Université Constantine 2
جامعة قسنطينة 2

Big Data & NoSQL

– Chapitre1–

Explosion des données et émergence du NoSQL

Dr. GHEMMAZ W
NTIC Faculty
Wafa.ghemmaz@univ-constantine2.dz

Etudiants concernés

Faculté/Institut	Département	Niveau	Spécialité
NTIC	TLSI	Master 1	SDSI

Plan

- Big Data
- Les nV des Big Data
- ACID Vs BASE
- Framework Big Data
- BDR Vs NoSQL

Plan

- Big Data
- Les nV des Big Data
- ACID Vs BASE
- Framework Big Data
- BDR Vs NoSQL


AI
BIG
DATA

Big Data

BiG

Problème de Big Data ?= Volume de données

Problème de Big Data

=

AUGMENTATION (exponentielle)
du volume de données

Big Data

BiG

Problème de Big Data

=

AUGMENTATION (exponentielle)
du volume de données

Le système doit s'adapter pour assurer la
scalabilité

Le système doit s'adapter pour assurer la scalabilité

Définition : Scalabilité (Scalability)

La possibilité pour les systèmes de traitement de données **d'augmenter leurs capacités de traitement** au fur et à mesure que **les données augmentent .**

Scalabilité

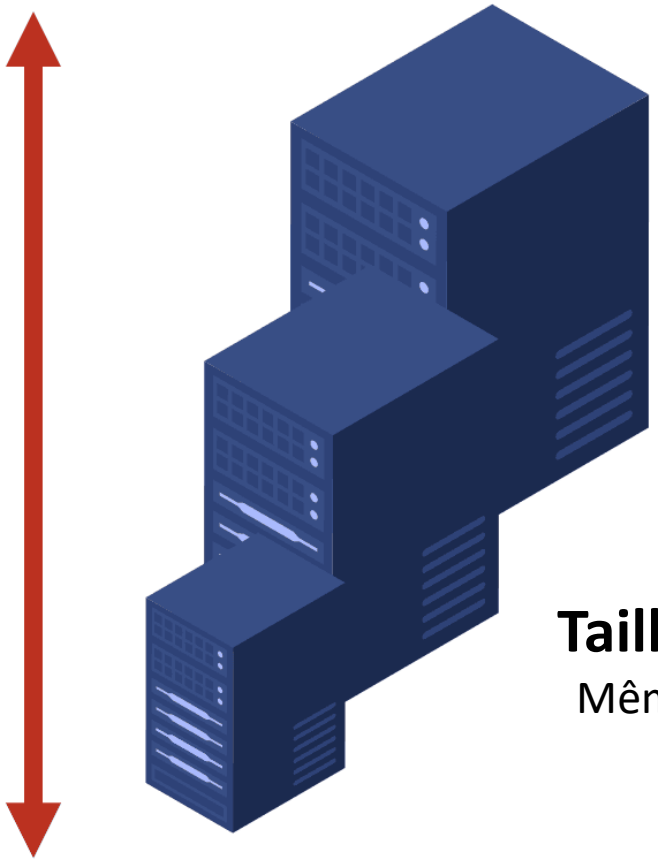
Système Classique Vs Système Big Data

Scalabilité: Système Classique

Vertical Scaling

Increase or decrease the capacity of existing services/instances.

