

BIG DATA

Osman AIDEL



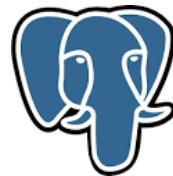
Responsable Bases De Données



- Mariadb



- PostgreSQL



- MySQL



- Oracle



- ElasticSearch



- MongoDB



- Neo4J



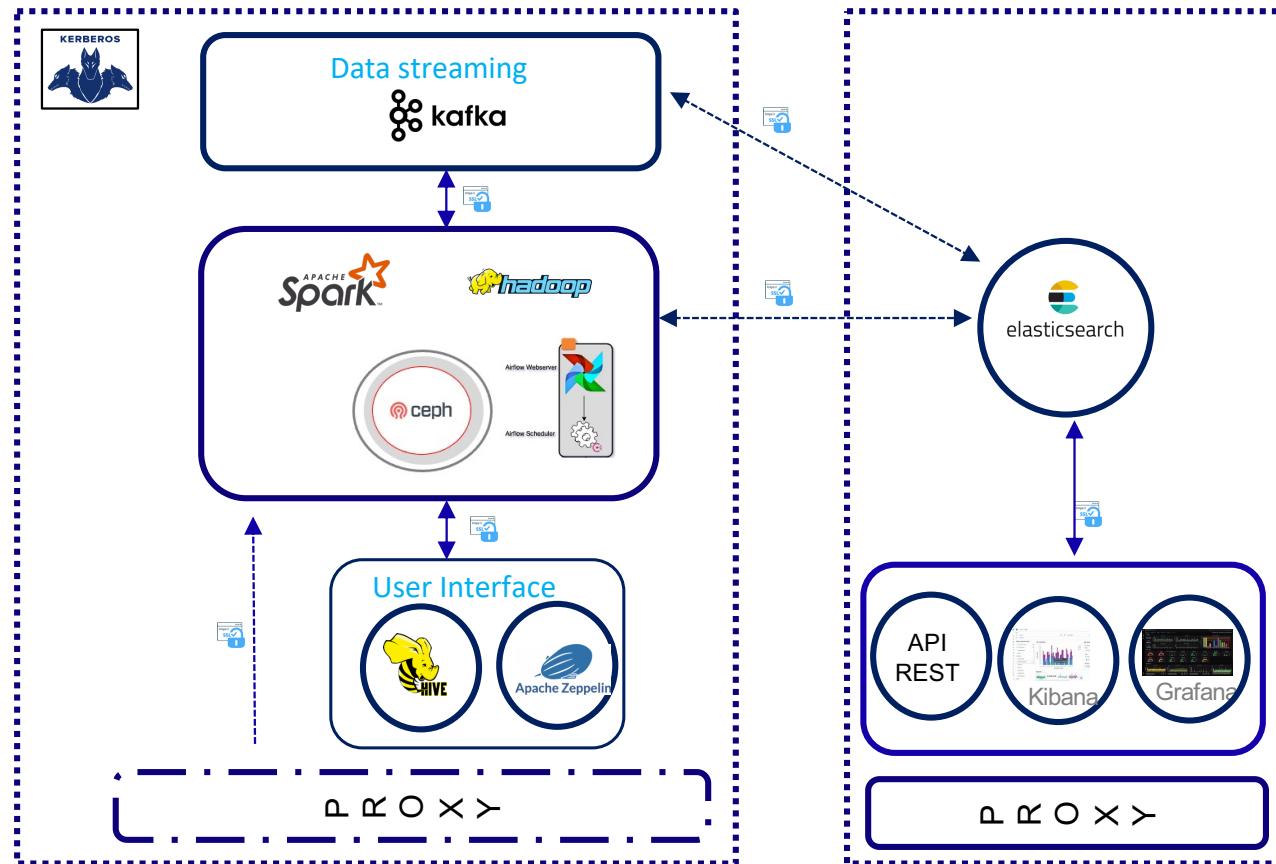
Plus de 1 000 bases de données :
Volumétrie > 50 To
32 serveurs de bases de données



Osman AIDEL

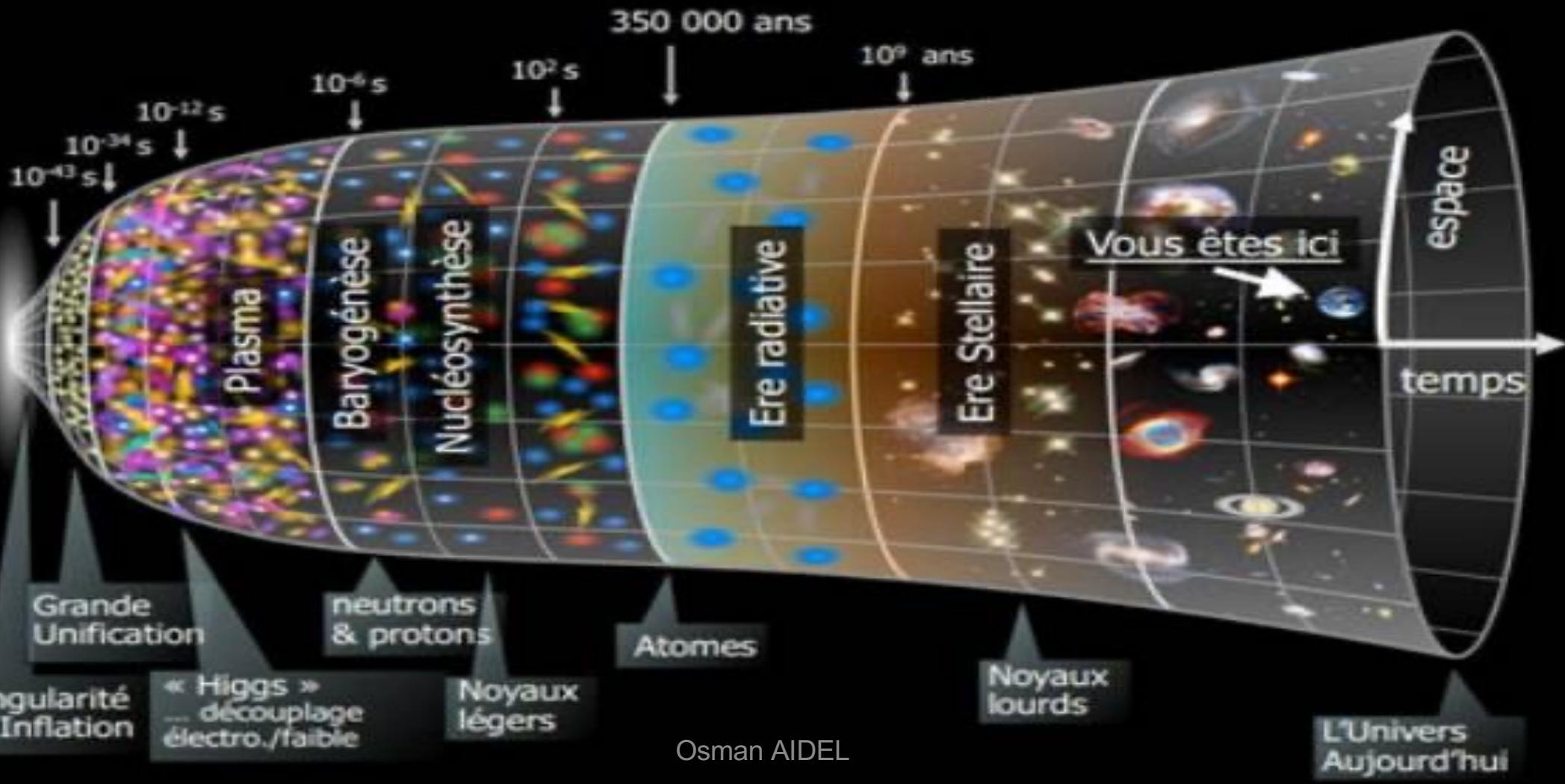


Data Platform



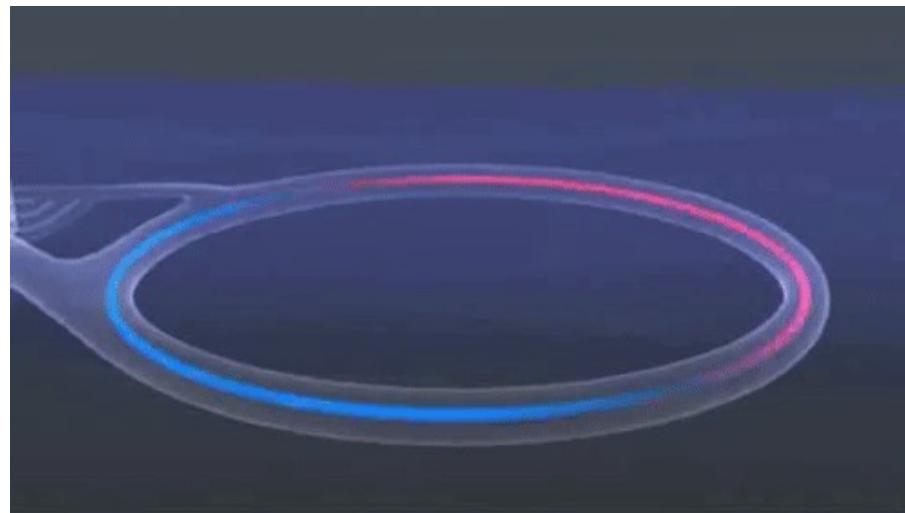
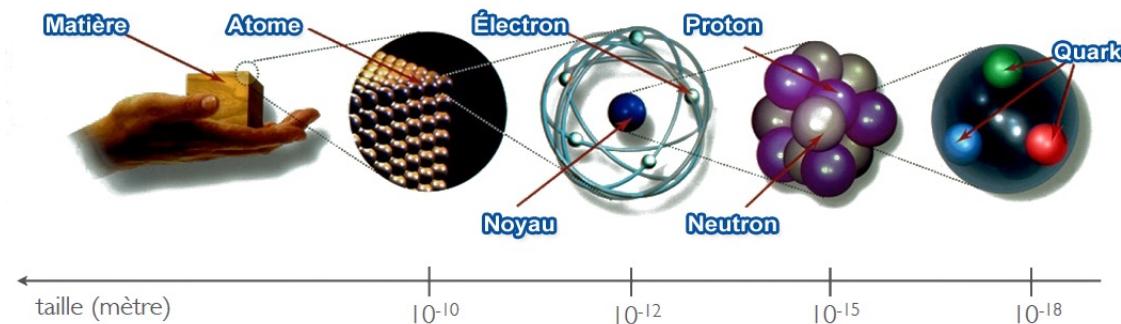
Osman AIDEL

