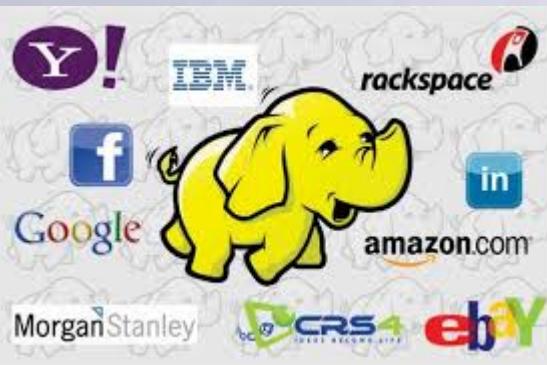
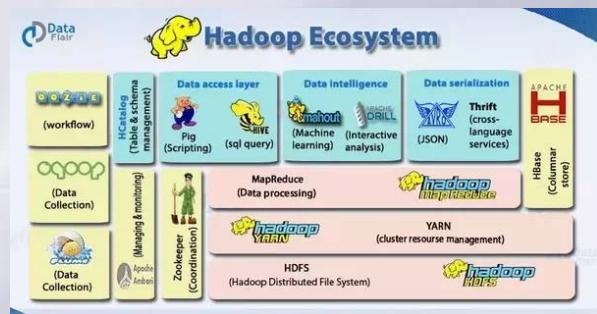


# Big Data - Gestion de données massives



# Laurent Lapasset

2024

# Pour me contacter

# Laurent Lapasset

[laurent.lapasset@recherche.enac.fr](mailto:laurent.lapasset@recherche.enac.fr)

ENAC, Bâtiment Z, Bureau 132



# Plan général

- 1. Mégadonnées : cours
- 2. Hadoop : cours
- 3. Spark : cours
- 4. Cloud : cours
- 5. Virtualisation : cours + TPs
- 6. Docker : cours + TPs
- 7. Kubernetes : cours
- 8. Open Stack : cours



- 1. Origines et définitions
- 2. Nouveaux rôles, métiers et organisations
- 3. Chaînes de traitements
- 4. Exemples de Data Lab
- 5. Modèles de données
- 6. Stockages distribués

# Big Data



# Changement de paradigme

**Le modèle de Génération/Consommation de la donnée a changé**

**Ancien modèle : Quelques compagnies génèrent des données, les autres sont des consommateurs de données**



**Nouveau modèle : nous sommes **tous** des générateurs de données, et nous sommes **tous** des consommateurs de données**



# Générateurs de mégadonnées et usage



**Mobiles**  
(tracer tous les objets tout le temps)



**Média et réseaux sociaux**  
(tous des générateurs de données)



**Instruments scientifiques**  
(collecter toute sorte de données)



**Réseaux de capteurs**  
(mesurer tout type de données)

# Volumes de données

## What Happens in an Internet Minute?

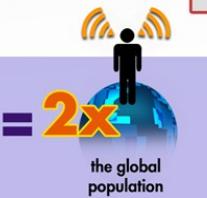


And Future Growth is Staggering

Today, the number of networked devices



By 2015, the number of networked devices



In 2015, it would take you 5 years



to view all video crossing IP networks each second

# La « donnée », une matière première

**“Data is a new class of economic asset, like currency and gold.”**

***Source: World Economic Forum 2012***

# Une définition de Big Data

- Le big data fait référence à l'explosion du volume de données informatiques qui transitent dans le monde.
- Ces données deviennent tellement volumineuses qu'elles ne peuvent plus être stockées dans des bases de données traditionnelles.
- Ce qui mène à l'utilisation de nouveaux outils informatiques pour les gérer et les traiter.

# D'autres définitions...

- « **Big Data** » est un volume massif de données structurées et non structurées qui est si important qu'il est difficile à traiter avec les techniques traditionnelles de bases de données et de logiciels
- « **Big Data** » sont des données dont le volume, la diversité et la complexité nécessitent de nouvelles architectures, modèles, techniques, algorithmes et analyses pour les gérer et en extraire de la valeur et des informations cachées...

# Caractérisations des données grande dimension

A quel moment on parle de données “grande dimensions” ?

Le big data peut être défini par ses trois caractéristiques majeures, toutes commençant par la lettre V :

- Volume
- Variété
- Vélocité

# AnalyseS des mégadonnées

- Reposent principalement sur la fouille de données (Data Mining) et sur l'analyse statistique.
- Analyse de mégadonnées stockées : Hadoop et Spark
- Analyse de flots de données : Spark Streaming, Storm
- Analyse de sons, images, textes et graphes
- Apprentissage automatique et modèles IA

# **Les métiers, les rôles**

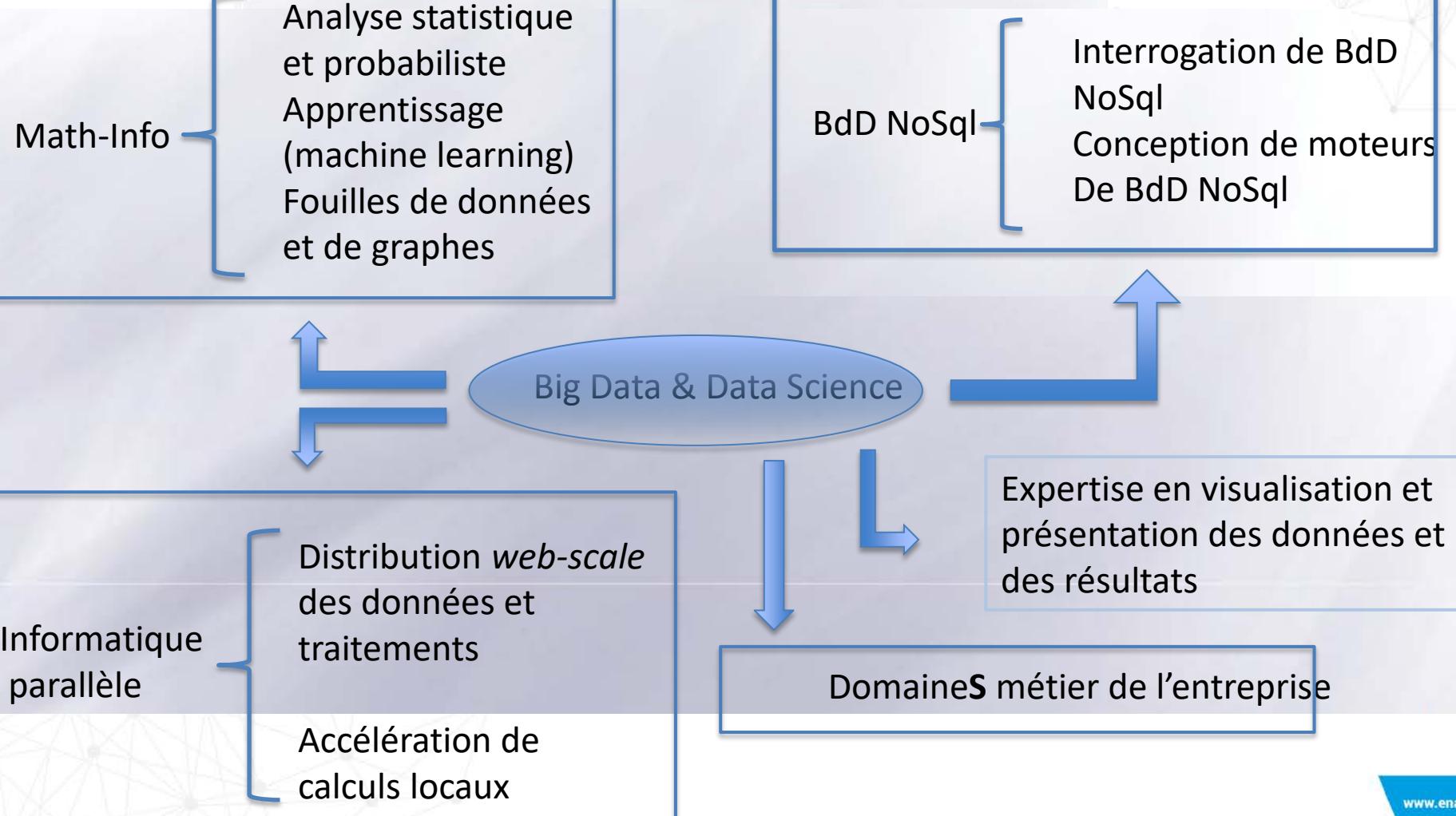
# La gouvernance des données

- **Une compréhension commune des données**
- **Amélioration de la qualité des données**
- **Une cartographie des données**
- **Une vue à 360° de chaque client et des autres entités commerciales**
- **Une uniformité de la conformité**
- **Une gestion améliorée des données**
- **Une facilité d'accès**

# De nouveaux rôles

- **Les directeurs des données ou CDO (Chief Data Officer)**
- **Les responsables de la protection des données ou DPO (Data Protection Officers)**
- **Les data architects**
- **Les data stewards**
- **Les ingénieurs de données et développeurs**
- **Les data scientists**
- **Les business analysts**
- **Utilisateurs métiers/data curators**

# Composition pluridisciplinaire du Big Data

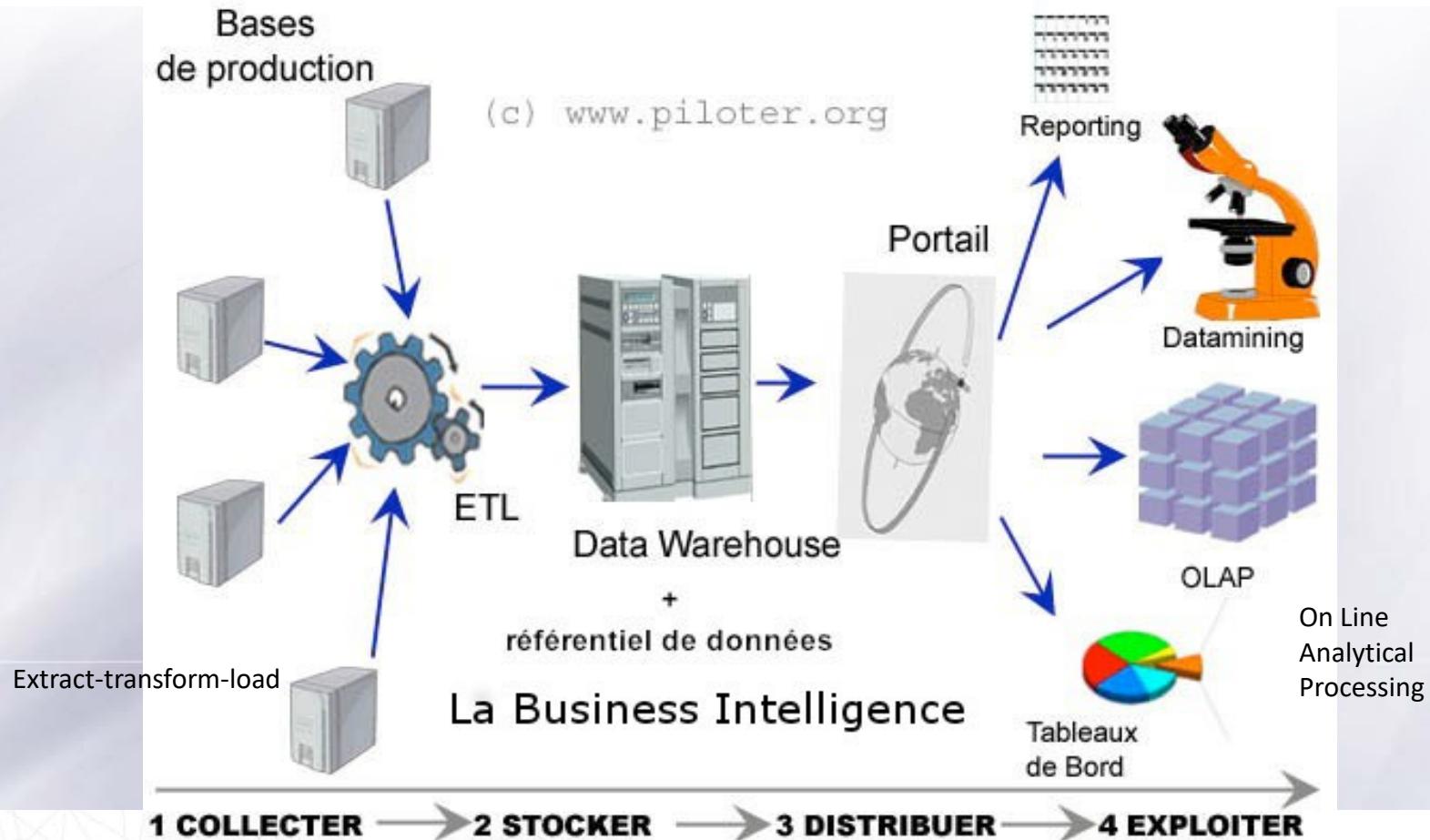


# **Chaine de traitement**

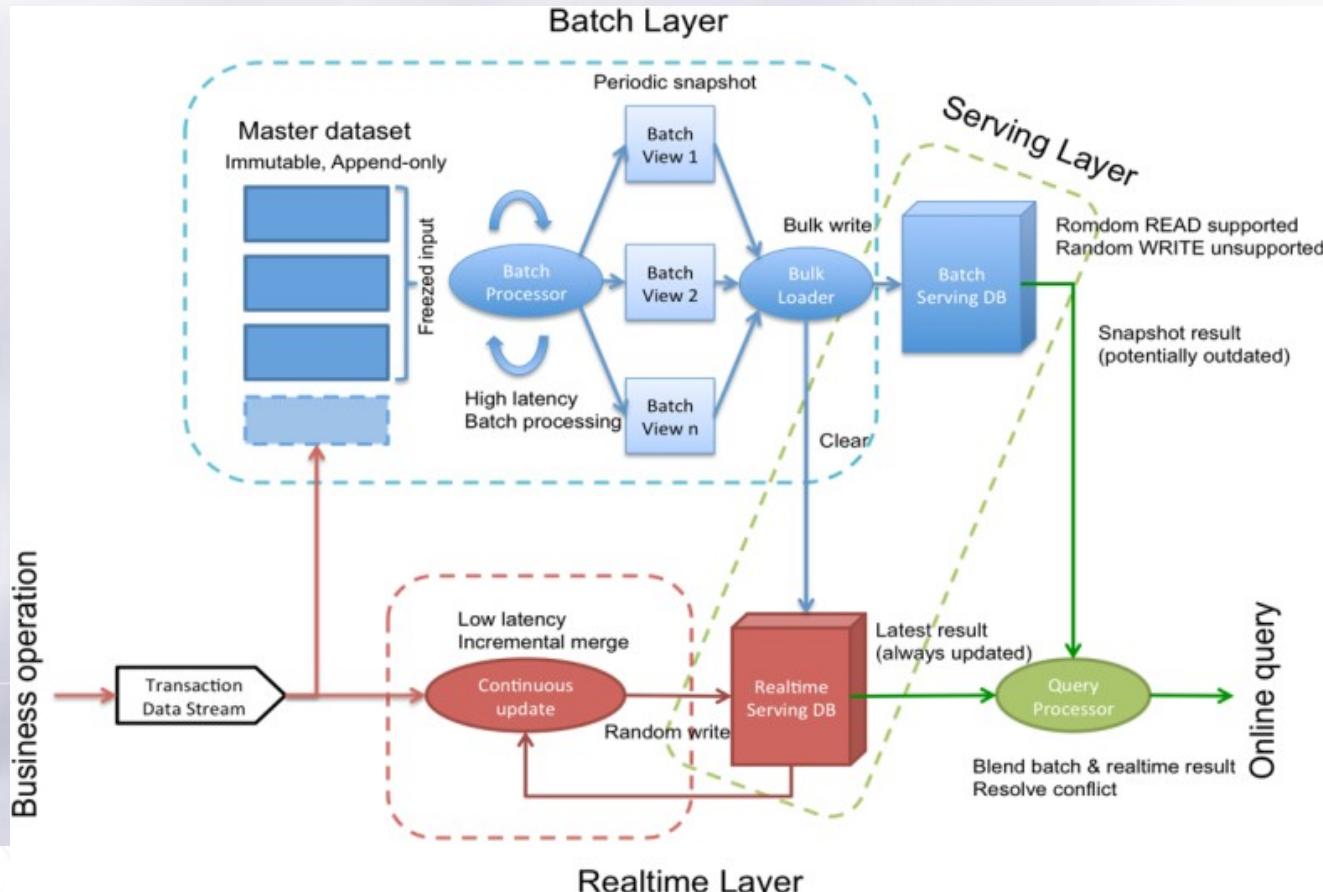
# Le traitement : pipelines

- **La collecte ou l'extraction d'ensembles de données brutes.**
- **La gouvernance des données.**
- **La transformation des données :**
  - La normalisation
  - Le dédoublonnage
  - La vérification
  - Le classement
  - Le partage des données

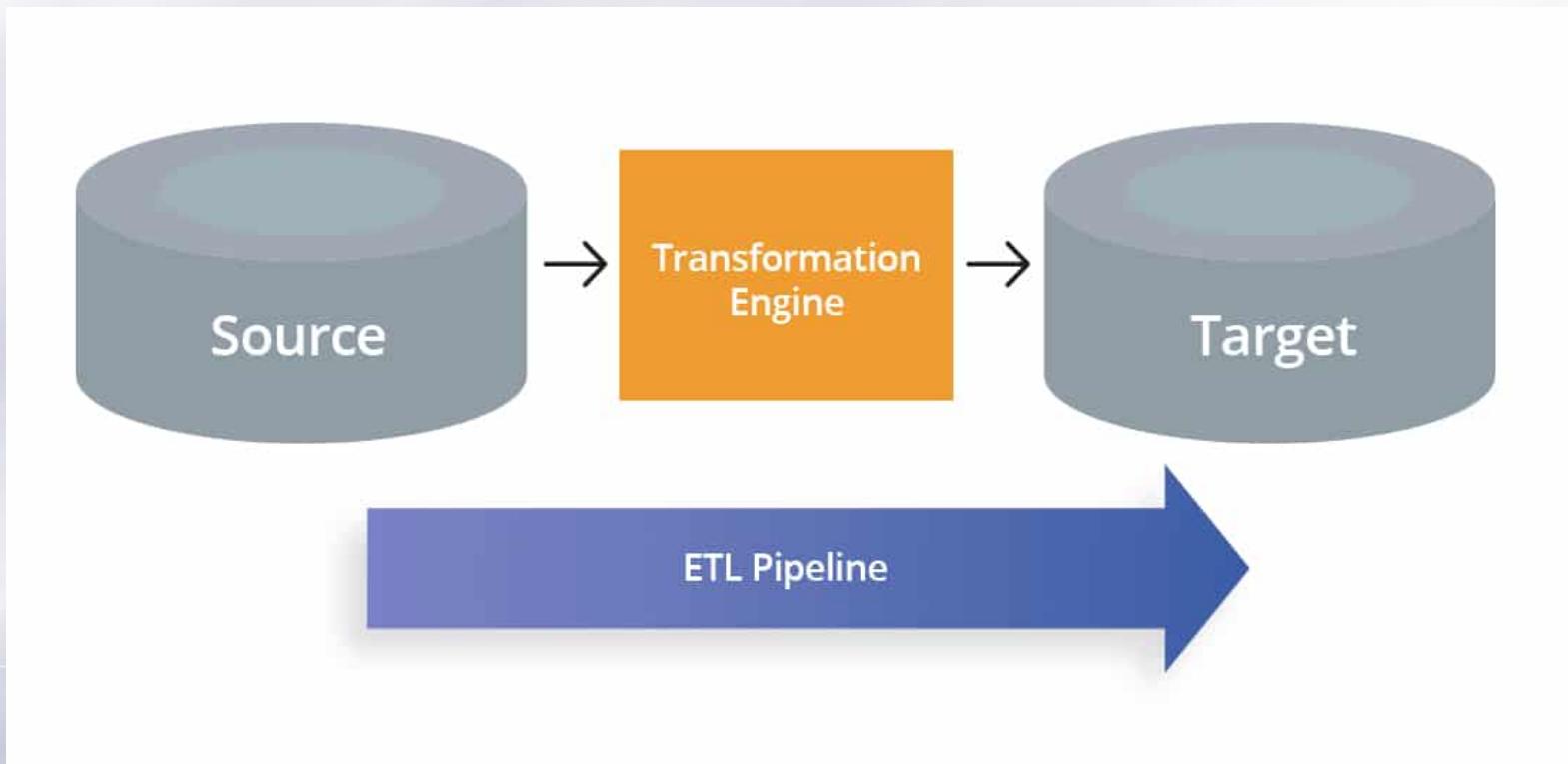
# Chaîne de traitement générique



# L'architecture Lambda



# Pipeline ETL



# Pipeline de données

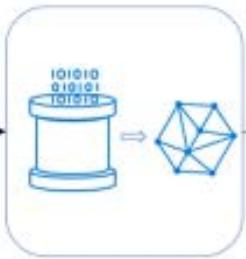


# Pipeline « Machine Learning »

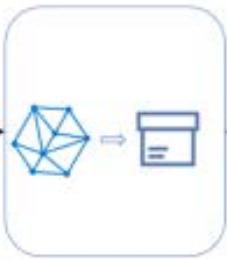
Prepare Data



Train Model



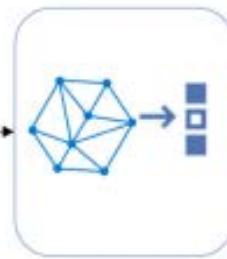
Package Model



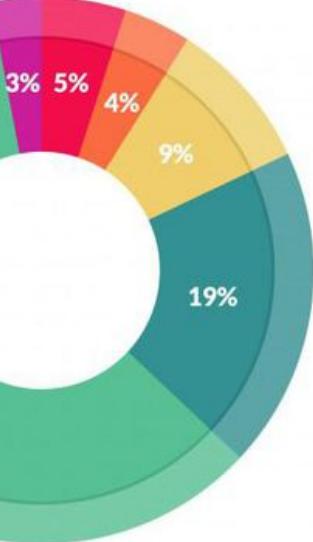
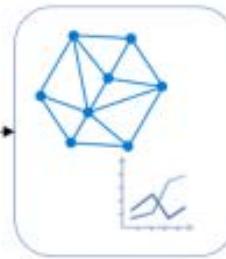
Validate Model



