Mémorandum

Mémorandum est un cabinet de **conseil en data stratégie**.

Nous intervenons en trois phases :

1. Réflexion sur l'usage de la donnée dans votre entreprise
2. Analyse de vos données
3. Industrialisation de solutions informatiques

Nous apportons :

- ➔ Une méthodologie mélant stratégie et technique.
- ➔ Des preuves de concepts "machine learning" avec les outils en pointe de la communauté open source
- ➔ Des méthodes agiles et de Lean Analytics qui garantissent des résultats adaptés

Chacune de nos missions s'accompagne d'une formation de nos clients à nos méthodes.

Romain Jouin - Associé
INT Management 2006 - Télécom Paris 2013
7 ans de commercial
25 ans d'informatique

Denis Oblin - Associé
Centrale 1994 - Télécom Paris 2013
10 ans de conseil en stratégie
7 ans en direction opérationnelle Groupama

Mémorandum a trois expertises majeures :

- ➔ Technique
  - ◆ Big Data
  - ◆ Machine Learning
- ➔ Fonctionnelle
  - ◆ Stratégie de la micro décision
  - ◆ Marketing
- ➔ Métier
  - ◆ Relation client
  - ◆ Force de vente

The buzz and the truth.


Big data a big deal in business

RIGHT STRATEGY:
It allows for better decision-making and marketing

DR AHMAD SHARAM

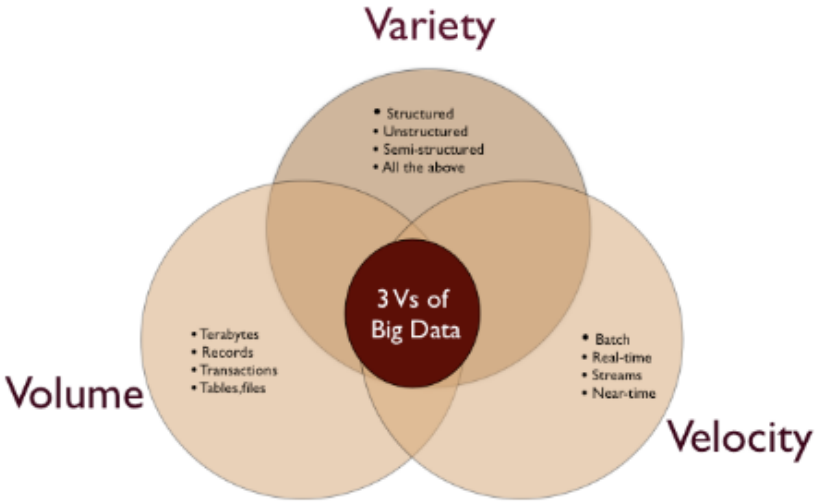
Big Data

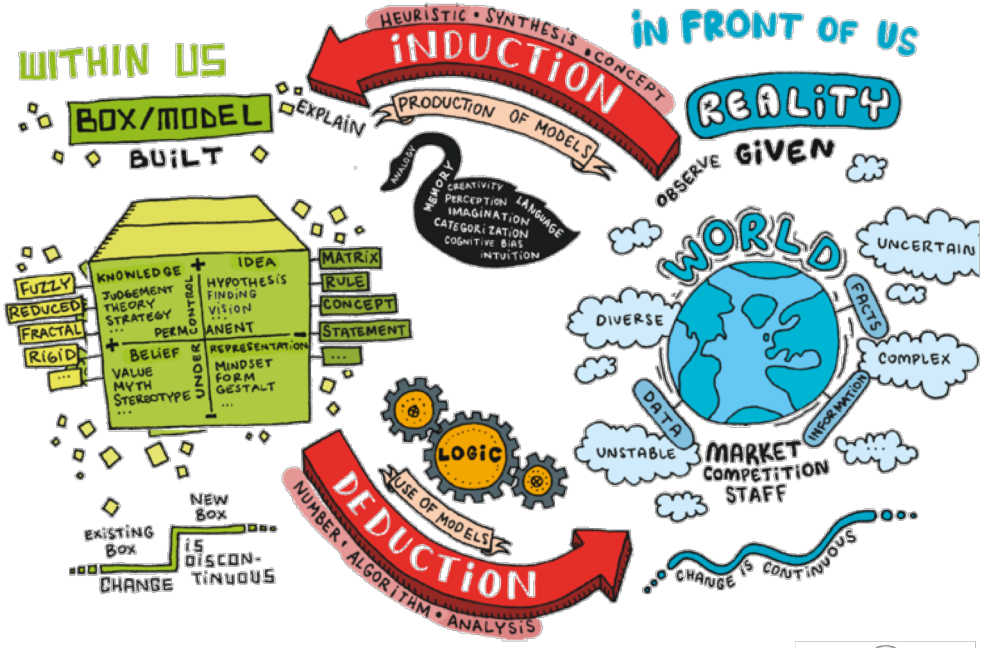
Big Data

3V

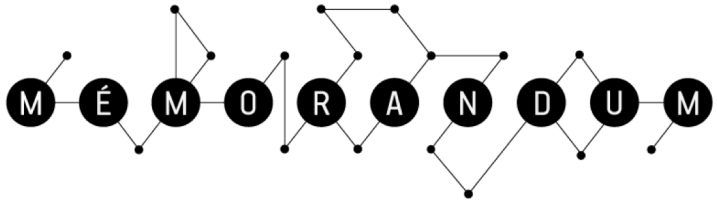
Machine Learning

0 V





© Luc de Brabandere 2012

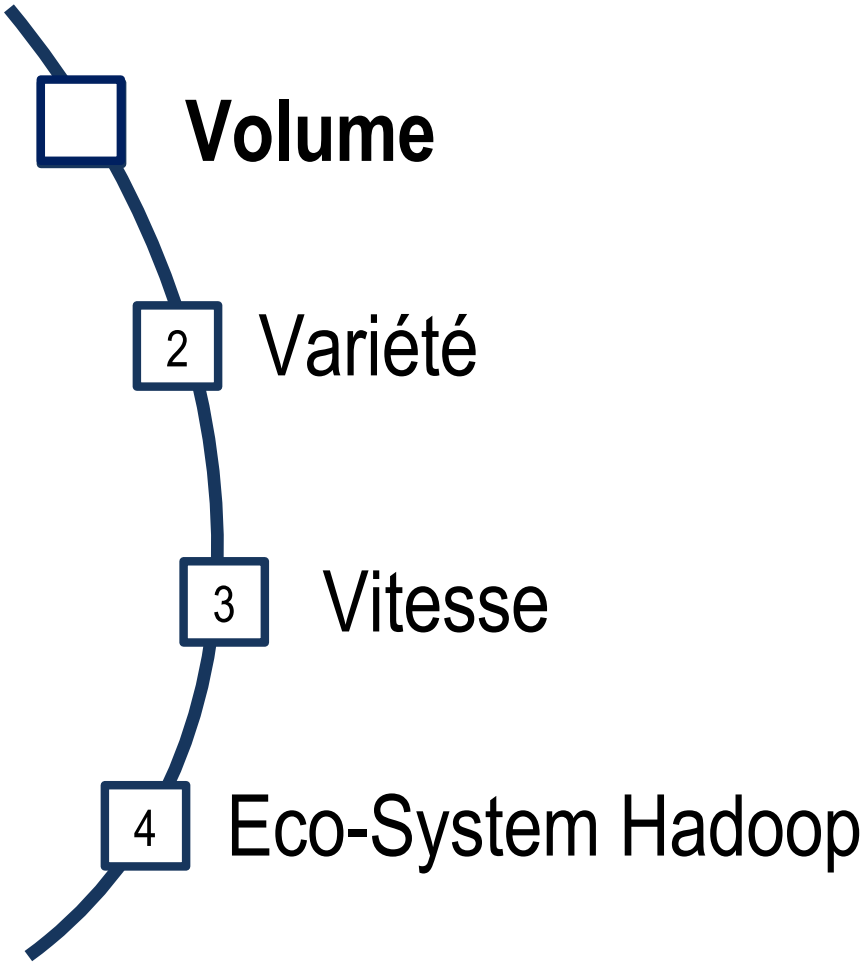


Introduction aux technologies Big Data

www.memorandum.pro

Conseil en stratégie Big Data

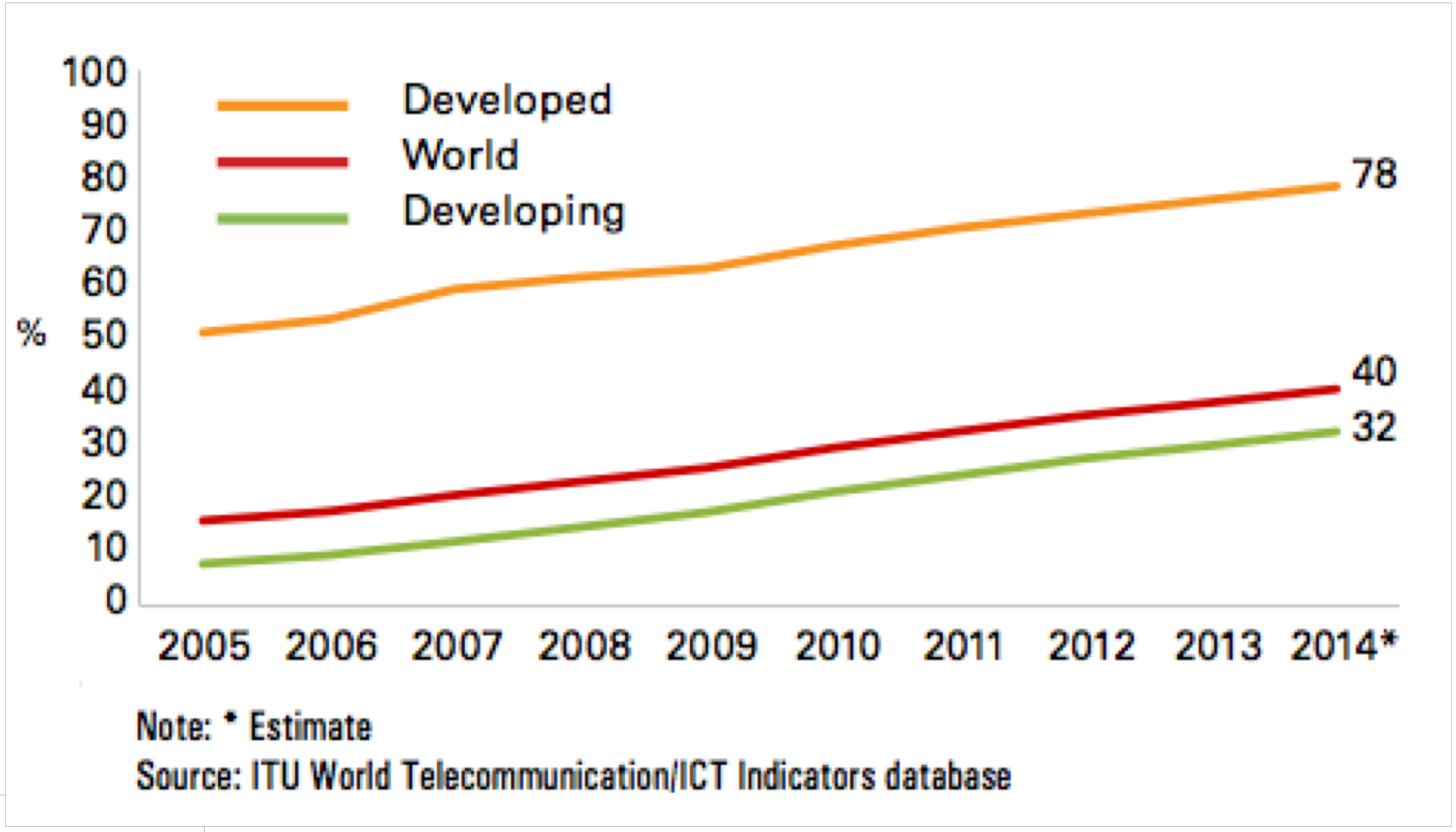
Cheminement du cours



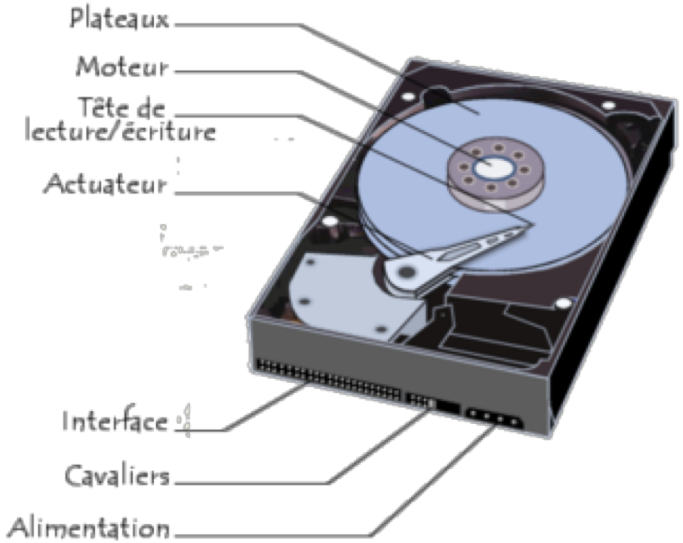
Cette année nous fêtons les 80 ans de l'informatique

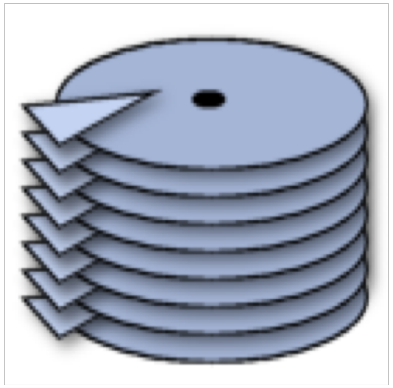
- 1934 : **Alan Turing**
- 1968 : Intel
- 1972 : Internet
- 1977 : **Oracle**
- 1992 : Internet = 1 million de PCs
- 1995 : **MySQL** / **PostGreSQL**
- 1996 : Internet = 36 millions de PCs
- 2000 : Internet = 360 millions de Pcs
- 2007 : Iphone
- *2015 : 2 milliards de smartphones*
- *2020 : 50 milliards d'objets connectés ?*

% de personnes se connectant à internet



Où vont les données ? Sur des disques !





La meilleure config sur Rue du Commerce : 60 To

➢Les interfaces de connexion au disque dur:

- IDE-ATA : 133 Mo/s (obsolète)
- SCSI : de 5 à 600 Mo/s (intelligent, plus rapide, standardisé)
- S-ATA : de 150 à 600 Mo/s (standard actuel)

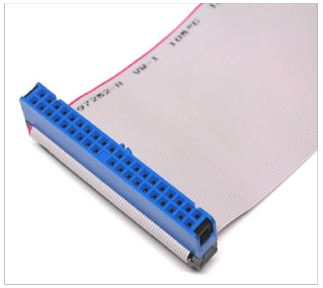
➢Carte mères : 500 eur

- 2 x SATA3 6.0 Gb / s
- 8 x SAS2/SATA3 6.0 Gb / s

➢Taille des disques : 6 To – 300 euros

➢Configuration : 60 To max / 3500 euros

=> Combien de temps pour tout lire ?



IDE-ATA

SCSI

S-ATA

IBM Benchmark (2011) : 480 disques !