Une Brève Histoire de l'Univers



Osman AIDEL

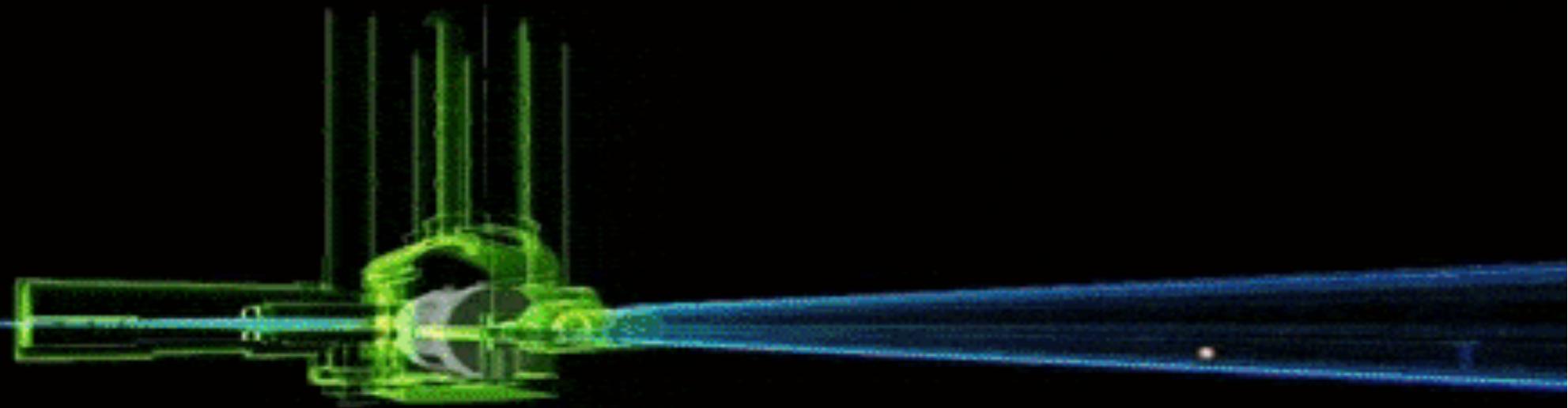
Les expériences Large Hadron Collider



Osman AIDEL

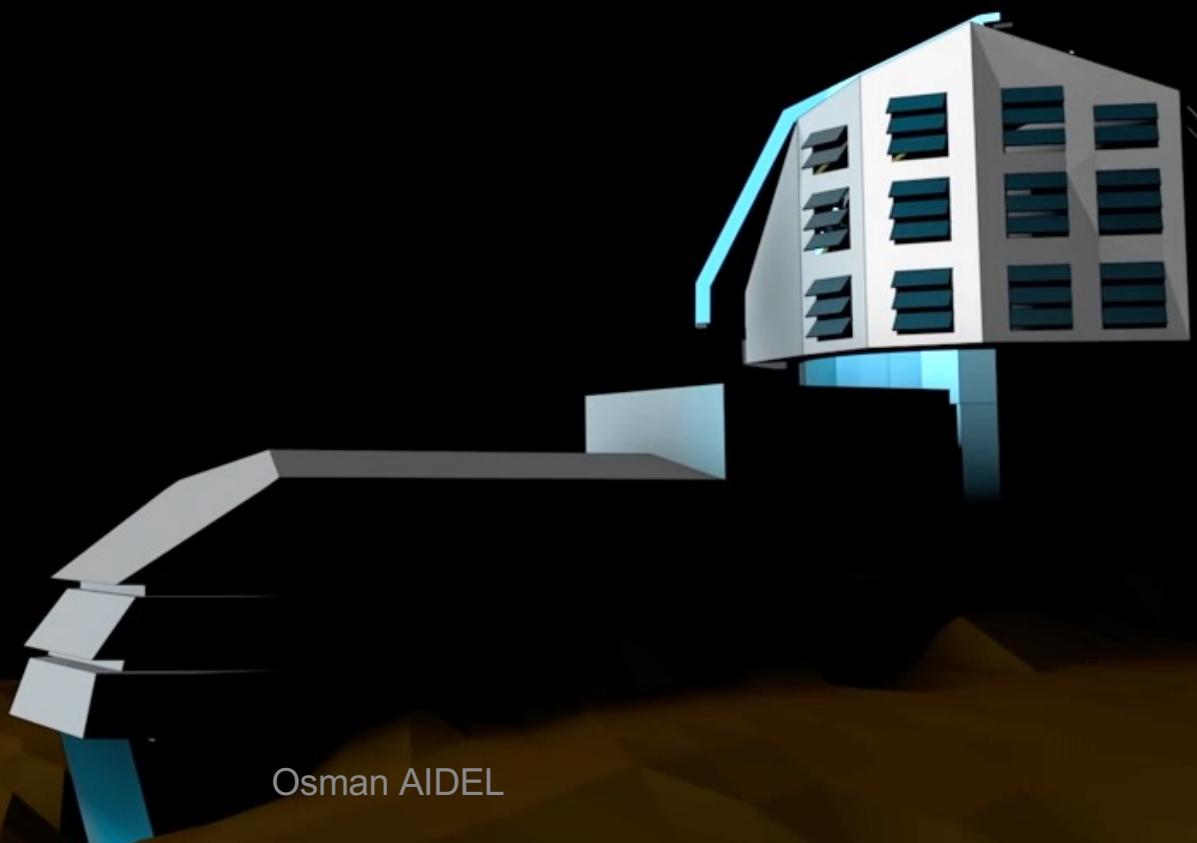
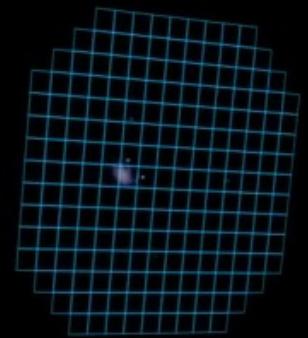


Les expériences LHC



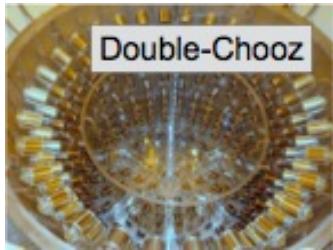
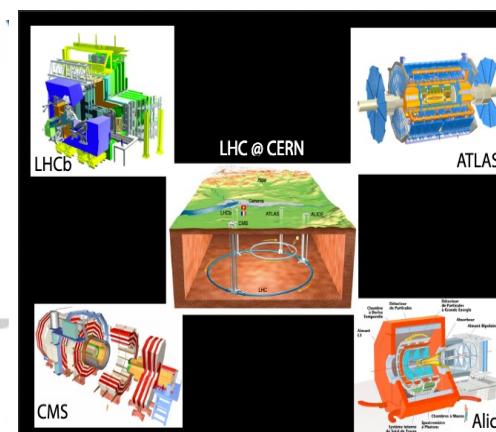
COLLISION EVENT IN THE ATLAS DETECTOR

Osman AIDEL



Osman AIDEL

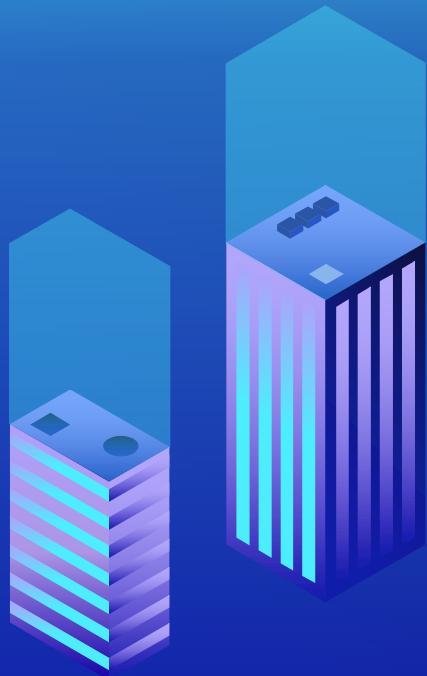
Les expériences IN2P3



Osman ADEL

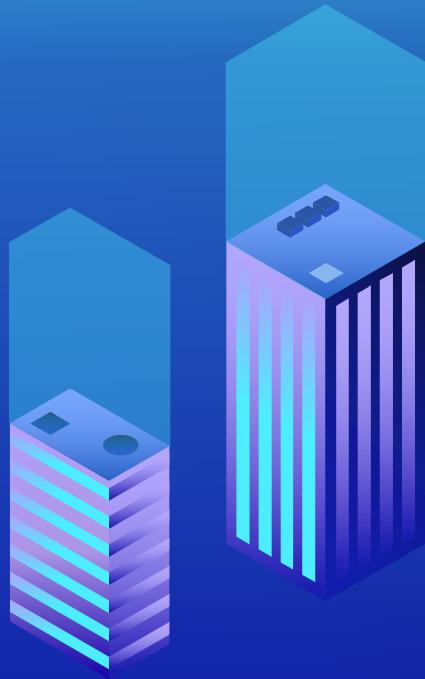


Programme



- 0 Pourquoi le BIG DATA
- 1 Environnement de travail
- 2 Le monde NoSQL
- 3 Les bases Graphe
- 4 Les bases Document
- 5 Les bases Clé-Valeur

Programme



- 5 Hadoop
- 6 L'écosystème Hadoop
- 7 SPARK

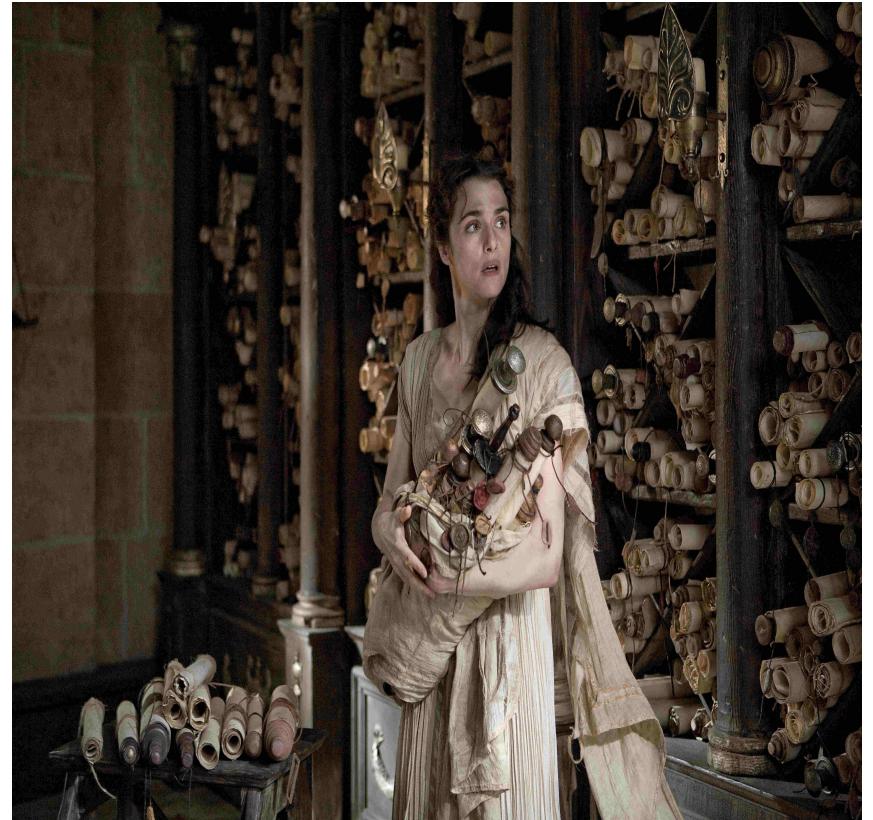


Introduction

- 1 – Le début de l'informatique
- 2 – L'évolution des technologies
- 3 – Pourquoi les architectures distribuées ?
- 4 – Les technologies emergentes

Qu'est-ce qu'une base de données ?