Deploy Model



Monitor Model



What data scientists spend the most time doing

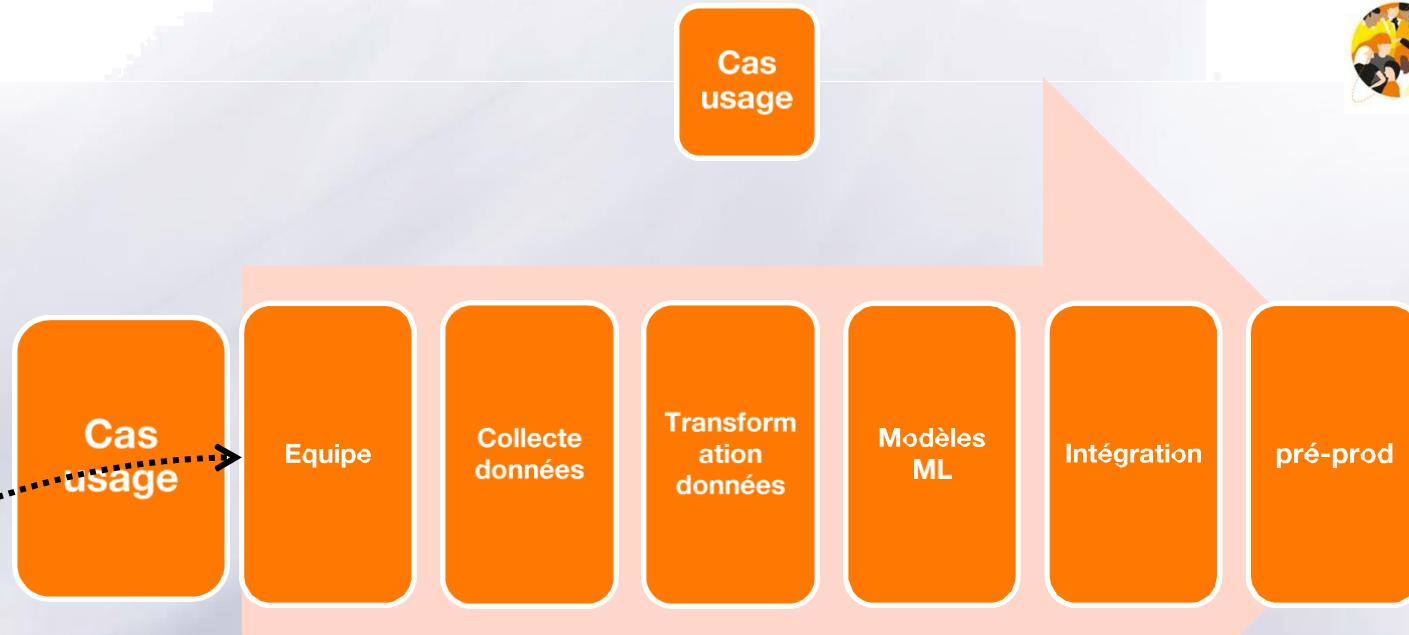
- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

# **Exemples de Data Lab**

# Projet Orange Labs

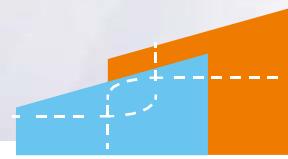


Orange Expert  
Security



Itérer !

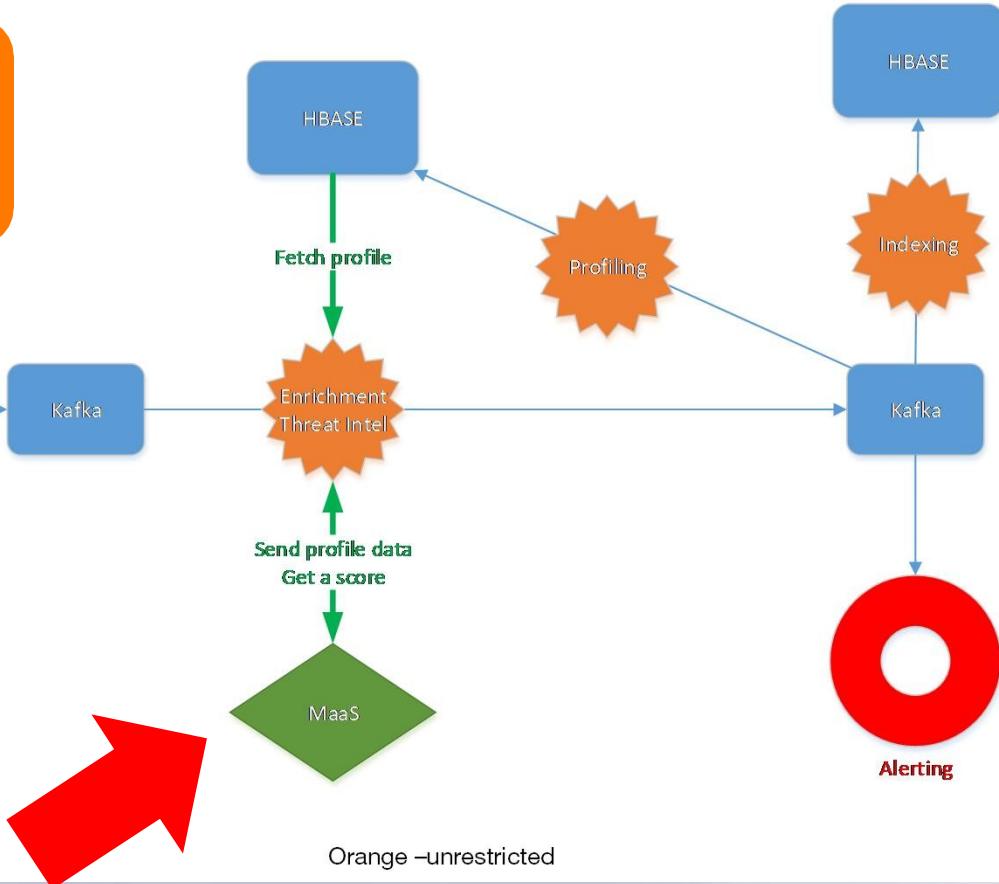
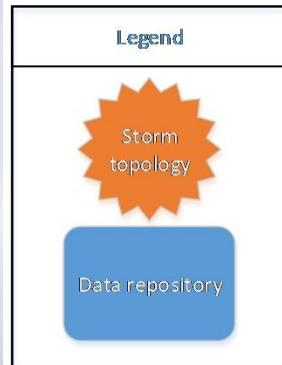
Orange –unrestricted



# Projet Orange Labs

## Integration Solution Open source

4 nœuds  
4x64Gb RAM  
4x10 cœurs  
12 To



Orange Expert  
Security

# Le site internet le plus visité au monde

The Facebook logo is displayed in its characteristic blue square with the word "facebook" in white lowercase letters.

**Facebook : l'un des plus grand cluster Hadoop du monde depuis 2010**

**Le réseau social s'adosse à une architecture LAMP : Linux, Apache, MySql et PHP**

**Il est utilisé pour sauvegarder son infrastructure MySql. Le cluster se « rapprocherait » des 100 pétaoctets.**

**~ 60 000 et 100 000 le nombre de serveurs de Facebook .**

# Exemple de stockage et gestion de mégadonnées

**facebook**

**Quelques chiffres sur MySQL :**

- **13 millions de requêtes par seconde en pic,**
- **38 Go/s de trafic MySQL en pic,**
- **temps de réponse moyen en lecture : 4 ms,**
- **temps de réponse moyen en écriture : 5 ms,**
- **450 millions de lignes lues par seconde en pic,**
- **3,5 millions de lignes modifiées par seconde en pic,**
- **5,2 millions d'I/O (disques) InnoDB par seconde.**

# Infrastructures optimisées

facebook

## Quelques chiffres intéressants sur MySQL :

- **Comme beaucoup d'autres, Facebook utilise une architecture LAMP (Linux / Apache / MySQL / PHP) :**
- **Linux : il s'agit d'une version de CentOS, avec un kernel 2.6 customisé,**
- **Apache 1.3,**
- **MySQL 5.1 (passage à la version 5.5 prévu à moyen terme),**
- **PHP5.**

## Sources :

- **A Day in the Life of Facebook:**

Operations [https://www.youtube.com/watch?v=T-Xr\\_PJdNmQ&ab\\_channel=O%27Reilly](https://www.youtube.com/watch?v=T-Xr_PJdNmQ&ab_channel=O%27Reilly)

- **MySQL Connect 2013 Keynote - MySQL at Facebook :**

[https://www.youtube.com/watch?v=S-KLVe4YSLY&ab\\_channel=MySQL](https://www.youtube.com/watch?v=S-KLVe4YSLY&ab_channel=MySQL)

- **MySQL to manage data :**

<https://www.facebook.com/watch/?v=695491248045>

# Data Innovation Lab R&D d'EDF

## Tout savoir sur IZIVIA en quelques mots

IZIVIA, filiale 100 % EDF, propose des solutions de recharge pour véhicules électriques à destination des collectivités, des syndicats d'énergie, des entreprises et des copropriétés. À ce titre, IZIVIA apporte son expertise à ses clients via une gamme d'offres complète : fourniture et installation de bornes de recharge, supervision et maintenance des infrastructures et offres de services. IZIVIA, en tant qu'opérateur de mobilité pour tous, propose un Pass et une application pour smartphone qui permettent de se recharger sur plus de 100.000 points de charge en France et en Europe.



Déploiement  
des bornes



Exploitation  
des bornes



Service  
pour les  
utilisateurs

## Le groupe EDF en quelques chiffres



\*source : edf.fr - juin 2020

# EDF R&D : attention « Produits d'appels » .



## INTEROPERABILITE via GIREVE ou en DIRECT OCPI



\* e- MIP: Protocole interne GIREVE sécurisé



**Accès à 100.000 points de charge en Europe  
Possibilité d'accords directs OCPI**

# Attention aux coûts «cachés» ...

https://console.cloud.google.com/billing/01CCD8-F38ABB-A7C113/reports;grouping=GROUP\_BY\_SKU;projects=izivia-business-intelligence?project=izivia-business-intelligence

Apps Apple importés dep... Télécharger to... Matrice de cor... http://www.ca... https://chrisal... Apprenez à uti... Aperçu des m... analyse-R Comment choi... Other bookmarks

Google Cloud Platform Search products and resources

Billing Reports Izivia PRINT SHARE

Overview Reports Account management

1–26 November 2020 (total cost) €1,844.77 ↑ 4,374.34% Includes €0.00 in credits €1,803.54 over 6–31 October 2020

November 2020 (forecasted total cost) €1,876.72 ↑ 4,443.02% Includes €0.00 in credits €1,835.41 over October 2020

Daily 1K

Cost trend

Nov 2 Nov 3 Nov 4 Nov 5 Nov 6 Nov 7 Nov 8 Nov 9 Nov 10 Nov 11 Nov 12 Nov 13 Nov 14 Nov 15 Nov 16 Nov 17 Nov 18 Nov 19 Nov 20 Nov 21 Nov 22 Nov 23 Nov 24 Nov 25 Nov 26 Nov 27 Nov 28 Nov 29 Nov 30

SKU	Service	SKU ID	Usage	Cost	Discounts	Promotions and others	Subtotal
Analysis	BigQuery	1DF5-1F98-1DD1	431.91 tebibyte	€1,835.91	—	—	€1,835.91
Static Ip Charge in Frankfurt	Compute Engine	DC8F-9C09-0D8C	327.49 hour	€3.34	—	—	€3.34
Active Storage	BigQuery	947D-3B46-7781	95.67 gibibyte month	€1.51	—	—	€1.51
E2 Instance Ram running in London	Compute Engine	5D70-7762-2DE7	343.28 gibibyte hour	€1.10	—	—	€1.10
E2 Instance Core running in London	Compute Engine	0F2A-2FA8-3F6A	42.91 hour	€1.03	—	—	€1.03
Standard Storage Europe Regional	Cloud Storage	A703-5CB6-E0BF	39.37 gibibyte month	€0.69	—	—	€0.69

Filters

Presets

Time range

Current month Usage data is available since January 2017

Group by SKU

Projects All projects (1)

Services All services (11)

SKUs All SKUs (122)

Locations Filter by location data like region and zone.

RESET CLOSE PANEL

# Ressources GCP et AWS EC2



Google Cloud Platform



GPU	Tesla K80(4vCPU)		Tesla P100(4vcpu)	
Hardware	GCP		GCP	
Ram	16Go		16Go	
SSD	30GO		30GO	
Type de modèle	MobileNet	VGG16	MobileNet	VGG16
Durée (mn)	50.82	52.79	41.98	43.36
Temp max. (°)	35° -> 48°		40°->53°	
Accuracy (Val)	0.9316	0.9332	0.9341	0.9423
SpeedUP [K80 vs P100] (mn)			+21.05%	+20.43%
durée utilisation service GCP	~ 5H00			
Billing GCP	10,42 USD			

GPU	GTX 1080	GTX 1060	Tesla K80	Tesla V100
Hardware	tour	portable	AWS	AWS
Ram	8Go	6Go	11.4Go	16Go
Durée (mn)	30.14	47.01	33.47	21.14
Temp max. (°)	56.00	65.00	x	x
Accuracy (Val)	0.9644	0.9641	0.9637	0.9643
SpeedUP [1080 vs 1060] (mn)		35.89%		
SpeedUP [K80 vs 1080] (mn)			-11.05%	
durée utilisation service AWS			~ 6H00	
Billing AWS	5.33 USD			
SpeedUP [V100 vs 1080] (mn)			29.86%	
durée utilisation service AWS			51 mn	
Billing AWS				3.56 USD
Stockage s3 per Day AWS	(216.4 MB)	1.03 USD		

Échec du démarrage de l'instance de VM "instance-eol-dl". Erreur : The zone 'projects/centering-badge-224319/zones/europe-west1-b' does not have enough resources available to fulfill the request. Try a different zone, or try again later.



# Le cluster FEAT (DTI/SIB)

## Composition matérielle



### Matériel

Un baie de 12 machines composées de :

- 2 orchestrateurs (NameNode + RessourceManager)
- 8 calculateurs (DataNode)
- 2 serveurs applicatifs (EdgeNodes)

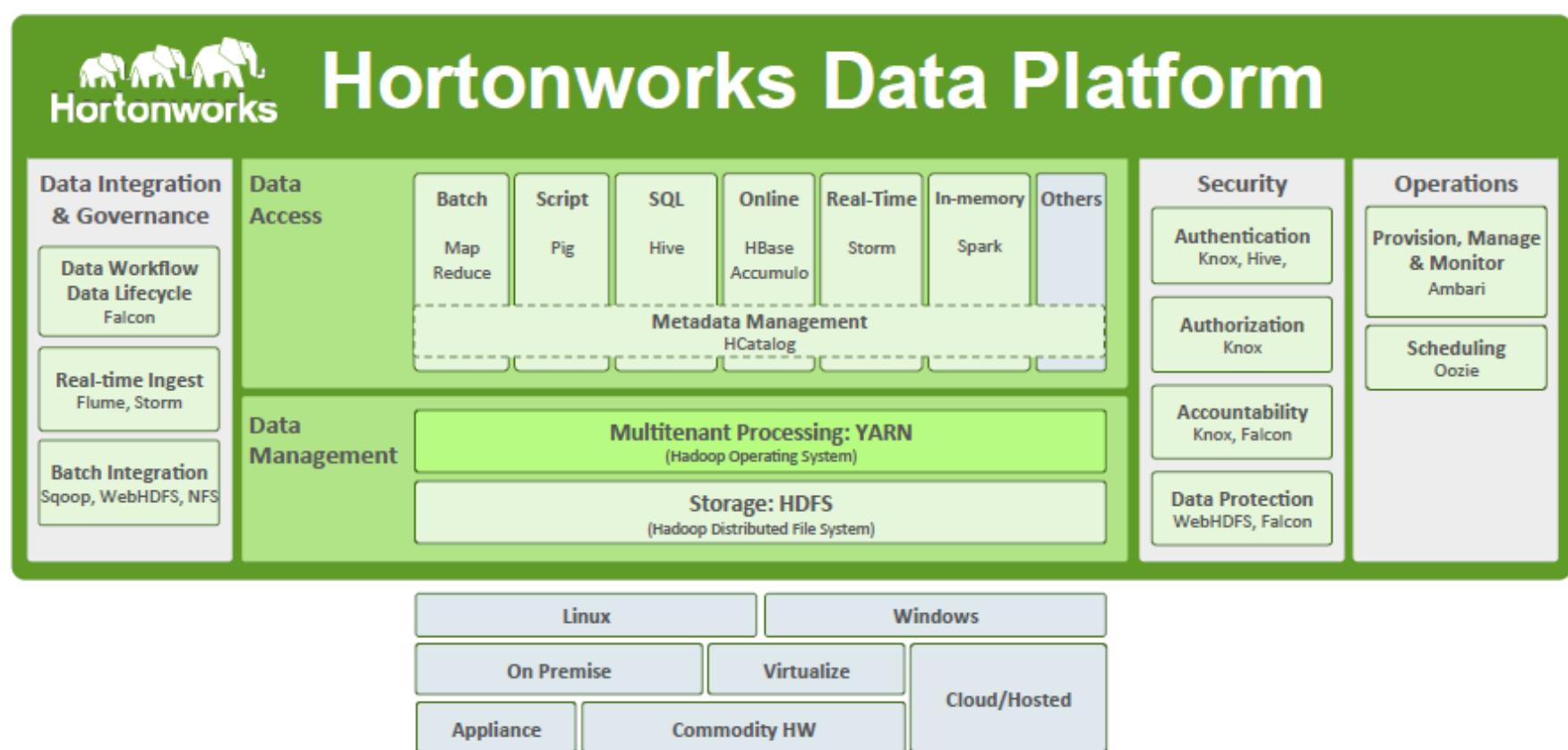
### Ressources

Un outil de stockage et traitement de données :