Priced Storage Configuration:
24 – 8 Gbps dual port FC HBAs
IBM System Storage DS8870 2 –SMP processing clusters Each cluster contains: 8 – processor cores 128 GB – processor memory *(256 GB total)*
16 – 8 Gb, 4 port SW FCP/FICON adapter pairs *(128 host port front-end connections, 32 used)*
8 – 4 port, 8 Gb FC-AL device adapter pairs *(4 adapter pair/cluster)* *(64 backend connections, 64 used)*
1 – Management Console *(internal laptop)*
1 – DS8870 Expansion Unit
10 – Disk Enclosure pairs *(48 disk drives per enclosure pair)*
480 – 146 GB, 15K RPM, 2.5” disk drives

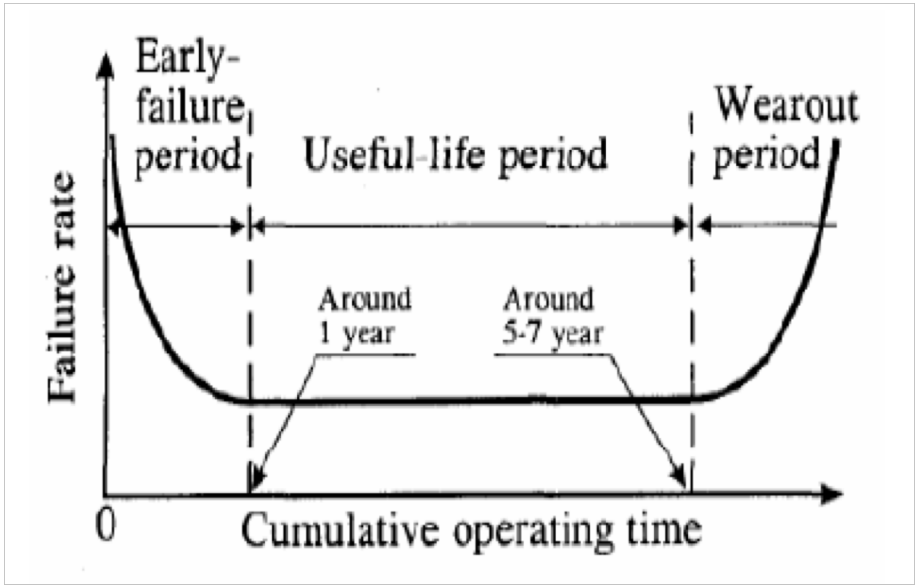
480 disques de 146 Gb : 71 Tb

15 K RPM
14 Gb / s

=> Combien de temps pour tout lire ?

=> A quel prix ?

MTBF : Mean Time Before Failure

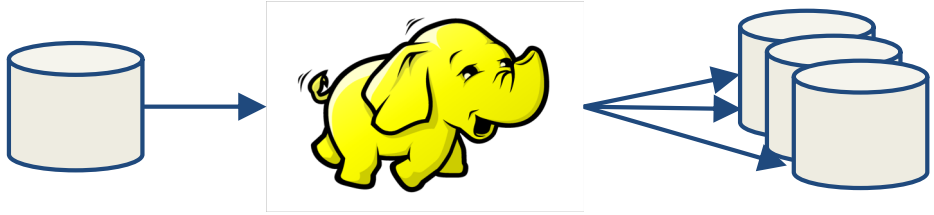


100 disques | MTBF 5 ans = 20 pannes / an
Soit un disque à changer toutes les 2 semaines.

Risques :

1. Coût
2. Instabilité du système
3. Perte d'information !

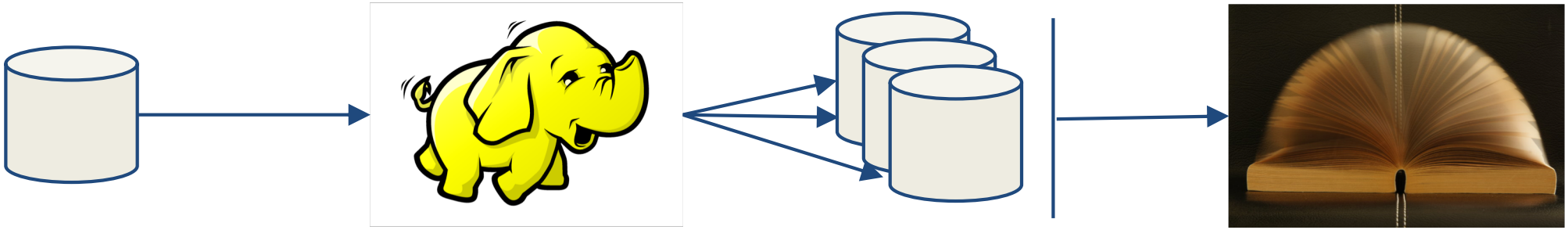
Hadoop résout le MTBF en déduplicant la donnée :
replication factor = 3



Si un disque tombe en panne, on retrouve l'information sur un des deux autres disques.

MTBF : Transformer un problème en atout

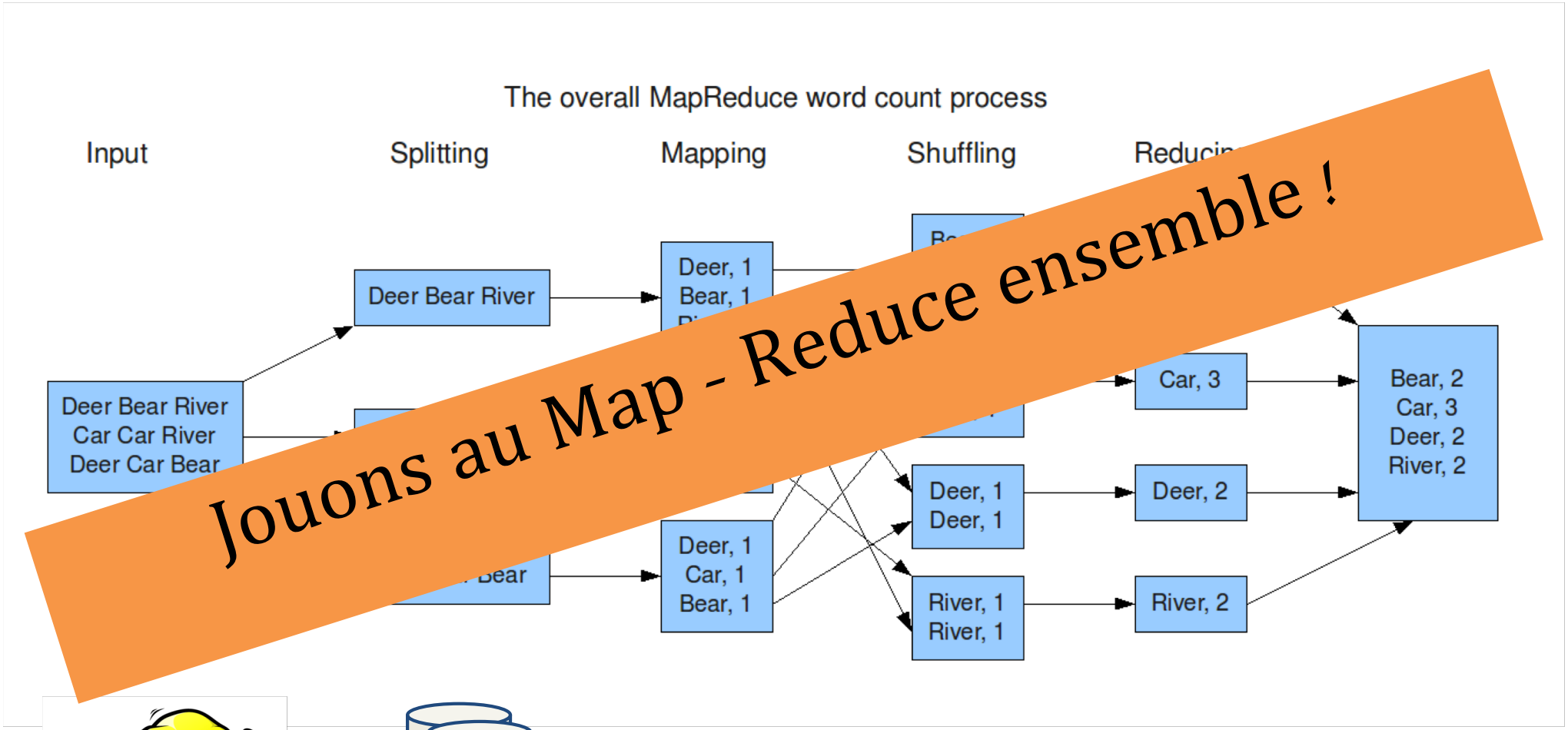
Lent (dédié au mode “batch”)



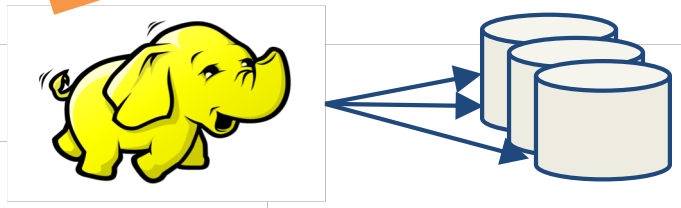
Pas cher
(12 Keur/machine)
mais redondant !

Rapide
Map (Sort) Reduce
mais complexe !

Map-Reduce Algorihtm



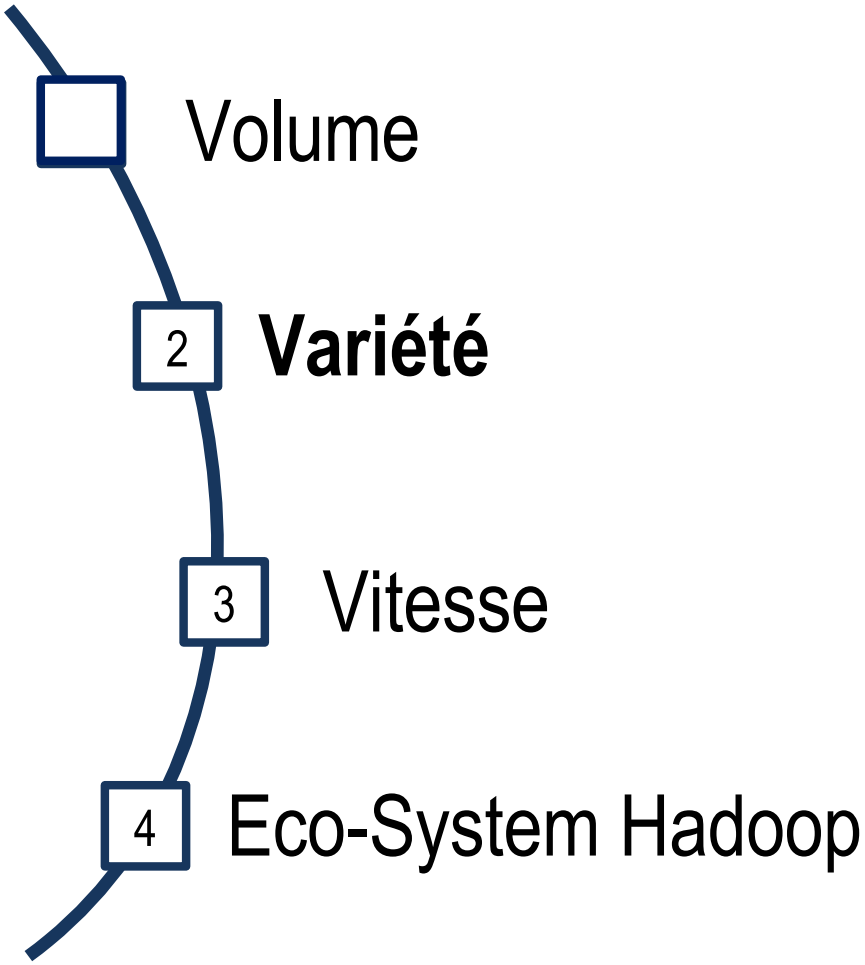
Jouons au Map - Reduce ensemble !



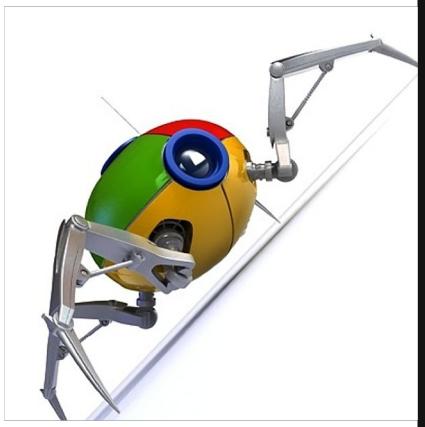
MAP (MAGIC SORT) REDUCE

(RESULT)

Cheminement du cours

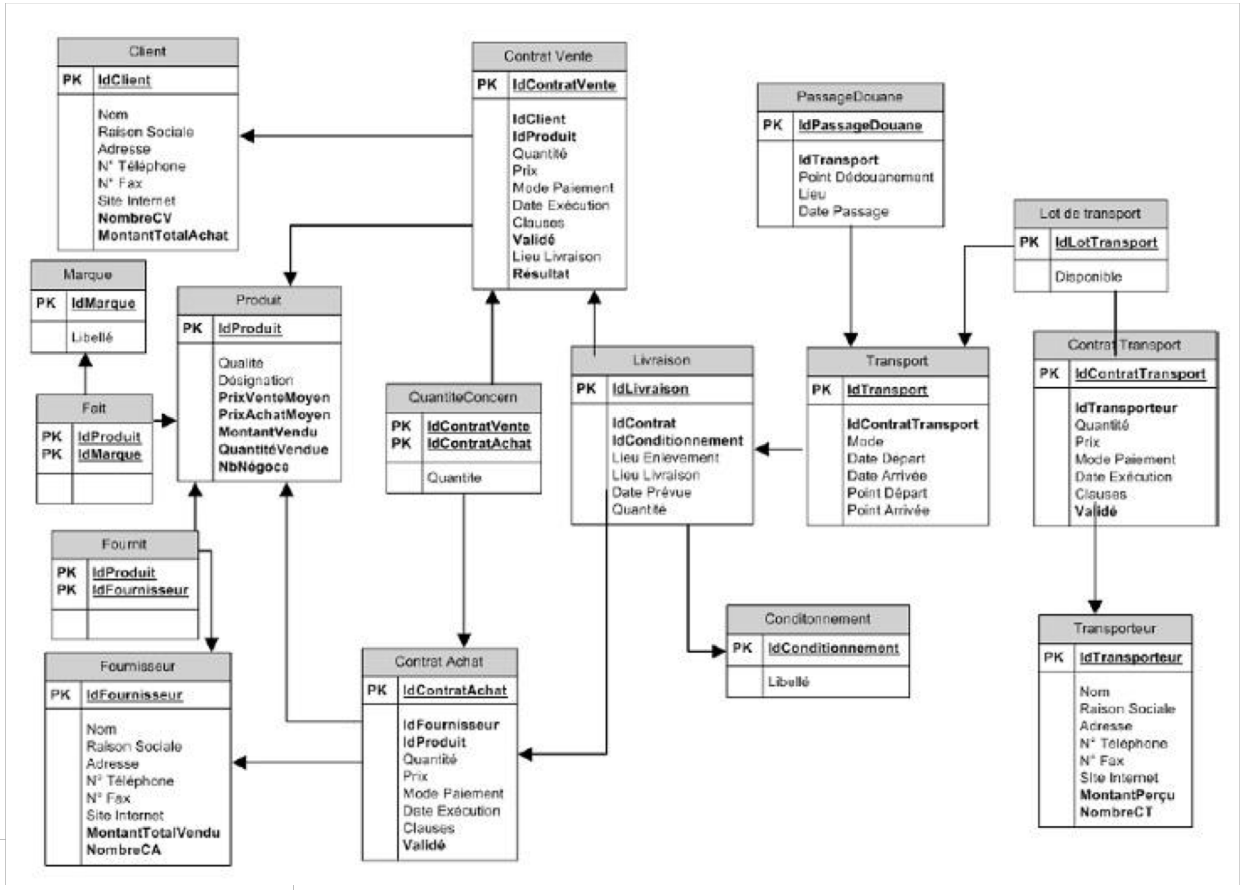


Des sources de plus en plus diverses

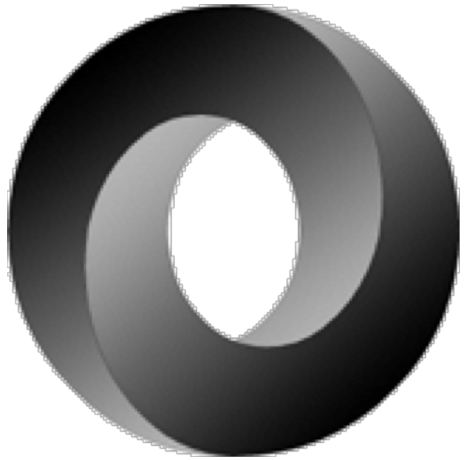


La fin d'un monde

1950 - 2009

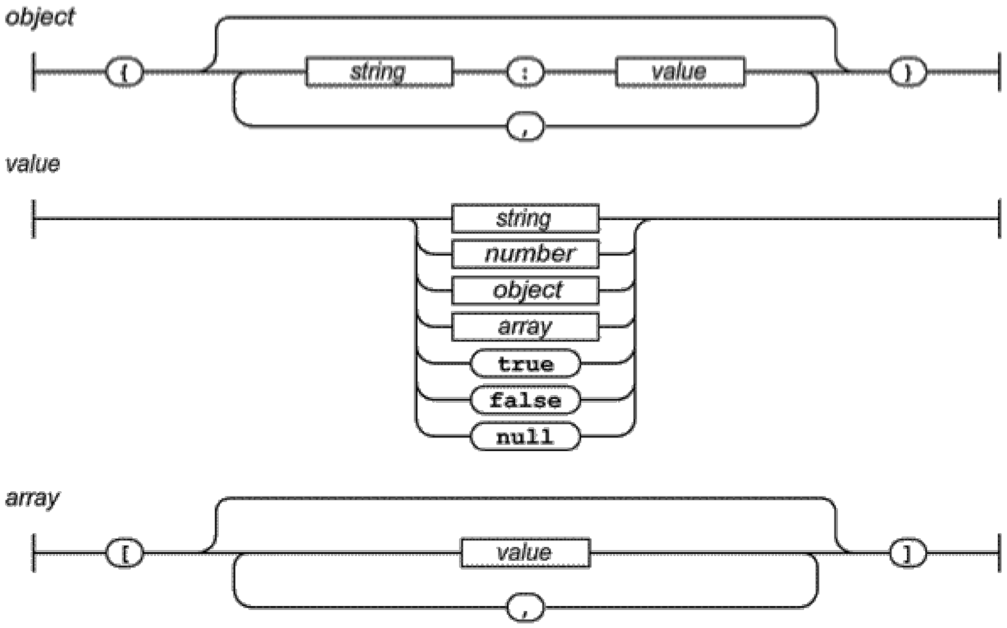


2015



JSON

Formalisme JSON :