- Centraliser les données
- Collecter et stocker les données
- Organiser les données pour faciliter la recherche



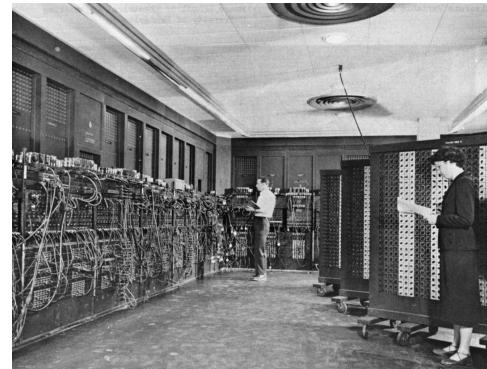
Osman AIDEL



Les défis technologiques

Le premier calculateur est né en 1946

- ENIAC : *Electronic Numerical Integrator And Computer*
- 400 flops
- 167 m²
- 30 Tonnes



1955 : Stockage par bande magnétique



Osman AIDEL

Et votre smartphone ???



Smartphone 2022 :

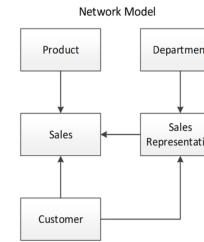
- Stockage 1To
- 400 Gflops

Première génération de bases de données : 1950 - 1970

- 1965 : Modèle hiérarchique



- 1969 : Modèle Réseau



Connu comme base de données de navigation. On navigue entre les objets par des liens/pointeurs.
Pas pratique à manipuler.



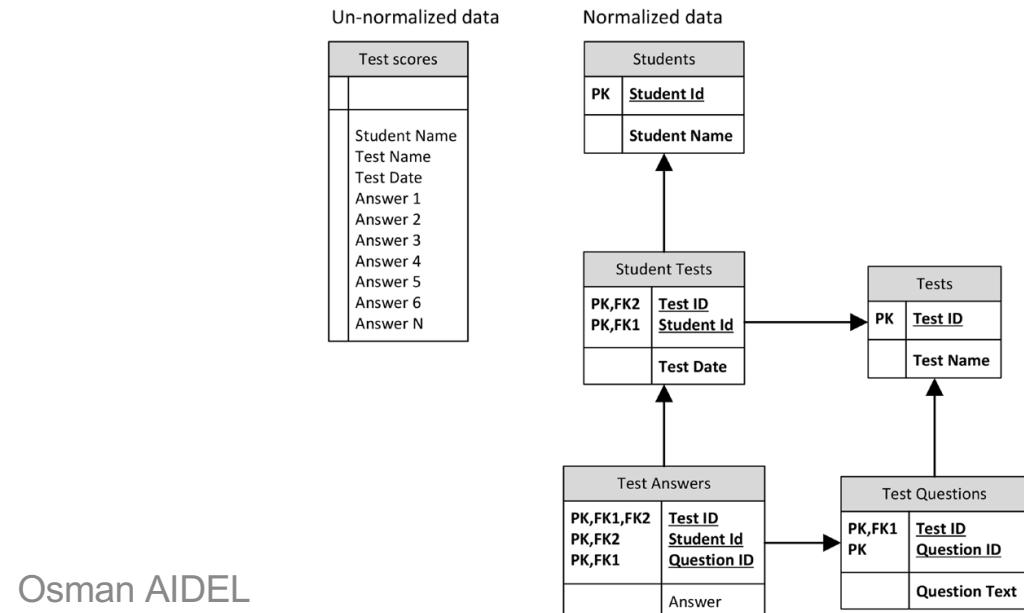
Seconde génération des bases de données

1970 : Le modèle relationnel a fait son apparition avec la publication du Dr Edgar Codd (chercheur IBM)

Un modèle reposant sur une théorie mathématique : l'algèbre relationnel.

Les données sont regroupées dans des tables à 2 dimensions avec des lignes et des colonnes.

Un modèle normalisé afin d'éviter les incohérences de données.



Osman AIDEL

Seconde génération des bases de données

Le modèle relationnel définit la manière d'organiser les données mais il ne définit pas la manière dont la base de données gère les accès concurrents.

Le Dr Jim Gray va proposer un modèle transactionnel où une transaction est définie comme un ensemble de modifications laquelle dispose des propriétés suivantes :

Atomicité : Une transaction est indivisible. Soit toutes les modifications dans la transaction sont appliquées soit aucune.

Cohérence : La base de données reste dans un état cohérent avant et après application d'une transaction.

Isolation : Plusieurs transactions peuvent être traitées simultanément.

Durabilité : Une fois une transaction appliquée (COMMIT), les changements sont conservés même en cas de panne.



Seconde génération des bases de données



Les bases de données relationnelles

Avantages :

La technologie est mature (création il y a plusieurs dizaines d'années) ce qui fait qu'aujourd'hui le SQL est un langage standard et normalisé.

On a une garantie que les transactions sont atomiques, cohérentes, isolées et durables – principe ACID (Atomic, Consistent, Independant, Durable).

La possibilité de mettre en œuvre des requêtes complexes (croisement multiple des données)
Du fait du nombres d'années d'existence, un large support est disponible et il existe également de fortes communautés.

Inconvénients :

La modification du modèle de données peut être couteuse.

L'évolution des performances est verticale (augmentation des ressources du serveur).

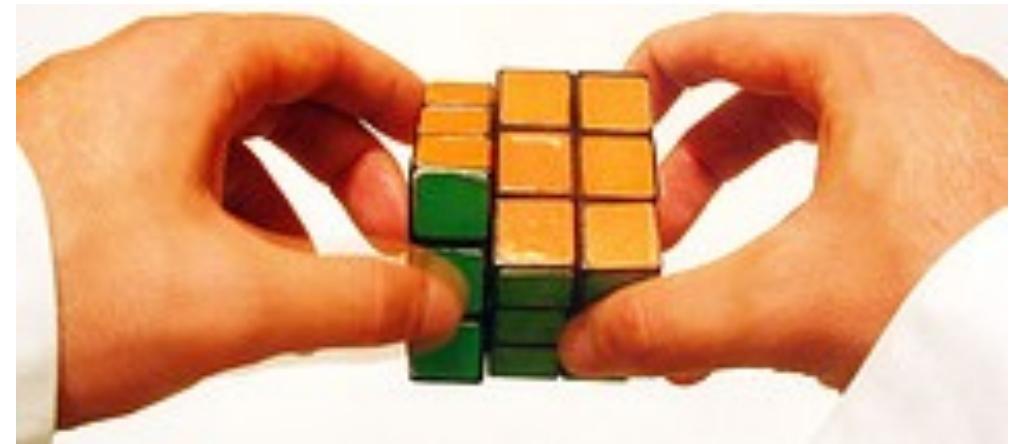
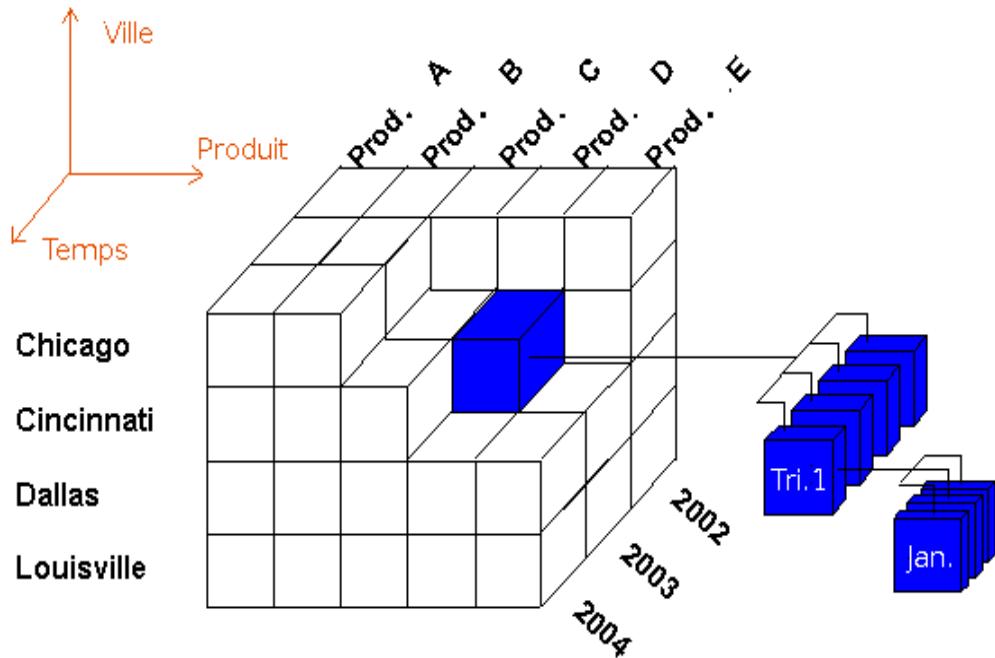
Sur un très grand volume de données (centaines-milliers de Tera octets) le modèle peut atteindre des limites en terme de performance

Pour certains éditeurs, le prix de licence est élevé.



