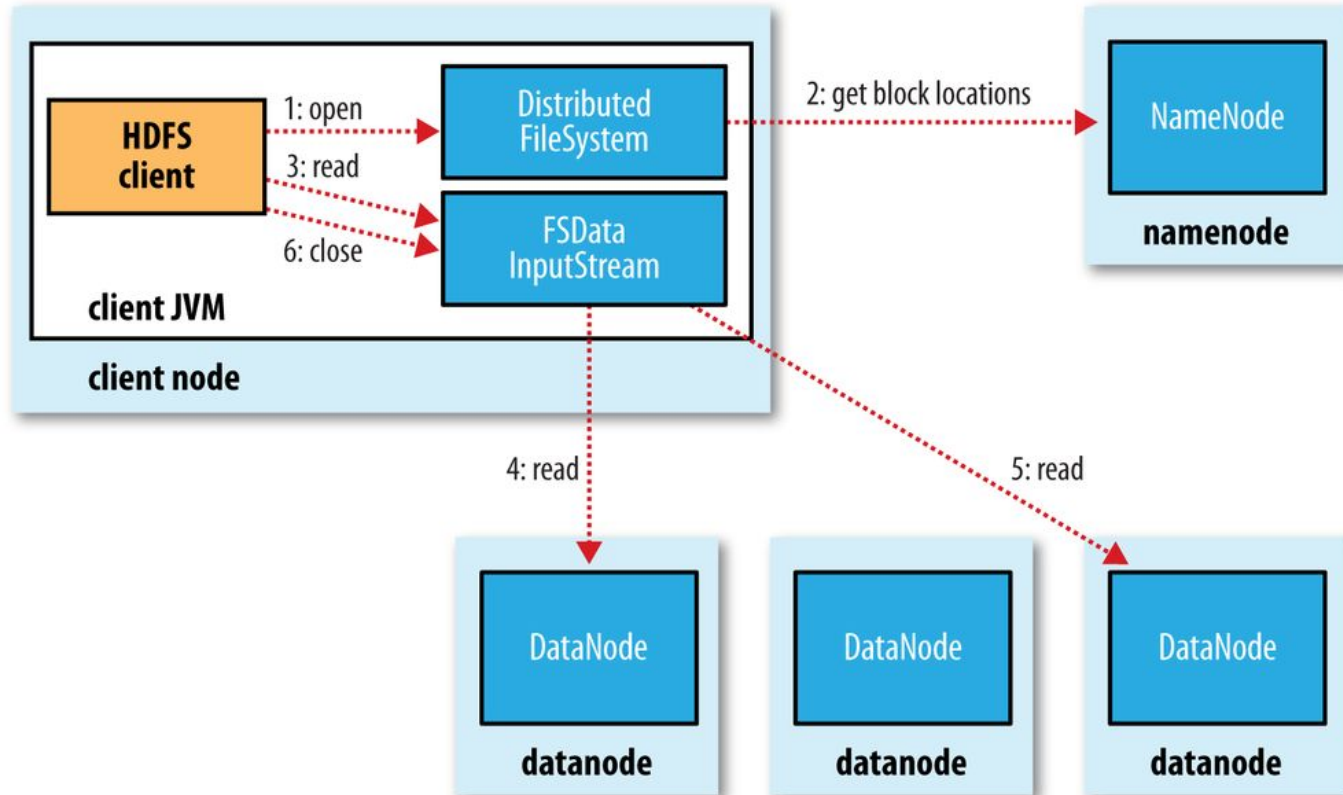


ii) HDFS proxies

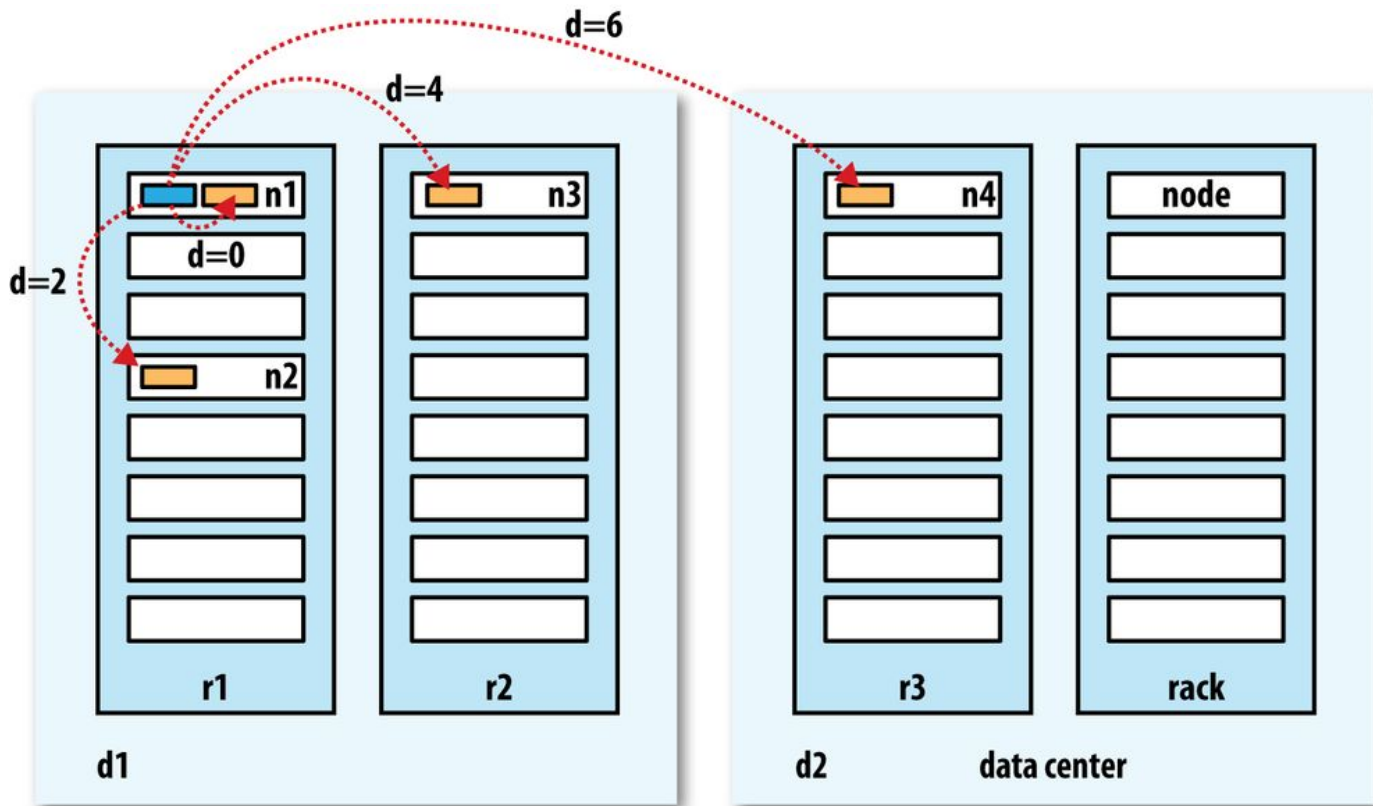


→ HTTP request - - - - - → RPC request ······ → block request

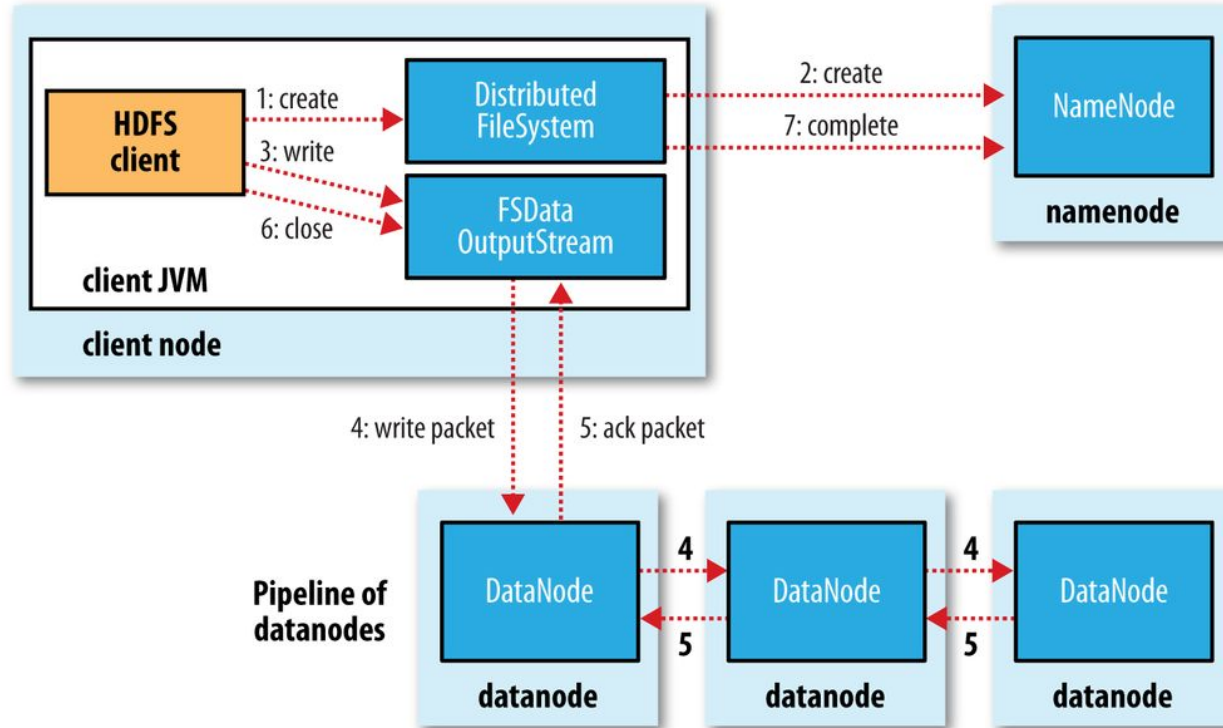
Lecture sur HDFS



Distance



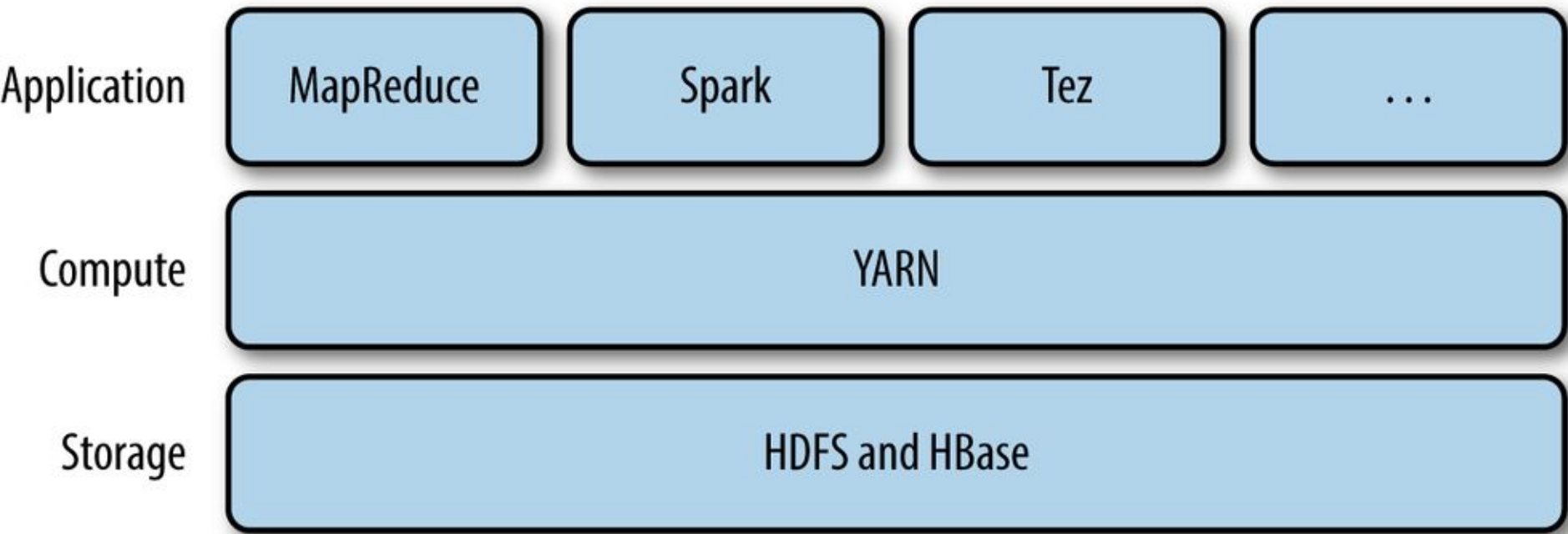
Ecriture sur HDFS



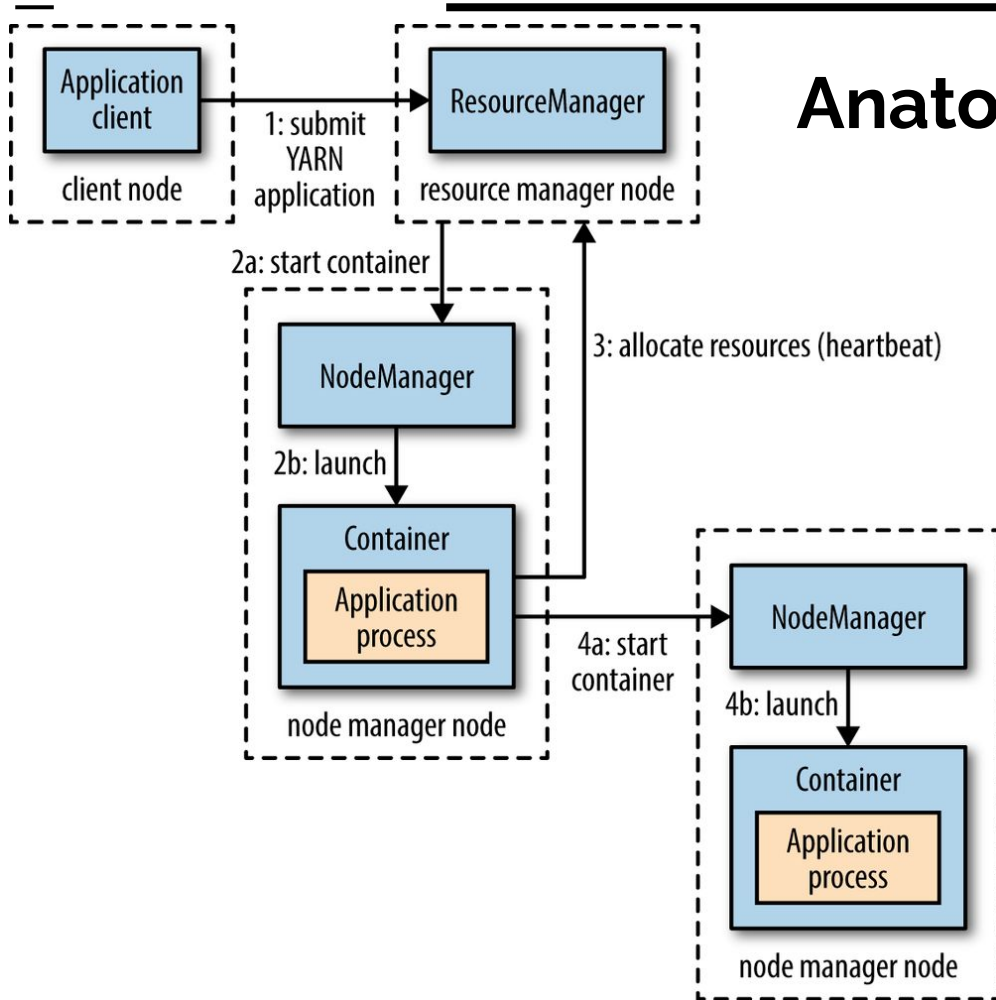
4- YARN

Yet Another **R**esource **N**egotiator