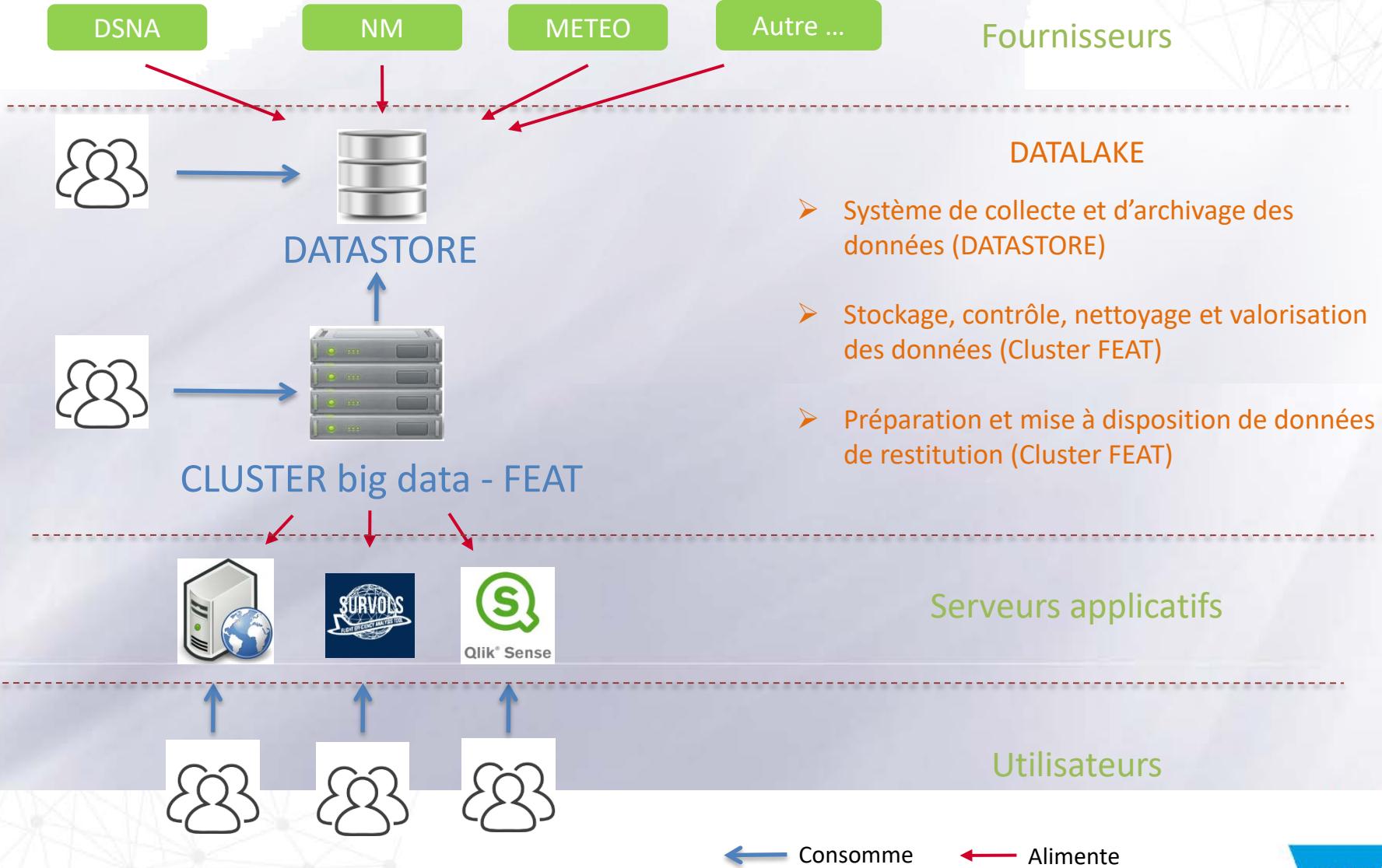
- Avec 150 To de ROM, 1 To de RAM, ~300vCPU
- Sécurisé, gouverné et scalable
- À disposition des développeurs, DataScientists et DataAnalysts

# Le cluster FEAT

## Composition technologique



# Le cluster FEAT

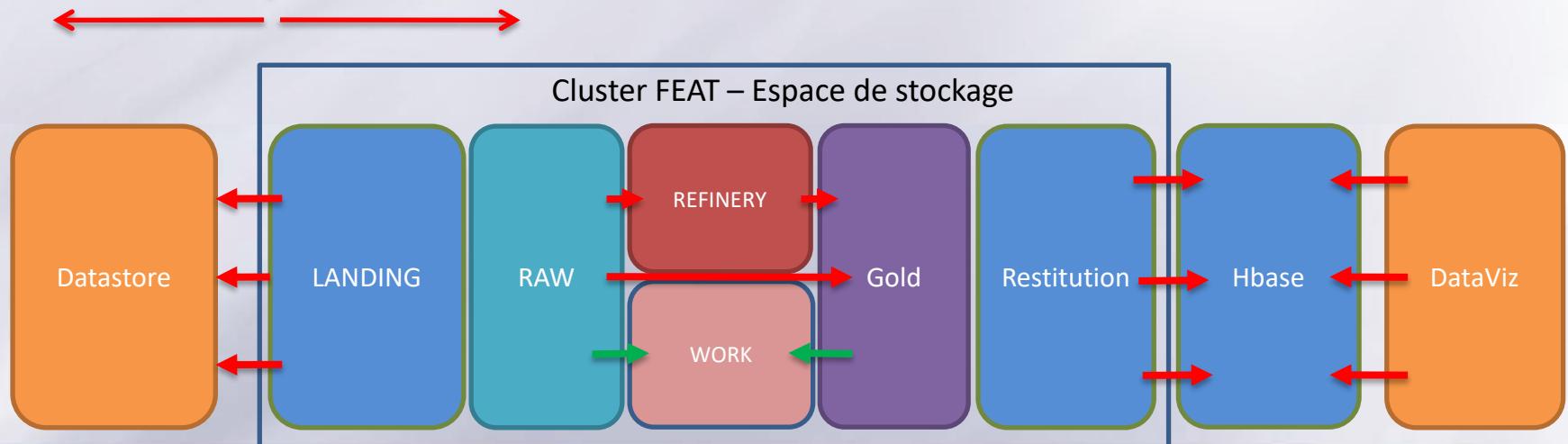


# Le cluster FEAT

## Fonctionnement

### Ingestion : (équipe dédiée)

Un projet commun permettant d'uniformiser les méthodes d'ingestions depuis des sources variées et de formater les données jusqu'en zone de données brutes (RAW)



→ Traitement

→ Expérimentation

### Projets : (Survols, MonCiel, ...)

Un ensemble de projets visant à nettoyer, corriger, agréger ou croiser les données pour en produire de la valeur ajoutée

# Le cluster FEAT

## Technologies

- Spark 2.3.2
- Zeppelin 0.8.0
- QlikSense

## Utilisation des ressources

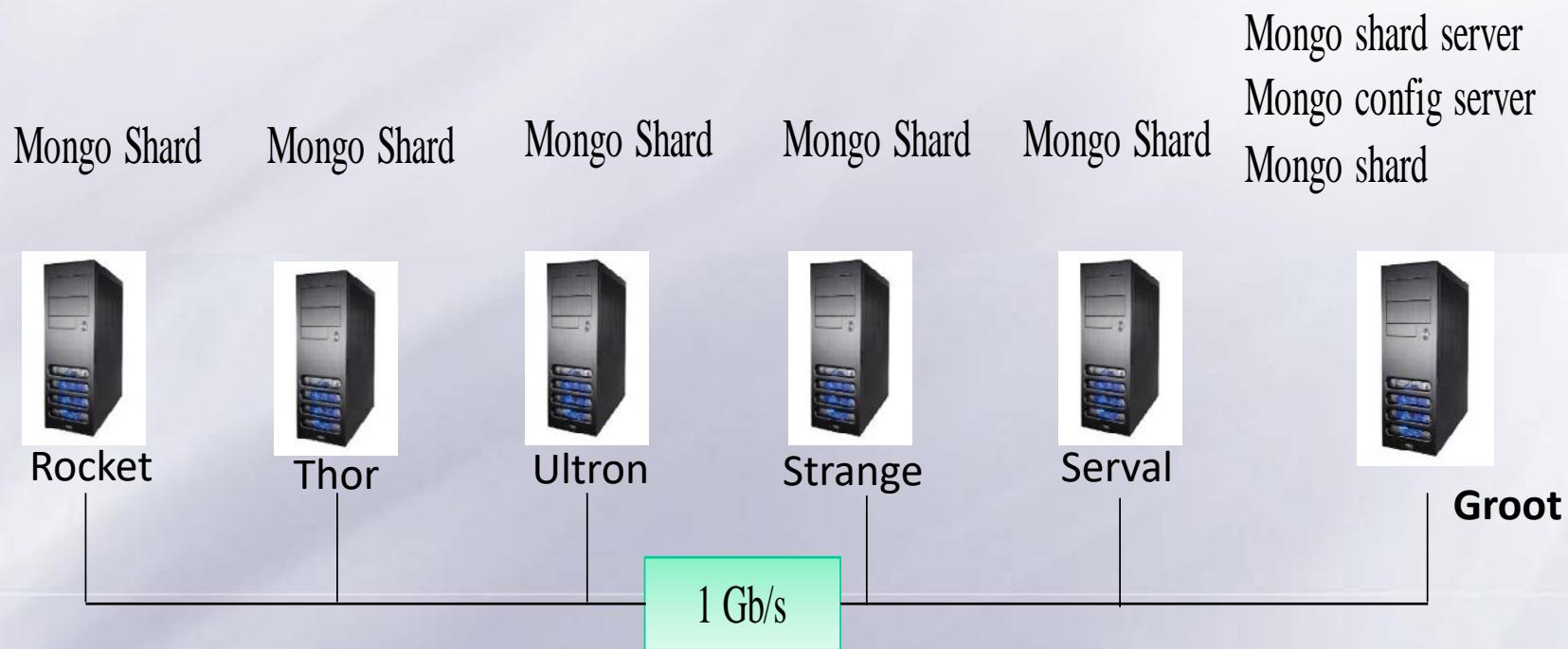
### Fouille et analyse de données

- Catalogue de données: Liste des données du cluster et de leur définition
- QlikSense: Outil de Business Intelligence permettant de créer des dashboard d'analyse et de visualisation de données (accès sous licence)
- Survols: Permet une visualisation des données de trajectoires et des indicateurs de Survols (accès libre sous réserve d'une création de compte)

### Manipulation de données

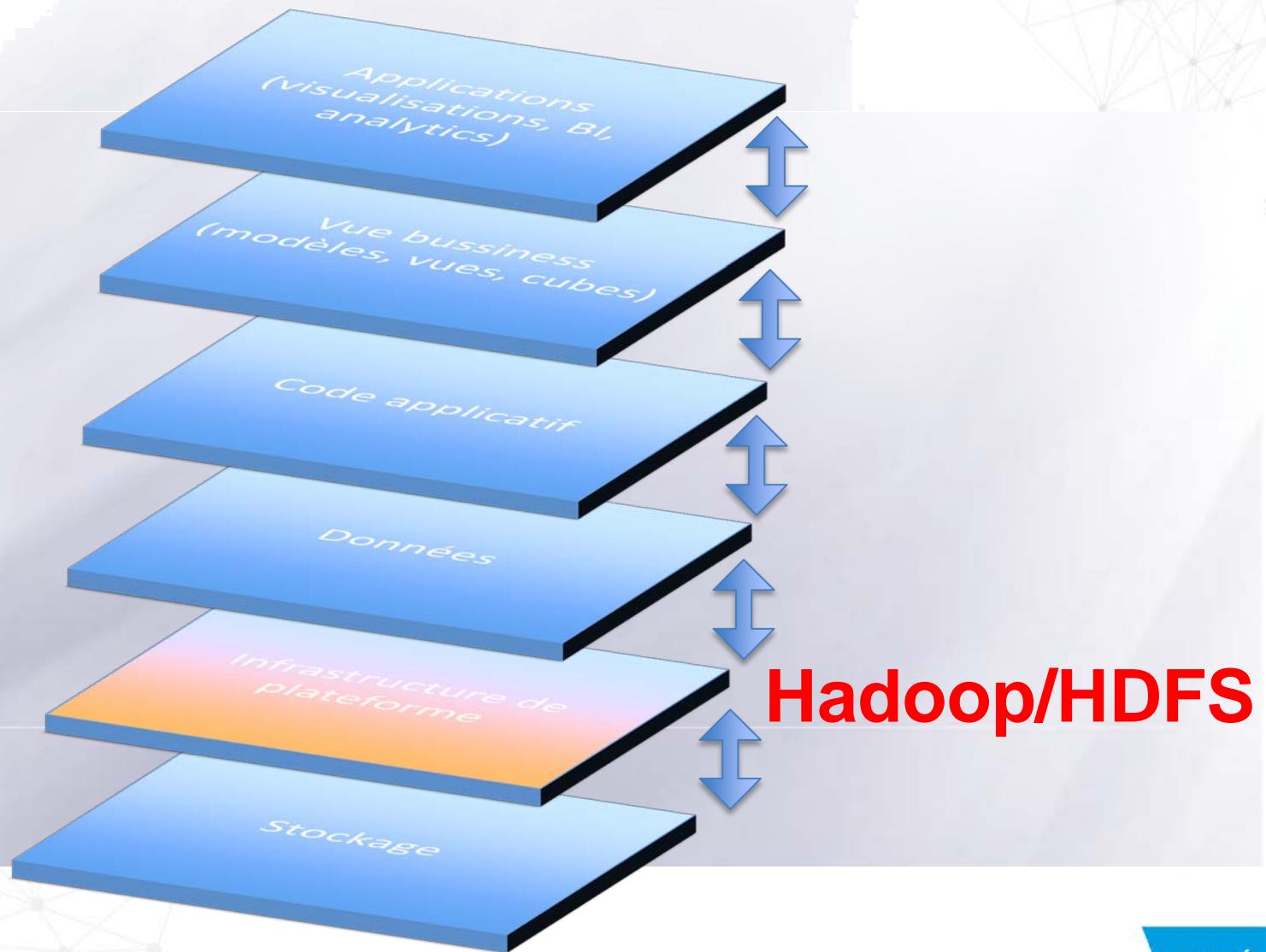
- Zeppelin: Notebook permettant d'utiliser Spark et d'accéder au données
- PySpark: Shell permettant d'écrire du Spark et de tester des algorithmes

# Cluster de recherche de l'ENAC

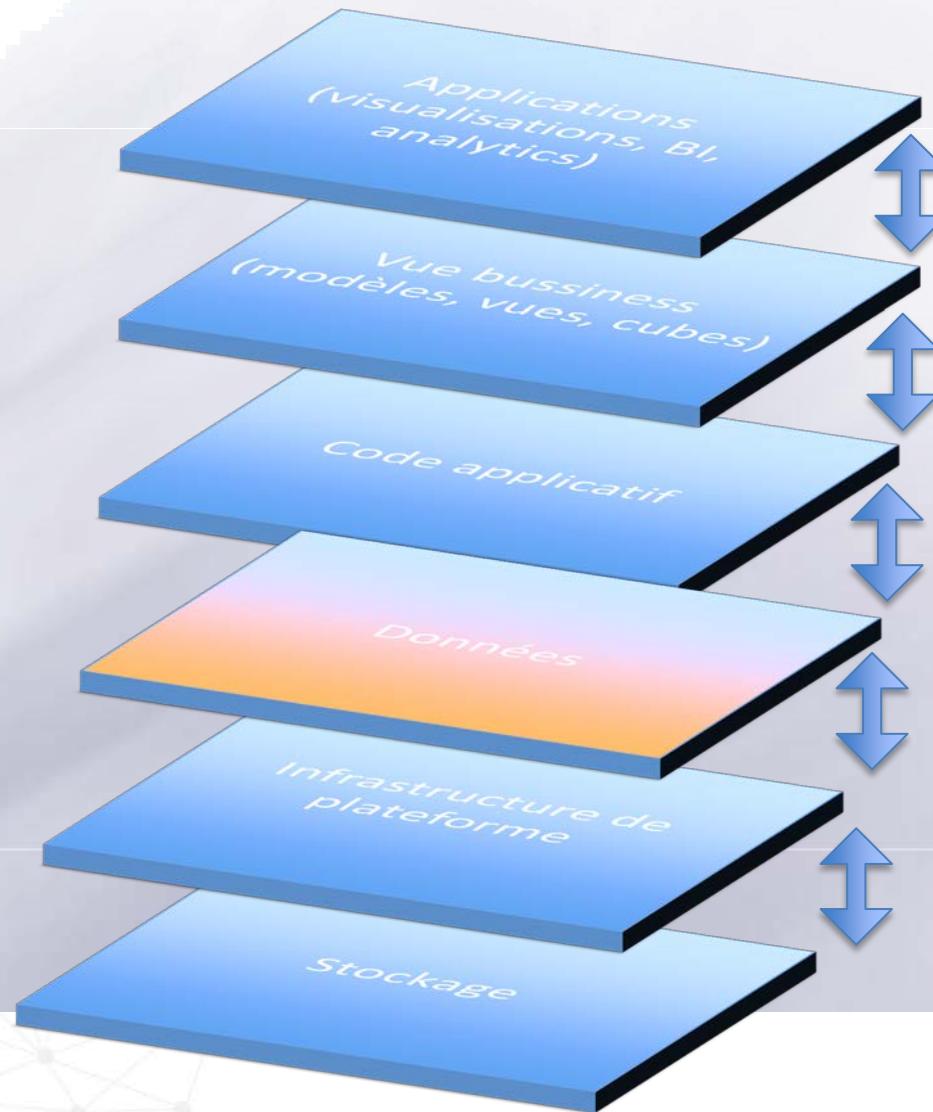


# Avec quels modèles ?

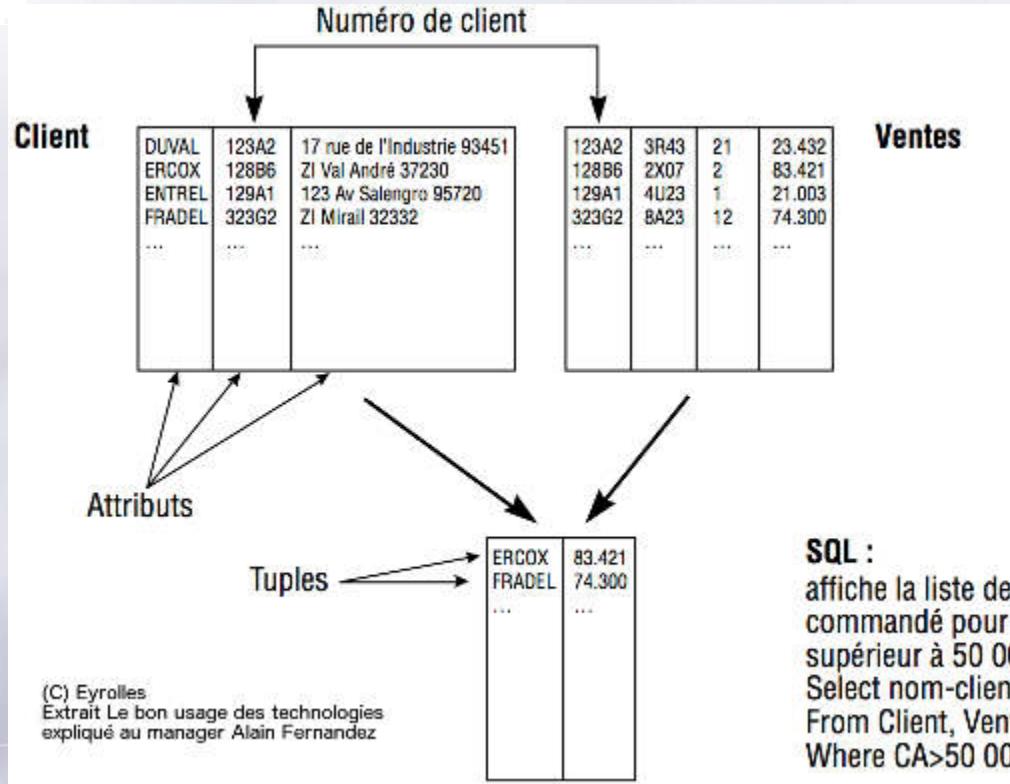
# Couche « Infrastructure de plateforme »



# Couche « Données »



# Le modèle relationnel (SGBDR)



(C) Eyrolles  
Extrait Le bon usage des technologies  
expliqué au manager Alain Fernandez

# Les bases de données NoSQL

**Not Only SQL : contourner les limites des bases de données relationnelles.**

**Le mode de stockage NoSQL**, permet de gérer de gros volumes de données qui peuvent être non structurées dans des serveurs de stockage **distribués** qui sont capables de monter en charge à moindre coût : **scalabilité**

# Les architectures distribuées

Besoin de répartir le stockage sur différentes machines, pour cela il existe deux types d'architectures : **distribuer** et **répartir** les données et les traitements le plus efficacement possible.