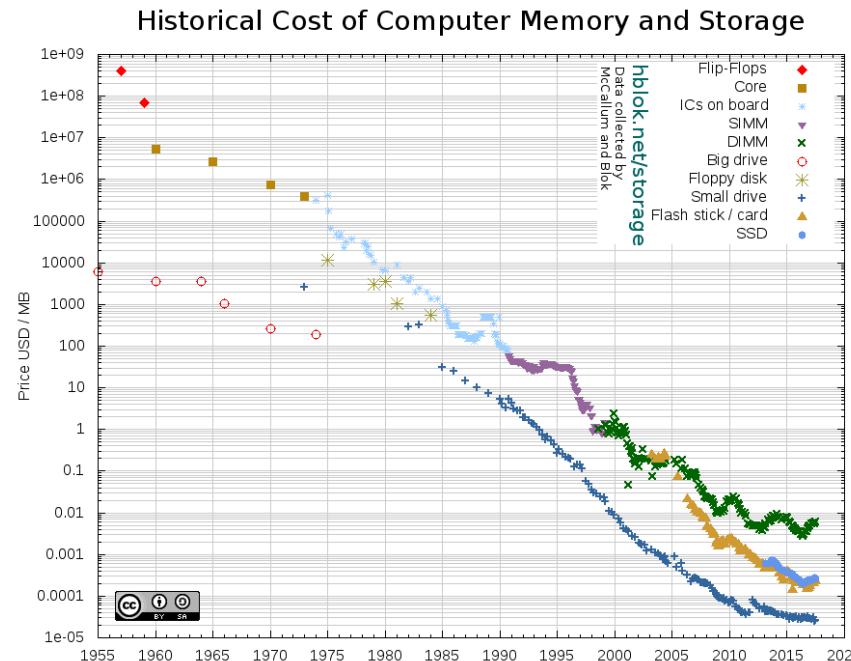
Seconde génération des bases de données

□ Les entrepôts de données (Dataware house)



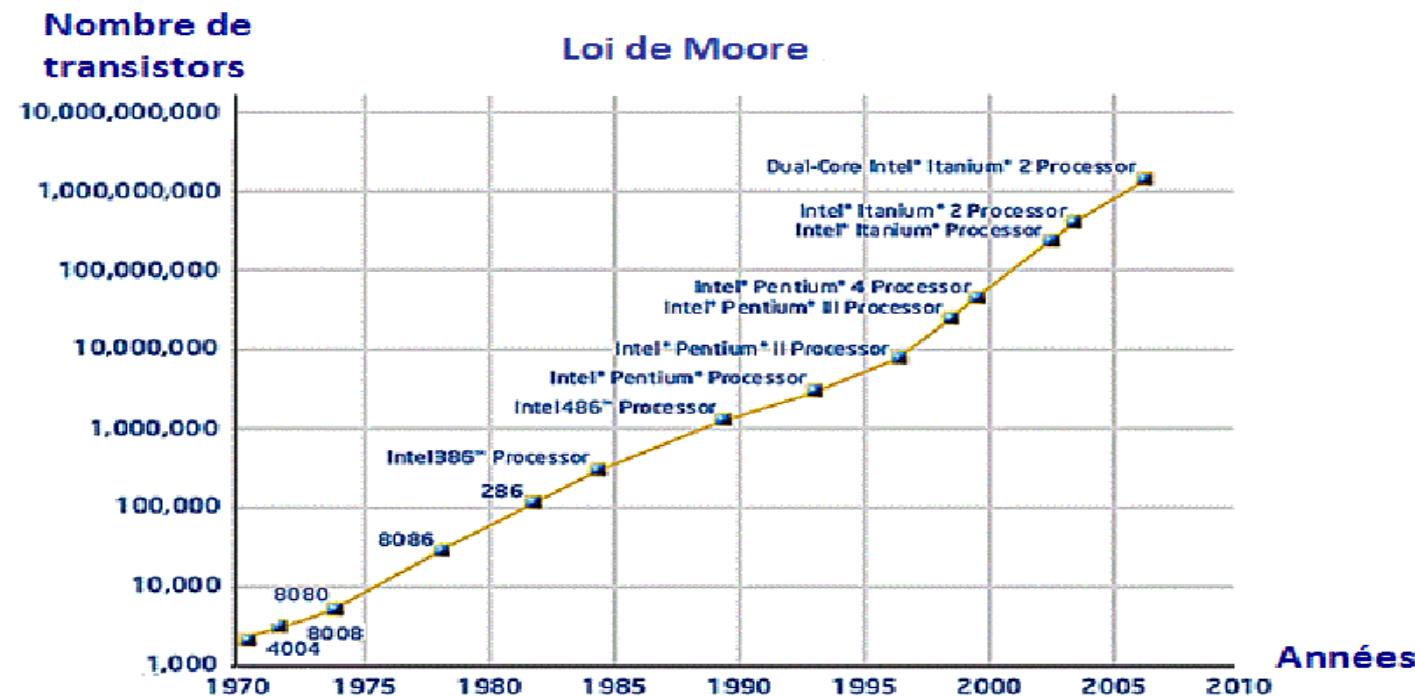
L'évolution matérielle

- La marginalisation du matériel converge avec les ambitions du BIG DATA.
- A partir des années 2000, le matériel devient accessible à tous.



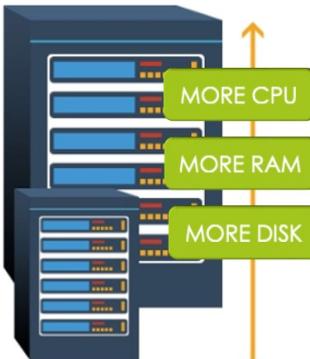
Osman AIDEL

L'évolution matérielle



Osman AIDEL

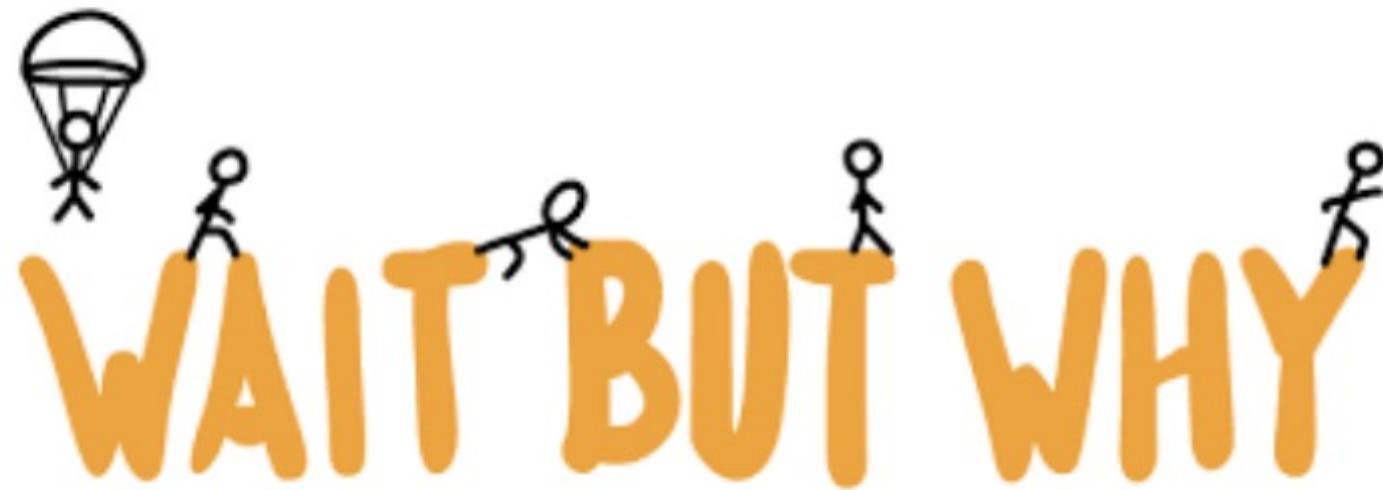
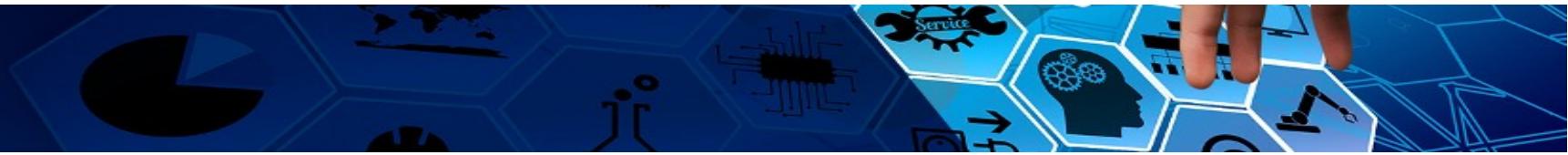
Le besoin



- Tous les géants d'internet tel que Google, Facebook, Amazon etc... sont confrontés aux problématiques de traitements de grosses volumétries.
- La plupart des données est accédée à partir de bases de données relationnelles.
- Malheureusement, plus le volume de données est important et plus les bases de données traditionnelles atteignent leurs limites et deviennent de plus en plus lentes.
- L'approche par extensibilité verticale (matériel plus performant) atteint ses limites et est très coûteuse.
- Les applications web passent à l'échelle mais les bases de données relationnelles restent le goulot d'étranglement.
- Pour répondre à la charge croissante des applications web, les SGBDR mettent en place des systèmes de replication pour les accès en lecture seulement. Les écritures sont toujours concentrées vers une seule machine.



Le besoin



Osman AIDEL

Le besoin : Forensic



Le besoin : Profiling



Osman AIDEL

Le besoin : Prédiction / IA



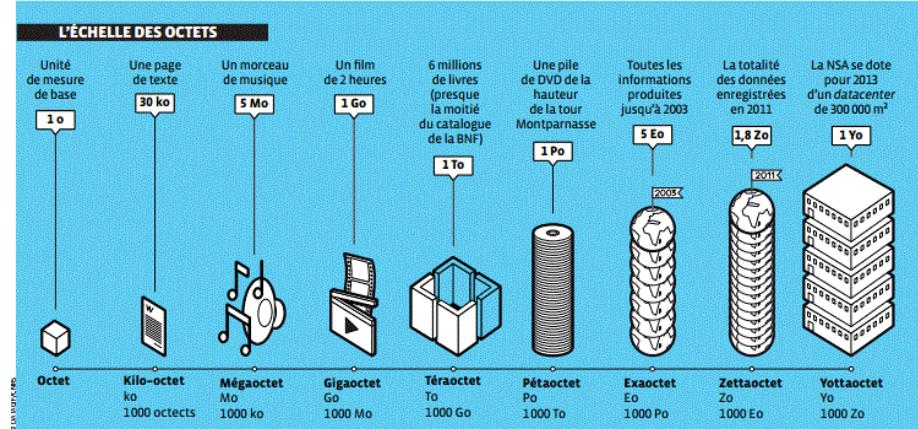
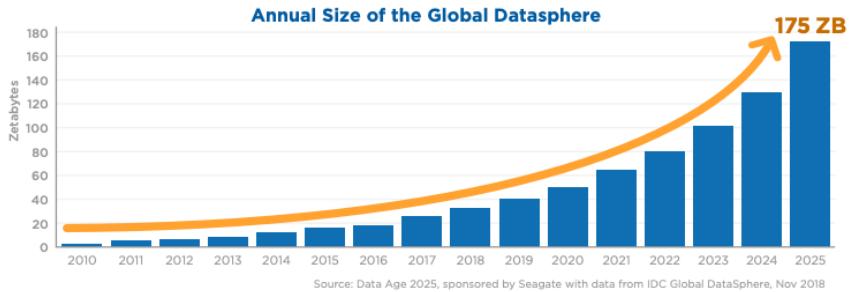
Osman AIDEL

Le besoin



Des volumétries astronomiques sont en jeux :

Figure 1 – Annual Size of the Global Datasphere



Pourquoi stocker autant de données ?