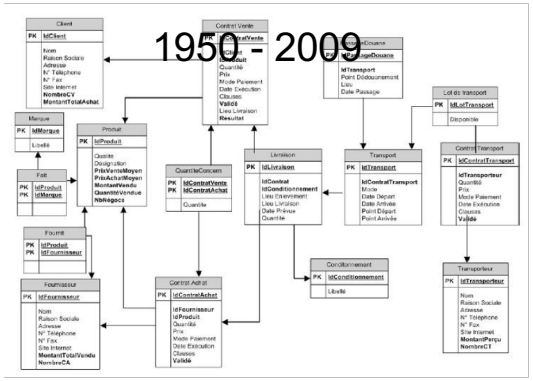


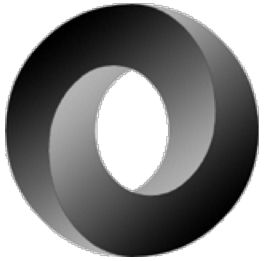

```
{
  "arguments" : { "number" : 10 },
  "url" : "http://localhost:8080/restty-tester/collection",
  "method" : "POST",
  "header" : {
    "Content-Type" : "application/json"
  },
  "body" : [
    {
      "id" : 0,
      "name" : "name 0",
      "description" : "description 0"
    },
    {
      "id" : 1,
      "name" : "name 1",
      "description" : "description 1"
    }
  ],
  "output" : "json"
}
```

Des caractéristiques différentes ...

1950 - 2009

2015





Online Transaction Processing

Transactions garanties
Lecture et écriture
Schéma défini

Banques / Systèmes de sécurité

Oracle / Access
PostGreSQL / MySQL

Online Analytical Processing

Pas de transactions
Principalement en lecture
Sans schéma

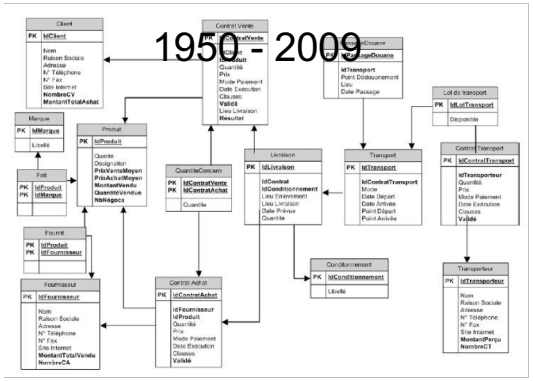
Sites Web / Applications non critiques

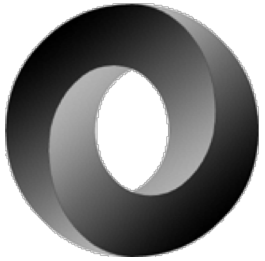
MongoDB / CouchDB
HBase / Cassandra

Pour de nouveaux usages :

1950 - 2009

2015





Online Transaction Processing

Transactions garanties
Lecture et écriture
Schéma défini

Banques / Systèmes de sécurité

Oracle / Access
PostGreSQL / MySQL

Online Analytical Processing

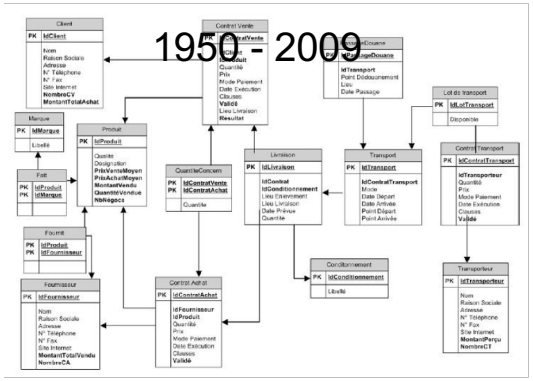
Pas de transactions
Principalement en lecture
Sans schéma

Sites Web / Applications non critiques

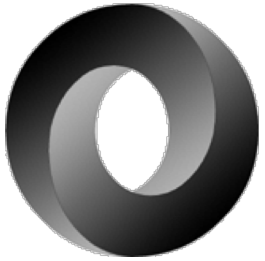
MongoDB / CouchDB
HBase / Cassandra

Un vocabulaire qui change :

1950 - 2009



2015



Online Transaction Processing

MCD / Schéma / Relationnel
UML / MERISE

Référentiel / dictionnaires de données
SQL Triggers

Silos / Logiciels / Licences / BI

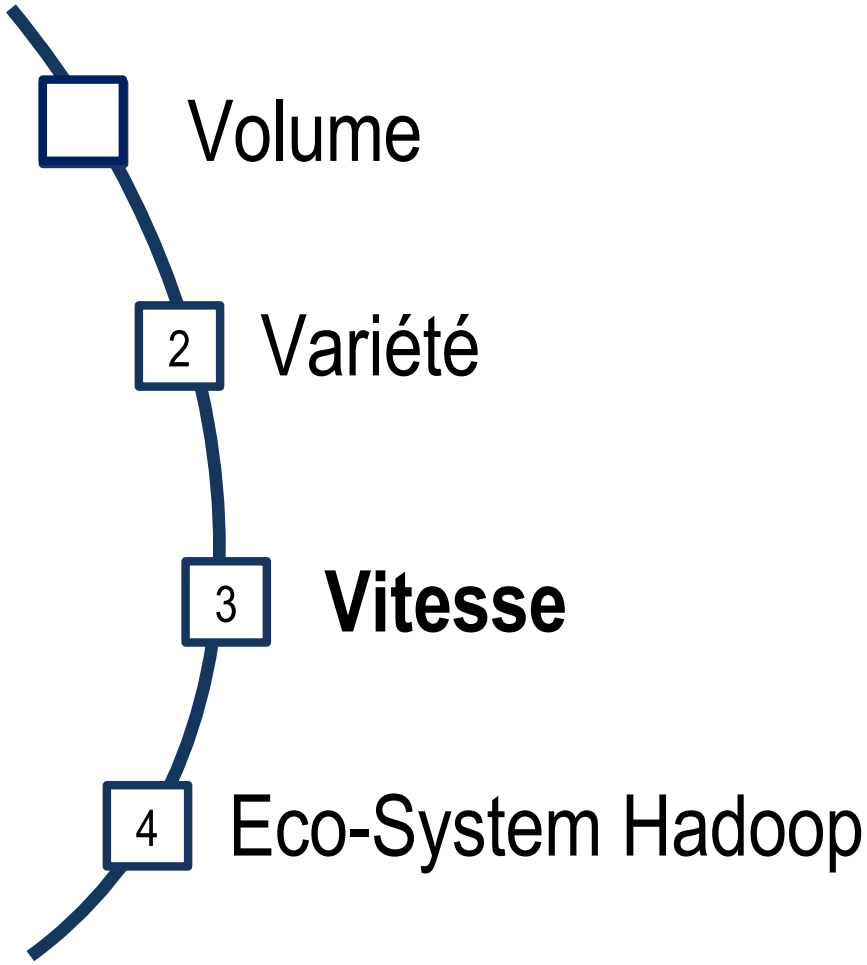
Online Analytical Processing

NoSQL / Schemaless

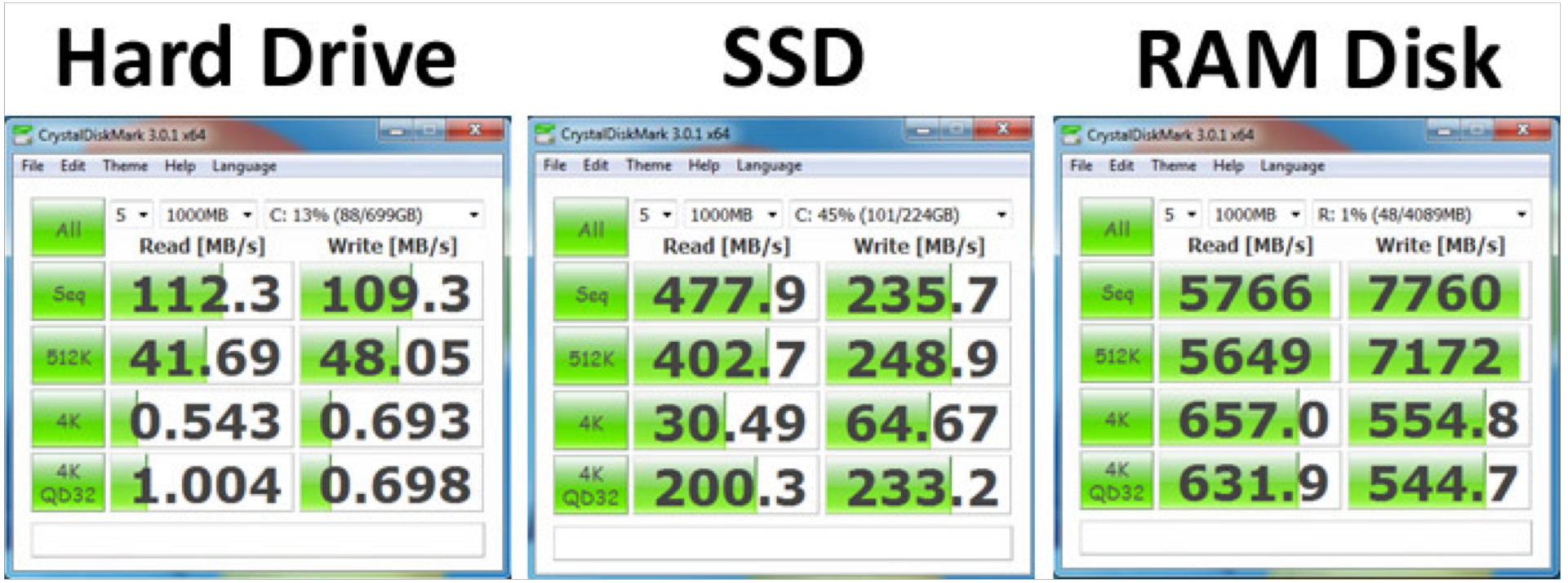
Machine Learning / Prédictif
Apprentissage (Non) Supervisé