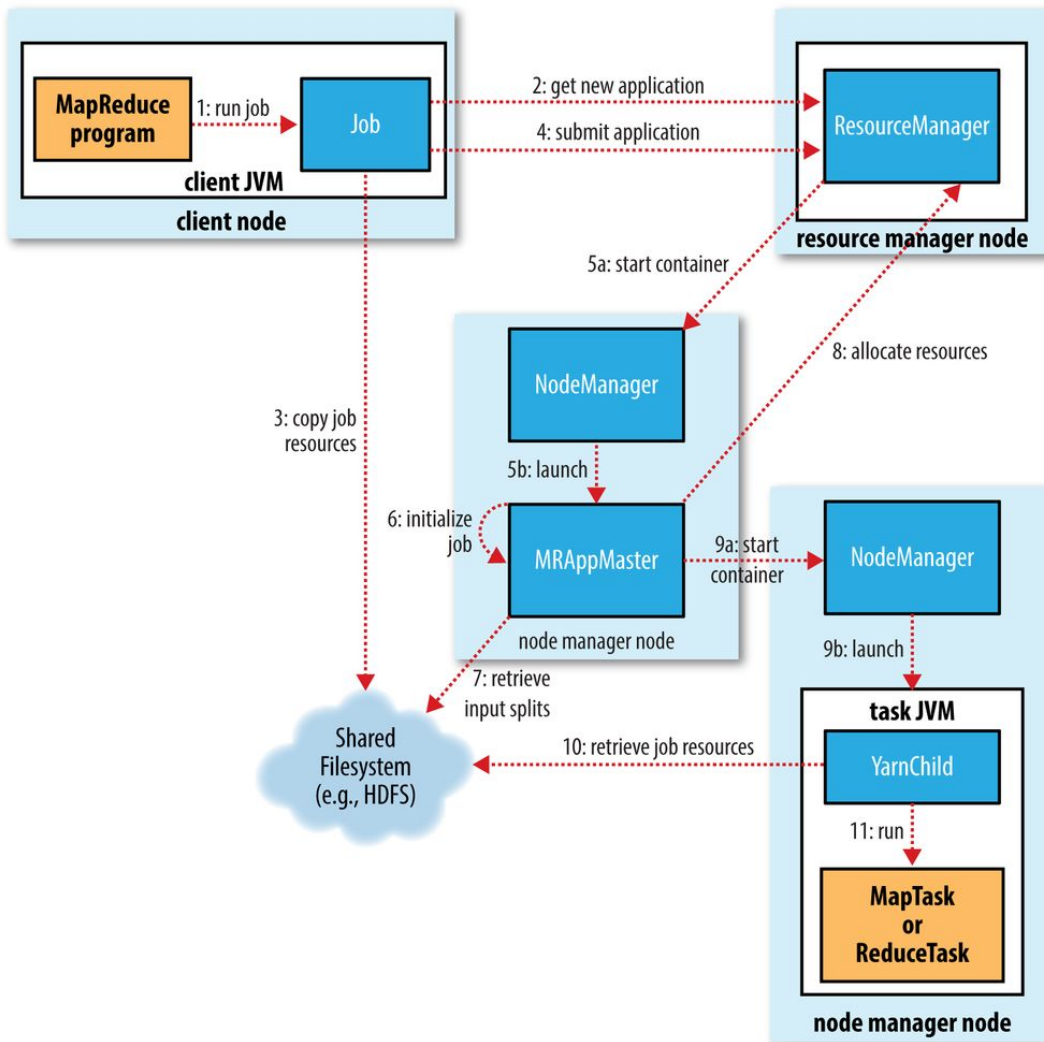
Architecture



Anatomie d'une application



6- Comment Hadoop execute MR



- Le client qui soumet le job MapReduce.
- Le gestionnaire de ressources YARN, qui coordonne l'allocation des ressources de calcul sur le cluster.
- Les gestionnaires de nœuds YARN, qui lancent et surveillent les conteneurs sur les machines du cluster.
- Le AppMaster de MapReduce, qui coordonne les tâches exécutant le job MapReduce. Le AppMaster et les tâches MapReduce s'exécutent dans des conteneurs qui sont planifiés par le gestionnaire de ressources et gérés par les gestionnaires de nœuds.
- Le système de fichiers distribué (HDFS), qui est utilisé pour partager les fichiers du job entre les autres entités

Merci

Introduction à l'Analyse des données

Intro

L'analyse de données,
qu'est-ce que c'est ?