Le savoir est dans les données et « Le savoir, c'est le pouvoir » (Dan Abnett).

Aider les compagnies / la science à mieux comprendre leur métier / notre univers.

A mieux comprendre les clients.

Découvrir de nouvelles logiques métiers.

Prendre de rapides et de bonnes décisions.

En 2019, les données ont une valeur non négligeable.

Pour certains, la croissance des données de l'entreprise est un facteur de développement.

Osman AIDEL

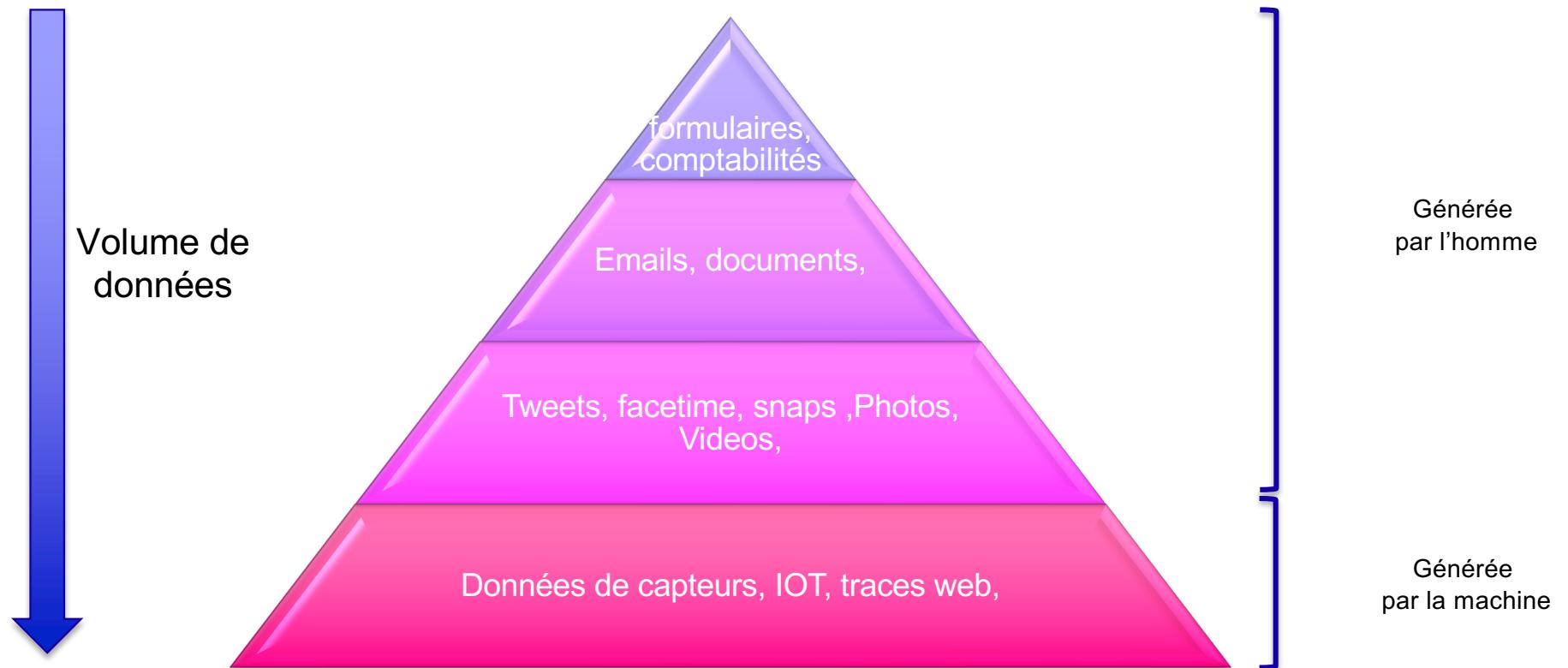


BIG DATA



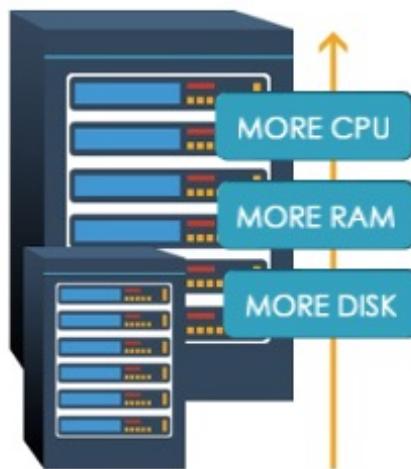
Osman AIDEL

Les défis technologiques

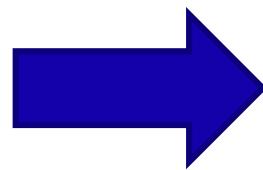


Les architectures distribuées

- Une alternative à cette approche est de basculer à une architecture horizontale.
- « Diviser pour mieux régner »



Vertical Scaling
(Scaling up)



Horizontal Scaling
(Scaling out)

modèle MPP (Massively parallel processing)
où
shared nothing.

Osman AIDEL

Les architectures distribuées

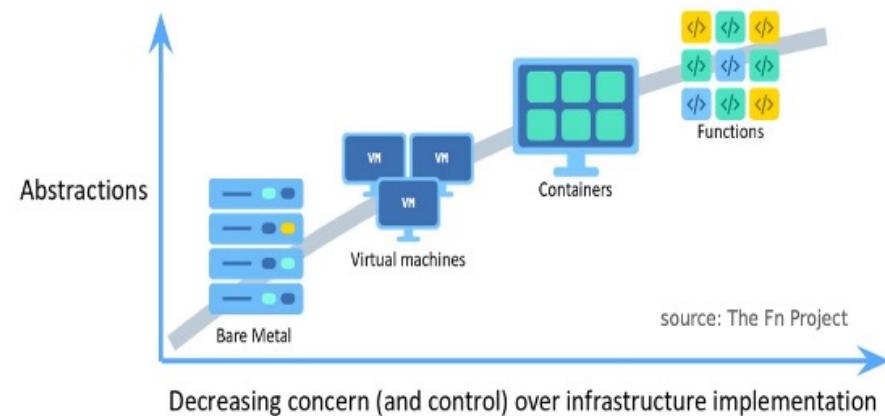
Les technologies de virtualisation offrent aux entreprises une flexibilité plus accrue dans l'architecture des systèmes informatique et la gestion des coûts.

L'internet favorise par son réseau la distribution des environnements informatique.

Avec le cloud, les environnements se dimensionnent à la demande et le nombre, le volume et la variété des données augmentent très rapidement.

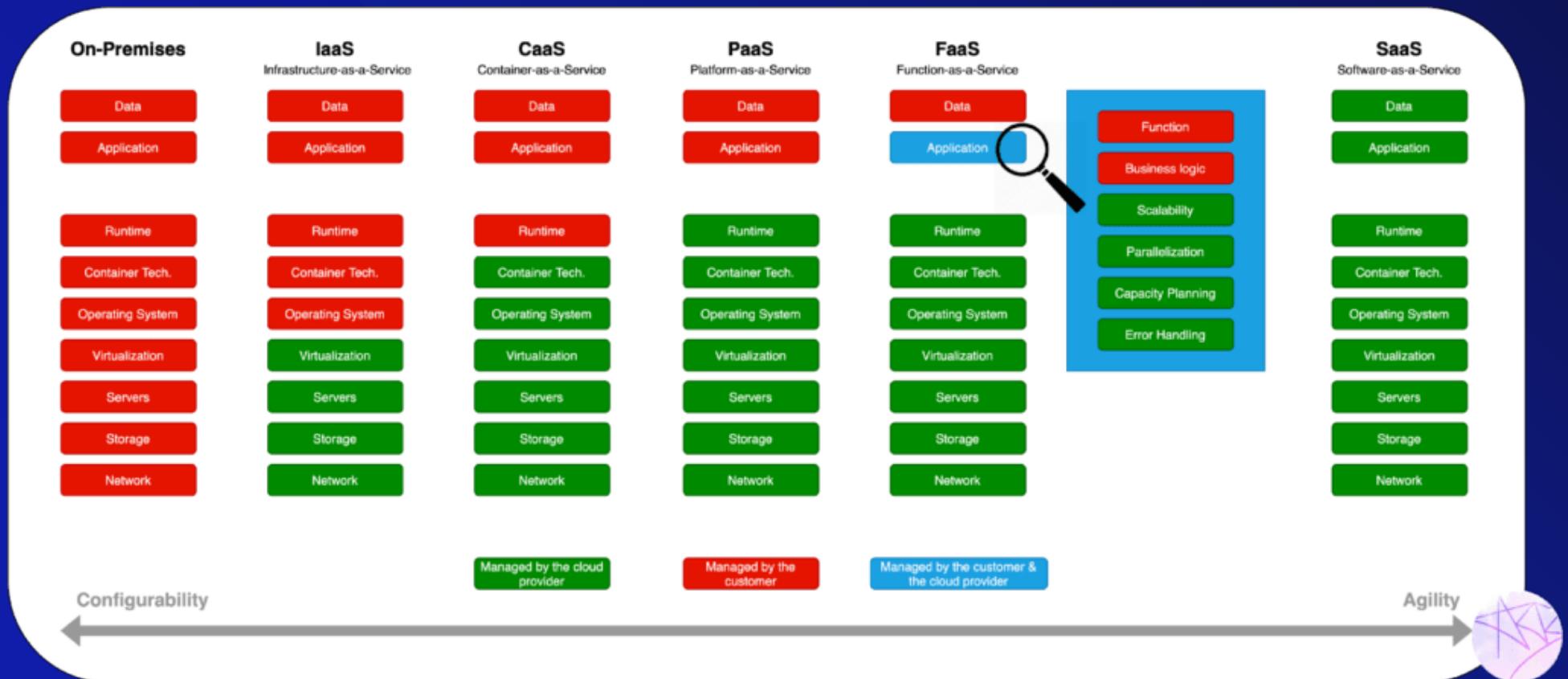
Abandon des applications monolithiques pour des applications plus nombreuses, interopérable, hétérogènes et extensibles.