- Architecture maître/esclave
- Architecture sans maître

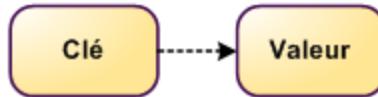
# Théorème de Brewer

- Ces nouvelles bases abandonnent la notion de durabilité, comme le cas des bases de données en mémoire vive, d'autres abandonnent la notion de transaction pour laisser place aux caractéristiques du théorème CDP (**Cohérence, Disponibilité, tolérance au Partitionnement**) établies par Éric Brewer en 2000, auxquelles répondent les moteurs NoSQL.
- Caractéristiques du théorème CDP :
  - **Cohérence** : même version des données sur tous les nœuds ;
  - **Disponibilité** : les données sont accessibles à tout instant, et une requête reçoit une réponse ;
  - **Tolérance au Partitionnement (pannes)** : le système doit pouvoir continuer à fonctionner et à répondre, même si un problème survient sur un nœud du cluster.

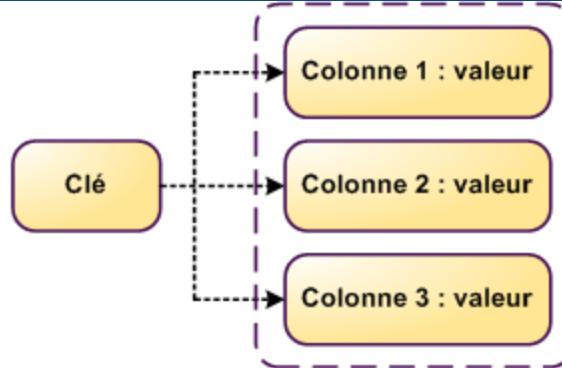
# Brewer et système à état partagé

- Dans un système à état partagé en réseau, seules deux des trois propriétés peuvent être satisfaites en même temps : un impact sur les choix du moteur NoSQL, pour répondre à des problèmes de montée en charge en fonction de la complexité des données.
- On peut combiner différents moteurs NoSQL c'est ce qu'on appelle « **persistence polyglotte** ».
- Sachant que dans un contexte big data la « scalabilité » est le facteur le plus important, on ne parle plus d'un système consistant, mais « éventuellement consistant » selon la base NoSQL choisie, et qui fait partie du concept « **BASE** » (**Basically Available, Soft state, Eventual Consistant**).

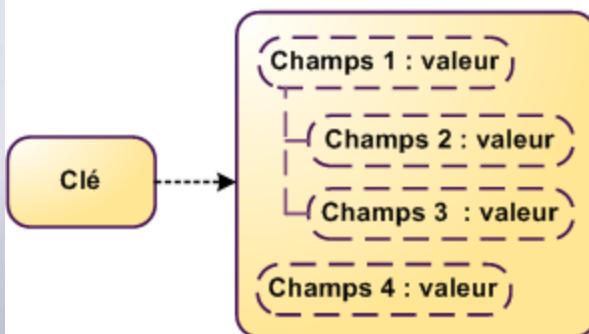
# Types de base de données NoSQL



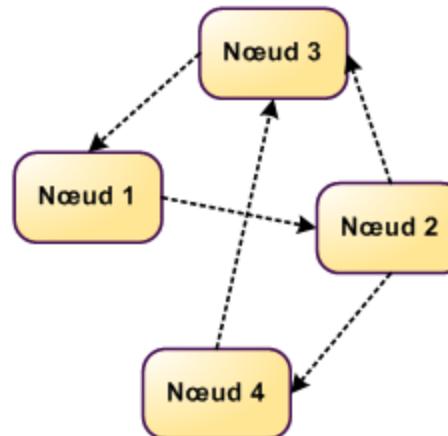
BDD Clé-Valeur



BDD Orientée colonnes

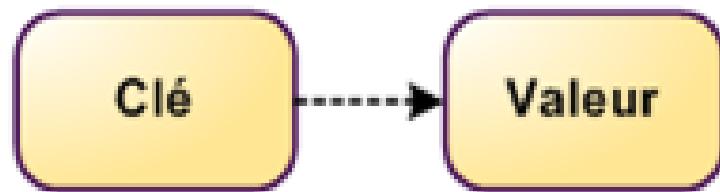


BDD Orientée document



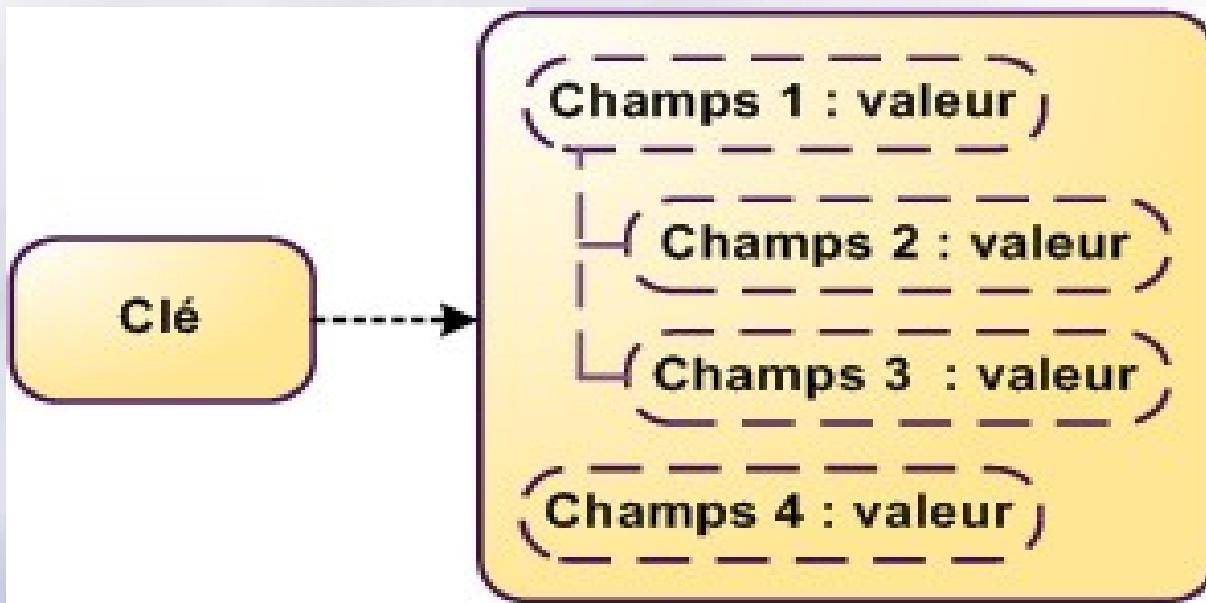
BDD Orientée graphe

# Entrepôt Clé/Valeur - ECV



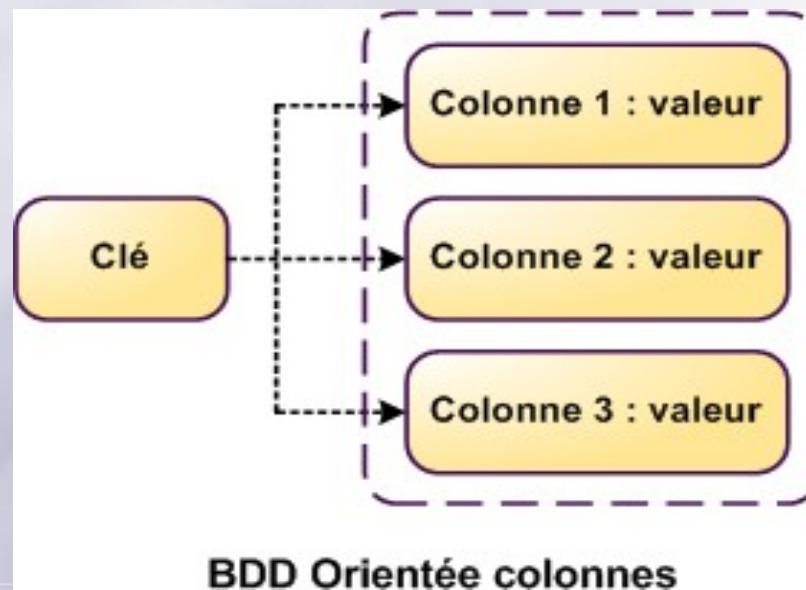
**BDD Clé-Valeur**

# Entrepôt de type Document

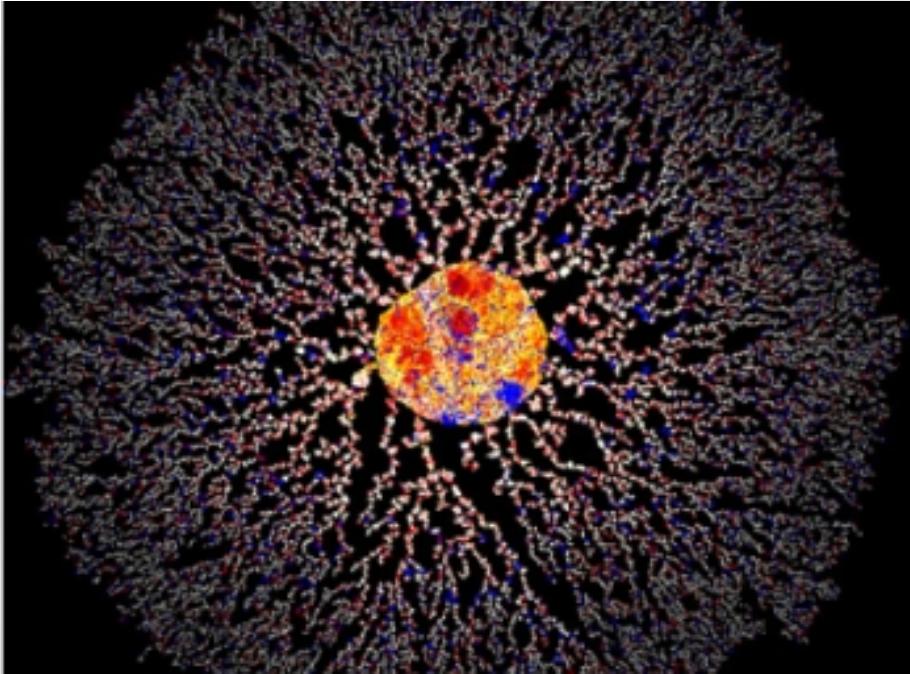


BDD Orientée document

# Entrepôt de type Colonne

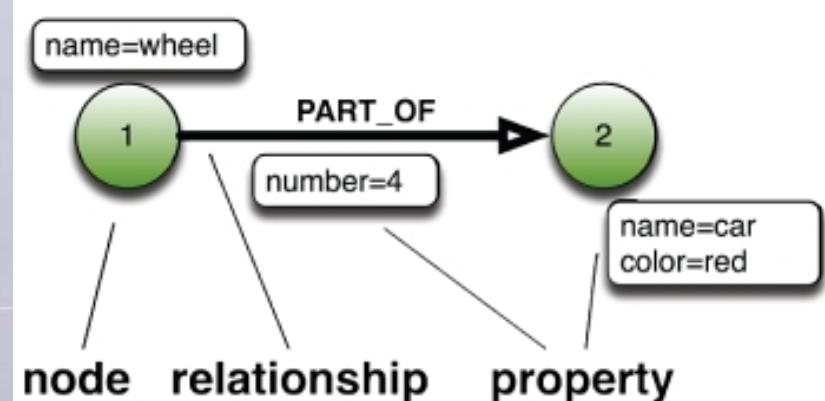


# Base de type Graphe



Exemple : Protein Homology Network

*Terminologie de base pour les graphes marqués-attribués :  
Nœud, Relation, Propriété*



# Les différents moteurs NoSQL

	Format de Stockage	Transactions	Relation	Haute Disponibilité	Fonctions avancées
Hbase	Colonnes	Non	Oui	Cluster	
Cassandra	Colonnes	Non	Oui	Cluster	Réplication Avancée
Redis	Clé/Valeur	Non	Non	Cluster	Communication Asynchrone
Riak	Clé/Valeur, Série temporelle			Cluster	Tolérance aux pannes
MongoDB	Document Format JSON	Non	Non	Cluster	Agrégation
OrientDB	Multi-Format	Oui	Oui	Cluster	Transactions et sécurité
Neo4j	Graphe	Oui	Oui	Cluster	Intégration avec d'autres solutions NoSQL

# Types de base de données NoSQL

documentaire

Graphe

Colonne

Key/value

Exemple :  
MongoDB  
CouchDB  
Cloudant

Exemple :  
Neo4j  
DEX

Exemple :  
Cassandra  
Hbase  
BigTable

Exemple :  
MemCached  
Redis  
Coherence

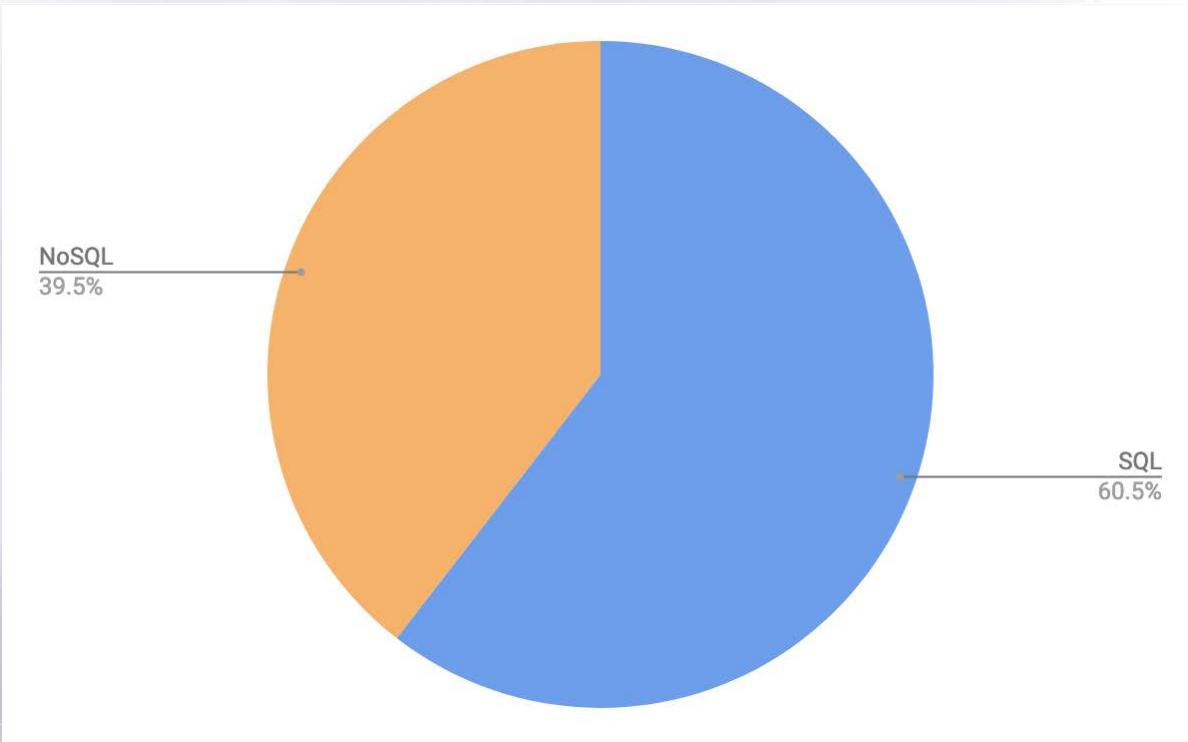
Stockage key/value

Stockage graphe

Stockage key/value

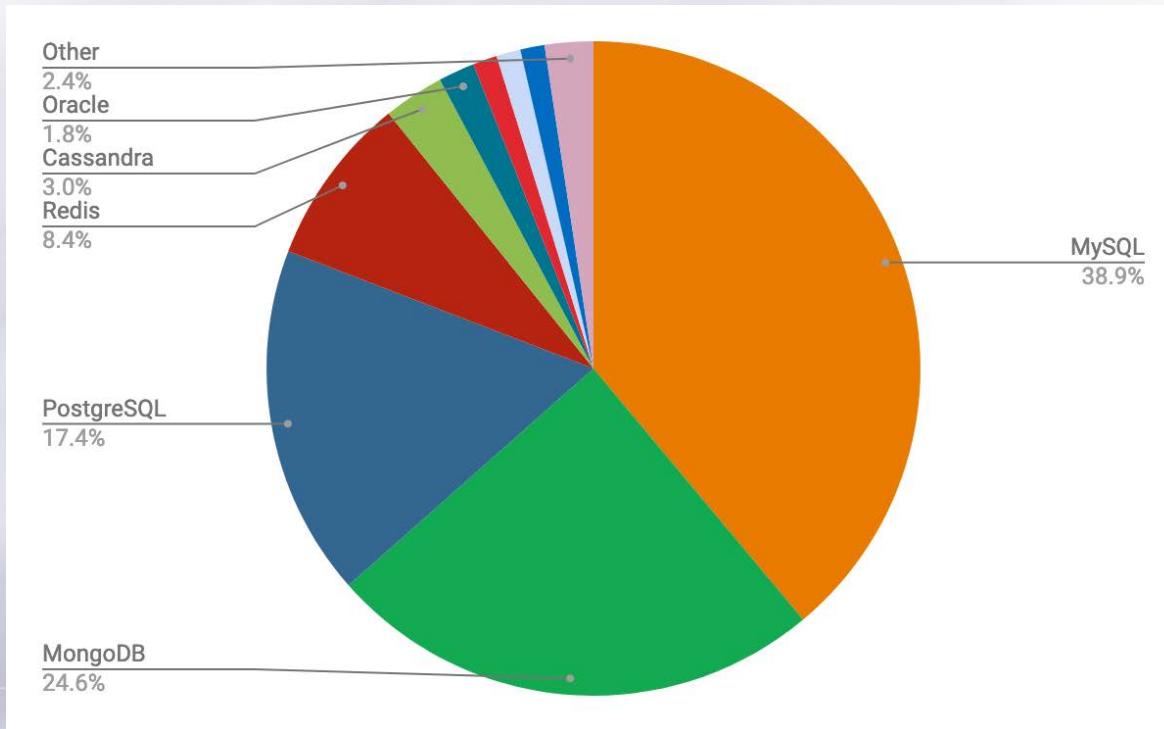
Stockage key/value

# SQL vs. NoSQL



Source : scalegrid.io, 2019

# Most Popular Databases

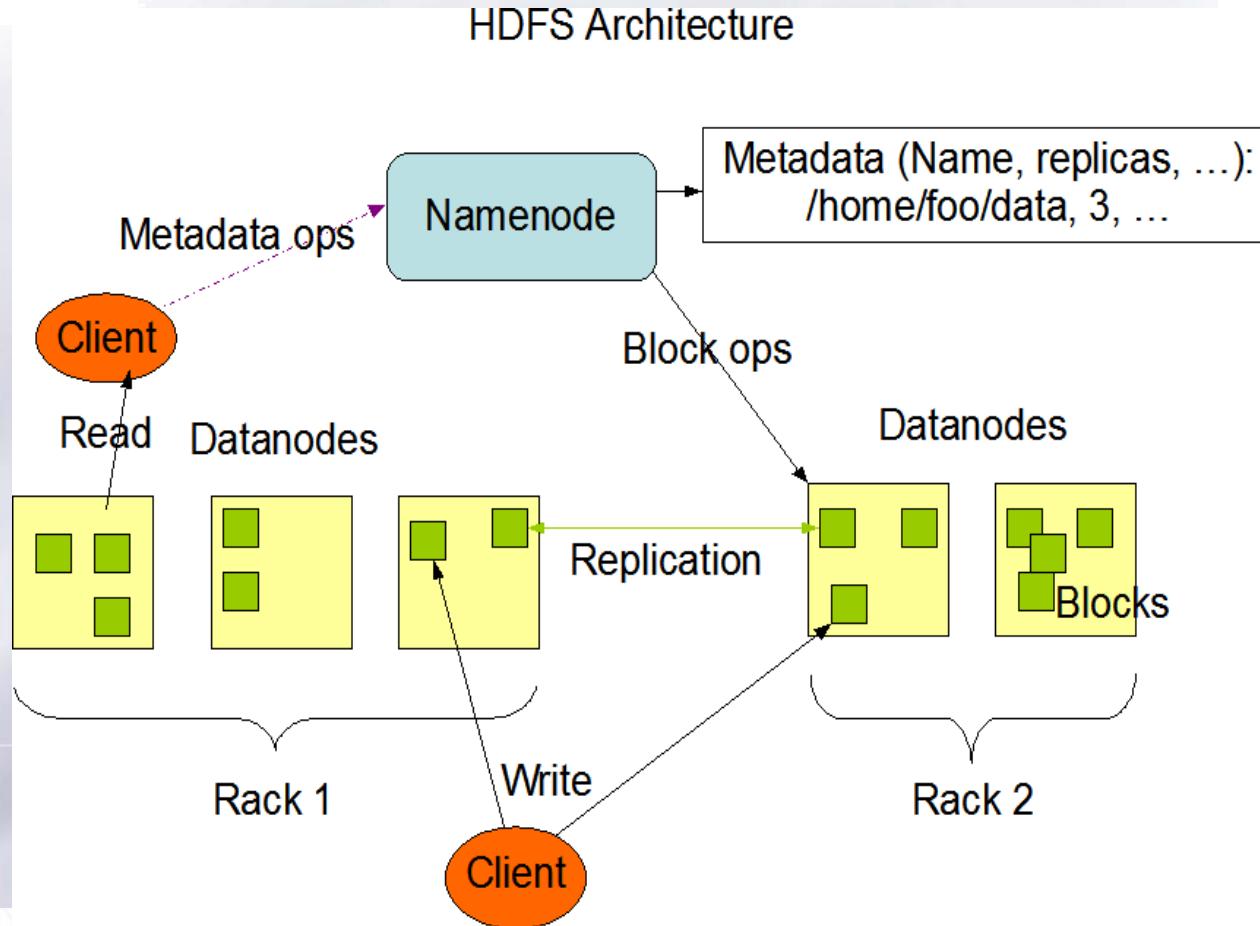


Source : scalegrid.io, 2019

**Merci de votre attention !**



# Architecture HDFS



# Architecture HDFS

"Je veux stocker 1152 Mo de données, avec une taille de block de 128 Mo et un facteur de réPLICATION de 3"