Partage / API / Open Sources /
Dashboard / Data Visualisation

Cheminement du cours

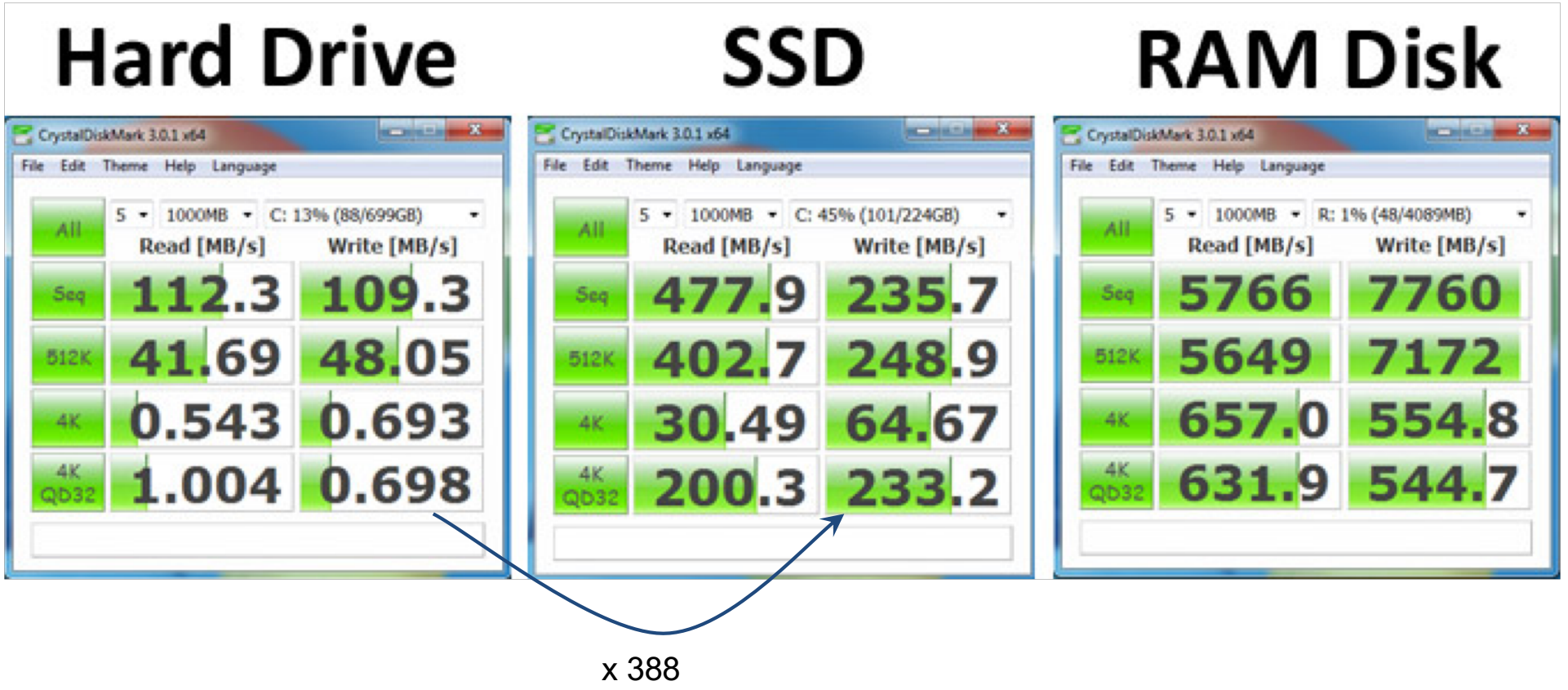


Support matters

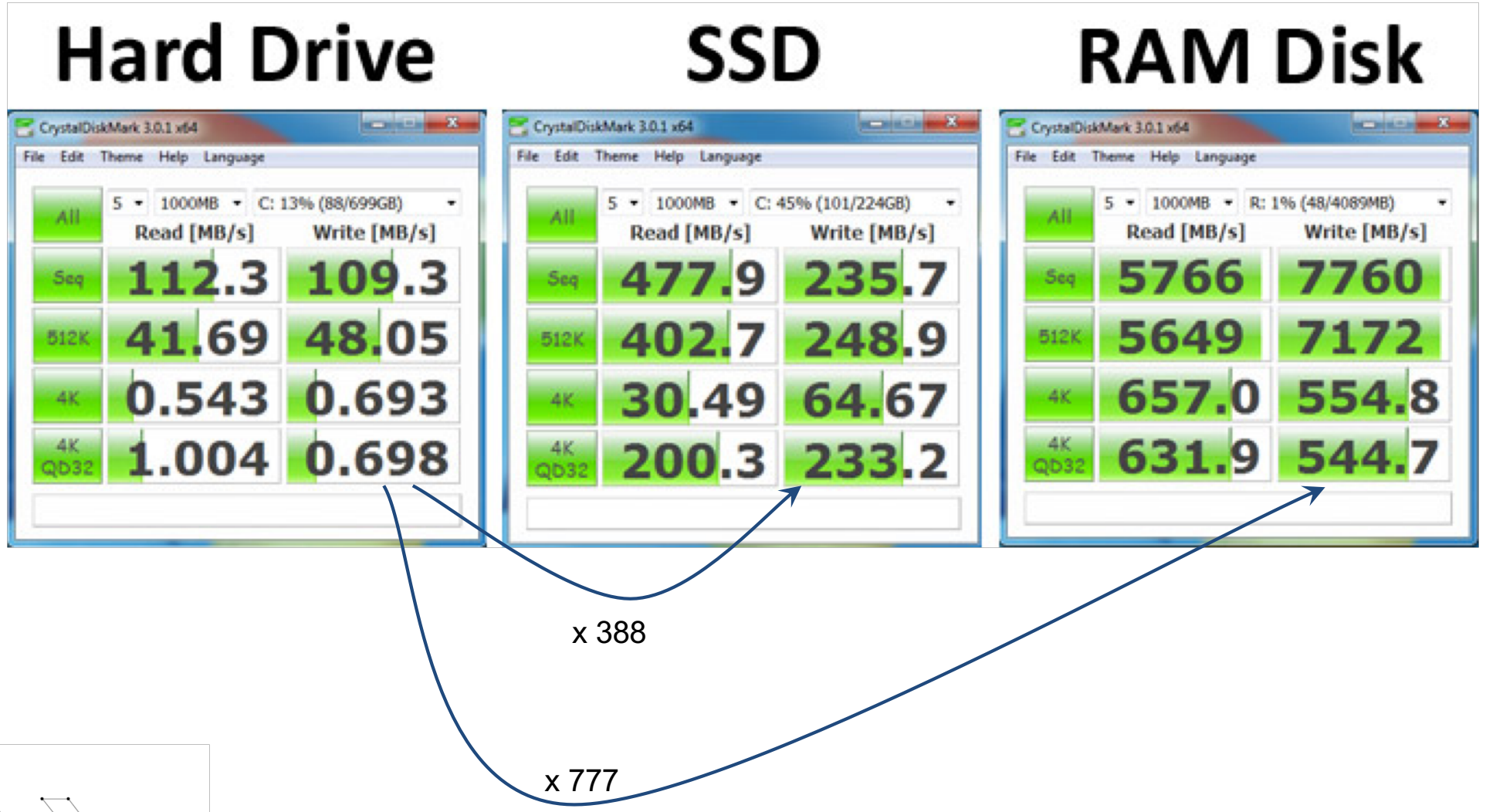


La RAM est jusqu'à 70 fois plus rapide que le disque en écriture
et 50 fois en lecture