Les applications sont à présent conçus nativement pour être scalables.



Osman AIDEL

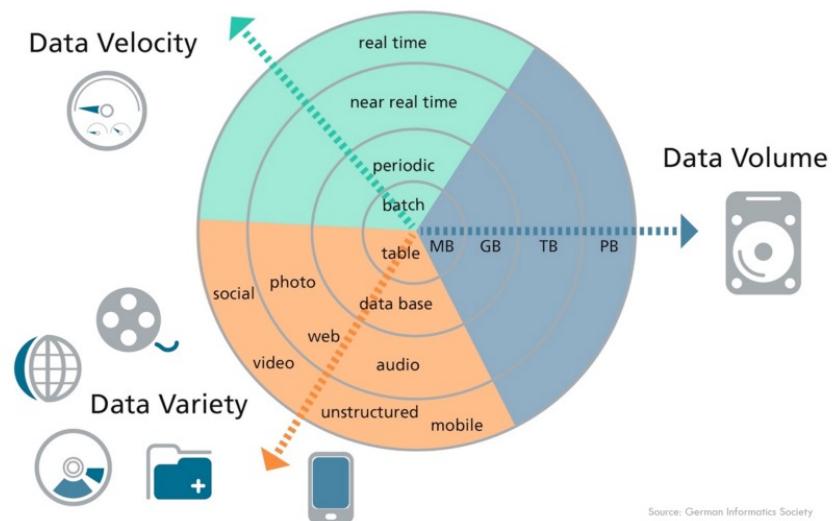
Les architectures distribuées



Osman AIDEL

La tendance BIG DATA

- Il n'y a pas de définition officielle du BIG DATA.
- BIG DATA fait référence aux situations où les ressources matérielles existantes ne sont pas suffisantes pour le traitement des données.
- La définition de BIG DATA selon Gartner est connue sous le nom de la règle des 3V :
 - Volume
 - Vélocité
 - Variété



La tendance BIG DATA

□ Volume

- Le volume réfère au montant de données générées où stockées.
- Le volume est généralement la principale composante qui détermine l'adoption de solutions BIG DATA d'où le nom.
- Le volume peut s'exprimer en quelques gros fichiers où des milliers de petits fichiers.
- L'idée est de traiter les données en un temps « raisonnable » quelque soit la quantité de données.

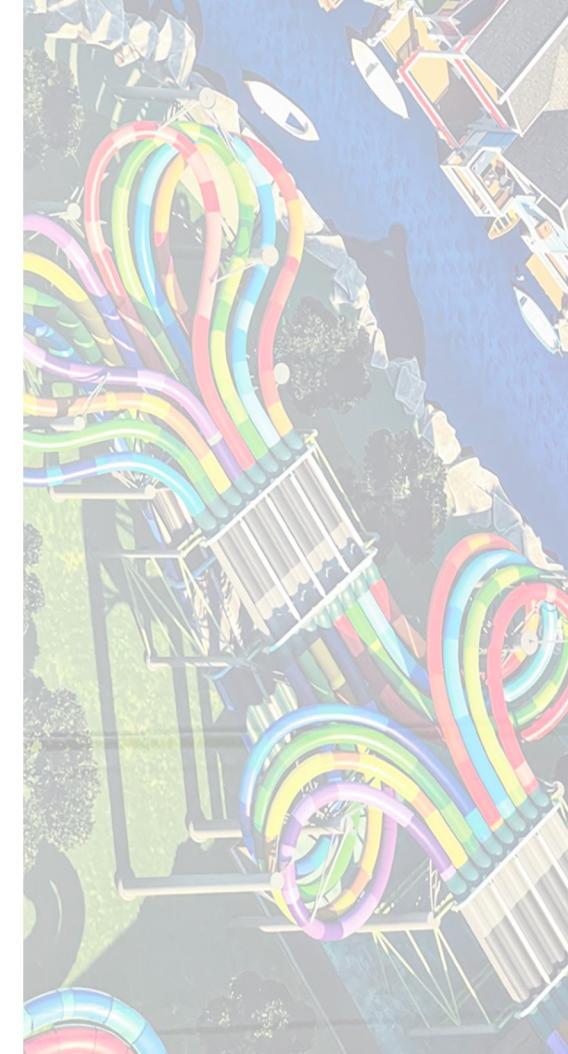
Osman AIDEL



La tendance BIG DATA

□ Vélocité

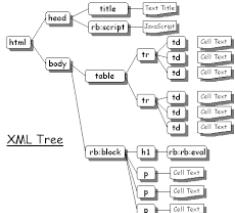
- **Vélocité** réfère à la vitesse à laquelle les nouvelles données sont générées.
- Les technologies de BIG DATA nous permettent aujourd'hui d'analyser les données alors qu'elles viennent d'être générées.
- Les données sont générées à grande vitesse et peuvent avoir des structures différentes.
- Ces aspects sont des défis majeurs qui couplés à des grosses volumétries rendent les systèmes de traitements de données existant inappropriés.
- Dans une base de données relationnelle, modifier la structure des données revient à reconcevoir le modèle de données.



Osman AIDEL

La tendance BIG DATA

Relational Table Model			
ID	Name	City	Country
1/1	Espen	Oslo	Norway
2/2	Harald	Munich	Germany
3/3	Sam	San Jose	USA



□ Variété

○ Données structurées

- Les données sont stockées sous une structure définie.
- Le type et la taille de chaque information sont connus.
- Les données peuvent être accédées et traitées à partir d'un langage.
- Exemple : les tables relationnelles.

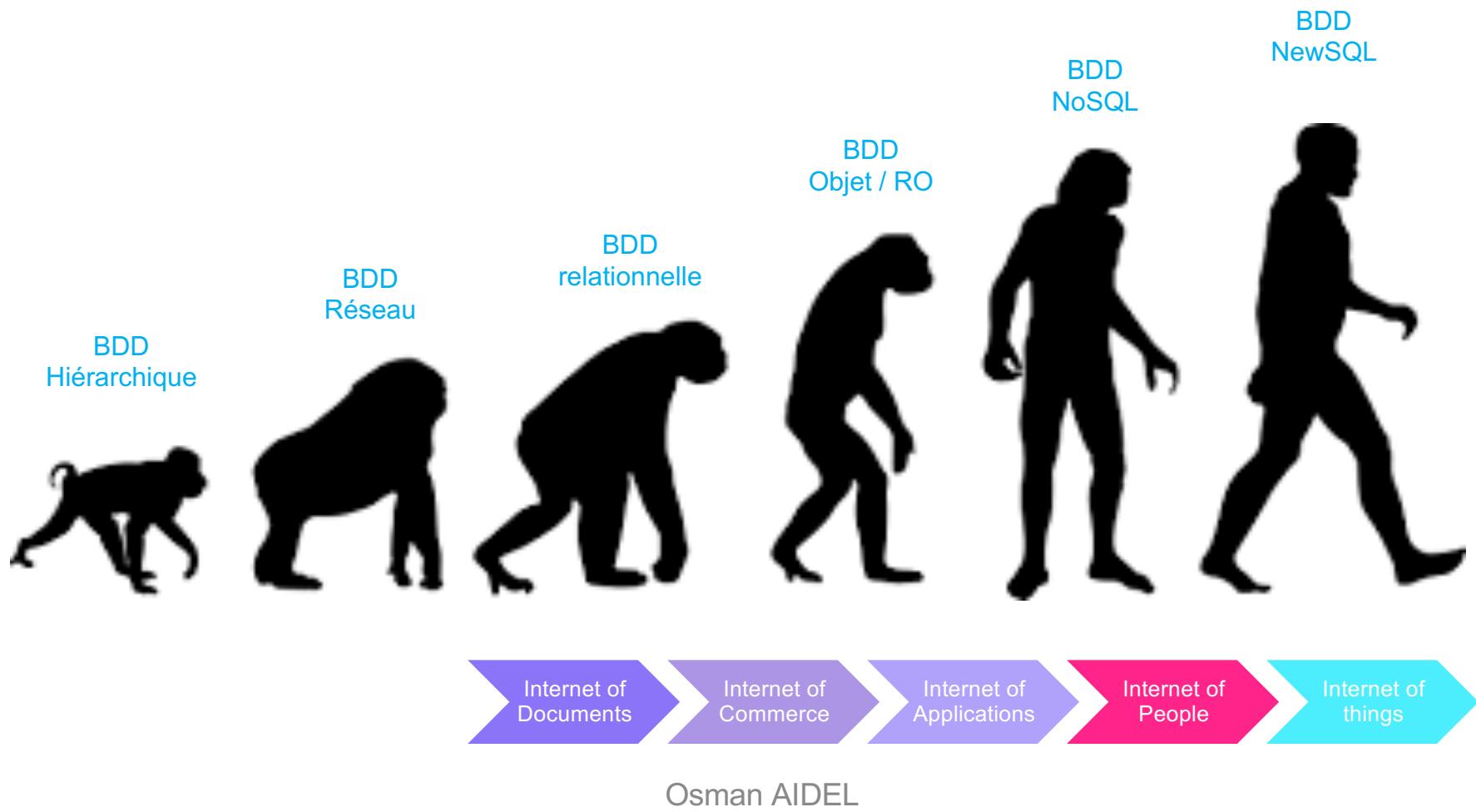
○ Données non-structurées

- Les données avec un format inconnu.
- Les données ne suivent aucune règle ou pattern.
- Il n'y a pas de langage qui permette de les exploiter.
- La taille d'une donnée est variable et peut être massive.
- Exemple : vidéo, image, un long texte ...

○ Données semi-structurées:

- Elles ont une structure.
- Les données sont basées sur une sémantique auto-descriptive.
- La même information peut être classifiée avec différentes tailles et types.
- Les données peuvent être accédées et traitées à partir d'un langage.
- Exemple : les données d'une station météo (XML ou JSON) ...