Question de Vocabulaire 1/2



Attention :

- ▶ historiquement : plusieurs
« point de départ »
 - ▶ domaine récent dont le vocabulaire n'est pas fixé
 - ▶ évolution rapide
 - ▶ domaine applicatif *versus* domaine de recherche
-

Question de Vocabulaire 2/2

Conséquence : on ne va pas parler que d'analyse de données

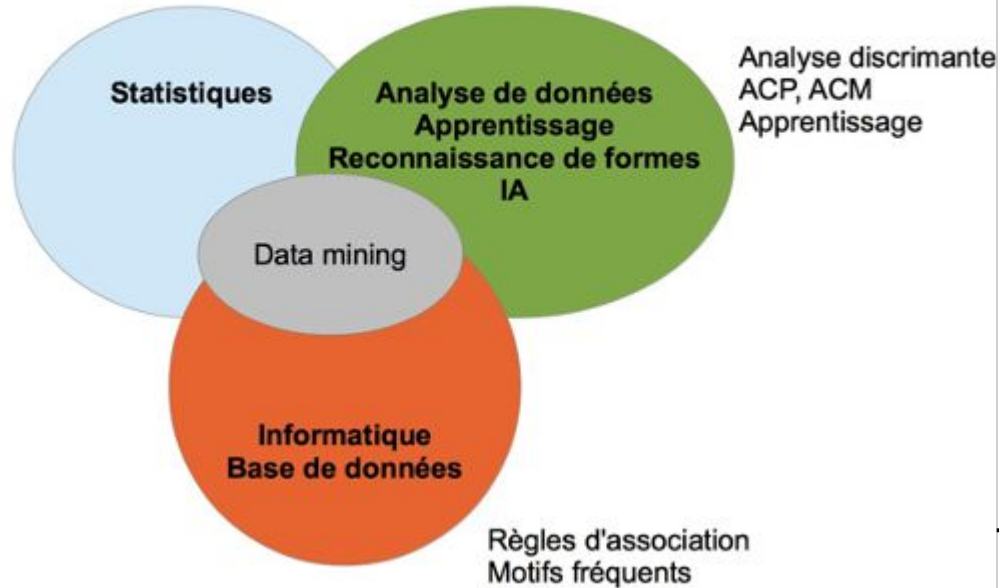
- ▶ reconnaissance des formes
(*pattern recognition*)
- ▶ Analytics
- ▶ apprentissage automatique
(*machine learning*)
- ▶ fouille de données (*data mining*)
- ▶ intelligence artificielle
- ▶ statistique
- ▶ ...

Analyse et Fouille des Données (AFD)

Rencontre de plusieurs disciplines

Régression

Maximum de vraisemblance, moindres carrés



Définitions

- **Extraction de connaissances à partir de données (KDD) :**
 - Cycle de découverte d'information regroupant la conception des grandes bases de données ou les entrepôts de données (data warehouses).
 - Ensemble des traitements à effectuer pour extraire de l'information aux données.
 - **L'analyse et la fouille de données** est un des traitements.
 - **Analyse et fouille de données = data mining**

Ensemble des techniques d'exploration de données permettant d'extraire des connaissances sous la forme de **modèles de description** afin de :

 - **Décrire** le comportement actuel des données.
 - Et/ou **Prédire** le comportement futur des données.
-

Dans ce cours

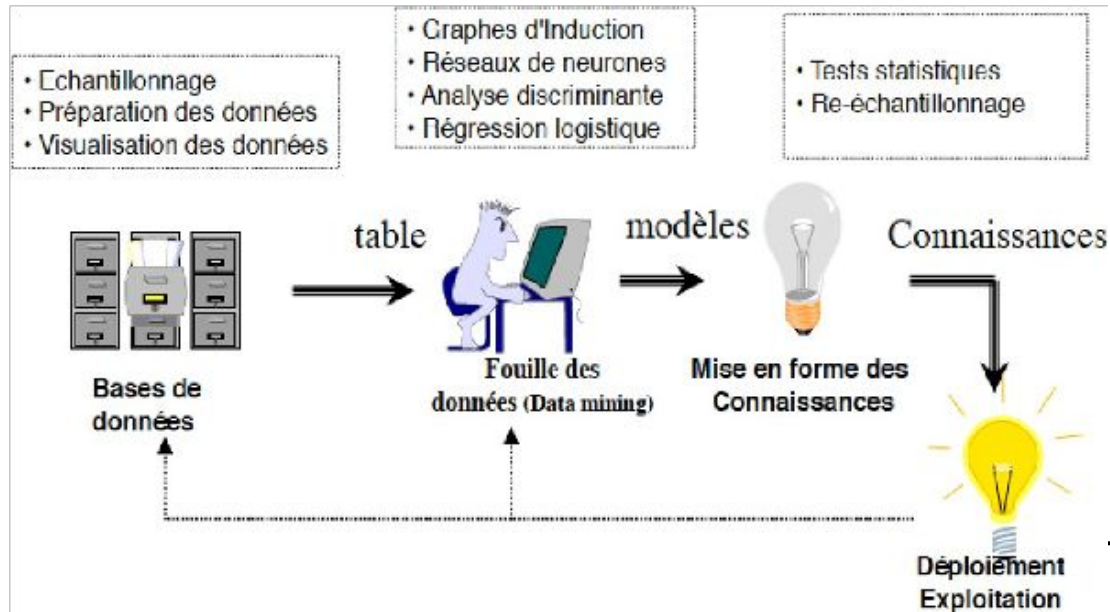
Définition (S. Tuffery)

L'AFD est l'ensemble des :

- algorithmes et méthodes
 - ... destinés à l'exploration et l'analyse
 - ... de (souvent) grandes bases de données informatiques
 - ... en vue de détecter dans ses données des règles, des associations, des tendances inconnues (non fixées a priori), des structures particulières restituant de façon concise l'essentiel de l'information utile
 - ... pour l'aide à la décision.
-

AFD: une démarche plus qu'une théorie

Processus ECD (extraction de connaissances à partir de données) ou KDD (Knowledge Discovery in Databases)



AFD: Pourquoi?

Un sujet d'actualité...