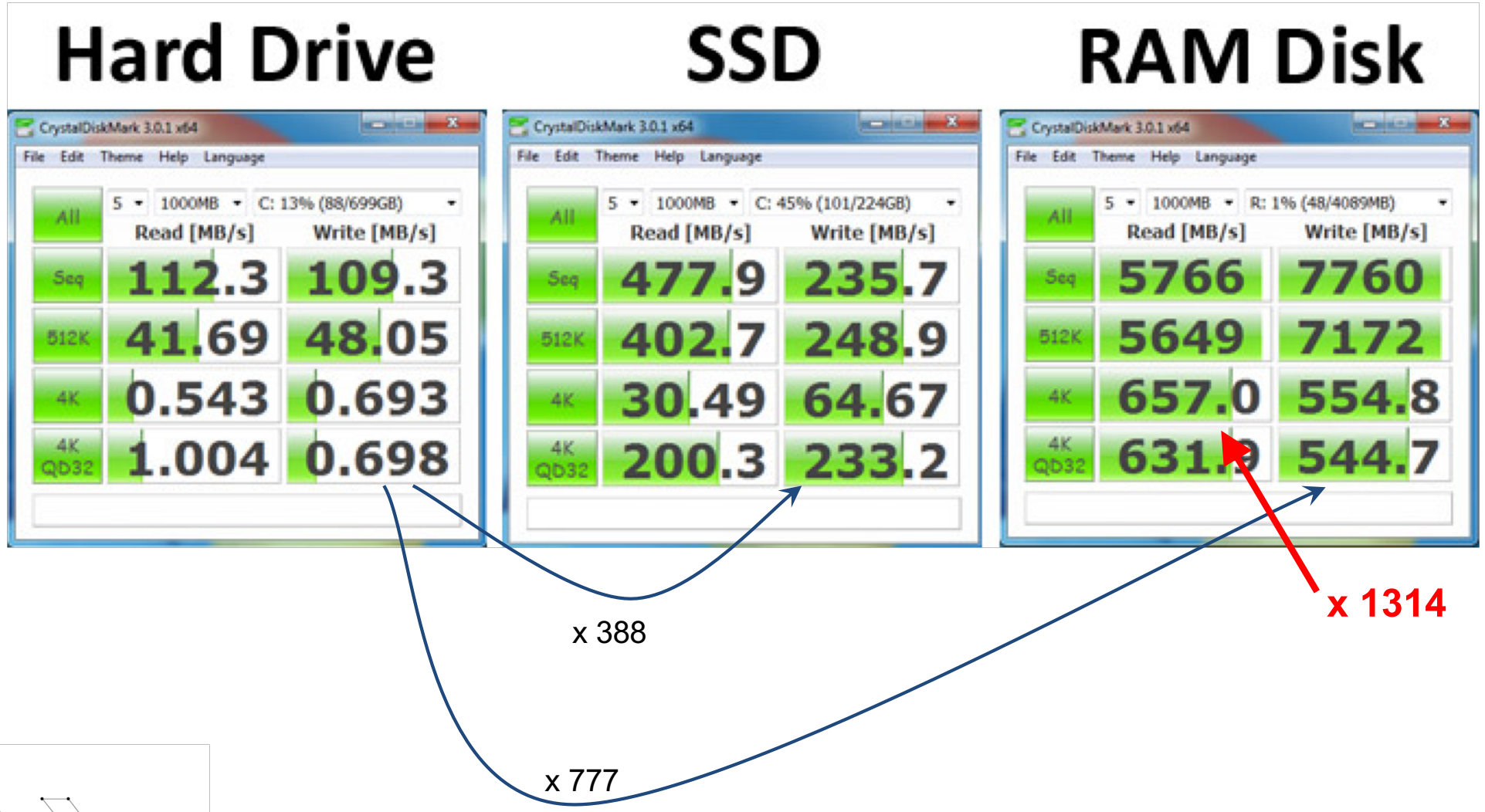
Support matters



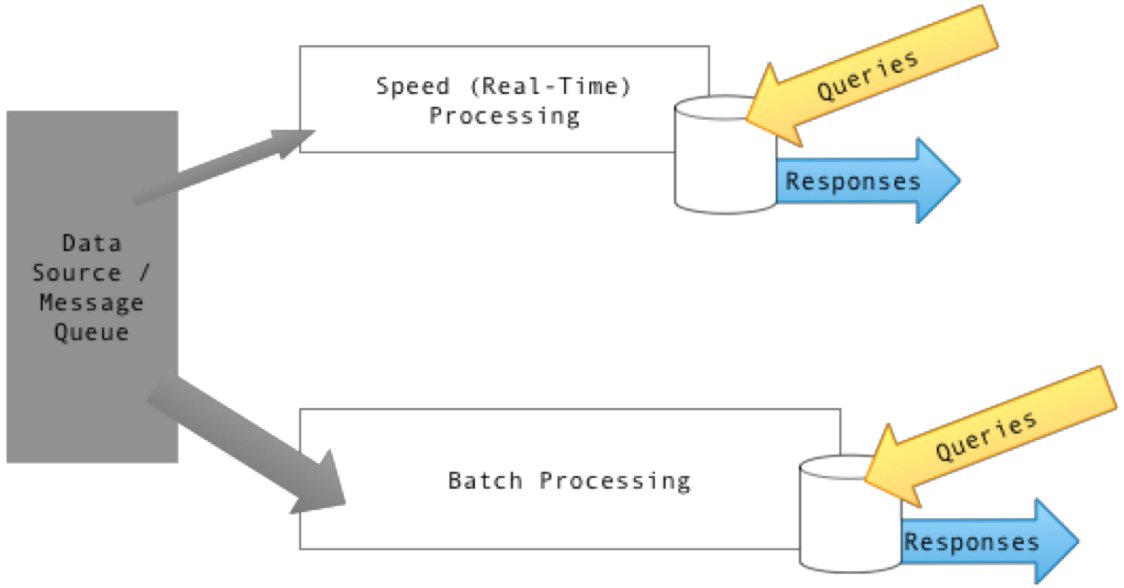
Support matters



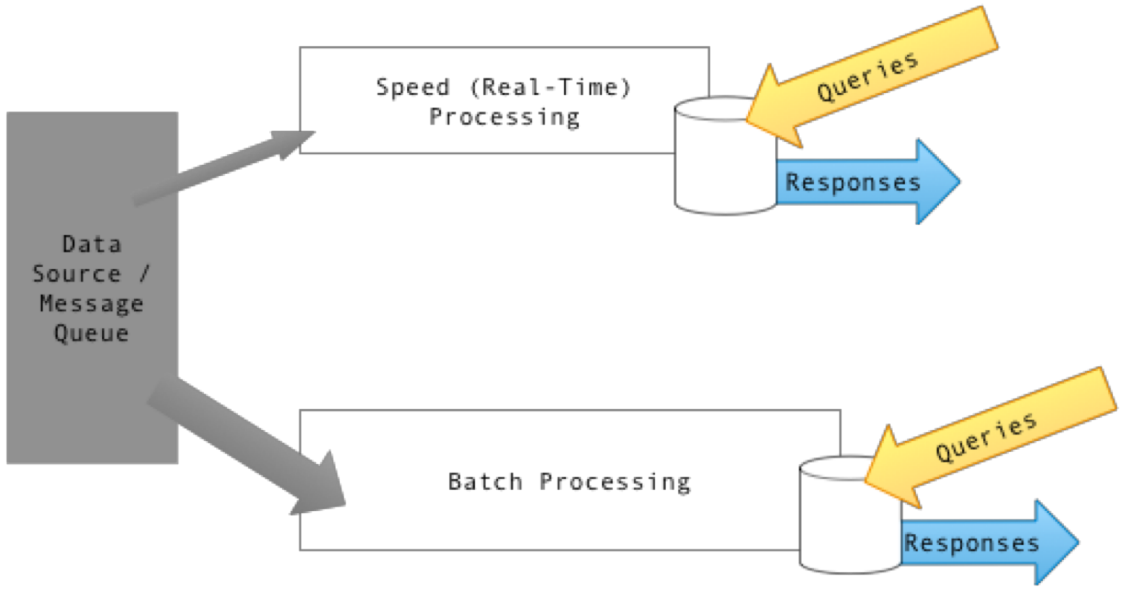
Support matters

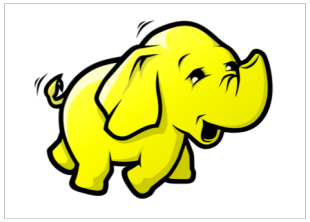


Lambda architecture

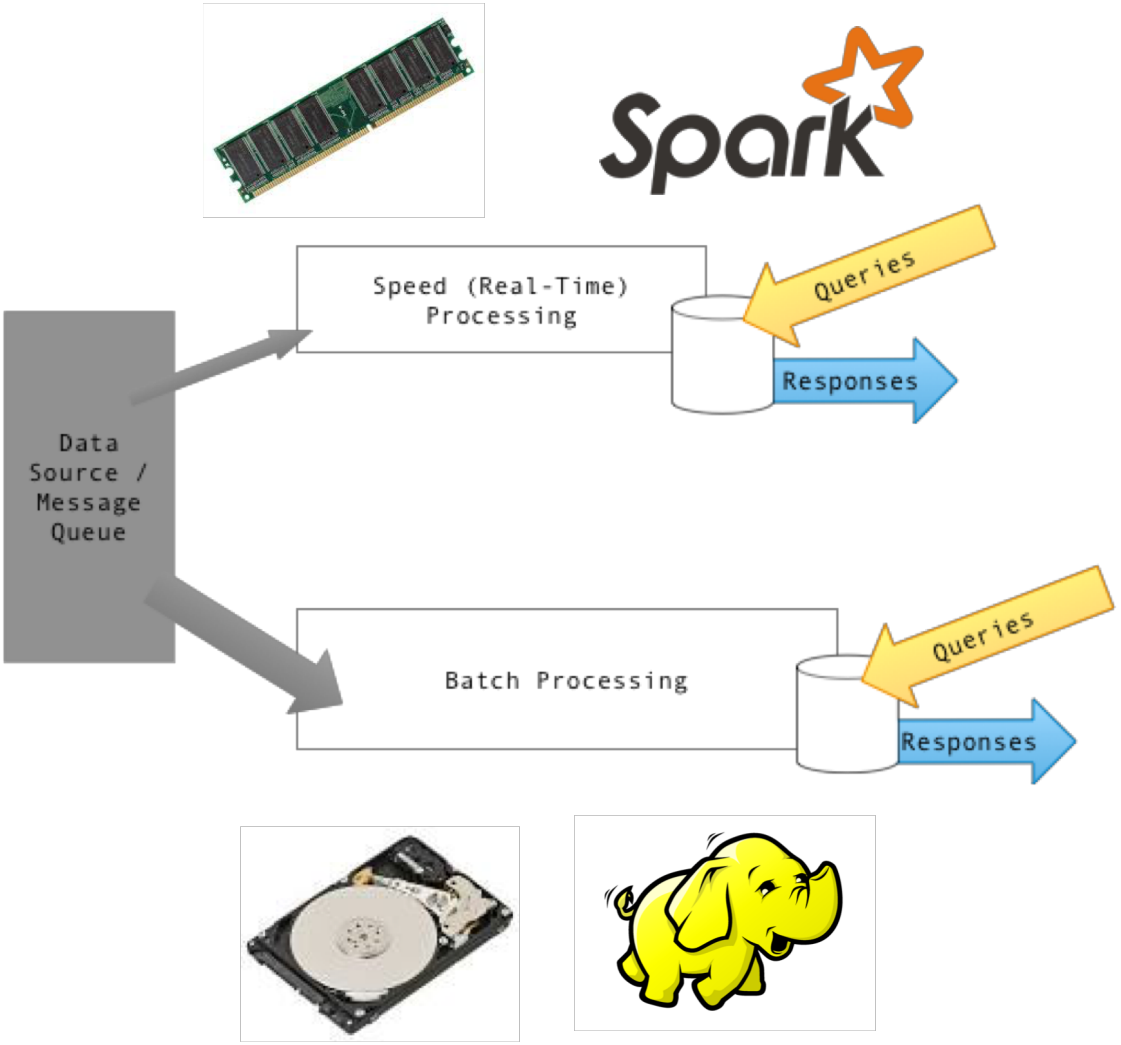


Lambda architecture

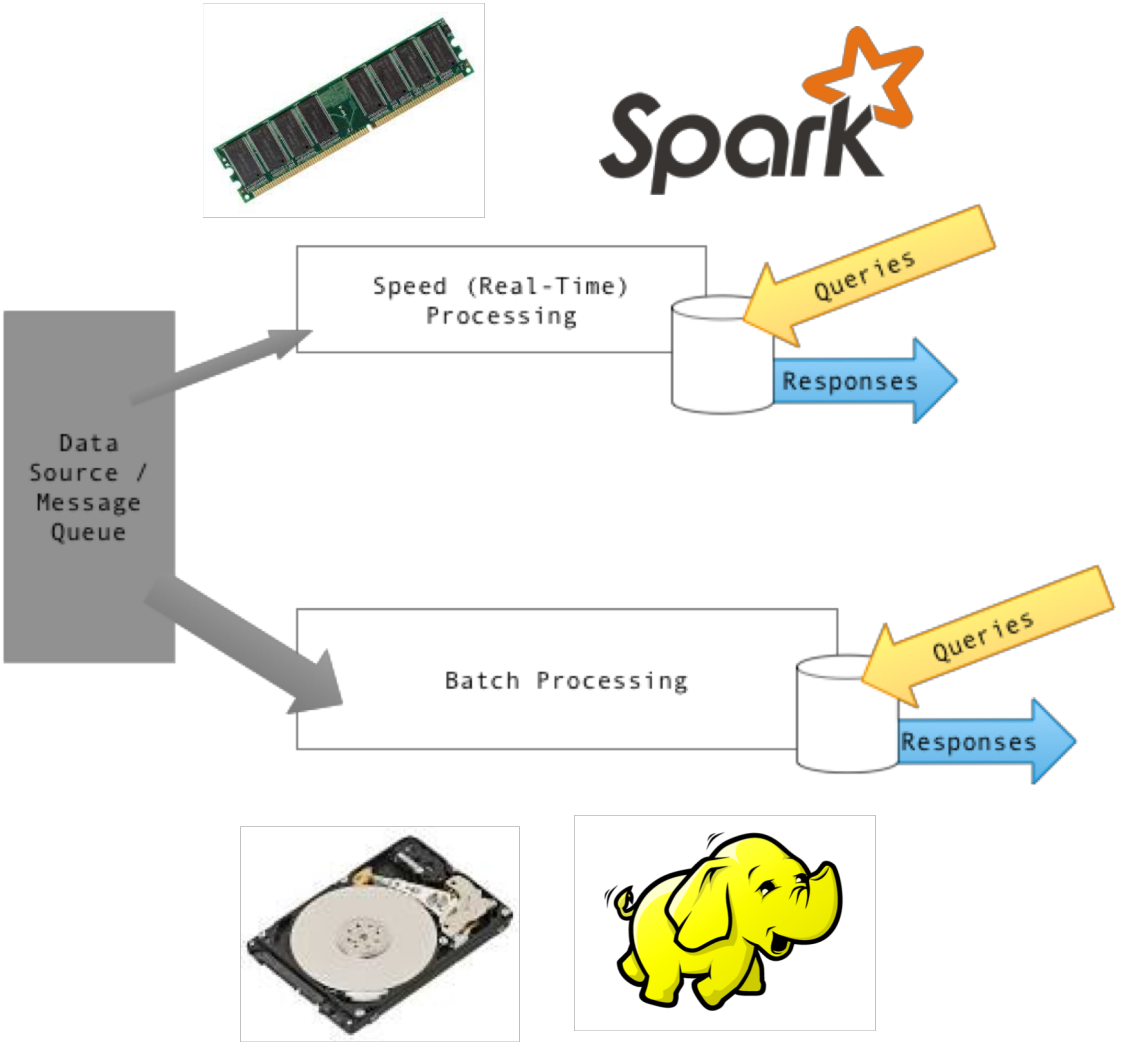


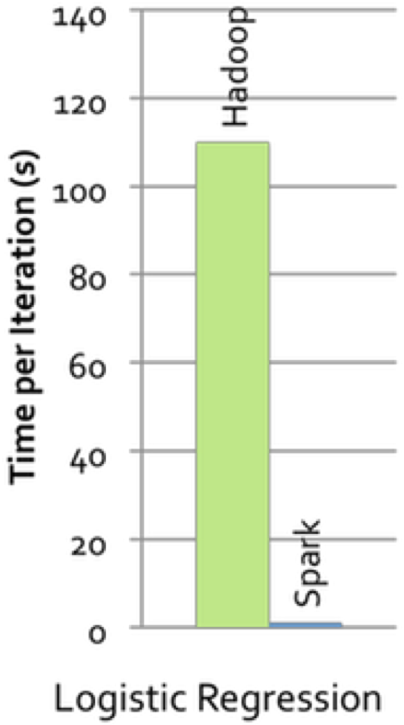


Lambda architecture

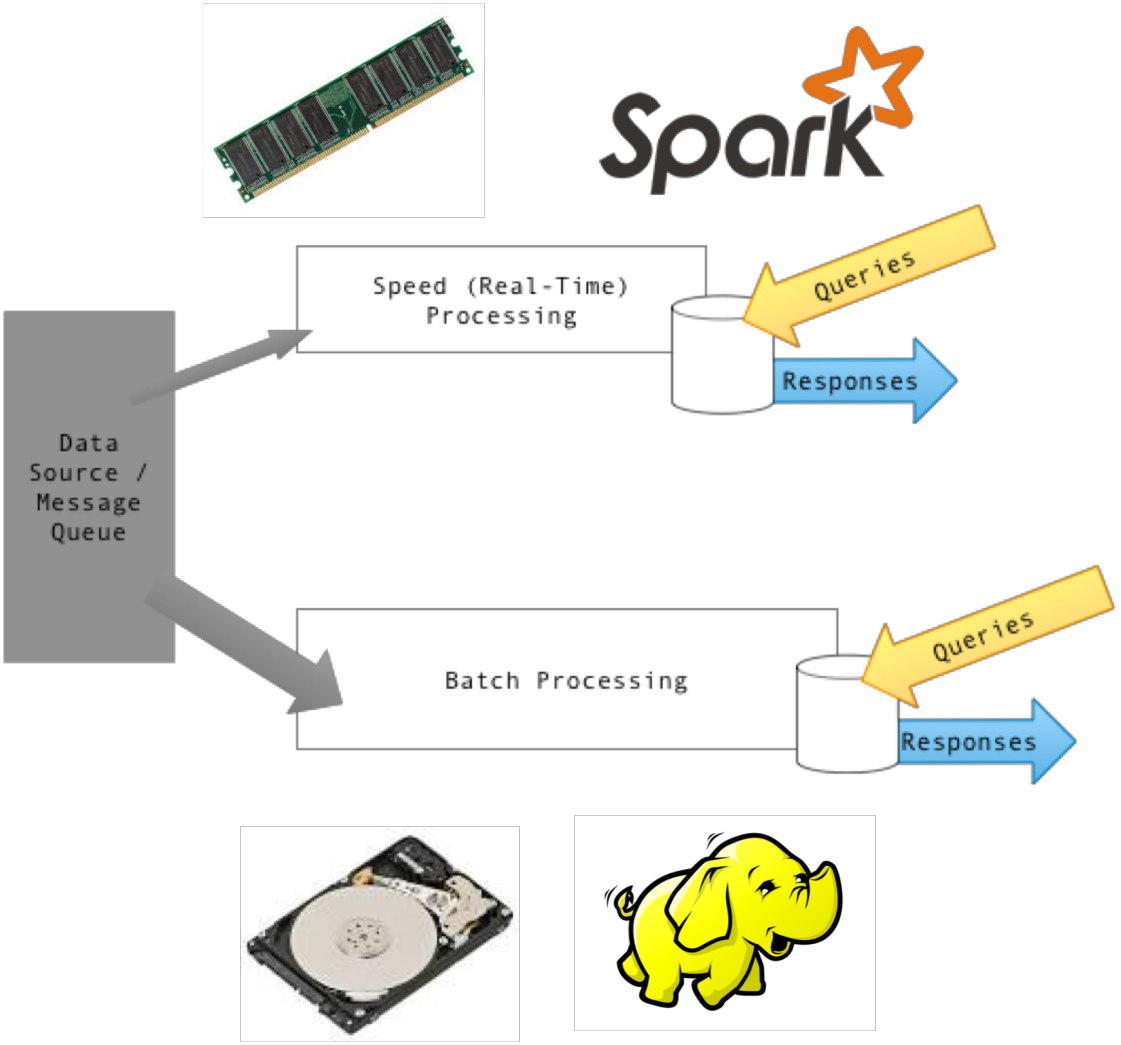


Lambda architecture



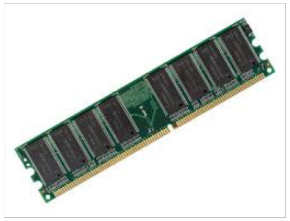


Lambda architecture



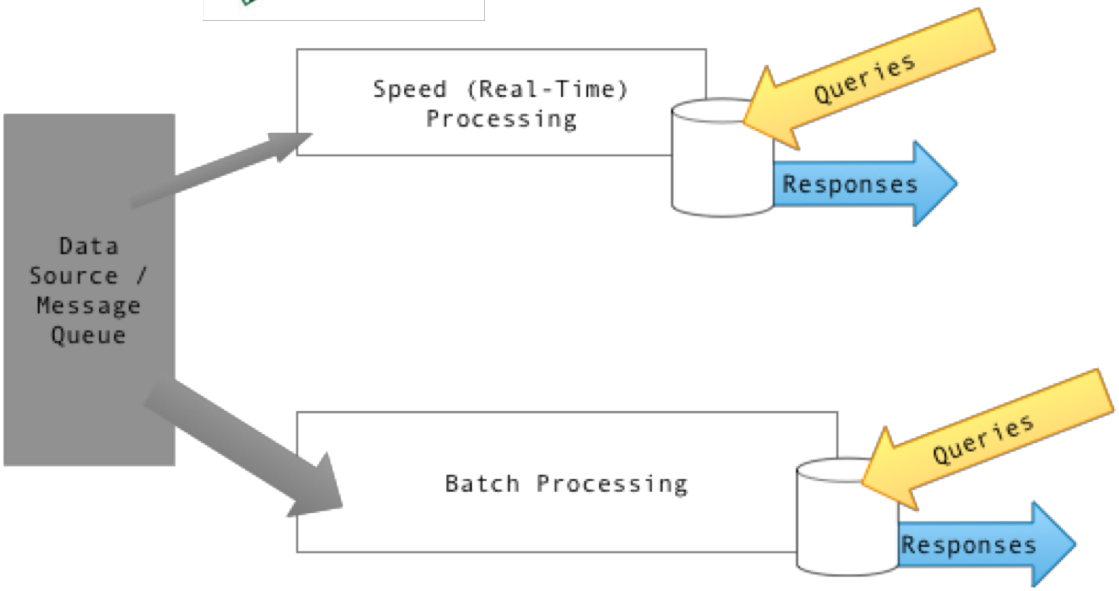
20 € / Tb

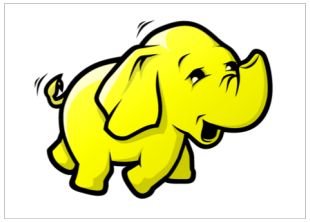
Lambda architecture



Spark

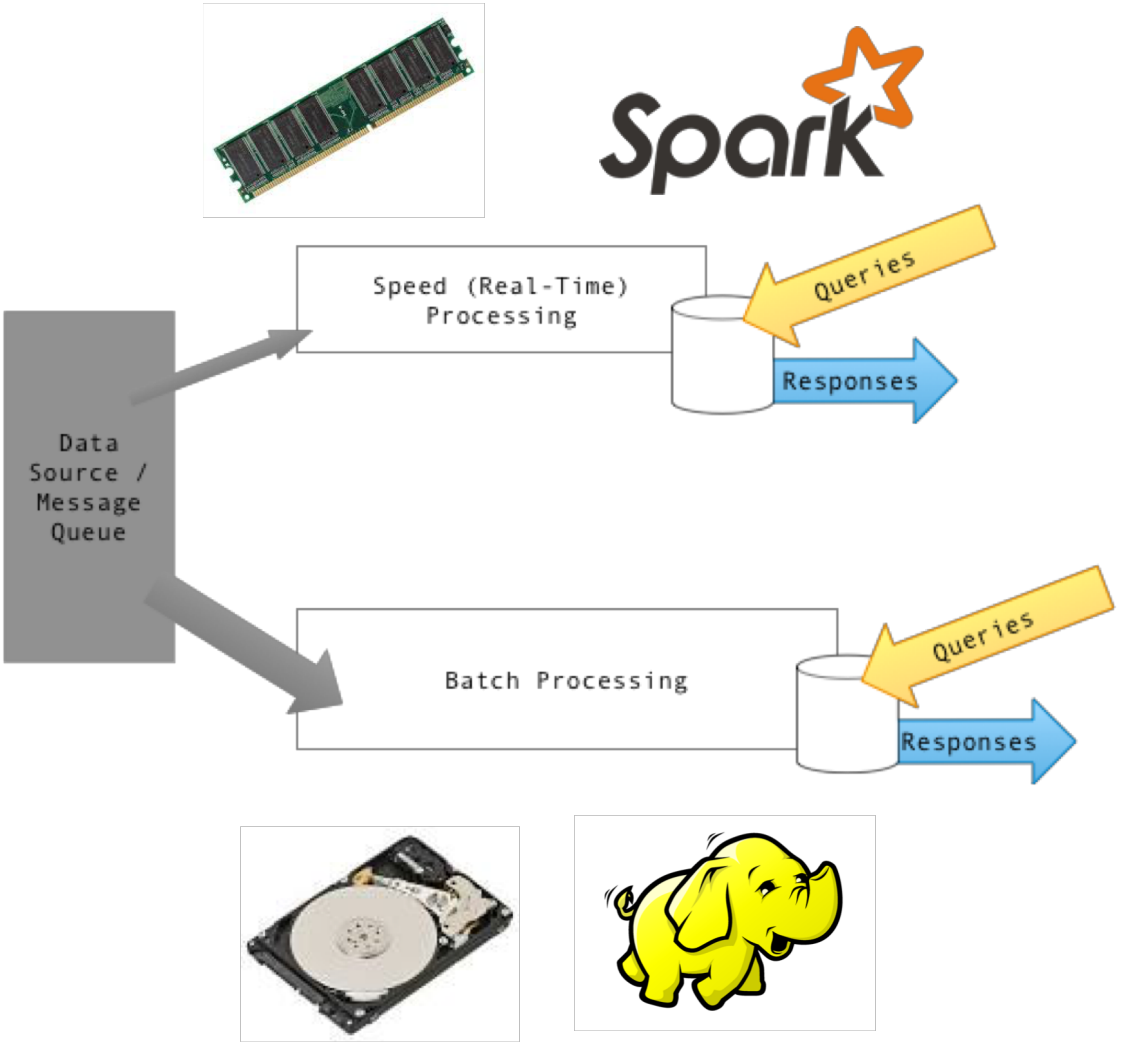
9 000 € / Tb



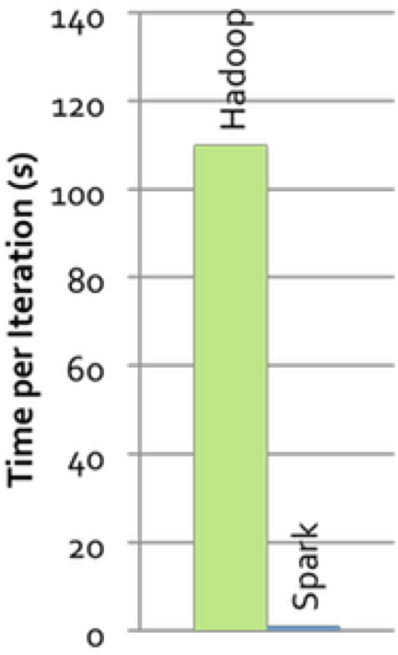


20 € / Tb

Lambda architecture

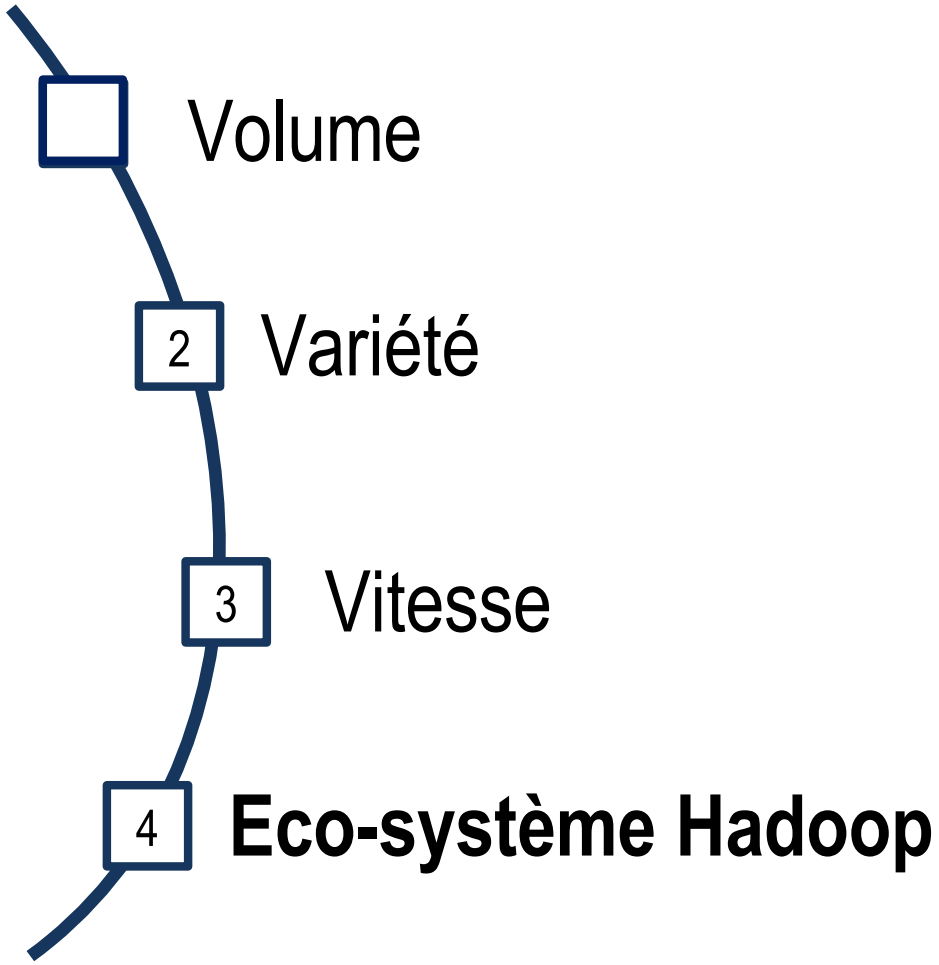


9 000 € / Tb

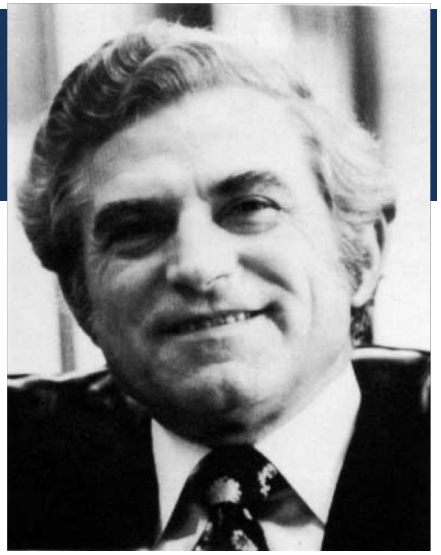


20 € / Tb

Cheminement du cours



Coordination complexe dans les Systèmes Parallèles



Loi d'Amdhal :