La tendance BIG DATA

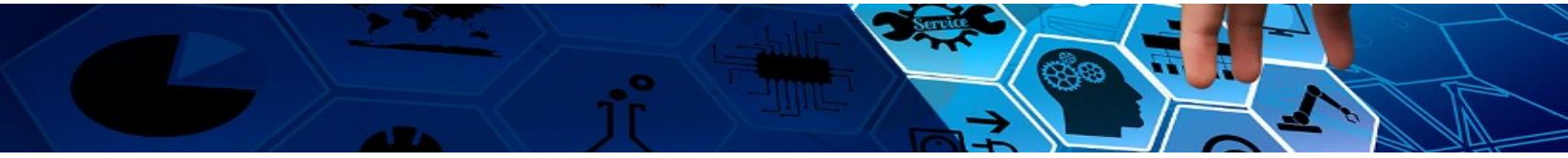


**What
the
.....?**



Osman AIDEL

Le NoSQL



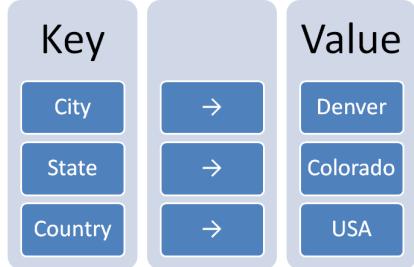
- Une nouvelle famille de SGBD capable de proposer :
 - Un modèle de données sans schéma/structure imposé.
 - Des temps de réponse très faible quelque soit la volumétrie.
 - Conçu principalement sur une architecture horizontale.
 - Des traitements complexes qui sont difficiles à réaliser par les SGBDR.
- Le nombre de SGBD NoSQL va exploser dans les années 2000.

Le NoSQL



Orienté Colonne

train6101	PARIS 06:07	VALENCE 08:19	AVIGNON 09:07	AIX 09:30	MARSEILLE 09:41
train2917		AVIGNON 11:30		11:59	MARSEILLE



Orienté Document

```
{  
    first_name: 'Paul',  
    surname: 'Miller',  
    cell: 447557505611,  
    city: 'London',  
    location: [45.123,47.232],  
    Profession: ['banking', 'finance', 'trader'],  
    cars: [  
        { model: 'Bentley',  
         year: 1973,  
         value: 100000, ... },  
        { model: 'Rolls Royce',  
         year: 1965,  
         value: 330000, ... }  
    ]  
}
```

Fields

String

Number

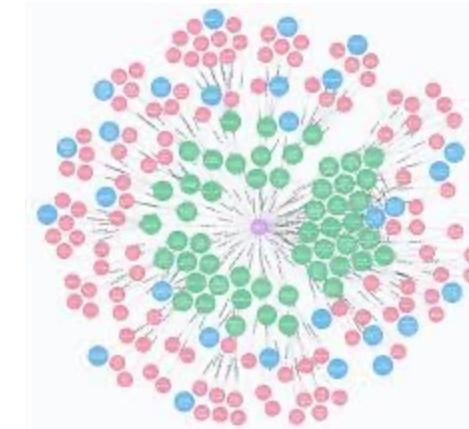
GeoCoordinates

Typed field values

Fields can contain arrays

Fields can contain an array of sub-documents

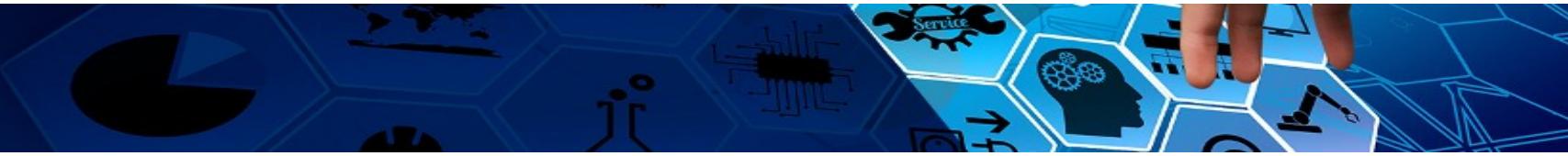
Orienté Graphe



Osman AIDEL



Le NewSQL



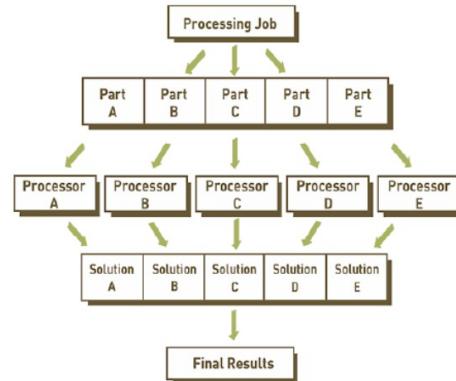
- « Many enterprise systems that handle high-profile data (e.g., financial and order processing systems) also need to be able to scale but are unable to use NoSQL solutions because they cannot give up strong transactional and consistency requirements. »
- **Google** se plaint que ses développeurs passent trop de temps à gérer l'intégrité des données dans le code applicatif.
- Nouvelles plateformes. Développements « from scratch »
- A partir de 2011, naissance des bases de données NEWSQL
 - Google Spanner (successeur de BigTable)
 - VoltDB
 - H-Store



Osman AIDEL

L'écosystème BIG DATA

- Pour les traitements de données complexes, des nouvelles plateformes émergent.
 - Naissance en 2004 de la première plateforme de traitement parallélisé Open-source : HADOOP
 - Avec un système de stockage distribué HADOOP Distributed File System très très similaire à GoogleFS
 - Un modèle de traitement basé sur le paradigme Map-Reduce.

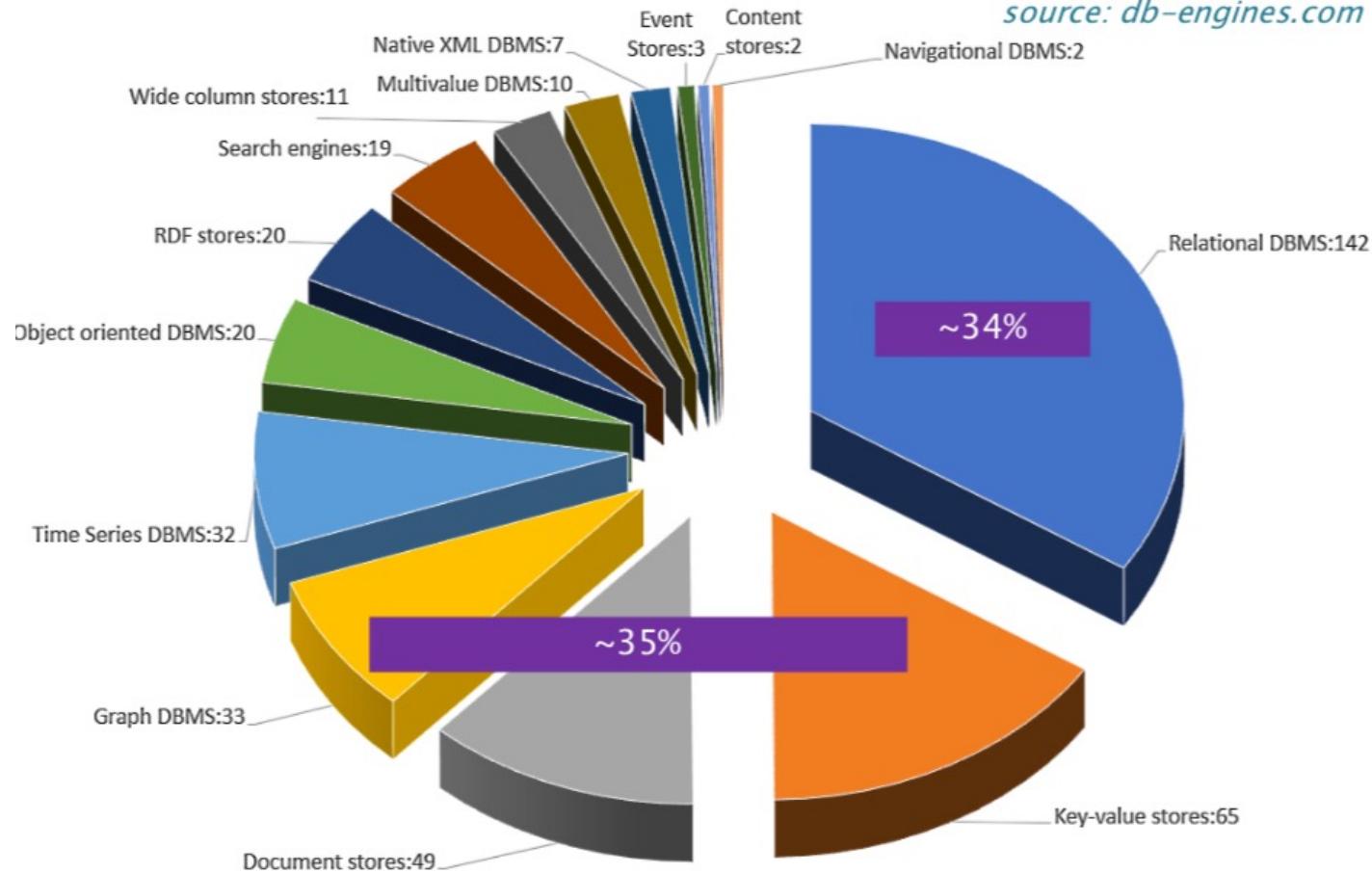


Osman AIDEL

Le NoSQL aujourd'hui (415 DB)