Client

Hadoop distributed file system (HDFS)

Noeud 1



Noeud 1

Données primaires  
B  
C  
D  
E  
F  
G  
H  
I

Données secondaires

Noeud 2



Données primaires  
E  
F  
A  
B  
C  
G  
H  
I

Données secondaires

Noeud 3



Données primaires  
G  
H  
I  
D  
E  
F  
A  
B  
C

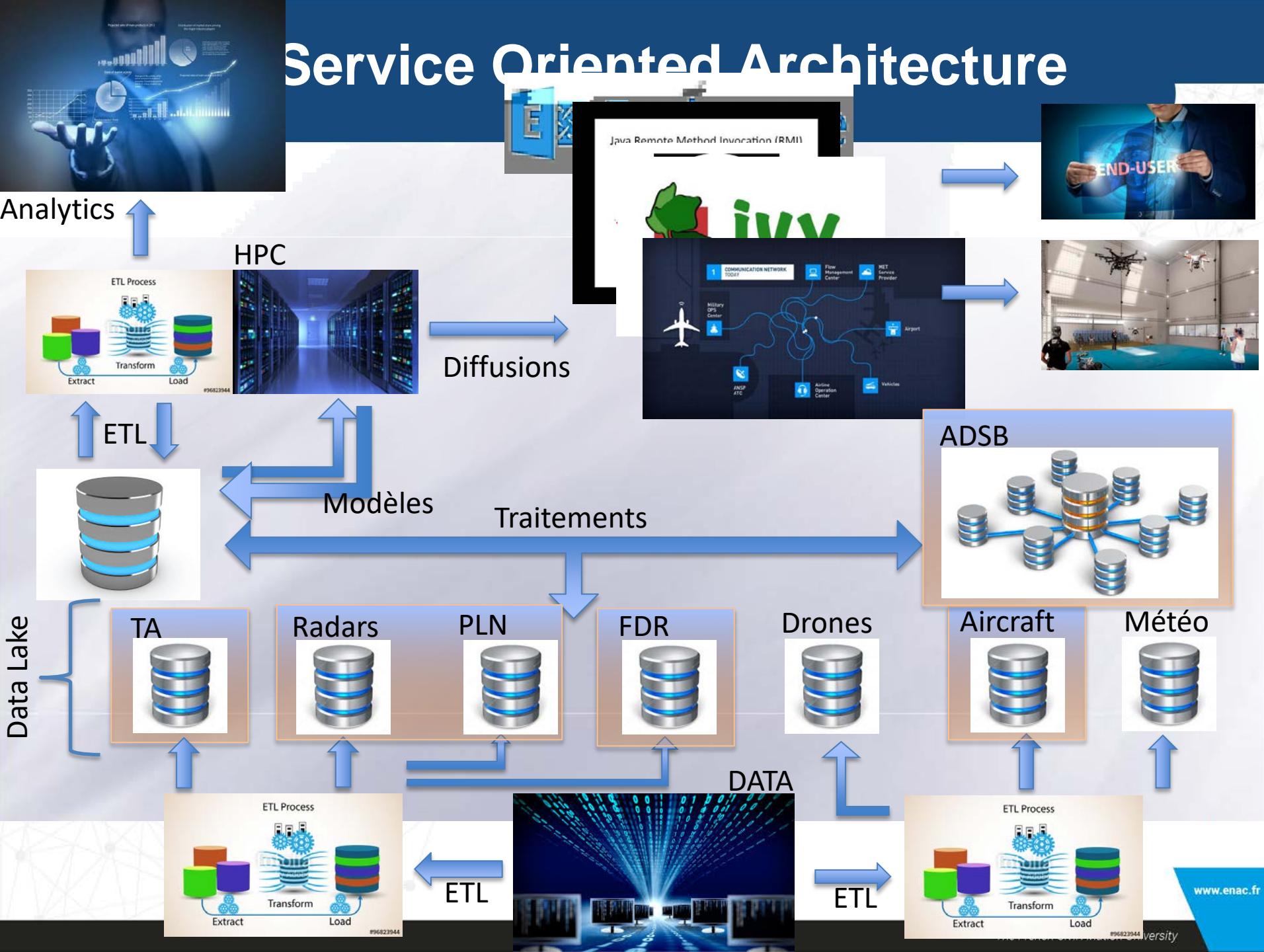
Données secondaires

## Répartition des données

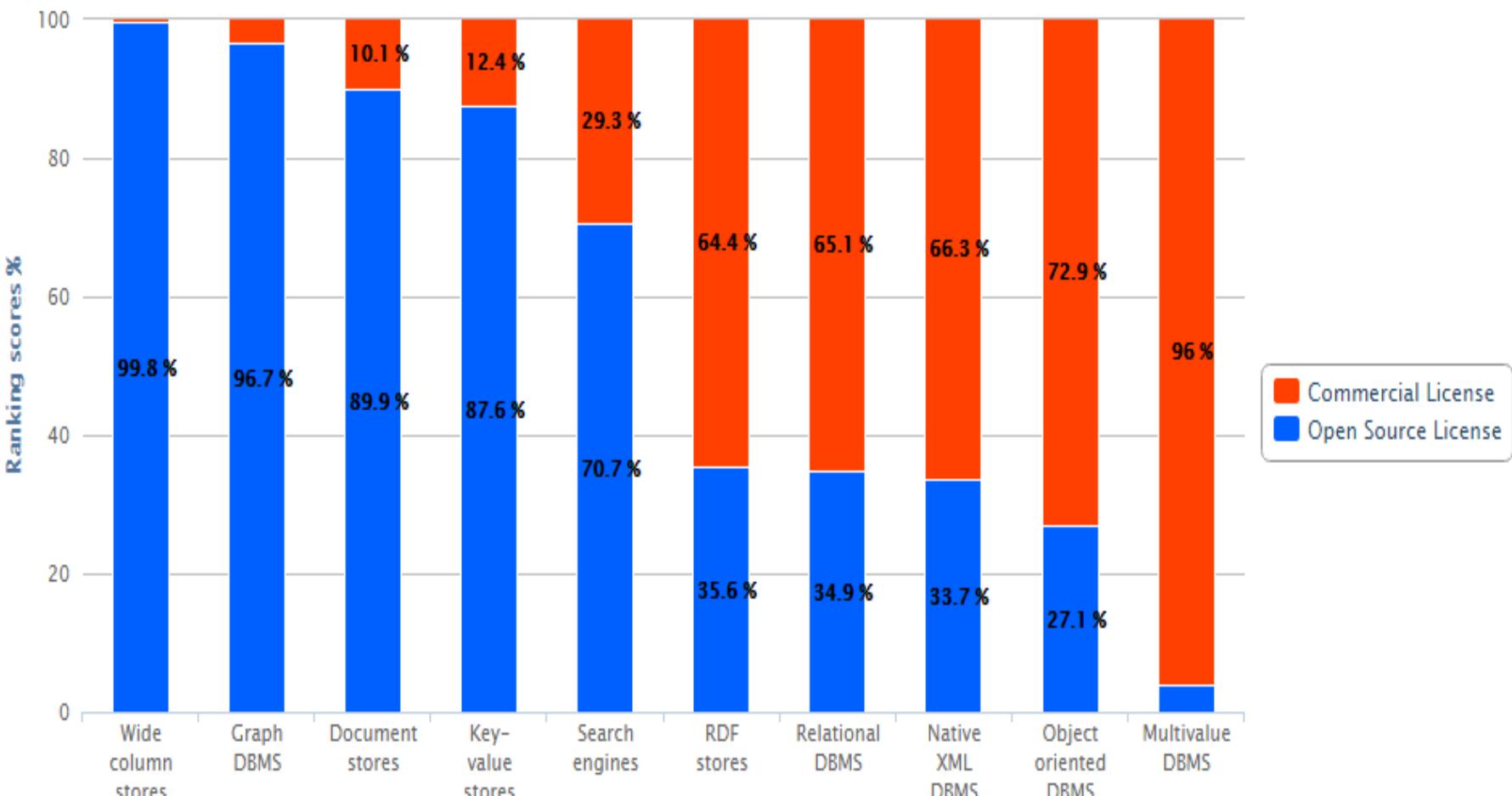
# Projet DGAC : Data Lab

- Radar primaire, secondaire, fusionné
- MLAT
- ADS-B (Flight Radar 24, OpenSky, etc.)
- Aéroports (Elvira)
- Structures de l'espace aérien (DDR2, CESNAC)
- Traffic aérien
- Structures avions : projet ANR HKUST
- Paramètres de vols : Corsair, Nouvelair, Aigle Azur, etc.
- Drones
- Données expérimentations : ACHIL par exemple
- Météo : METAR, Météo France, NASA, etc.

# Service Oriented Architecture

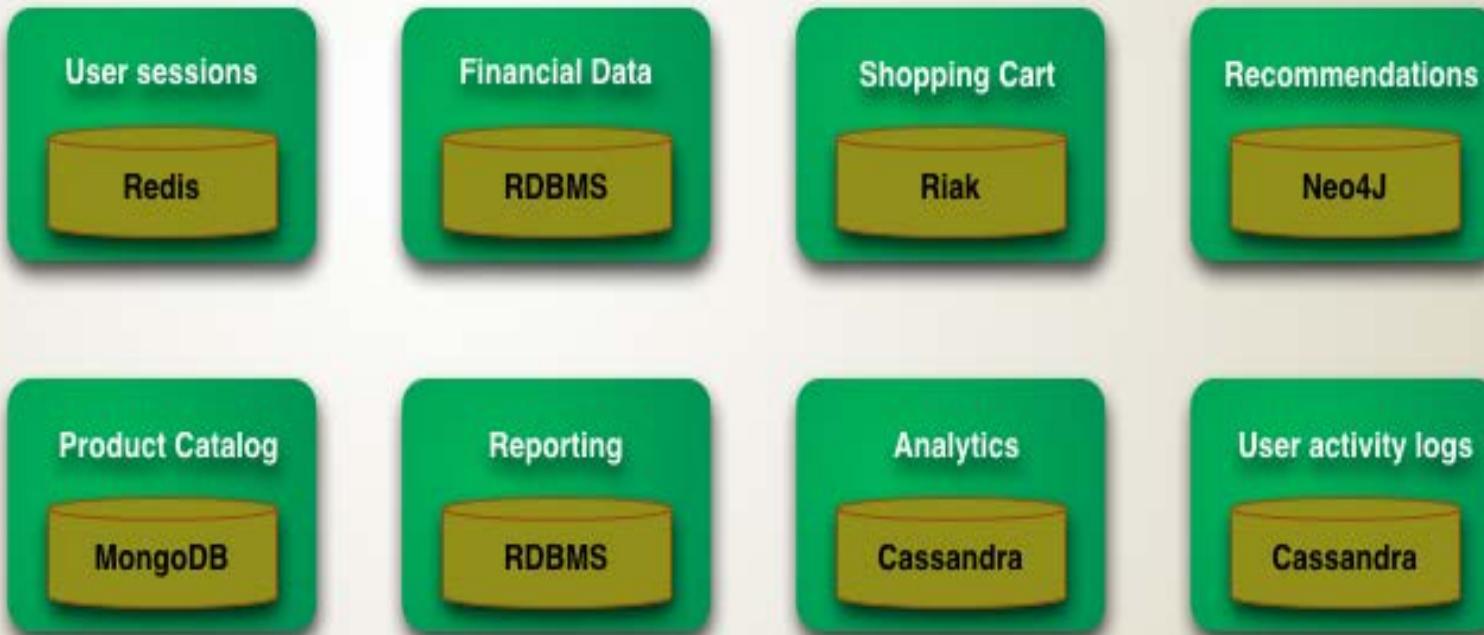


# Classement d'utilisation des différentes bases de données

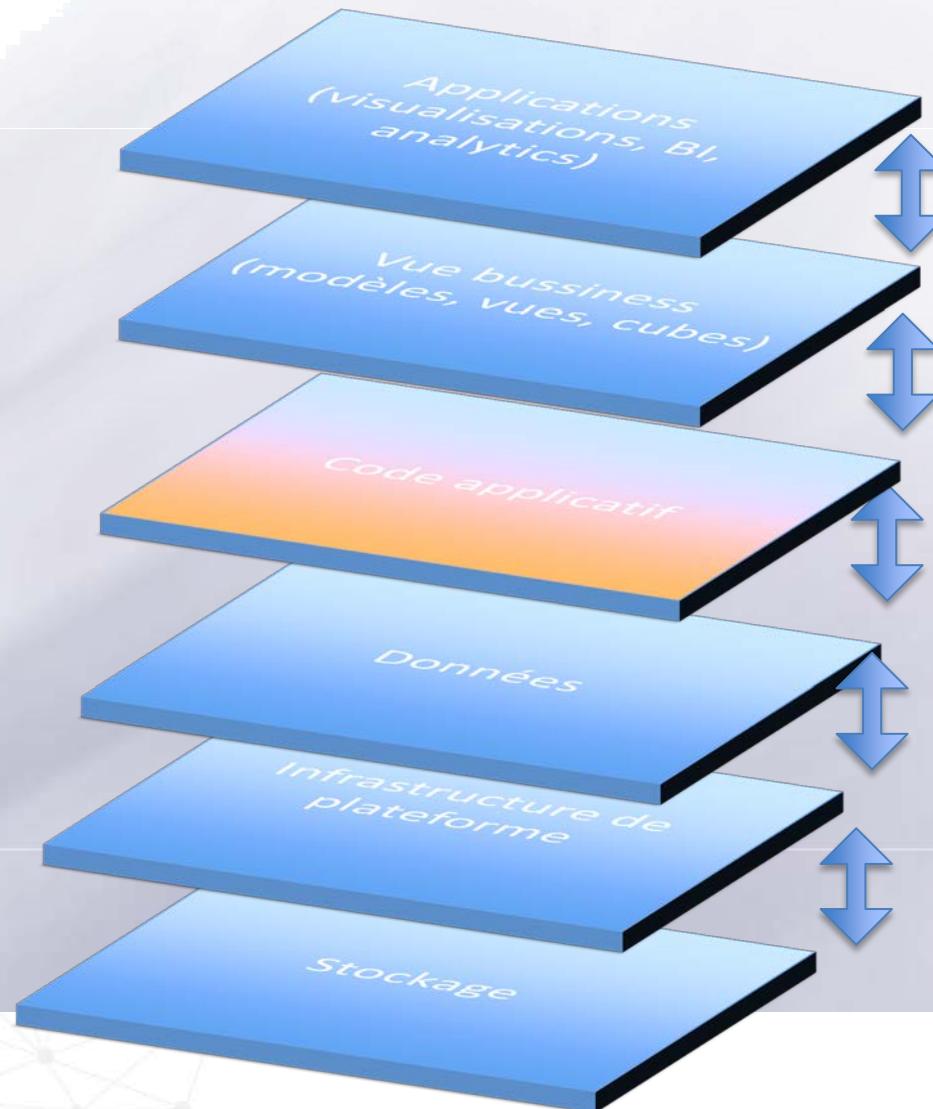


# Exemple d'utilisation de la persistance polyglotte

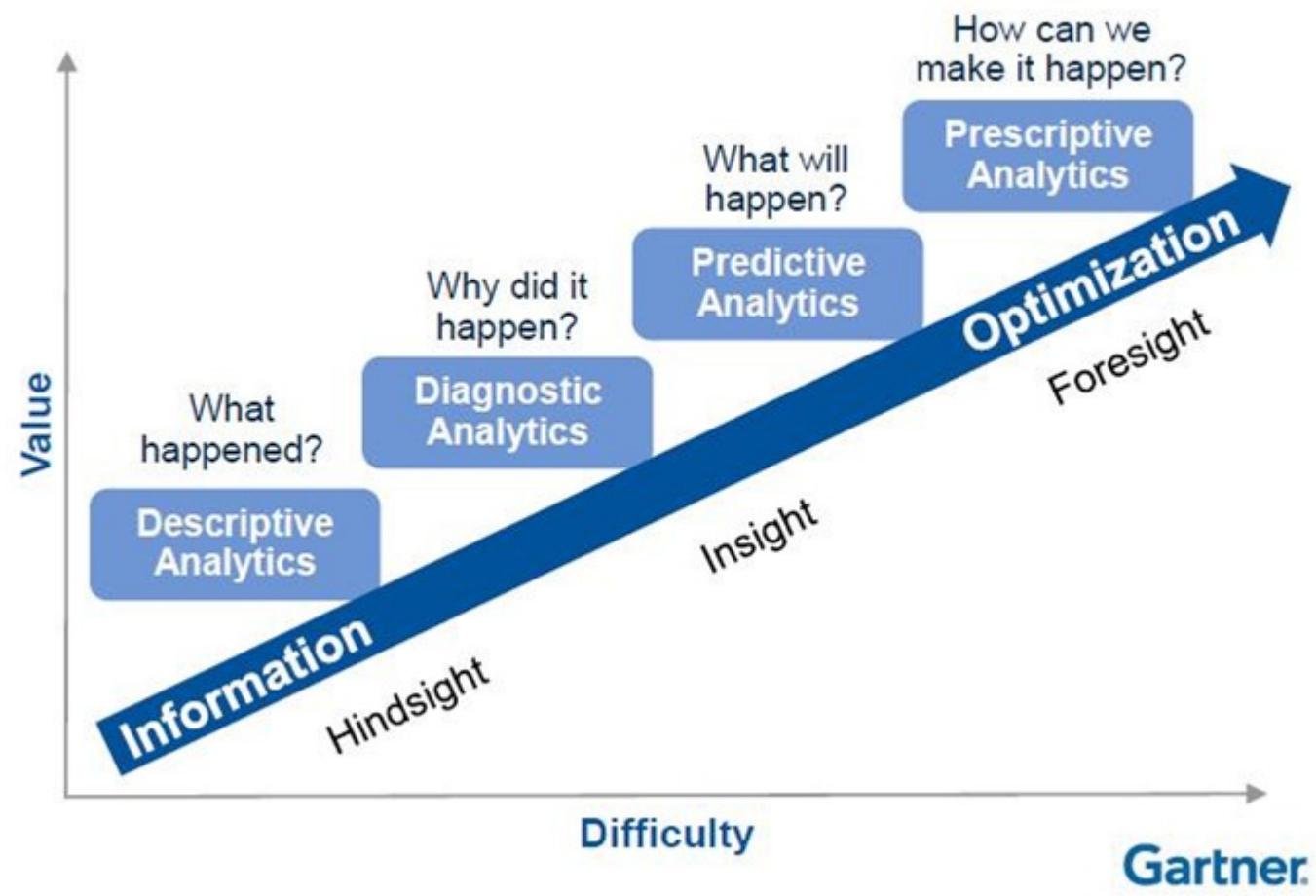
## Speculative Retailers Web Application