Soit un programme :

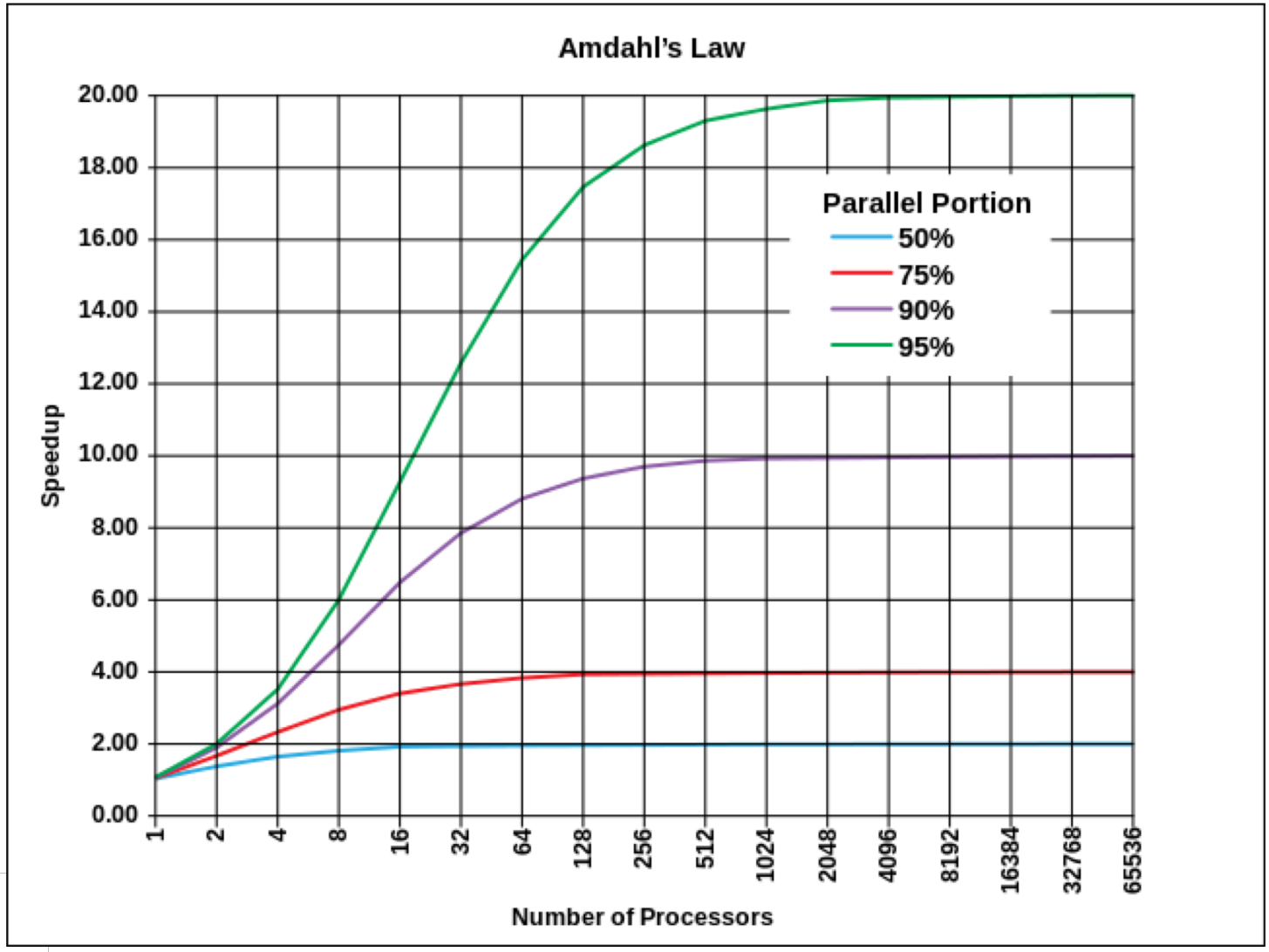
- P = % possible en parallèle sans synchronisation (dans [0, 1[ )
- N = nombre de processeurs
- 1 = durée nécessaire pour effectuer l'algorithme avec un processeur

Gain possible en temps :

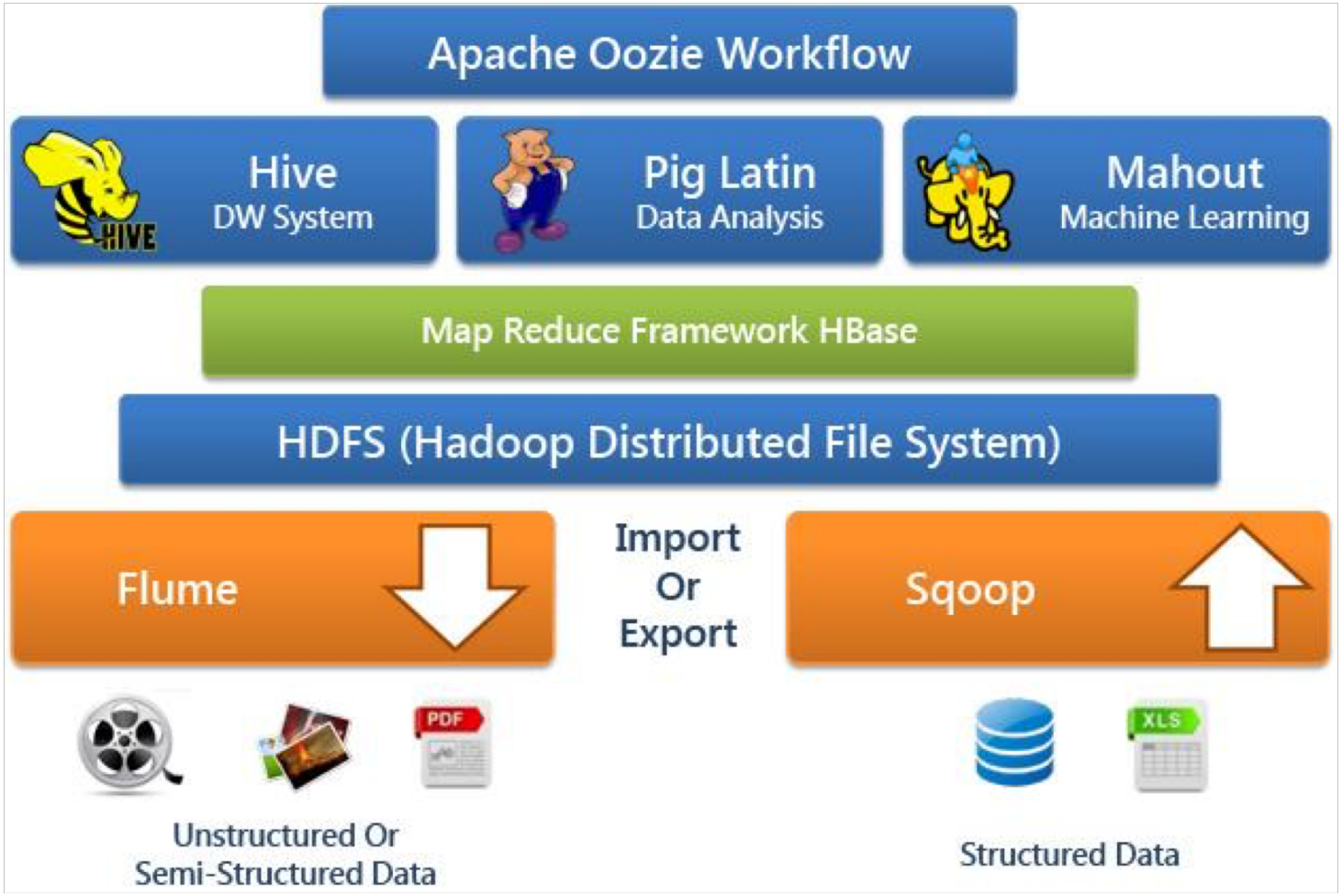
- Gain sur P = P/N < P < 1
- Partie sur laquelle on ne peut rien gagner : 1 – P < 1
- Durée nécessaire : (1-P) + P/N <1
- Accélération possible : 1 / Durée nécessaire > 1

- Exemples :
  - 95% parallélisable : P = 0.95
  - 100 processeurs : N = 100
  - Accélération = 1 / (0.05 + 0.95/100) = 16,8 fois plus rapide
  - Avec 50 processeurs : 1 / (0.05 + 0.95/10) = 14,5 fois plus rapide

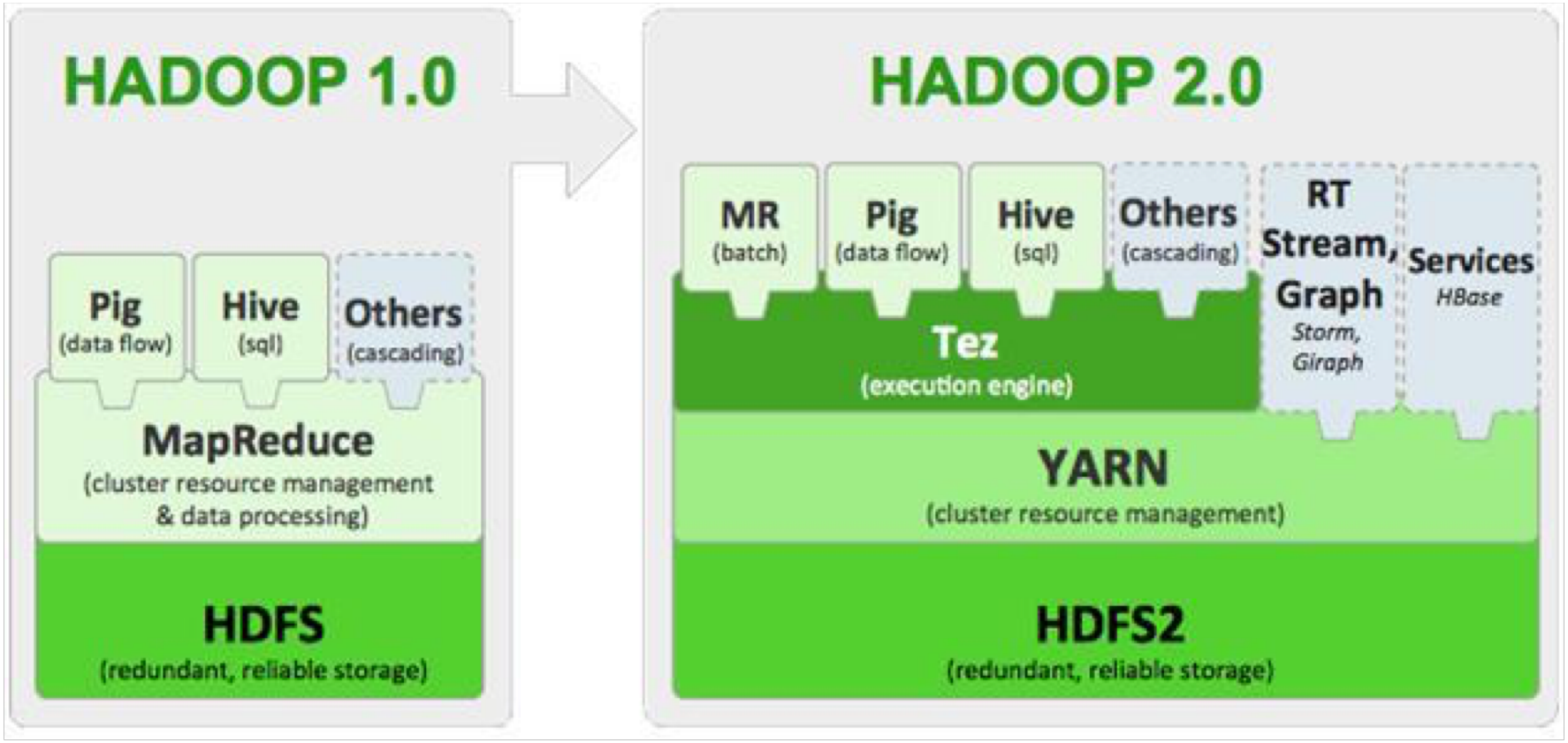
Rendez vos codes Parralélisables !