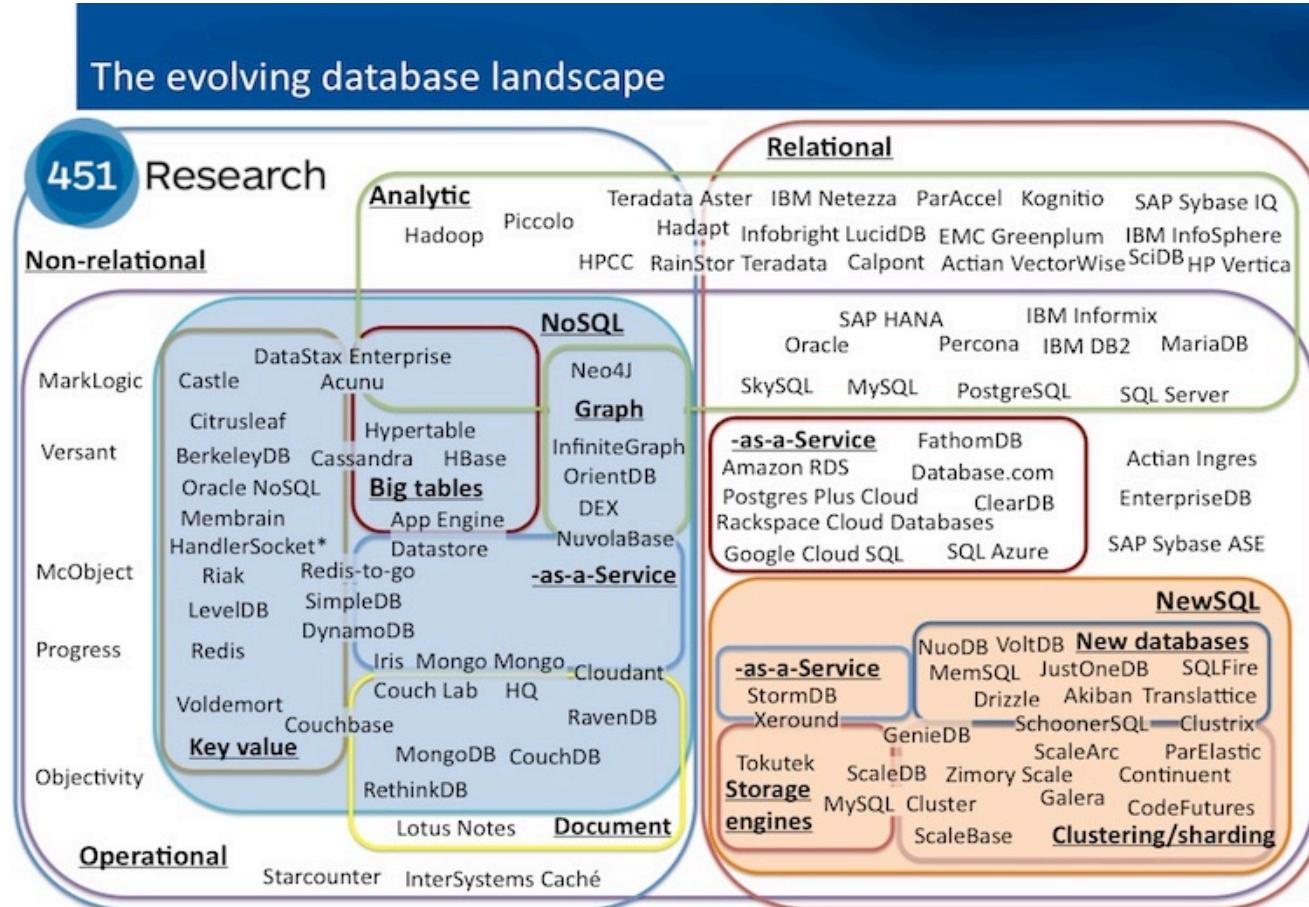
source: db-engines.com



Osman AIDEL



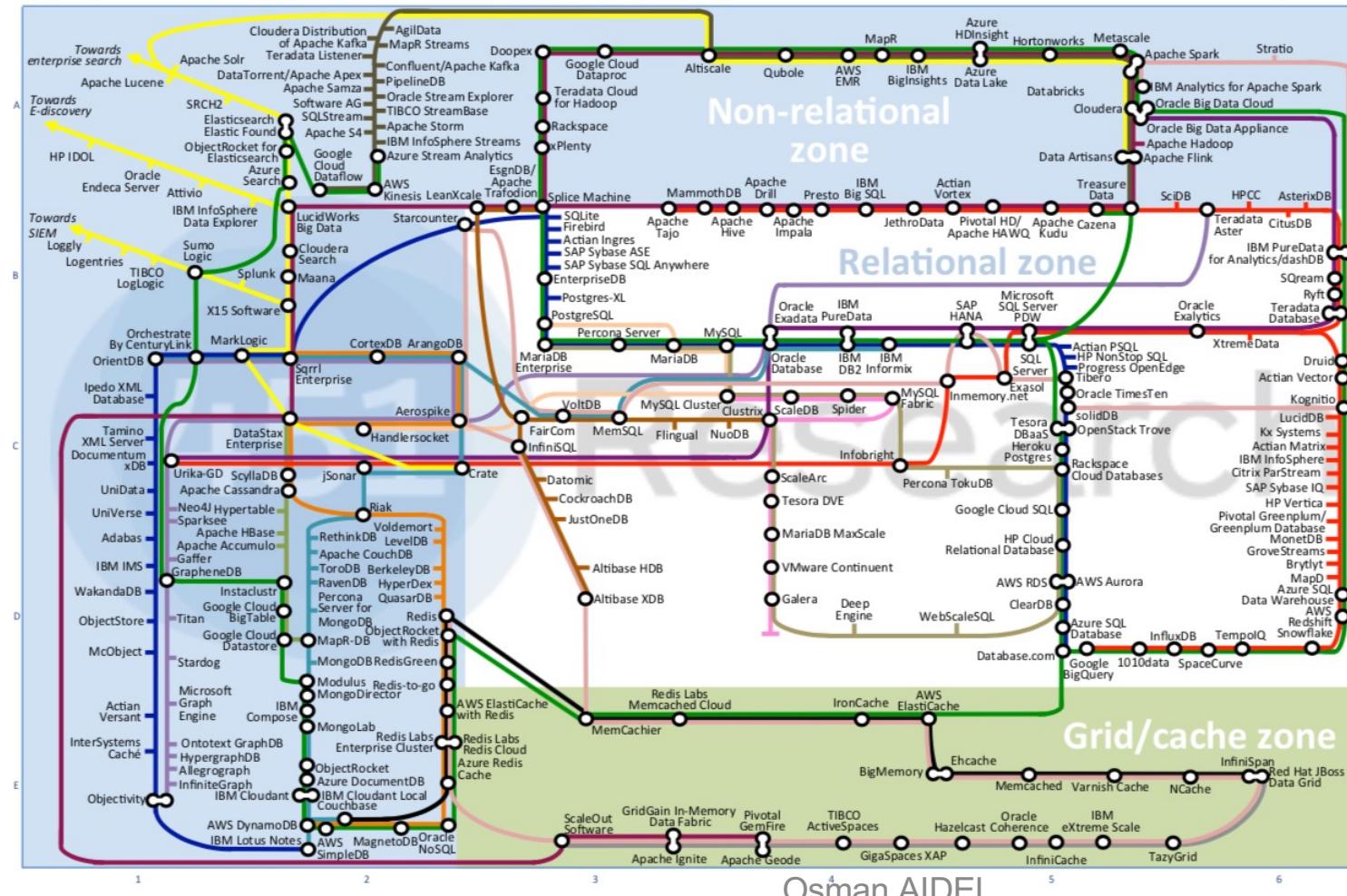
Vision des bases de données en 2012



© 2012 by The 451 Group. All rights reserved

Osman AIDEL

Vision des plateformes de données en 2016



451 Research

A Data Platforms Map

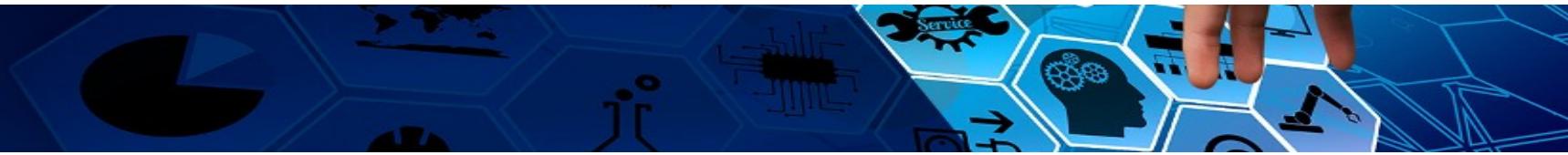
January 2016

B	Key:	General purpose
		Specialist analytic
		-as-a-Service
		BigTables
		Graph
		Document
		Key value stores
		Key value direct access
C	Hadoop	
	MySQL ecosystem	
	Advanced clustering/sharding	
	New SQL databases	
	Data caching	
	Data grid	
D	Search	
	Appliances	
	In-memory	
	Stream processing	

[https://451research.com/
state-of-the-database-landscape](https://451research.com/state-of-the-database-landscape)

© 2016 by 451 Research LLC.
All rights reserved

Le futur ?



Big Data Storymap

EMC²

CURRENT STATE

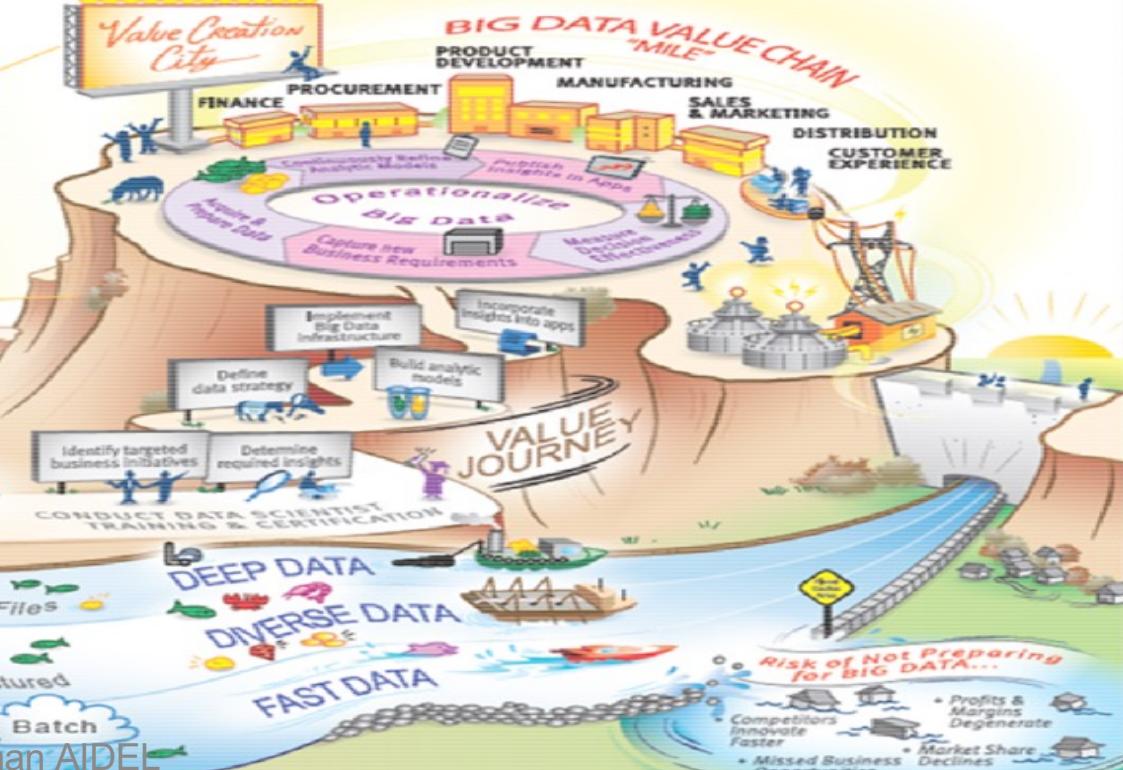
BI/DW CHALLENGES

- Batch oriented, inflexible, high latency
 - Brittle & labor intensive
 - Performance & scalability challenged
 - Aggregated, silo'd, structured data



FUTURE STATE

- Big Data analytics continuously drive business innovations
 - Real-time insights optimize operations and boost profits
 - Unprecedented customer insights improve products and user experience



Big Data concepts by William Schmarzo
Visualization concept by Matt A. Larson and Glenn Steinhardt
<http://GLENNSMC.COM> | <https://www.linkedin.com/in/glensteinhardt/>