L'exploitation des données est importante car c'est :

- ▶ **méthode scientifique** ⇒ nécessité de savoir exploiter des données
 - ▶ c'est la base de la méthode scientifique (observations → lois/règles)
 - ▶ les données (et leur exploitation) au cœur de beaucoup d'avancées récentes
 - ▶ **source de revenus**
 - ▶ modèle économique des entreprises du web (Google, Facebook, Amazon, ...)
 - ▶ fournisse un service gratuit
 - ▶ seule « valeur » : capacité à exploiter les données collectées
 - ▶ **nouvelle « approche de programmation »**
 - ▶ « rêve » de l'intelligence artificielle : l'ordinateur qui apprend
 - ▶ il y a des algorithmes que l'on ne peut pas/sait pas formaliser
-

AFD: Pourquoi?

⇒ intérêt de savoir utiliser des méthodes statistiques pour exploiter de grandes masses de données aussi bien d'un point de vue économique (facebook, google, ...) que scientifique (CERN et autre).

Facebook's Data Science Group

“... on any given day, a team member could author a multistage processing pipeline in Python, design a hypothesis test, perform a regression analysis over data samples with R, design and implement an algorithm for some data-intensive product or service in Hadoop, or communicate the results of our analyses to other members of the organization”

AFD: Pourquoi?



Carte de crédit

- ▶ tous les achats sont enregistrés
- ▶ détection des fraudes/comportement à risque
- ▶ ciblage
- ▶ accord de prêt
- ▶ ...

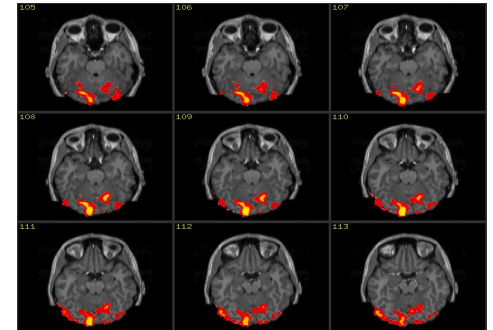
Navigation Web

- ▶ historique de la navigation
 - ▶ ciblage/marketing
 - ▶ optimisation des sites / du trafic
-

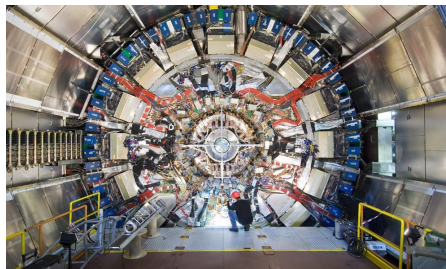
AFD: Pourquoi?

fMRI

- ▶ functional Magnetic Resonance Imaging
- ▶ variation de pression sanguine en réponse à des stimuli
- ▶ brain computer interface



Big Science



- ▶ détecteur ATLAS du CERN
 - ▶ 40M événements par secondes, 25Mo par événement
 - ▶ 1Po de données générées par secondes à analyser
 - ▶ même situation en biologie, astronomie,
- ...

AFD: Pourquoi?



- ▶ tous les textes et discussion du parlement européen sont disponibles...
 - ▶ ...avec leur traduction/interprétation
 - ▶ **corpus parallèle** : les phrases sont **alignés**
 - ▶ utilisable pour apprendre :
 - ▶ des dictionnaires
 - ▶ des systèmes de traduction automatique
 - ▶ des mémoires de traduction
 - ▶ ⊕ analyse « politique » des données
-

AFD: Pourquoi?



AFD: Pourquoi?

- ▶ historique des achats
 - ▶ historique des passages de frontières
 - ▶ liste des appels
 - ▶ analyse de la circulation routière
 - ▶ test A/B pour choisir les prix
 - ▶ pollution
 - ▶ données médicales
 - ▶ ...
-

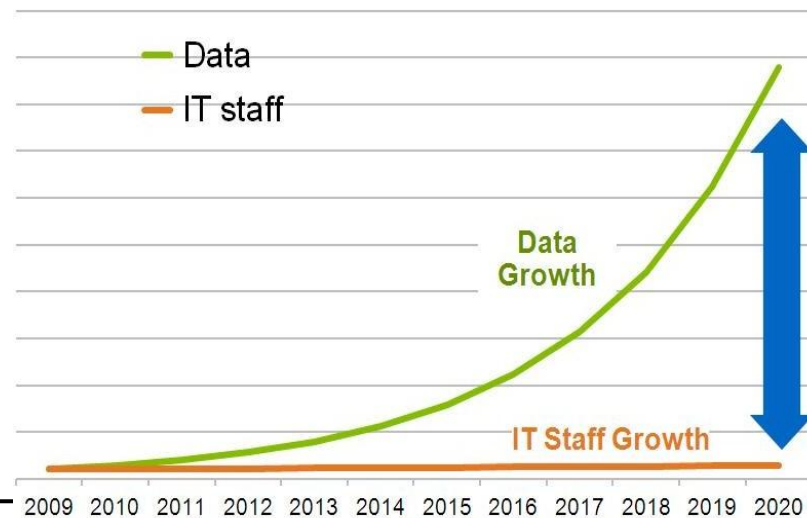
AFD: Pourquoi?

Le fossé des données (data gap)

Une grosse quantité de données qui n'est jamais analysée
⇒ mettre en place des mécanismes d'analyse automatique.

Big Data

The Data Management Gap



AFD: Composants de base

Grande quantité de données + algorithmes efficaces

Un domaine qui s'appuie sur :

- ▶ **La disponibilité de grandes quantités de données**

- ▶ Si ensemble trop petit, les structures peuvent ne résulter que du hasard.
- ▶ On peut espérer qu'un gros volume de données représente bien l'univers (échantillon).

- ▶ **Des algorithmes sûrs et efficaces**

- ▶ Algorithmes sûrs : fondés théoriquement, corrects.
 - ▶ Efficaces en temps et en espace.
 - ▶ Résultats interprétables.
 - ▶ Paramètres ajustables facilement et rapidement.
-

AFD: un exemple

Issu du livre de Adriaans and Zantige (d'après B. Espinasse)

- ▶ Un éditeur vend 5 sortes de magazines : sport, voiture, maison, musique, cinéma.
- ▶ Il veut étudier ses clients pour découvrir de nouveaux marchés ou vendre plus à ses clients habituels.

Quelques questions

1. Combien de personnes ont pris un abonnement à un magazine de cinéma cette année ?
 2. A-t-on vendu plus d'abonnement de magazines de sport cette année que l'année dernière ?
 3. Est-ce que les acheteurs de magazines de musique sont aussi amateurs de cinéma ?
 4. Quelles sont les caractéristiques principales des lecteurs de magazine de cinéma ?
 5. Peut-on prévoir les pertes de client et prévoir des mesures pour les diminuer ?
-

AFD: Exemple

1: Combien de personnes ont pris un abonnement à un magazine de cinéma cette année ?