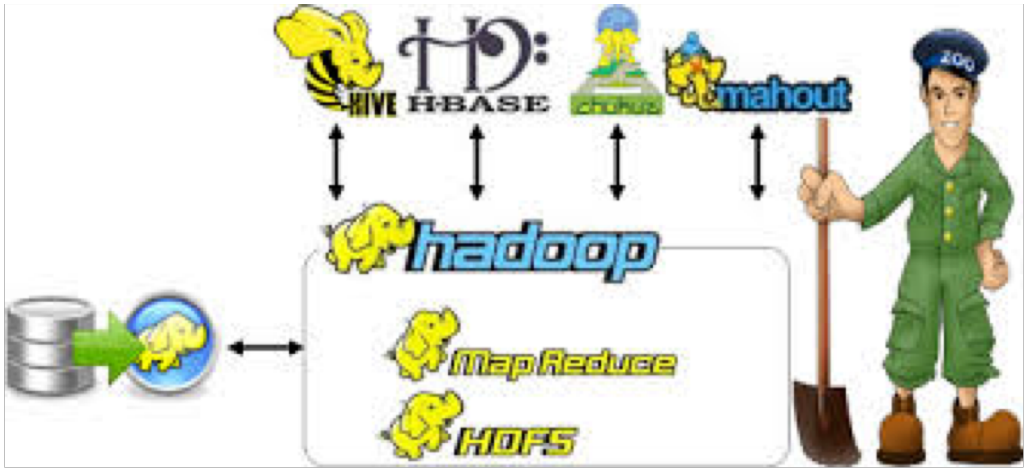
HADOOP 1.0



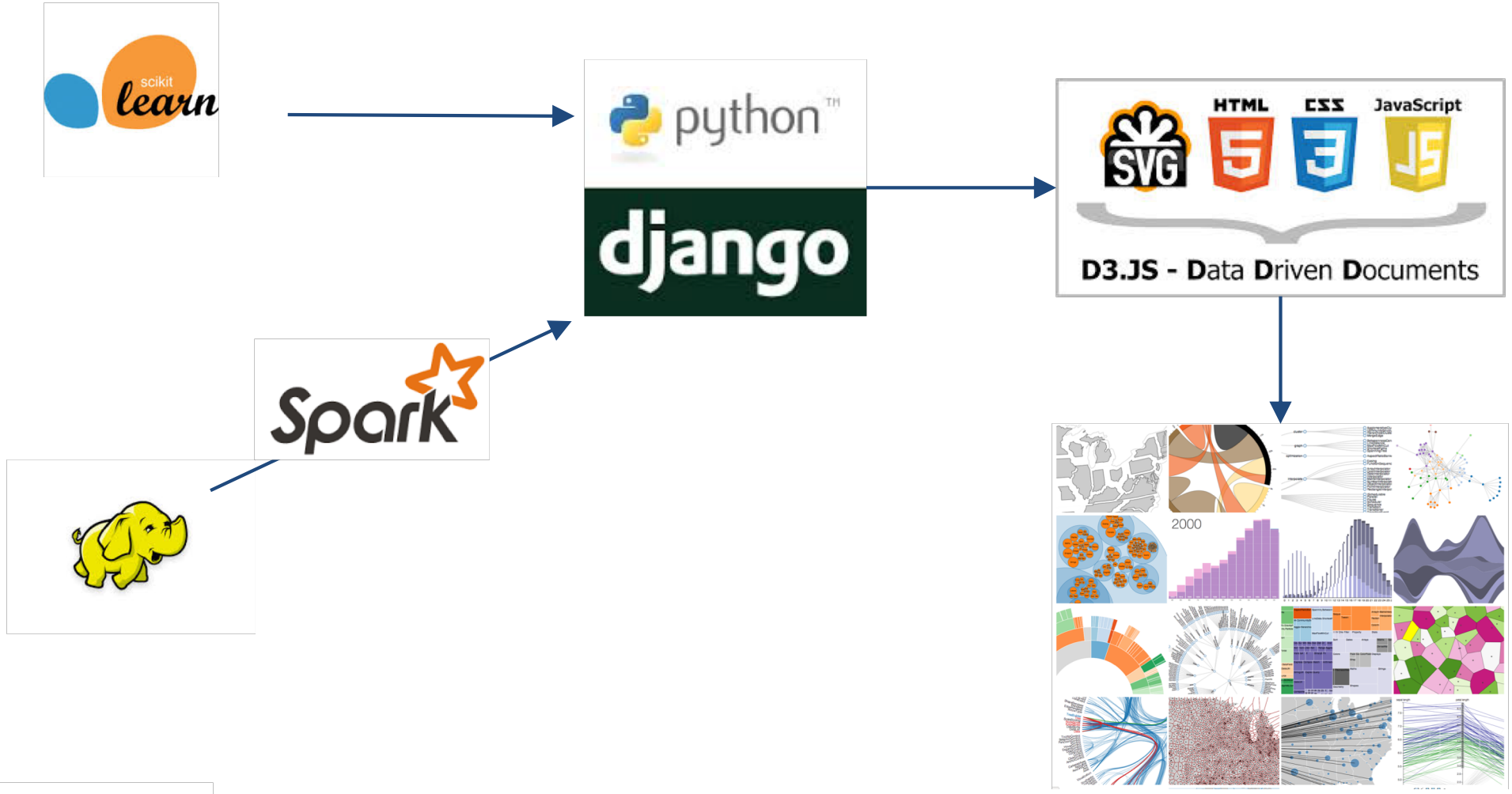
HADOOP 2.0



Zookeeper



Front - End



D3JS

Data-Driven Documents

D3.js is a JavaScript library for manipulating documents based on data. **D3** helps you bring data to life using HTML, SVG, and CSS. D3's emphasis on web standards gives you the full capabilities of modern browsers without tying yourself to a proprietary framework, combining powerful visualization components and a data-driven approach to DOM manipulation.

Présentation de Memorandum.pro

Romain Jouin

- Fondateur cabinet Mémorandum
- Alcatel-Lucent, Toshiba Services
- ESCP, Télécom Paris

COMPÉTENCES

- Développement informatique
- Développement commercial
- Stratégie Big Data
- Analyse de données

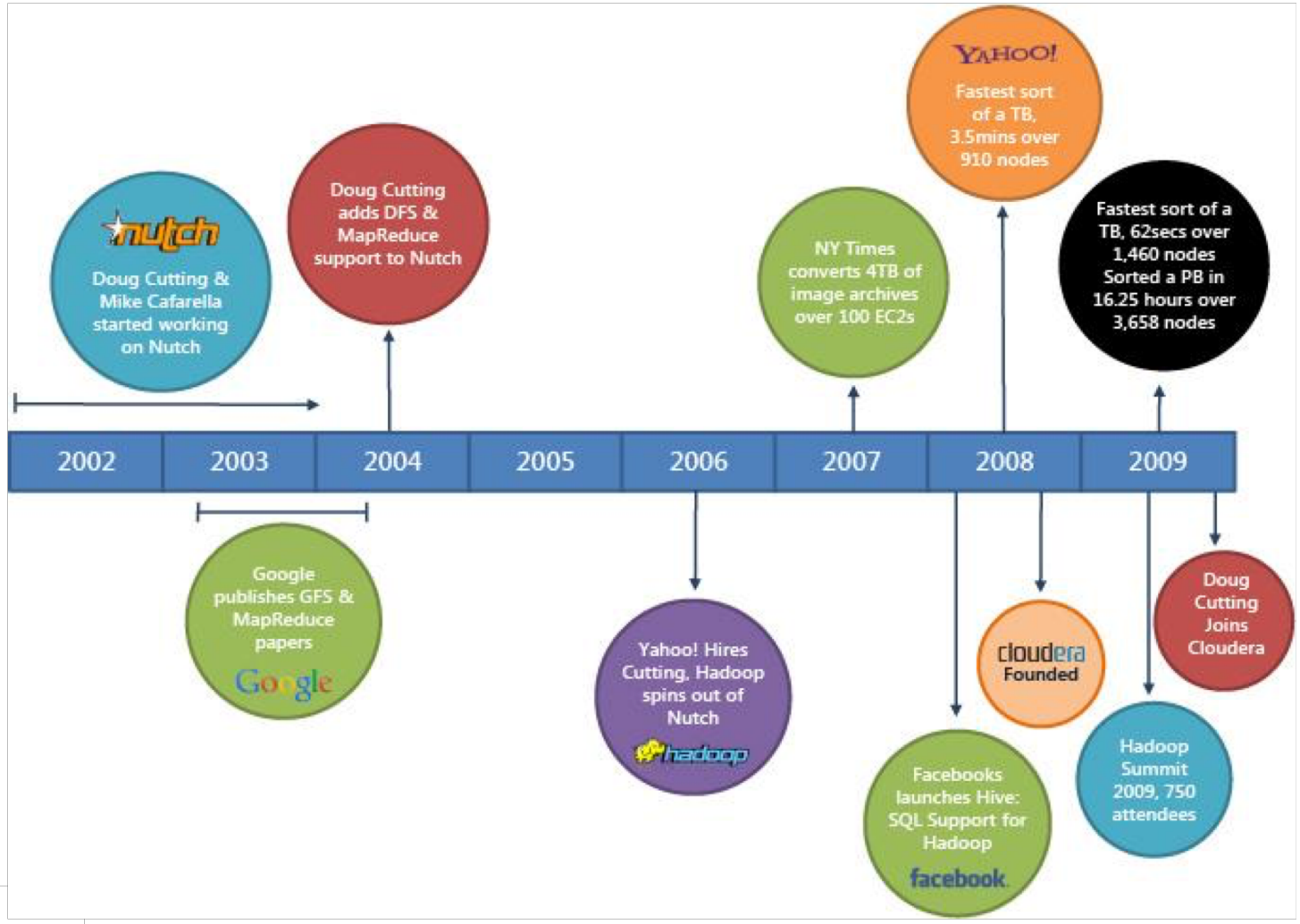
SELECTION DE PROJETS RECENTS

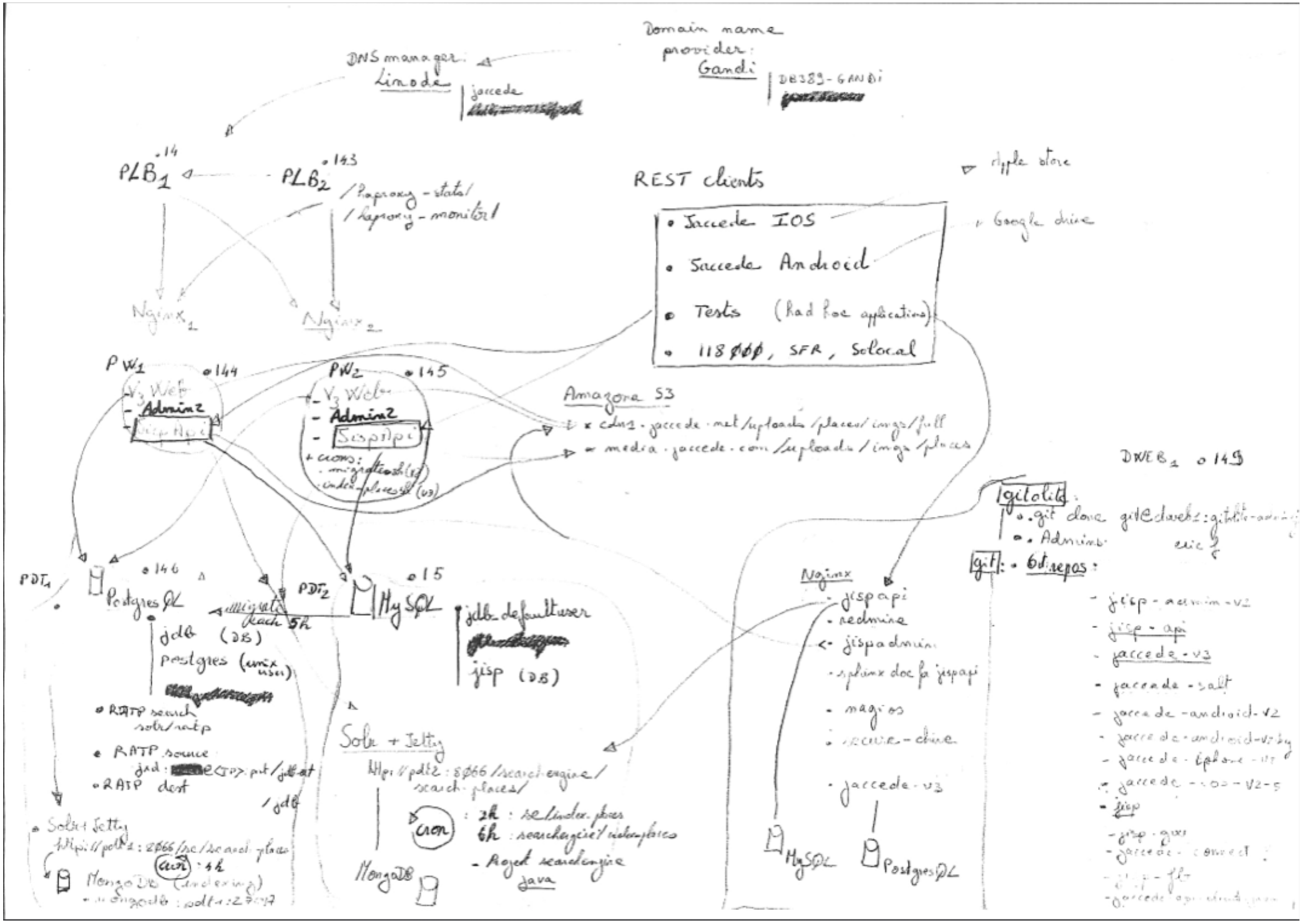
- Jaccede.com – gestion d'infrastructure
- EDF – E-reputation
- Toshiba –1ère plateforme de Cloud Computing
- Alcatel – Développement commercial Ex-URSS

MÉMORANDUM

- **Cabinet de conseil en Stratégie Big Data**
- **Expertise Usages et Applications Big Data**
- **Missions de conseil et formations**
- **Développement logiciel**
- **Gestion d'industrialisation Big Data**

History




Domain name provider: Gandi
DB389-GANDI
DNS manager: Linode
jaccede
PLB1 .14
PLB2 .143
/haproxy-stats/
/haproxy-monitor/
Nginx1
Nginx2
REST clients
• Jaccede IOS
• Saccede Android
• Tests (Rad for applications)
• 118 000, SFR, Solocal
Apple store
Google drive
PW1 .144
V3 Web
- Admin2
- Sisp Api
PW2 .145
V3 Web
- Admin2
- Sisp api
+ crons:
Amazone S3
• cdn1.jaccede.net/uploads/placeur/imgs/full
• media.jaccede.com/uploads/imgs/places
DWEB1 .149
gitolite:
• .git done
• Admin
git: • 6 dépôts:
- jisp-admin-v2
- jisp-api
- jaccede-v3
- jaccede-salt
- jaccede-android-v2
- jaccede-android-v2-by
- jaccede-iphone-v2
- jaccede-ios-v2-s
- jisp
- jisp-gui
- jaccede-connect
- jaccede-api-cloud-gun
PDT1
Postgres QL .146
jdb (DB)
postgres (unix user)
PDT2
MySQL .15
jdb_defaultuser
jisp (DB)
• RATP search solr/ratp
• RATP source jdd: [redacted]/jdbat/jdb
• RATP dest /jdb
Solr + Jetty
http://pdt2:8966/search-engine/search-places/
cron: 2h: reindex-places
6h: searchengine/reindexplaces
- Projet searchengine Java
MongoDB
Nginx
- jispapi
- redmine
- jispadmin
- sphinx doc fo jispapi
- magios
- jaccede-v3
MySQL
PostgresQL

La base :

CONTACT

romain.jouin@memorandum.pro

06.52.86.87.30

www.memorandum.pro

Conseil en stratégie Big Data

The buzz and the truth.