# Couche « Données »

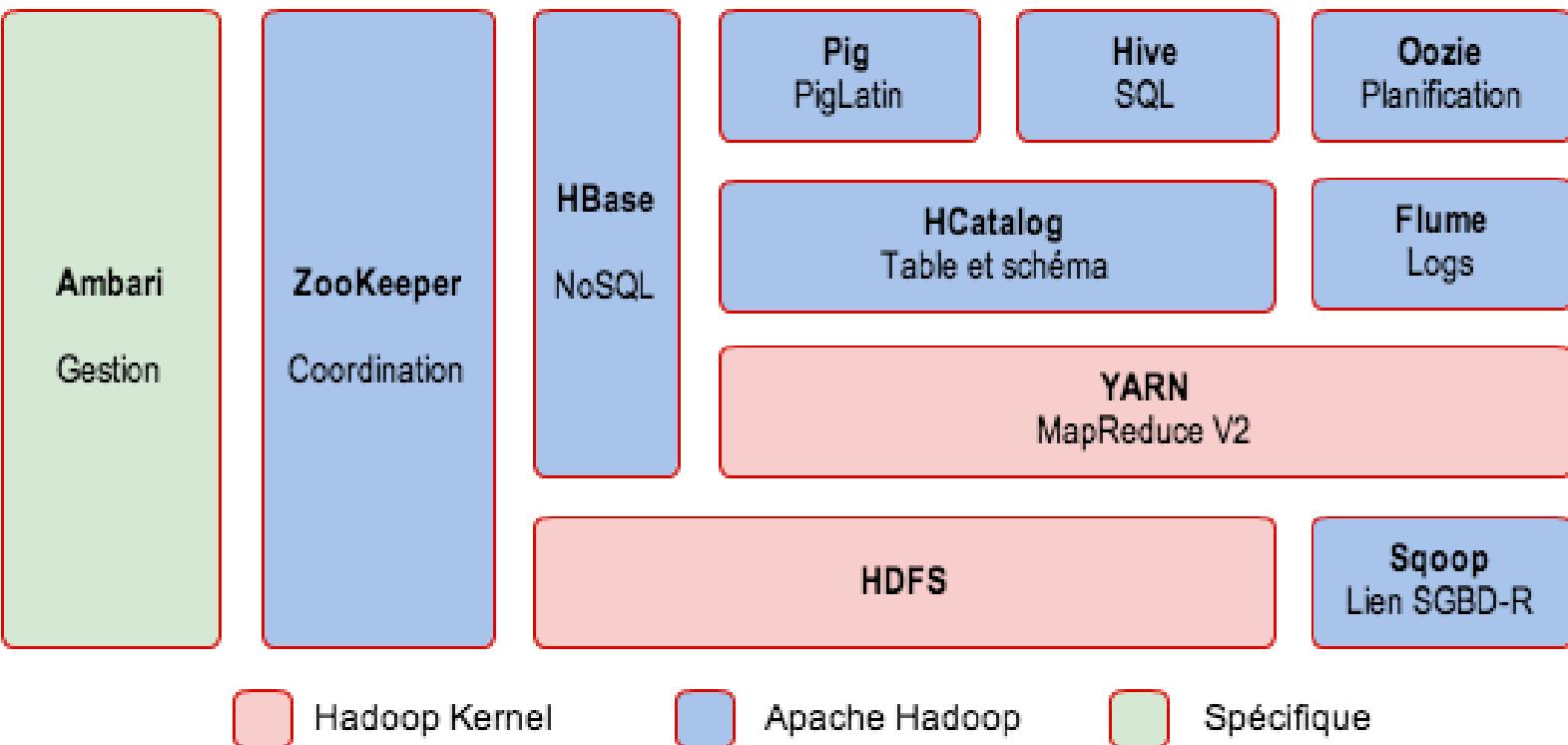


# De l'information à l'optimisation



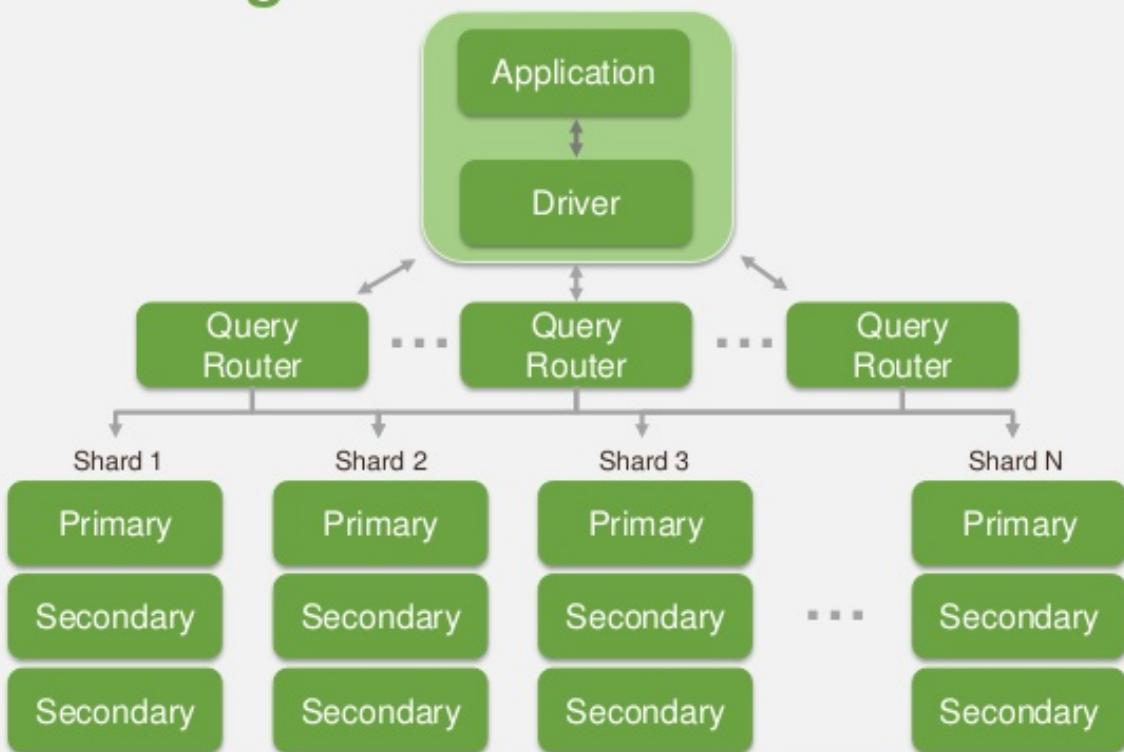
# Hortonworks

## HortonWorks Hadoop Platform (HDP)

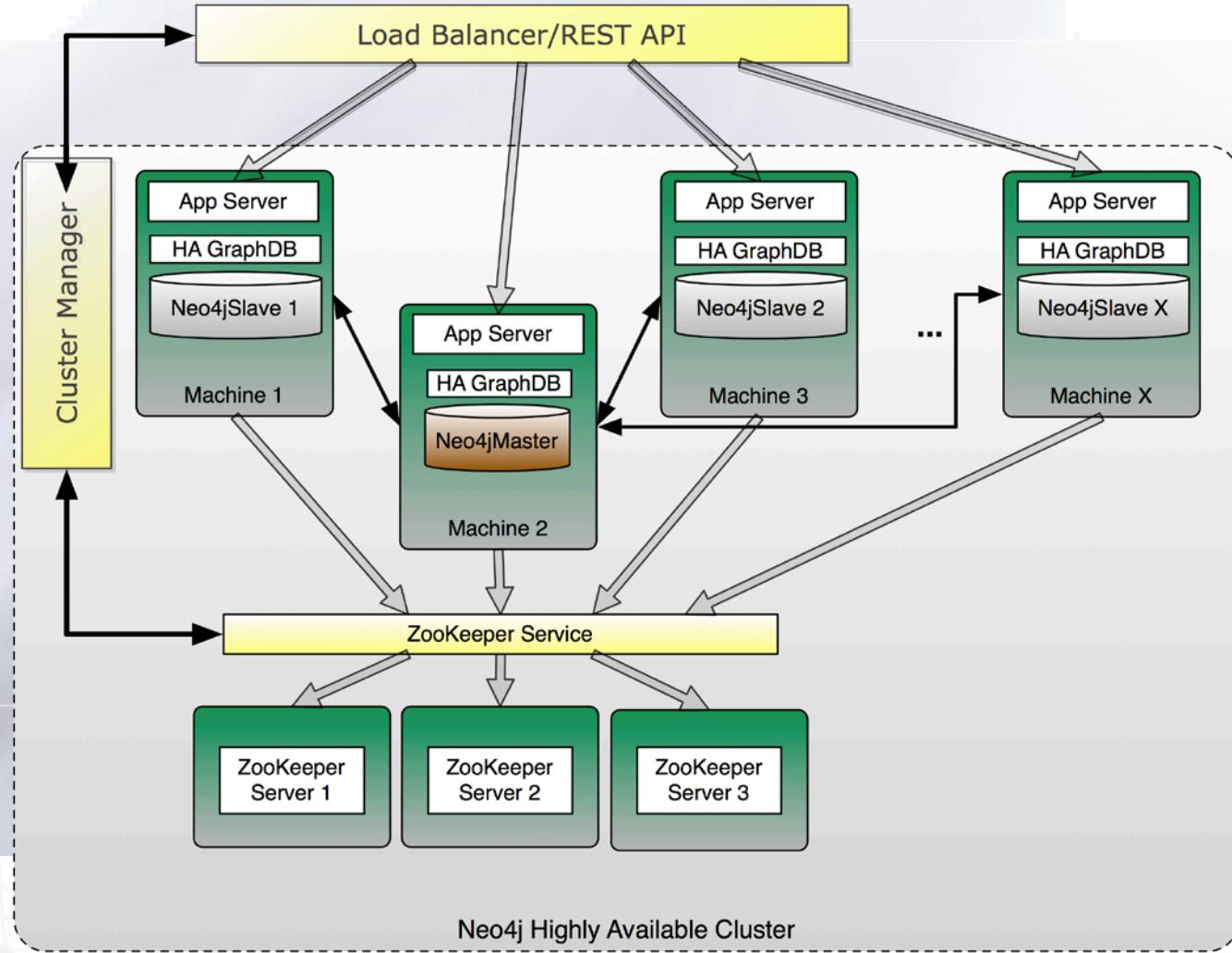


# MongoDB

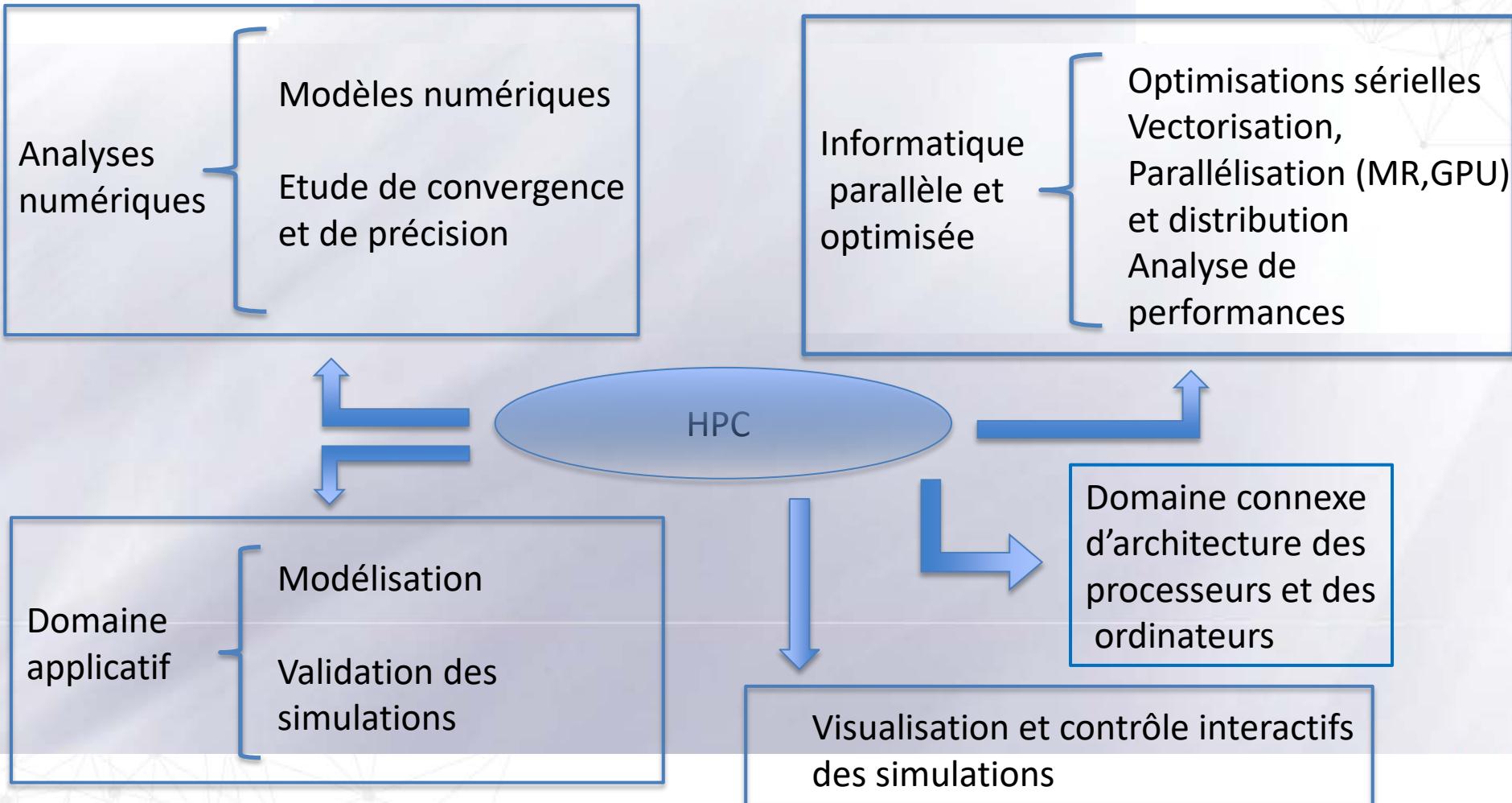
## Sharding Overview



# Neo4-J



# Composition pluridisciplinaire du HPC



# Les différents moteurs NoSQL

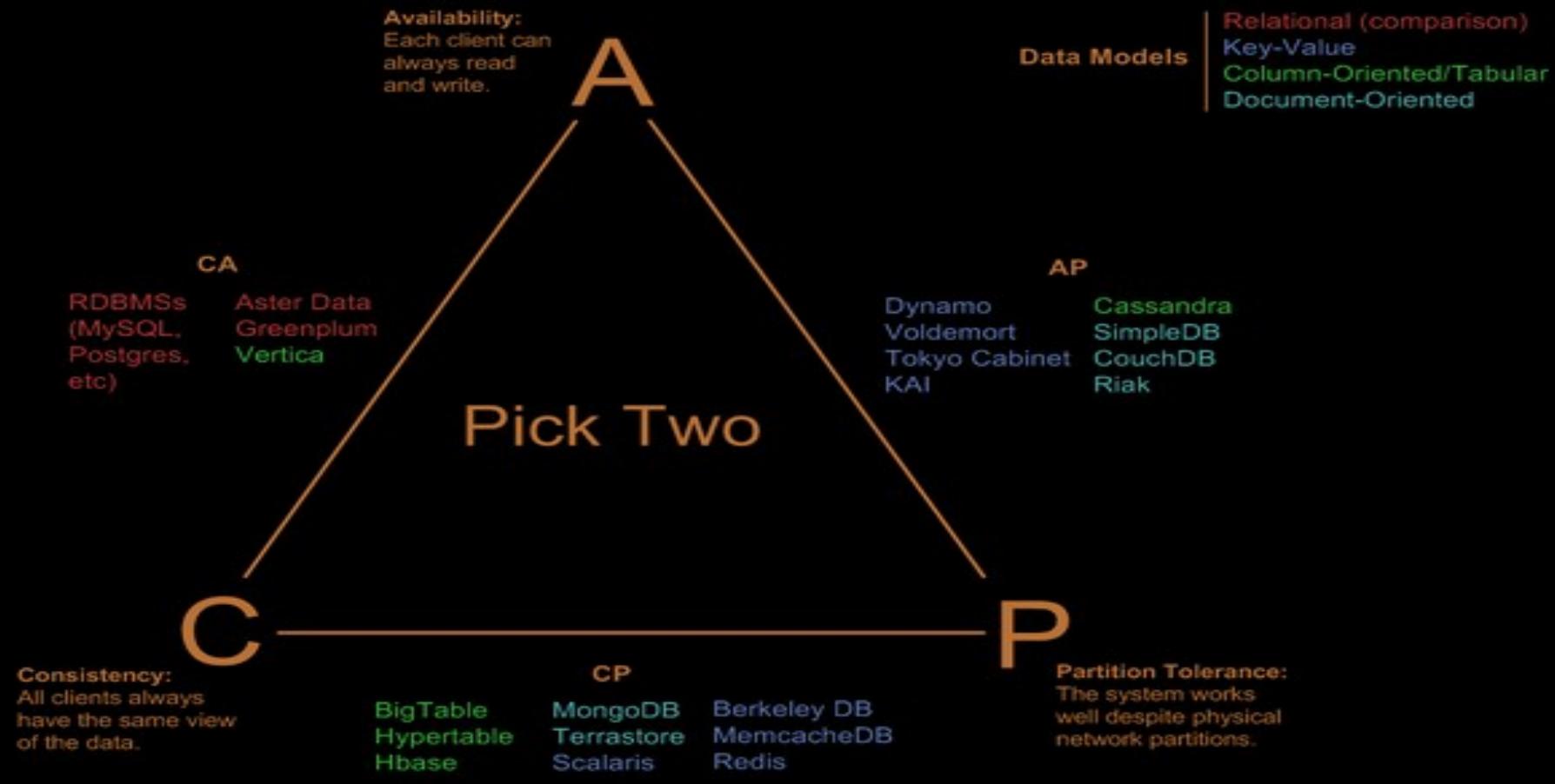
	Format de Stockage	Transactions	Relation	Haute Disponibilité	Fonctions avancées
Hbase	Colonnes	Non	Oui	Cluster	
Cassandra	Colonnes	Non	Oui	Cluster	Réplication Avancée
Redis	Clé/Valeur	Non	Non	Cluster	Communication Asynchrone
Riak	Clé/Valeur, Série temporelle			Cluster	Tolérance aux pannes
MongoDB	Document Format JSON	Non	Non	Cluster	Agrégation
OrientDB	Multi-Format	Oui	Oui	Cluster	Transactions et sécurité
Neo4j	Graphe	Oui	Oui	Cluster	Intégration avec d'autres solutions NoSQL