Requête SQL à partir des données opérationnelles suffit si les tables concernées ont été suffisamment indexées.

2: A-t-on vendu plus d'abonnement de magazines de sport cette année que l'année dernière ?

- ▶ Nécessite de garder toutes les dates de souscription, même pour les abonnements résiliés.
 - ▶ Requêtes multidimensionnelles de type OLAP.
-

AFD: Exemple

3 : Est-ce que les acheteurs de magazine de musique sont aussi amateurs de cinéma ?

- ▶ Exemple simplifié de problème où l'on demande si les données vérifient une règle.
 - ▶ Réponse formulée par une valeur estimant la probabilité que la règle soit vraie.
 - ▶ Utilisation d'outils statistiques.
-

AFD: Exemple

4. Quelles sont les caractéristiques principales des lecteurs de magazine de cinéma ?

Question plus ouverte, il s'agit de trouver une règle et non plus de la vérifier ou de l'utiliser.

5 : Peut-on prévoir les pertes de client et prévoir des mesures pour les diminuer ?

Question ouverte : il faut disposer d'indicateurs comme durée d'abonnement, délai de paiement, ...

C'est pour ce type de questions que sont mis en oeuvre les outils d'analyse et de fouille de données

AFD: Les données

Les données peuvent être vues comme une collection d'objets (enregistrements) et leurs attributs.

- ▶ Un attribut est une propriété et ou une caractéristique de l'objet.
- ▶ Un ensemble d'attributs décrit un objet.

Attributes

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Objects

AFD: Les données

- ▶ La valeur d'un attribut est un nombre ou un symbole.
- ▶ Ne pas confondre attribut et valeur

Types

- ▶ Quantitative (numérique, exprime une quantité)
 - ▶ Discrète (ex : nombre d'étudiants dans un cours) ou continue (ex : longueur)
 - ▶ Echelle proportionnelle (chiffre d'affaires, taille), ou échelle d'intervalle (température, QI)
 - ▶ Qualitative
 - ▶ Variable ordinale (classement à un concours, échelle de satisfaction client)
 - ▶ Variable nominale (couleur de yeux, diplôme obtenu, CSP, sexe)
 - ▶ Les **modalités** d'une variable sont l'ensemble des valeurs qu'elle prend dans les données
ex : les modalités de notes sont $\{0, 1, 2, \dots, 20\}$ les modalités de couleur sont $\{\text{bleu, vert, noir, ...}\}$
-

AFD: Les données disponibles

- ▶ Transactions.
- ▶ Bases de données des entreprises.
- ▶ Téléphone portable.
- ▶ Satellites : espace et la terre.
- ▶ Données temporelles : cours de la bourse, météo.
- ▶ Génomique.
- ▶ Données du web.
- ▶ Données textuelles.
- ▶ ...



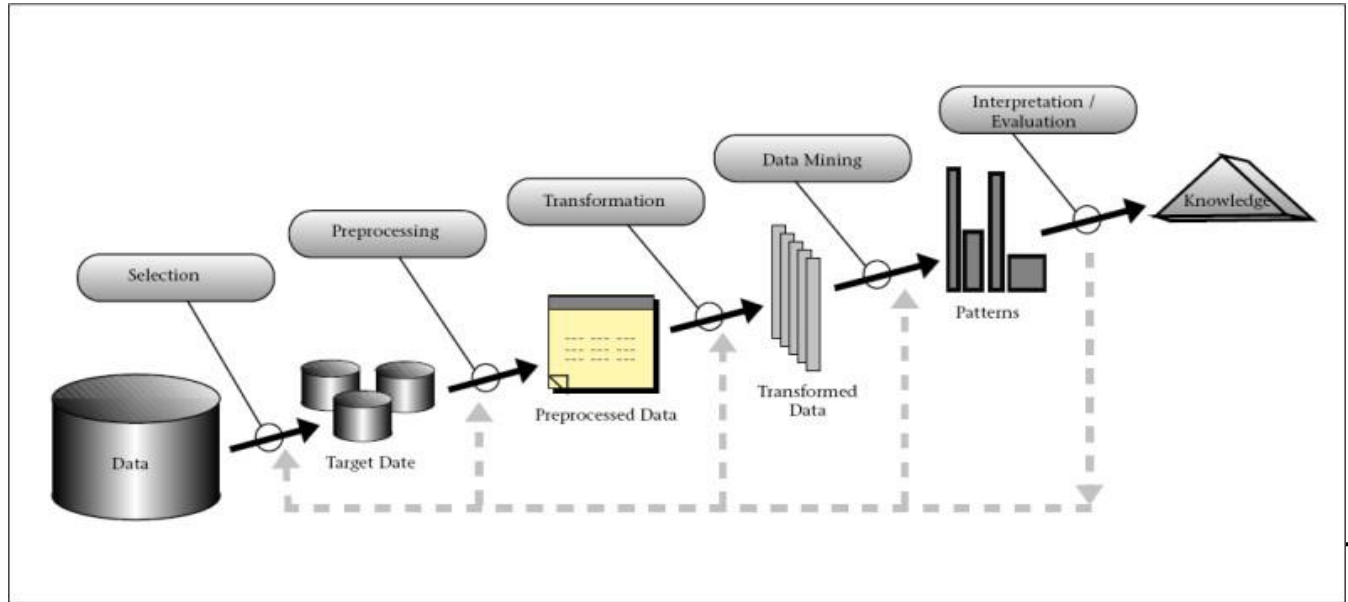
AFD: Types de connaissances extraites

Connaissances sous la forme de modèles de description permettant de

- ▶ **décrire le comportement actuel** des données et/ou
 - ▶ **prédire le comportement futur** des données.
 - ▶ **Analyses**
 - e.g. distribution du trafic routier en fonction de l'heure
 - ▶ **Règles**
 - e.g. si un client a acheté un produit alors il sera intéressé par un autre.
 - ▶ **Attribution de scores de qualité**
 - e.g. score de fidélité au client
 - ▶ **Classification d'entités**
 - e.g. mauvais payeurs.
-

Processus ECD (extraction de connaissances à partir de données)

Un processus découpé en 5 étapes, une dernière étape étant l'utilisation du modèle.



Processus ECD (extraction de connaissances à partir de données)

Un déroulement non linéaire

- On constate souvent à l'étape de validation que :
 - les performances obtenues sont insuffisantes.
 - les utilisateurs du domaine jugent l'information inexploitable.
 - ...
- Il faut donc :
 - Choisir une autre méthode de fouille.
 - Remettre en cause l'étape de transformation.
 - Enrichir les données

Dans un projet d'ECD, le temps passé à l'étape de fouille de données ne représente souvent que 20% du temps.