Big data a big deal in business

RIGHT STRATEGY:
It allows for better decision-making and marketing

DR AHMAD SHARAM

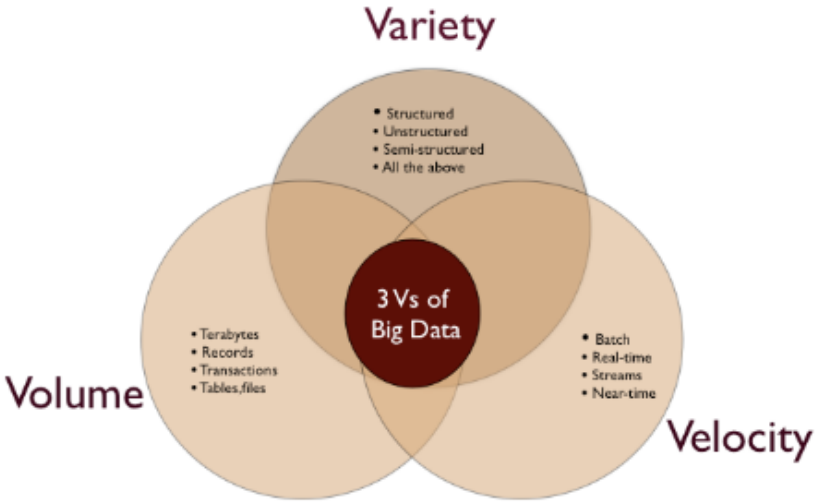
Big Data

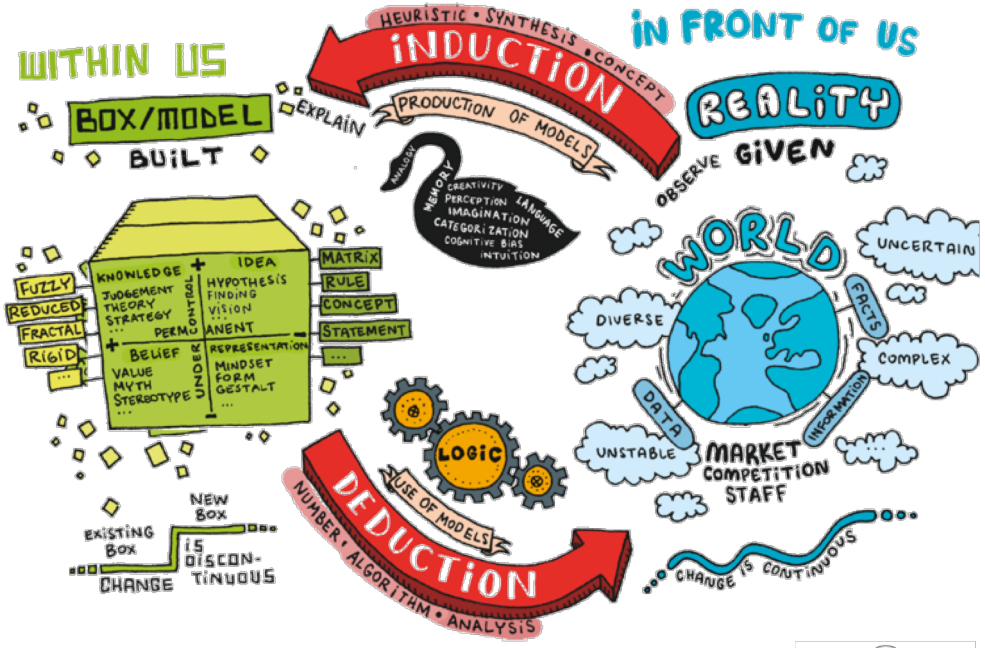
Big Data

3V

Machine Learning

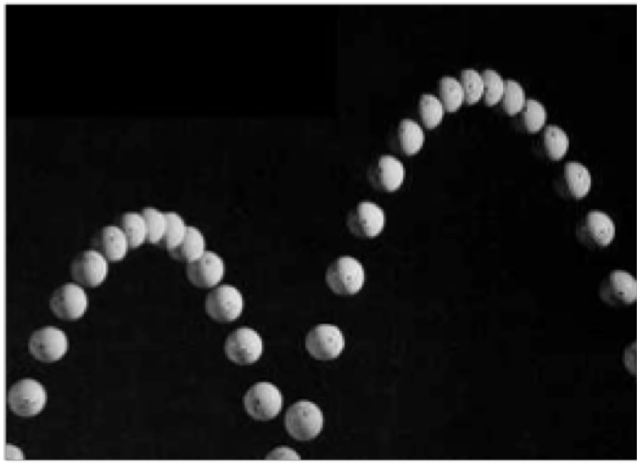
0 V

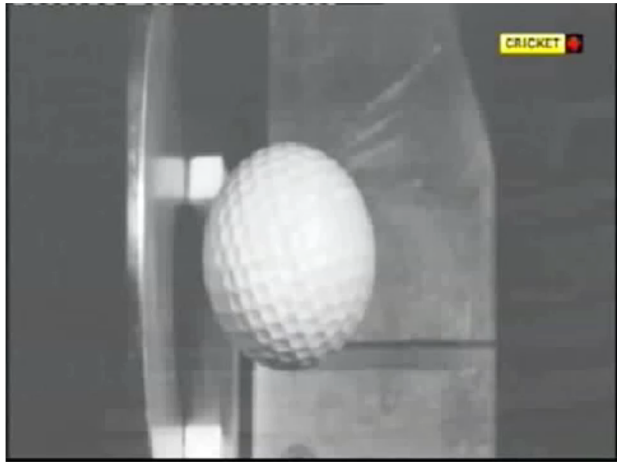




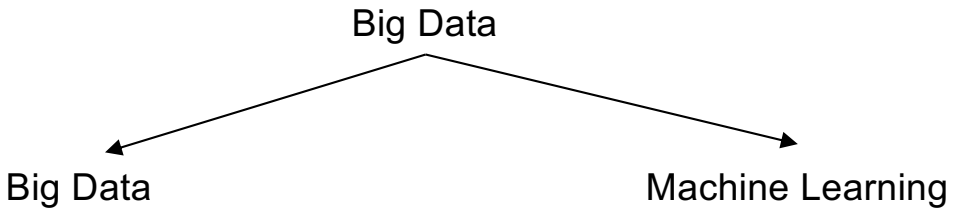
© Luc de Brabandere 2012

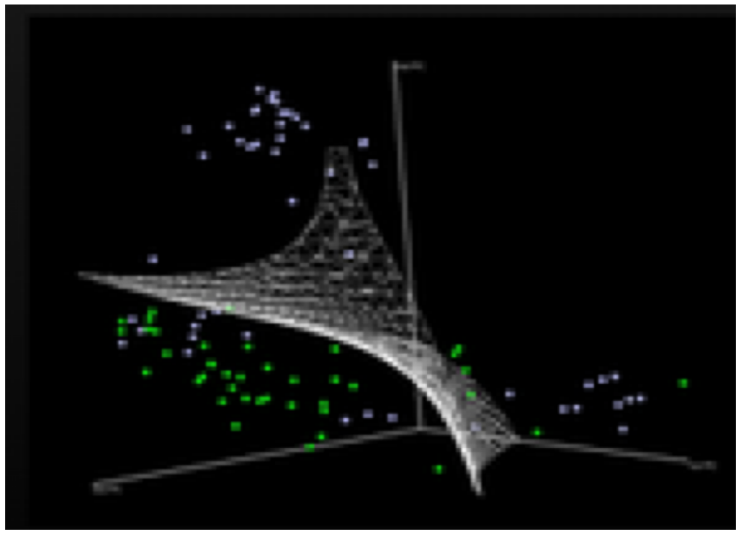
Connaissons-nous le monde ?



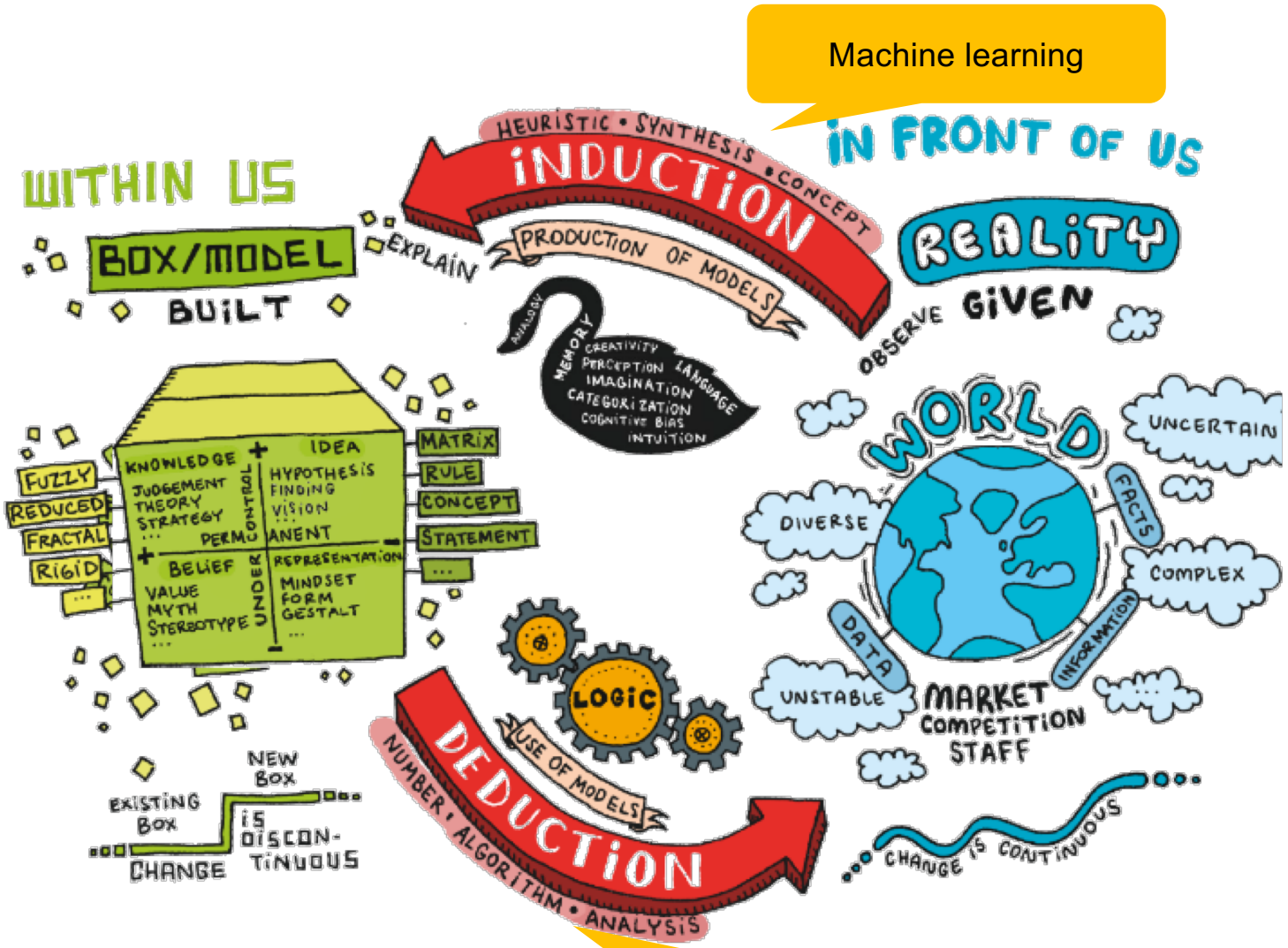


The buzz and the truth.





Datascience : ni plus ni moins ce que fait votre cerveau tous les jours



Chacune de vos données embarque deux leviers de valeur

Une donnée parmi d'autres :
« Le client X a contacté le service client »

L'information pour elle-même

Déclencher l'action suivante :
- Répondre au client

Mettre à jour la rémunération variable du conseiller,

Alimenter les reportings

→ **Domaine connu**

La donnée n'a été initialement produite que pour cela. Après usage cette donnée est aujourd'hui un déchet

Le contexte d'autres événements

Reconstituer le contexte d'autres événements de l'entreprise, tous interdépendants à des degrés divers :

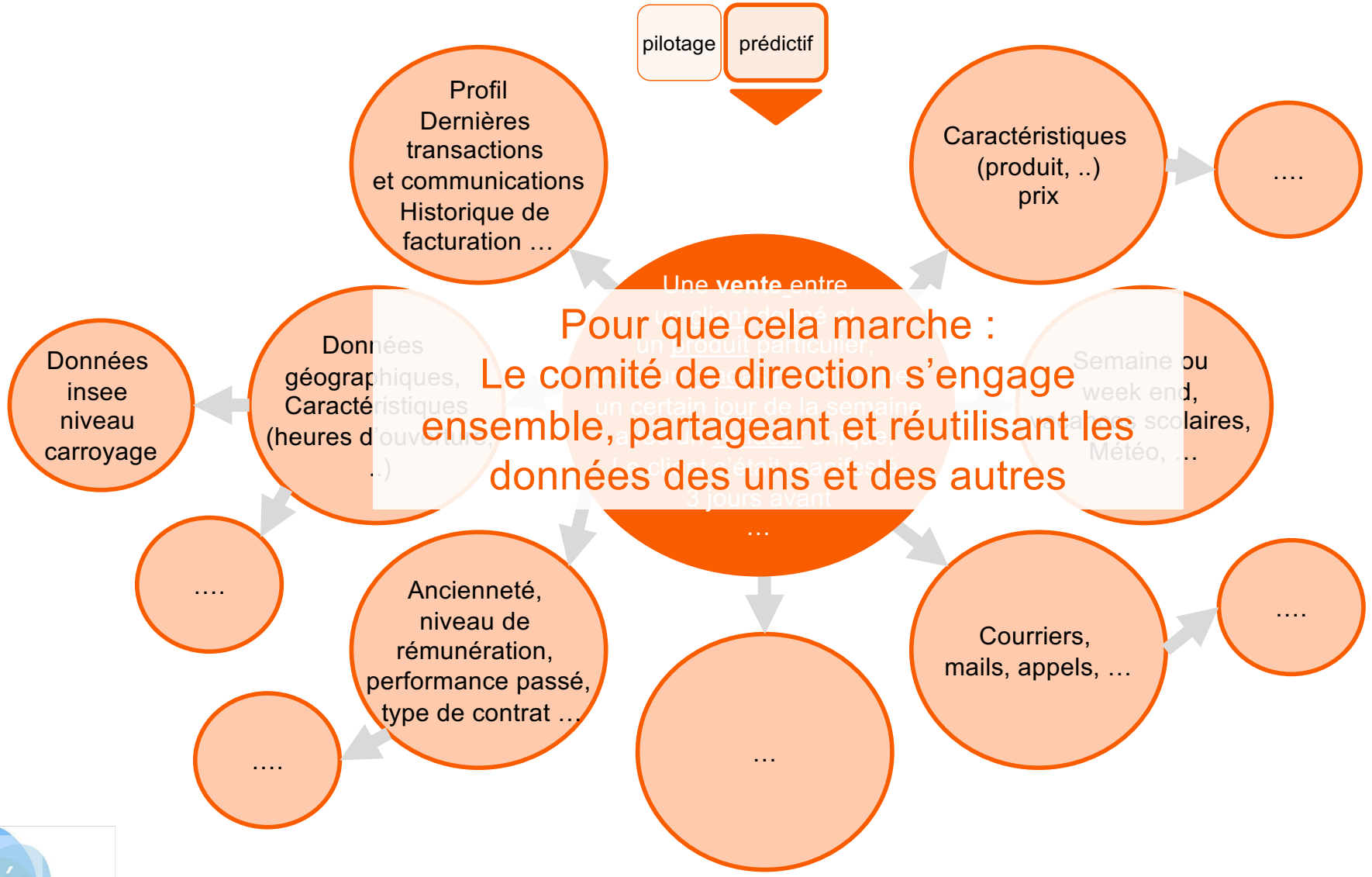
Une vente s'est faite
entre un client donné
et un produit particulier,
dans un magasin spécifique,
avec un vendeur unique,
un certain jour de la semaine
Le client s'était **manifesté 3 jours avant pour dire xxx**
….

→ **Domaine à découvrir !**

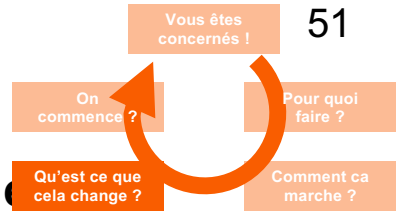
Un exemple de préoccupation

Une donnée sortie de son contexte qui éclaire le sujet

L'approche *machine learning* permet de reconstituer un contexte global



Pour que cela marche :
Le comité de direction s'engage ensemble, partageant et réutilisant les données des uns et des autres



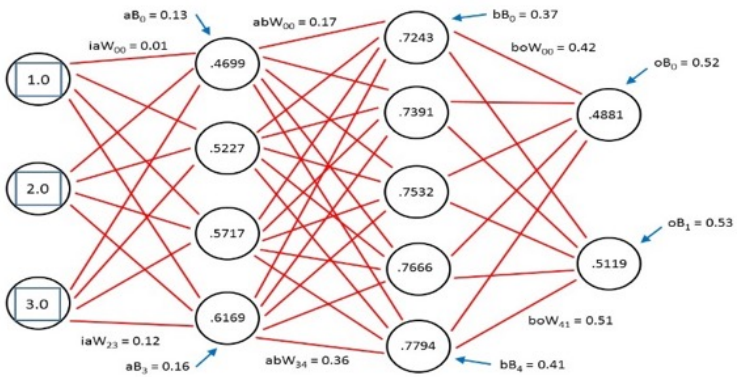
Contexts are so various that only a computer can handle th

hier

Marketing → Décision de campagne → Choix de période → Choix de créa → Choix de cible → **Client**

Aujourd'hui

Marketing



Client X
Client Y
Client Z
Client T
Client U
Client v
Client w
Client X

Client …

Valeur des données : grandes familles d'usages

Processus Internes	Optimiser ses processus	Innover dans ses démarches	Briser les silos
Business Model	Découvrir des opportunités	Dématérialiser son offre	Monétiser ses données
Relation Clients	Conseiller ses clients	Ajuster ses prix	Recommander des produits

Valeur des données : Exemples

Processus Internes	Maintenance prédictive Ville de rouen	Alsace Géolocalisation appels Conception chimique	Conception de produits – BMW So Local
Business Model	SANTEN FONCIA	Michelin loue ses pneus Rolls Royce ses moteurs	Tickets de caisse RATP Open Data Soft
Relation Clients	So Local M6 Allociné	SNCF Site web prédiction	Mister Auto Amazon

Valeur des données : Exemples

Processus Internes	Diminuer ses coûts	Améliorer sa compétitivité	Améliorer la cohésion interne
Business Model	Créer de nouveaux business	Vendre des services récurrents	Valoriser son patrimoine data
Relation Clients	Fidéliser ses clients	Améliorer sa marge	Augmenter son CA