# Traitement par lots et temps réel

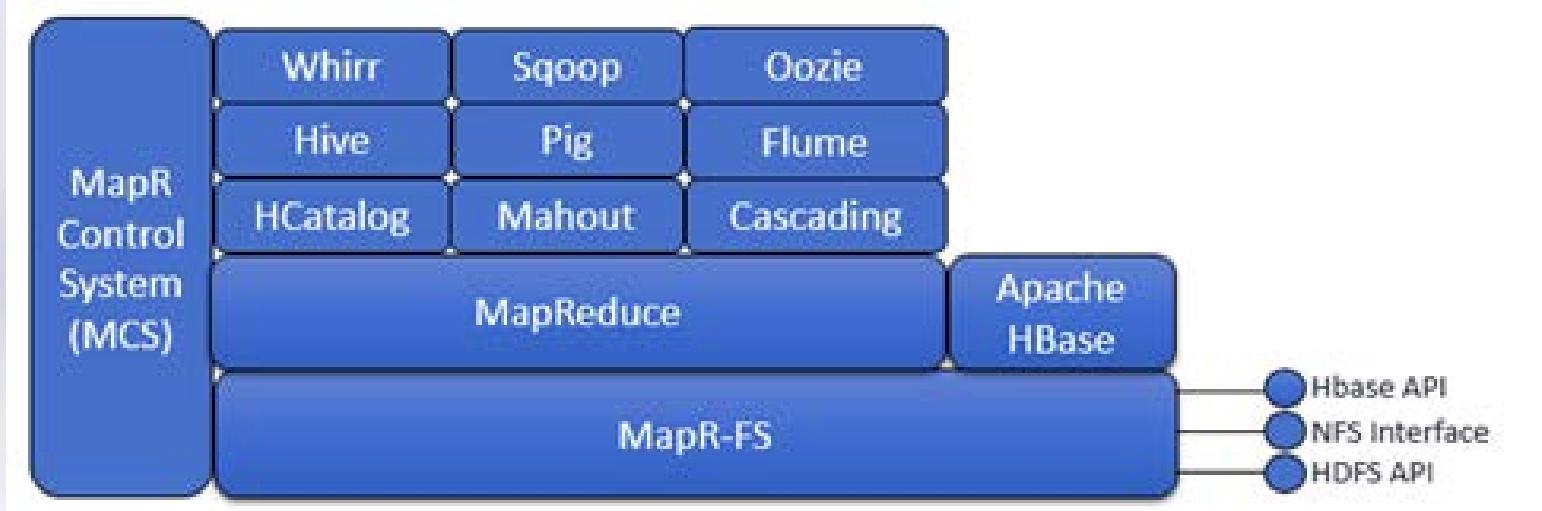


# Les architectures distribuées

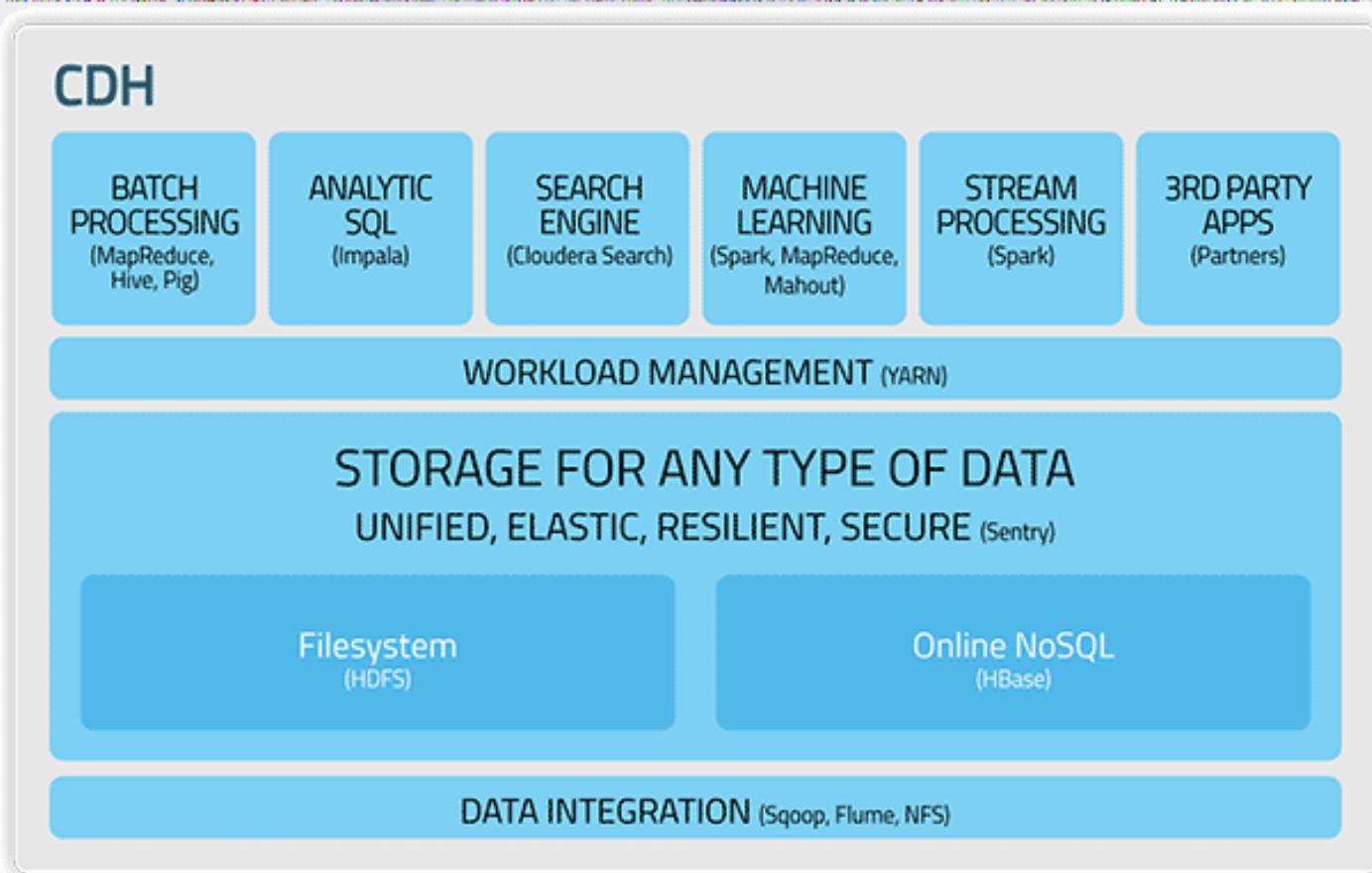
## Visual Guide to NoSQL Systems



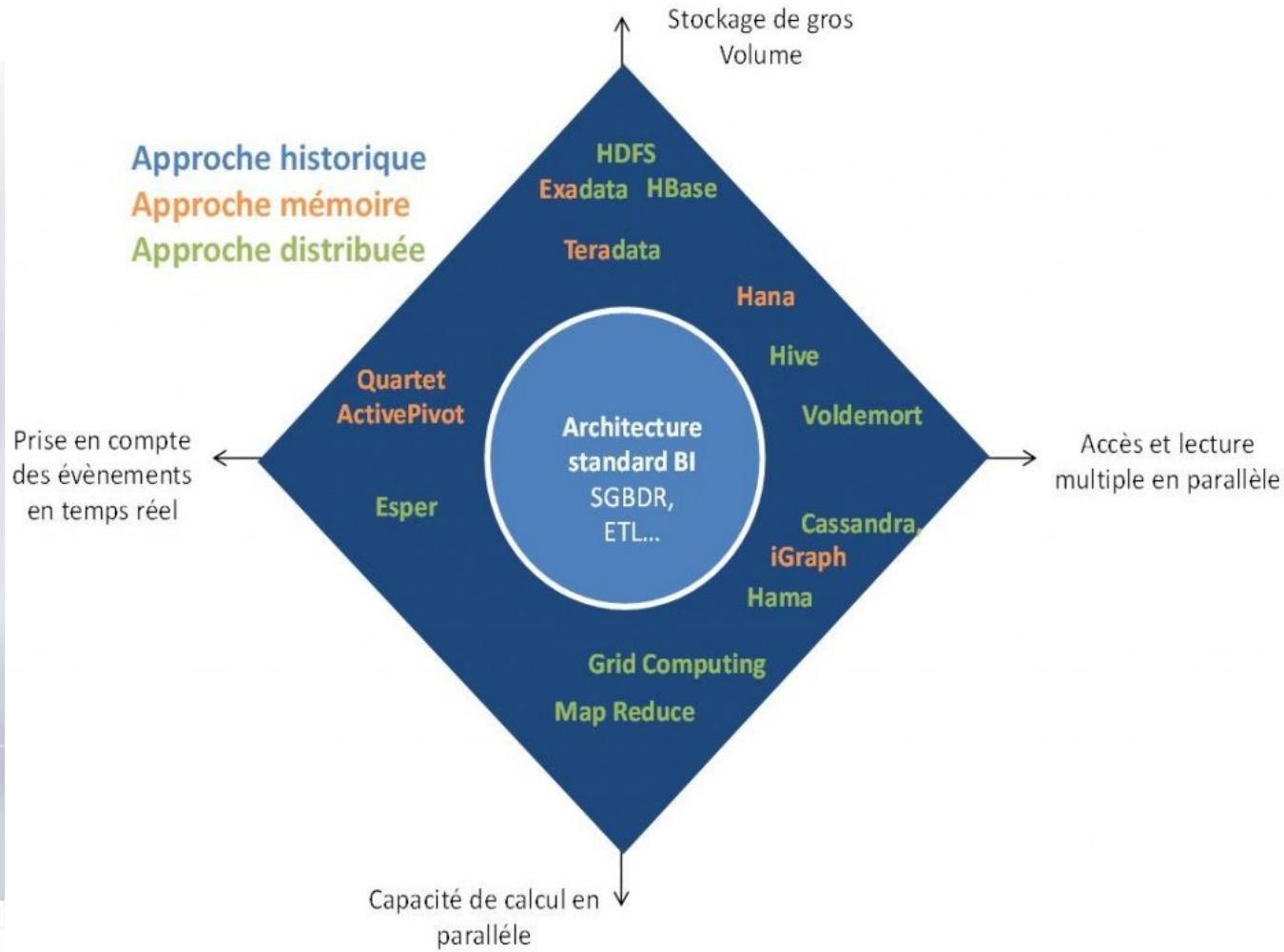
# MapR



# Cloudera ou CDH

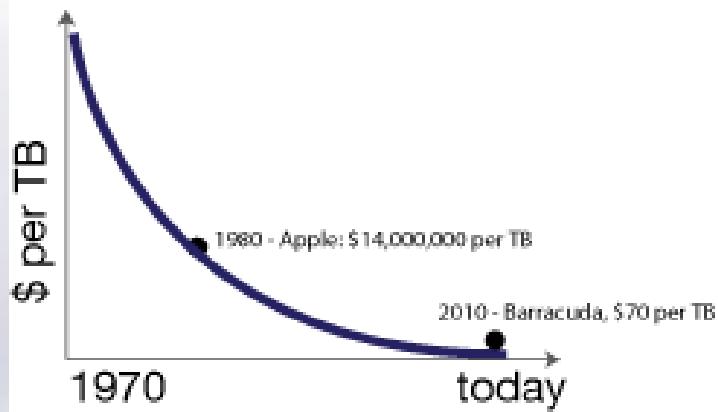


# Cartographie des solutions pour construire une architecture décisionnelle

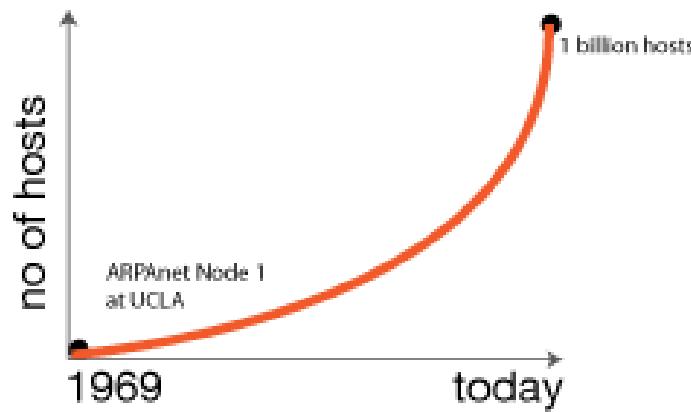


# Evolution des composants d'un serveur

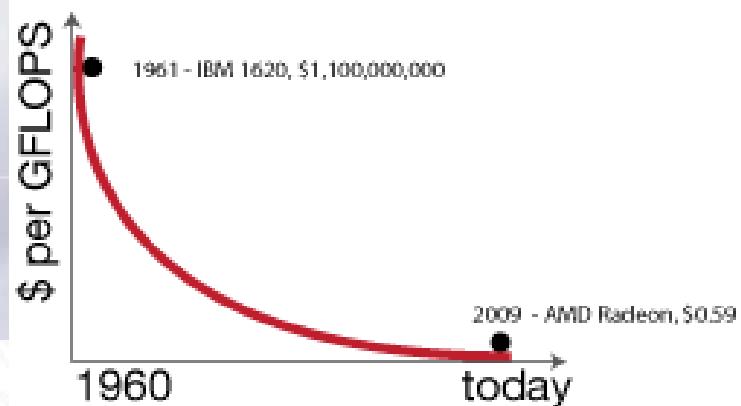
## Storage



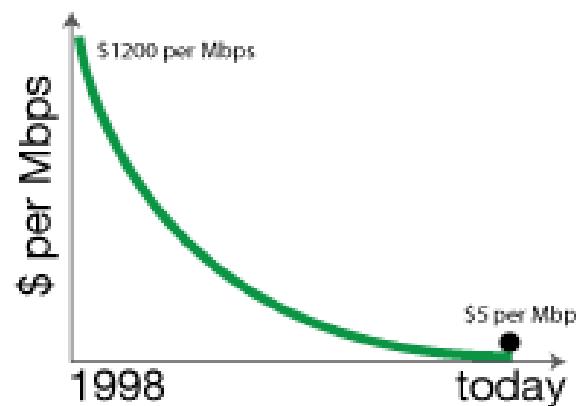
## Network



## CPU



## Bandwidth



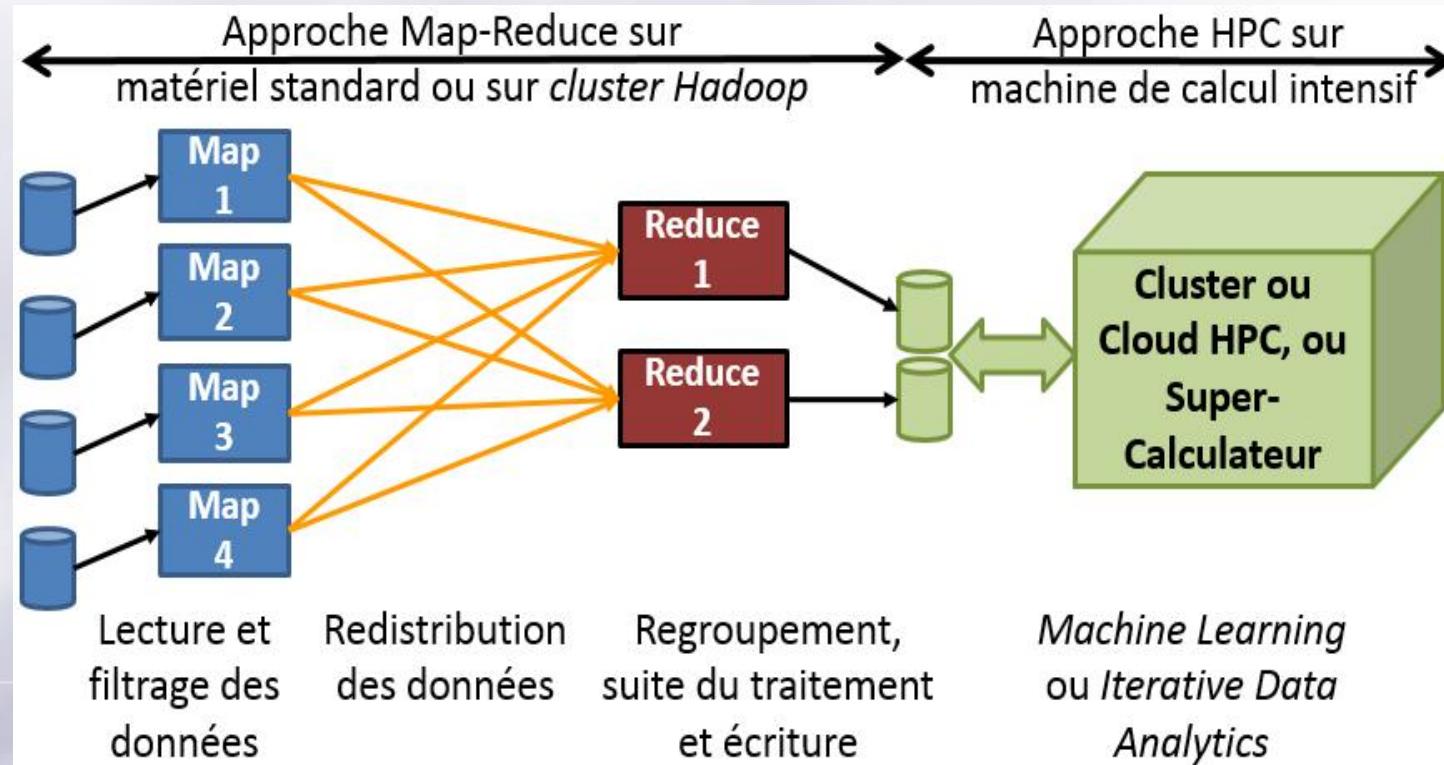
# Le goulot d'étranglement

La **capacité du débit du disque dur** est un problème qui n'a pas de solution technique aujourd'hui, « la capacité de stockage des disques a augmenté de 100 000, le débit lui n'a augmenté que de 100 ».

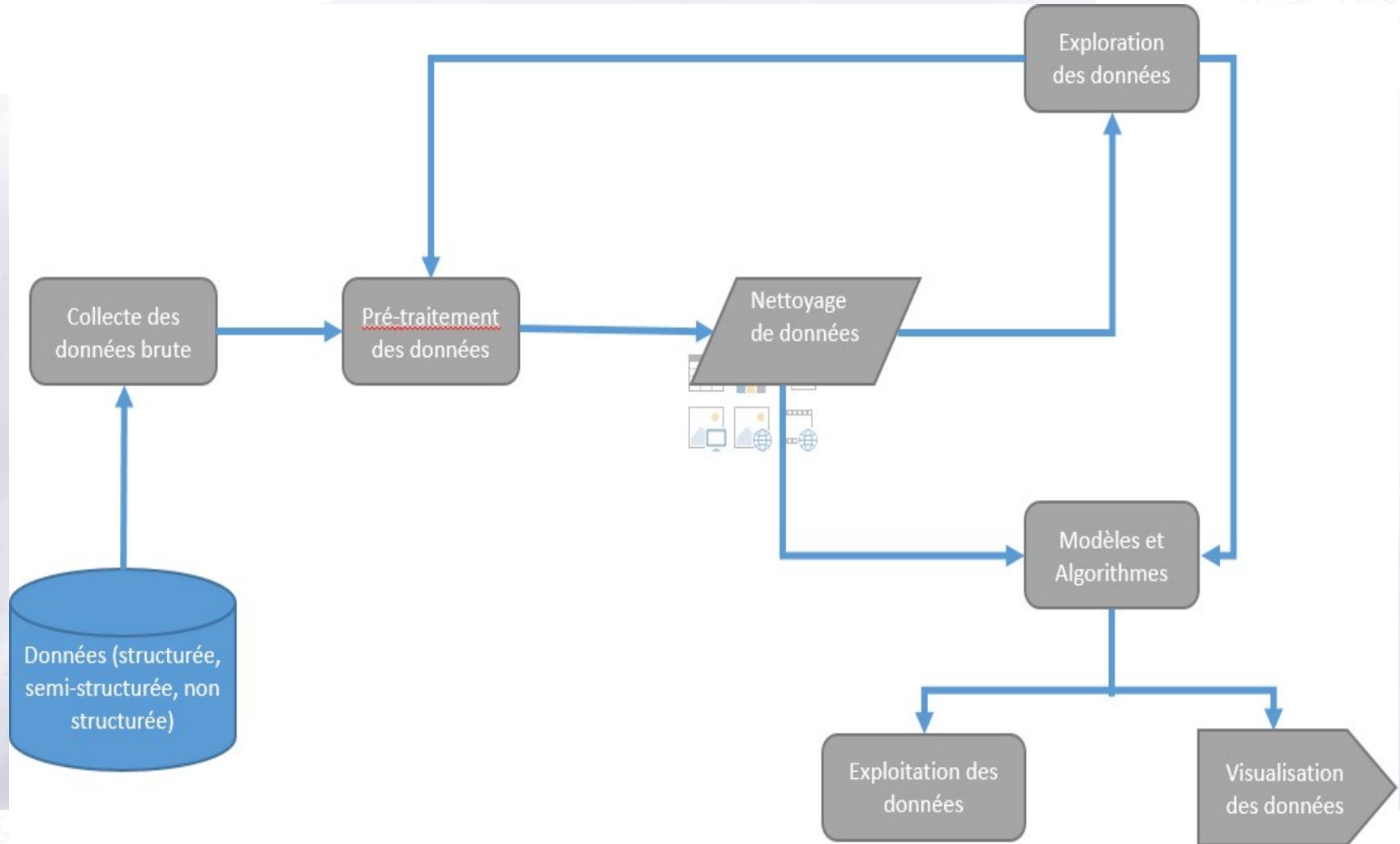
Ce problème est le goulot d'étranglement « bottleneck » des architectures actuelles.