Étape de sélection des données

Elle consiste à

- ▶ Obtenir des données en accord avec les objectifs de l'ECD.
- ▶ Ces données proviennent le plus souvent (mais pas toujours) de bases de production ou d'entrepôts.
 - ▶ Par l'utilisation d'outils de requêtage (SQL, OLAP, ...).
 - ▶ Copie sur une machine adéquate (pour pouvoir les modifier et pour des questions de performance)
- ▶ Structuration des données en champs typés.

Client	Nom	Adresse	Date d'abonnement	Magazine
23134	Bémol	Rue du moulin, Paris	7/10/2006	Voiture
23134	Bémol	Rue du moulin, Paris	12/5/2006	Musique
23134	Bémol	Rue du moulin, Paris	25/7/2005	BD
31435	Bodinoz	Rue Verte, Nancy	11/11/1111	BD
43342	Airinair	Rue de la source, Brest	30/05/2005	Sport
25312	Talonion	Rue du Marché, Paris	25/02/2007	NULL
43241	Manvussa	NULL	14/4/2006	Sport
23130	Bémolle	Rue du moulin, Paris	11/11/1111	Maison

Étape de prétraitement

Elle consiste à

- Nettoyer les données
 - Corrections des doublons, des erreurs de saisie.
 - Contrôle sur l'intégrité des domaines de valeurs :
détection des valeurs aberrantes.
 - Détection des informations manquantes
 - Enrichissement des données
-

Étape de prétraitement

Corrections des doublons, des erreurs de saisie

Client	Nom	Adresse	Date d'abonnement	Magazine
23134	Bémol	Rue du moulin, Paris	7/10/2006	Voiture
23134	Bémol	Rue du moulin, Paris	12/5/2006	Musique
23134	Bémol	Rue du moulin, Paris	25/7/2005	BD
31435	Bodinoz	Rue Verte, Nancy	11/11/1111	BD
43342	Airinair	Rue de la source, Brest	30/05/2005	Sport
25312	Talonion	Rue du Marché, Paris	25/02/2007	NULL
43241	Manvussa	NULL	14/4/2006	Sport
23130	Bémolle	Rue du moulin, Paris	11/11/1111	Maison

Étape de prétraitement

Intégrité de domaine

Client	Nom	Adresse	Date d'abonnement	Magazine
23134	Bémol	Rue du moulin, Paris	7/10/2006	Voiture
23134	Bémol	Rue du moulin, Paris	12/5/2006	Musique
23134	Bémol	Rue du moulin, Paris	25/7/2005	BD
31435	Bodinoz	Rue Verte, Nancy	11/11/1111	BD
43342	Airinair	Rue de la source, Brest	30/05/2005	Sport
25312	Talonion	Rue du Marché, Paris	25/02/2007	NULL
43241	Manvussa	NULL	14/4/2006	Sport
23130	Bémol	Rue du moulin, Paris	11/11/1111	Maison

Étape de prétraitement

Information manquante

- ▶ Cas où les champs ne contiennent aucune donnée.
- ▶ Parfois intéressant de conserver ces enregistrements car l'absence d'information peut être informative (e.g. fraude).

Client	Nom	Adresse	Date d'abonnement	Magazine
23134	Bémol	Rue du moulin, Paris	7/10/2006	Voiture
23134	Bémol	Rue du moulin, Paris	12/5/2006	Musique
23134	Bémol	Rue du moulin, Paris	25/7/2005	BD
31435	Bodinoz	Rue Verte, Nancy	NULL	BD
43342	Airinair	Rue de la source, Brest	30/05/2005	Sport
25312	Talonion	Rue du Marché, Paris	25/02/2007	NULL
43241	Manvussa	NULL	14/4/2006	Sport
23134	Bémol	Rue du moulin, Paris	NULL	Maison

Étape de prétraitement

Enrichissement

- ▶ Recours à d'autres bases de données souvent pour ajouter de nouveaux champs en conservant le même nombre d'enregistrements.
- ▶ Plusieurs difficultés :
 - ▶ Relier les données, parfois hétérogènes, entre elles.
 - ▶ Introduction de nouvelles valeurs manquantes et/ou aberrantes.

Client	Date naissance	Revenus	Propriétaire	Voiture
Bémol	13/1/50	20 000	Oui	Oui
Bodinoz	21/5/70	12 000	Non	Oui
Airinair	15/06/63	9 000	Non	Non
Manvussa	27/03/47	15 000	Non	Oui

Étape de transformation (codage et normalisation)

Une étape très dépendante du choix de l'algorithme de fouilles de données utilisés.

- ▶ Regroupements.
 - ▶ Cas où les attributs prennent un très grand nombre de valeurs discrètes (e.g. adresses que l'on peut regrouper en 2 régions (Paris- Province))
 - ▶ Attributs discrets.
 - ▶ Les attributs discrets prennent leurs valeurs (souvent textuelles) dans un ensemble fini donné (e.g. colonne magazine de l'exemple).
 - ▶ Deux représentations possibles : représentation verticale ou représentation horizontale ou éclatée (plus adaptée à la fouille de données)
 - ▶ Changements de types pour permettre certaines manipulations comme par exemple des calculs de distance, de moyenne (e.g. date de naissance)
 - ▶ Uniformisation d'échelle.
 - ▶ Certains algorithmes sont basés sur des calculs de distance entre enregistrements :
 - ▶ Variations d'échelle selon les attributs peuvent perturber ces
-

Étape de transformation (codage et normalisation)

Représentation horizontale

Client	Magazine
23134	Voiture
23134	Musique
23134	BD
31435	BD
43342	Sport
43241	Sport
23134	Maison

Représentation éclatée

Client	Sport	BD	Voiture	Maison	Musique
23134	0	1	1	1	1
31435	0	1	0	0	0
43342	1	0	0	1	0
43241	1	0	0	1	0

Étape de transformation (codage et normalisation)

Client	Sport	BD	Voiture	Maison	Musique	DN	Rev	Prop	Voit	PP	DA
23134	0	1	1	1	1	50	20	oui	oui	1	4
31435	0	1	0	0	0	30	12	non	oui	0	null
43342	1	0	0	1	0	37	9	non	non	1	5
43241	1	0	0	1	0	53	15	non	oui	null	4

Etape de transformation sur l'exemple

- Avec :
 - DN : date de naissance
 - Rev : revenus
 - Prop : Propriétaire
 - Voit : possède une voiture
 - PP : Paris ou province
 - DA : date d'abonnement
-

Étape de fouille de données

Etape :

- ▶ Au coeur même du processus ECD.
- ▶ Difficile à mettre en oeuvre.
- ▶ Coûteuse.
- ▶ Aux résultats devant être interprétés et relativisés.