Pour contourner ce problème, il faut soit minimiser l'utilisation du disque dur (utilisation de la mémoire) soit paralléliser le débit (architecture distribuée), afin qu'il soit acceptable, en utilisant les nouveaux outils.

# Architecture mixte d'analyse de données à haute performance



# Gestion du cycle de vie des données



# Critères du choix d'une architecture big data

Le type d'analyse ou de traitement

Méthodologie de traitement

Fréquences des données

Types des données

Format du contenu

Sources des données

L'utilisateur final de la donnée

Matériel

# Le système de stockage de fichier HDFS

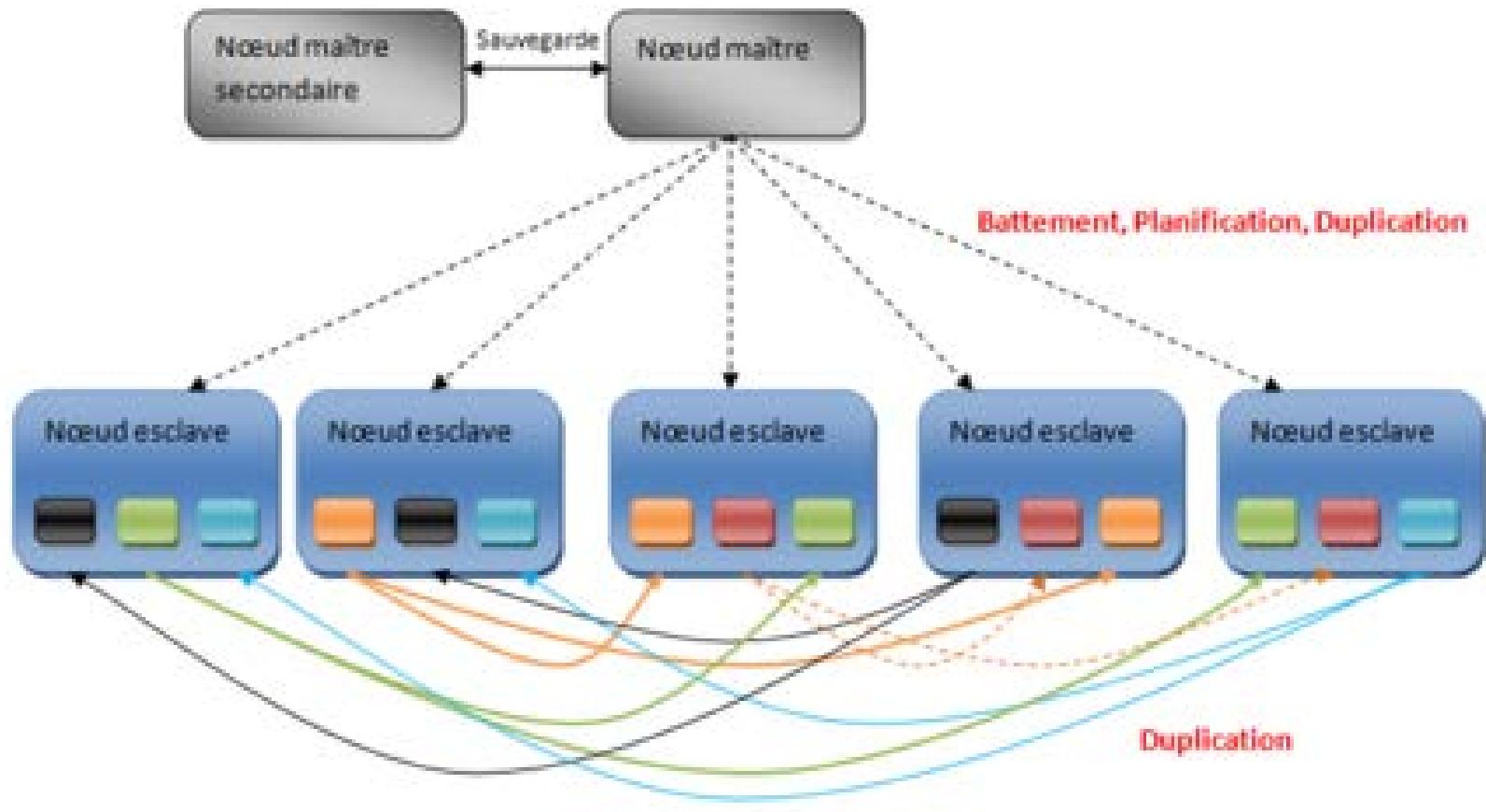
## Objectifs :

- Tolérant aux pannes d'une manière native (Fault tolerant)
- "Scalable"
- Modèle d'accès immuable
- Déplacer les calculs vers les données
- Simple à mettre en place en assurant la portabilité sur différentes plateformes.

## Fonctionnalités :

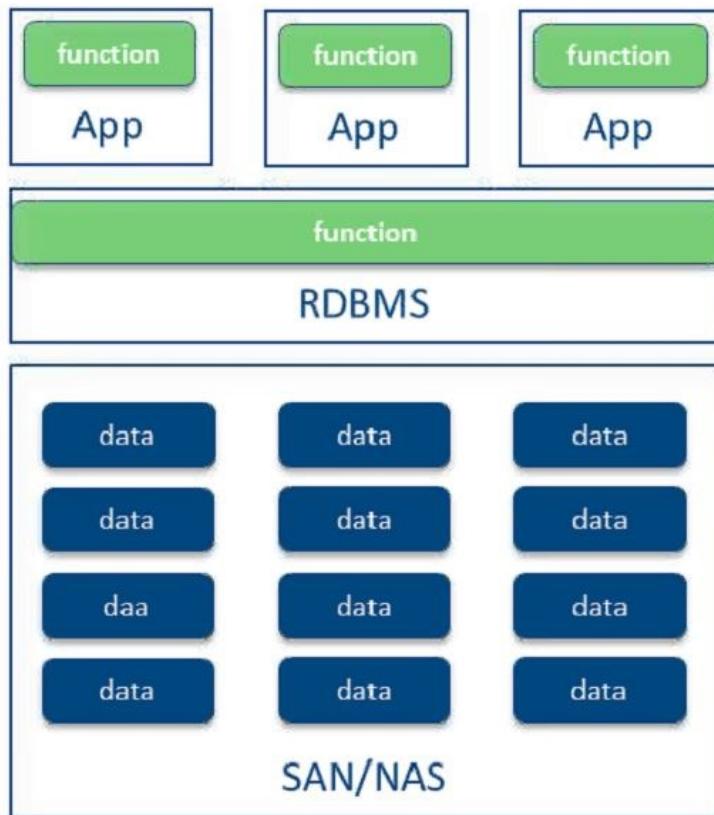
- Gestion des fichiers par blocs
- Réplication et distribution
- Gestion des droits
- Accès aux données en continu (Streaming)
- Stockage des grands jeux de données

# Une solution technique : Hadoop

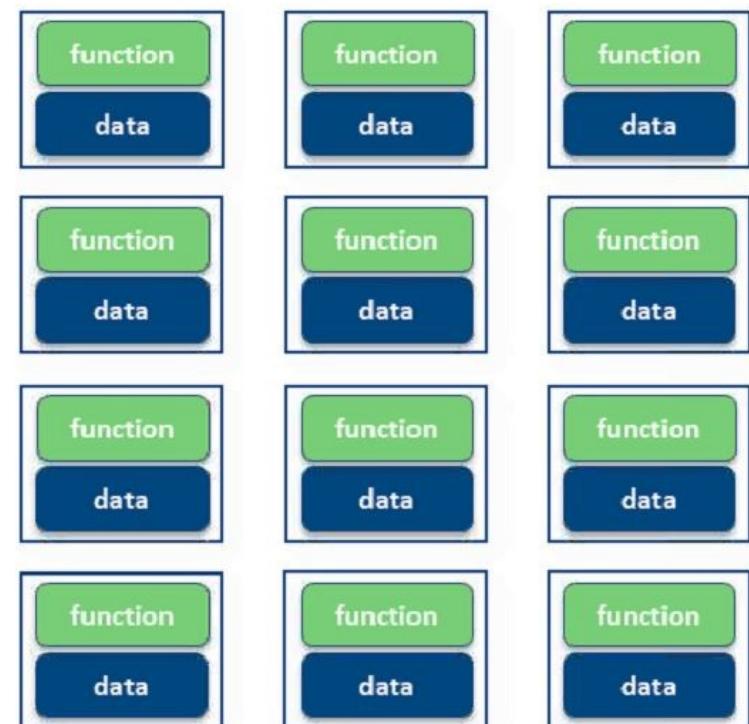


# Architecture HDFS

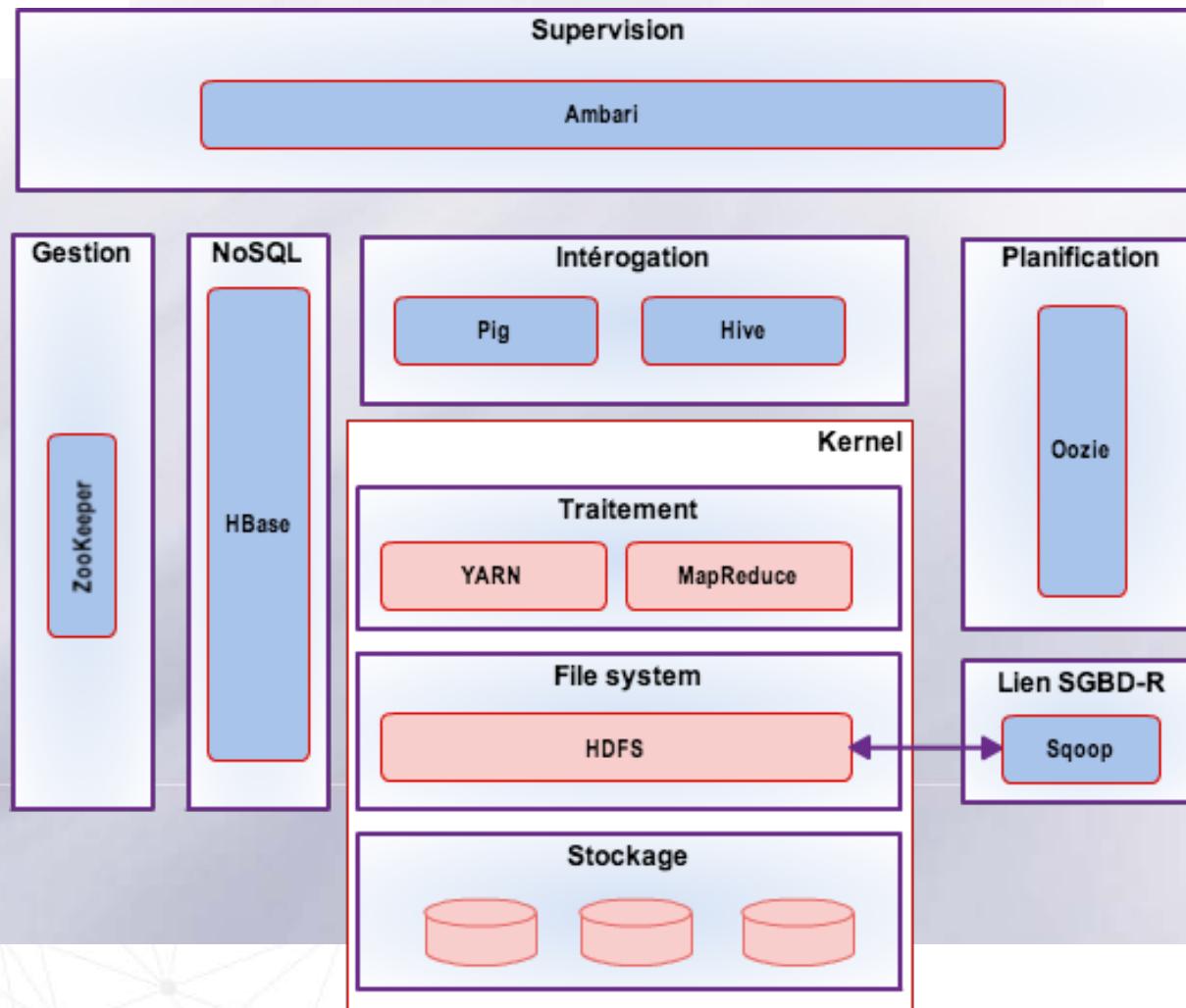
Traditional Architecture



Hadoop



# Vue d'ensemble de la plate forme Hadoop



**Données de grandes dimensions,  
Valeurs, Documents, Colonnes et  
Graphes, Cloud, HPC, etc**

**Ce n'est que le début de l'histoire .... !!**

# Volumétries des données

**8 bits : octet**

**$10^3$  bits – Un kilobit**

**$10^6$  bits – Un mégabit**

**$10^9$  bits – Un gigabit**

**$10^{12}$  bits – Un térbabit**

**$10^{15}$  bits – Un pétabit**

**$10^{18}$  bits – Un exabit**

**$10^{21}$  bits – Un zettabit**

**$10^{24}$  bits – Un yottabit**

# Ordres de grandeur de volumes de données

- **7 bits** – Taille des caractères dans la table des caractères ASCII.
- **32 bits** – Taille des adresses de l'IPv4, le protocole Internet actuel.
- **128 bits (16 octets)** – Taille des dresses de l'IPv6, le protocole Internet émergeant.
- **1 288 bits (210 bits, 128 octets)** – Capacité maximale approximative d'une carte à bande magnétique standard.
- **4 704 bits (588 octets)** – Longueur de frame non compressée d'un canal unique dans un fichier standard audio MPEG (75 frames par seconde et par canal), avec une qualité moyenne de 8-bit échantillonné à 44 100 Hz (ou 16-bit échantillonné à 22 050 Hz).
- **1 978 560 bits** – Un fax d'une page, en résolution standard noir et blanc ( $1\ 728 \times 1\ 145$  pixels).
- **11 796 480 bits** – Capacité d'une disquette 3.5", familièrement connue comme 1,44 mégaoctet mais actuellement  $1,44 \times 1000 \times 1\ 024$  octets.
- **$5,45 \times 10^9$  bits (650 mébioctets)** – Capacité d'un compact disc habituel.
- **$6,4 \times 10^9$  bits** – Capacité du génome humain, 3,2 milliards de paires de bases (chaque paire compte pour 2 bits de donnée).

# Ordres de grandeur de volumes de données

- **$4,04 \times 10^{10}$  bits (4,7 gigaoctets)** – Capacité d'un DVD simple face, simple couche.
- **$2,16 \times 10^{10}$  bits (2,7 gigaoctets)** – Taille de la Wikipédia anglaise sans les images (Compressée, elle fait 1,1 gibioctets).
- **$1,46 \times 10^{11}$  bits (17 gigaoctets)** – Capacité d'un DVD double face, double couche DVD.
- **$2,15 \times 10^{11}$  bits (25 gigaoctets)** – Capacité d'un disque Blu-ray simple face, simple couche de 12-cm
- **$1 \times 10^{13}$  bits (1,25 téraoctets)** – Capacité de la mémoire fonctionnelle d'un être humain, selon Raymond Kurzweil dans The Singularity Is Near, p. 126.
- **$1,5 \times 10^{14}$  bits (18,75 téraoctets)** – La quantité d'informations dans la Bibliothèque du Congrès aux États-Unis, si elle était entièrement numérisée
- **25 To** : The Millennium Simulation Project : 10 milliards de particules pour retracer l'évolution de la matière sur un cube de 2 milliard al (<https://wwwmpa.mpa-garching.mpg.de/galform/virgo/millennium/>)
- **117 To** : Plateforme Galactica, une capacité de stockage à l'échelle des études astronomiques
- **$1.6 \times 10^{18}$  bits (200 pétaoctets)** – La quantité totale de matériel imprimé dans le monde.
- **$2,36 \times 10^{21}$  bits (295 exaoctets)** – Évaluation de la quantité d'information qui a été stockée entre 1986 et 2007

# Ordres de grandeur de volumes de données

- **$2,36 \times 10^{21}$  bits (295 exaoctets)** – Évaluation de la quantité d'information qui a été stockée entre 1986 et 2007
- **1,3 zettaoctets** – Prévision (par une étude de [Cisco](#)) du seul trafic Internet annuel mondial en 2016
- **$1,52 \times 10^{22}$  bits (1,9 zettaoctets)** – La quantité d'information qui a été diffusée (télévision, GPS, Internet, ...) en 2007
- **$1,8 \times 10^{22}$  bits (2,25 zettaoctets)** – Évaluation de la quantité d'information qui peut être stockée dans un 1 [gramme d'ADN](#)

# Glossaire

ERP : en anglais pour Enterprise resource Planning.

CRM : en anglais pour Customer Relationship Management.

SCM : en anglais pour Supply Chain Management.

DW : en anglais pour Data Warehouse.

BI: Business Intelligence.

ETL: Extract Transform Load.

OLTP : On Line Transactional Processing.

OLAP : On Line Analytical Processing.

Datamart : magasin de données.

Batch : traitement par lots.

Cluster : ensemble de machines indépendantes situées dans un réseau dans le même espace géographique.

DHT : Distributed Hash Table.

Appliance : matériel spécifique.

MPP : Massive Parallel Processing, Traitement Massivement Parallèle, chacun avec son CPU, mais partage la mémoire.

SSD : Solid State Disk.

# Glossaire

CPU : Central Processing Unit.

RAM : Random Access Memory pour mémoire vive.

DCC : Distributed Computing Cluster.

Grid Computing (GRID) : un DCC s'exécutant sur Internet.

Stream processing : traitement de flux.

# Pour aller plus loin ...

- Pavlo A., Paulson E., Rasin A., Abadi D. J., Dewitt D. J., Madden S., and Stonebraker M., *A Comparison of Approaches to Large-Scale Data Analysis,* "Proceedings of the 2009 ACM SIGMOD, <https://web.archive.org/web/20090611174944/http://database.cs.brown.edu:80/sigmod09/benchmarks-sigmod09.pdf>
- Becla, J., et al. 2006, *Designing a multi-petabyte database for LSST*, <http://arxiv.org/abs/cs/0604112>
- Becla, J., & Wang, D. L. 2005, *Lessons Learned from Managing a Petabyte*, downloaded from <https://web.archive.org/web/20110604223735/http://www.slac.stanford.edu/pubs/slacpubs/10750/slac-pub-10963.pdf> on 2007-11-25.
- Bell, G., Gray, J., & Szalay, A. 2005, *Petascale computations systems: Balanced cyberinfrastructure in a data-centric world*, <http://arxiv.org/abs/cs/0701165>
- Duellmann, D. 1999, *Petabyte Databases*, ACM SIGMOD Record, vol. 28, p. 506, <https://web.archive.org/web/20071012015357/http://www.sigmod.org/sigmod/record/issues/9906/index.html#TutorialSessions>
- Hanushevsky, A., & Nowak, M. 1999, *Pursuit of a Scalable High Performance Multi-Petabyte Database*, 16th IEEE Symposium on Mass Storage Systems, pp. 169–175, <http://citeseer.ist.psu.edu/217883.html>.
- Shiers, J., *Building Very Large, Distributed Object Databases*, downloaded from <https://web.archive.org/web/20070915101842/http://wwwasd.web.cern.ch:80/wwwasd/cernlib/rd45/papers/dbprog.html> on 2007-11-25.