Approche traditionnelle

1. Regarder, explorer.
 2. Etablir une hypothèse, un modèle.
 3. Essayer de contredire ou de vérifier de modèle.
-

Étape d'évaluation et de validation

Deux modes de validation

- ▶ Par statistique et / ou.
 - ▶ Par expertise.
-

Typologie des méthodes de fouilles de données

Typologie selon l'objectif

- **Classification** : examiner les caractéristiques d'un objet et lui attribuer une classe.
e.g. diagnostic ou décision d'attribution de prêt à un client.
 - **Prédiction** : prédire la valeur future d'un attribut en fonction d'autres attributs.
e.g. prédire la qualité d'un client .
 - **Association** : déterminer les attributs qui sont corrélés.
e.g. analyse du panier de la ménagère
 - **Segmentation** : former des groupes homogènes à l'intérieur d'une population.
-

Typologie des méthodes de fouilles de données

Typologie selon le type de modèle obtenu

- Modèles prédictifs.
 - Utilisent les données existantes et des résultats connus sur ces données pour développer des modèles capables de prédire les valeurs d'autres données. e.g. *Prédire les clients qui ne rembourseront pas leur crédit.*
 - Utilisés principalement en classification et prédiction.
 - Modèles descriptifs.
 - Proposent des descriptions de données pour aider à la prise de décision.
 - Souvent en amont de la construction de modèles prédictifs.
 - Utilisés principalement en segmentation et association.
-

Typologie des méthodes de fouilles de données

Typologie selon le type d'apprentissage utilisé

- Apprentissage supervisé : fouille supervisée
 - Processus qui prend en entrée des exemples d'apprentissage contenant à la fois des données d'entrée et de sortie.
 - Les exemples d'apprentissage sont fournis avec leur classe.
 - But : classer correctement un nouvel exemple.
 - Utilisés principalement en classification et prédiction.
 - Apprentissage non supervisé : fouille non supervisée
 - Processus qui prend en entrée des exemples d'apprentissage contenant que des données d'entrée
 - Pas de notion de classe
 - But : regrouper les exemples en paquets (clusters) d'exemples similaires.
 - Utilisés principalement en segmentation et association.
-

Classification

Examiner les caractéristiques d'un objet et lui attribuer une classe (un champ particulier à valeurs discrètes).

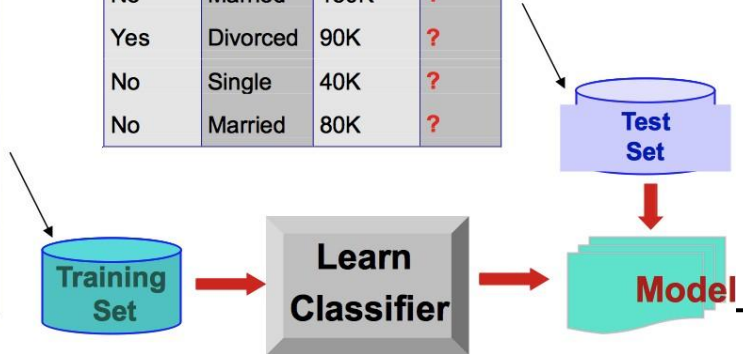
- ▶ Etant donnée une collection d'enregistrements (**ensemble d'apprentissage**).
 - ▶ Chaque enregistrement contient un ensemble d'attributs et un de ces attributs est sa classe.
 - ▶ Trouver un modèle pour l'attribut classe comme une fonction de la valeurs des autres attributs
 - ▶ But : permettre d'assigner une classe à des enregistrements inconnus de manière aussi précise que possible.
 - ▶ **Un ensemble de test** est utilisé pour déterminer la précision du modèle.
-

Classification : Exemple

categorical
categorical
continuous
class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



AFD: Exemple d'application

Marketing direct

- ▶ But : réduire le coût du mailing en ciblant un ensemble de consommateurs qui achètent vraisemblablement un nouveau téléphone portable.
 - ▶ Approche :
 - ▶ Utiliser des données pour un produit similaire.
 - ▶ On sait quels consommateurs ont acheté. La décision (Achat - Pas achat) est l'attribut classe.
 - ▶ Collecter diverses informations sur ce type de consommateurs.
 - ▶ Cette information représente les entrées du classifieur.
-

Segmentation

- ▶ Détection de fraudes à la carte bancaire à l'aide des transactions et d'informations sur le porteur du compte.
 - ▶ Détection de désabonnement à l'aide des données sur d'autres consommateurs présents ou passés.
 - ▶ Catalogage du ciel : classification des objets du ciel à l'aide d'images.
-

Segmentation

Former des groupes homogènes à l'intérieur d'une population

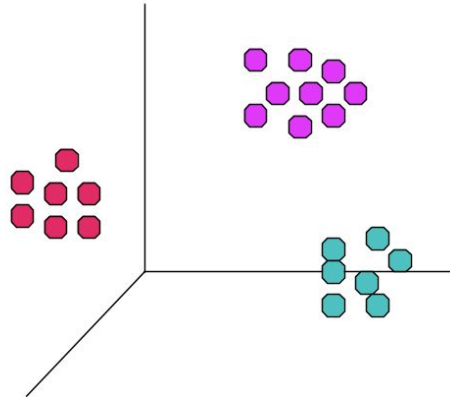
- ▶ Etant donné un ensemble de points, chacun ayant un ensemble d'attributs, et une mesure de similarité définie sur eux, trouver des groupes tels que :
 - ▶ Les points à l'intérieur d'un même groupe sont très similaires entre eux.
 - ▶ Les points appartenant à des groupes différents sont très dissimilaires.
 - ▶ Le choix de la mesure de similarité est important.
-

Segmentation

☒ Euclidean Distance Based Clustering in 3-D space.

Intracluster distances
are minimized

Intercluster distances
are maximized



Segmentation

- ▶ Segmentation de marchés .
- ▶ Segmentation de documents.
- ▶ ...



Association

Entrée : Un ensemble de tickets de caisse

- ▶ Une observation = un caddie, un ticket de caisse.
- ▶ Non prise en compte de la fréquence des produits.
- ▶ Un grand nombre de produits, un grand nombre de caddies (petit sous ensemble de l'ensemble de produits).

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Sortie : Des règles

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Association : exemples d'application

- ▶ Marketing et promotions sur des produits.
 - ▶ Gestion du supermarchés : rayonnage.
 - ▶ Inventaire.
 - ▶ ...
-

Résumé



Masse de données
(corpus)



- connaissances
- informations
- prédictions