Scale - Up



Taille de ressource est limitée

Même avec TOP machine en terme de capacité

Scalabilité: Système Big Data

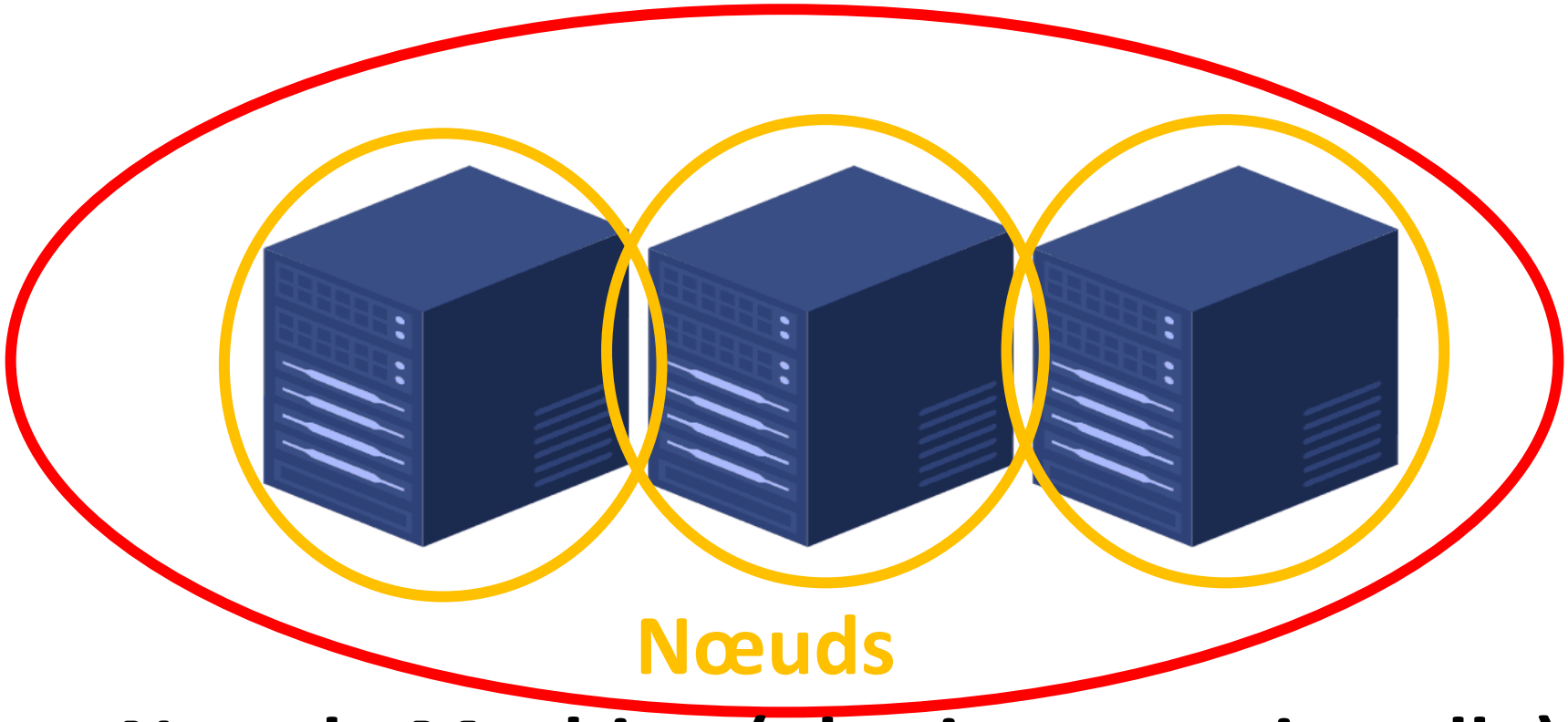
Horizontal Scaling

Add more resources like virtual machines to your system to spread out the workload across them.

Scale - Out

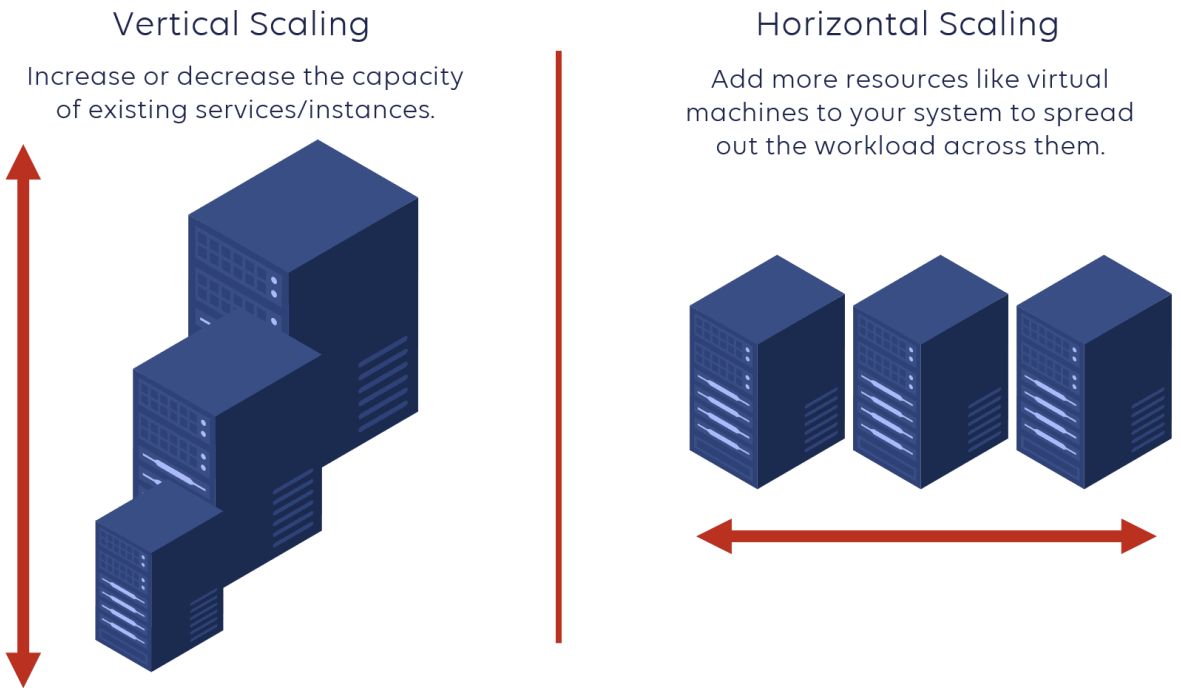
Scalabilité: Système Big Data

Cluster= Ensemble de nœuds



Nœud= Machine (physique ou virtuelle)

Scalabilité



Taille

Taille de ressource est limitée

Même avec TOP machine en terme de capacité

Taille de ressource est illimitée

Pas forcement d'ajouter des machines très puissante

Pas forcement des machines avec les mêmes capacités

Colocalité

Systèmes Distribués Vs Système Big Data

Architecture N-tiers Vs Architecture Big Data

Colocalité

Systèmes Distribués Vs Système Big Data

Architecture N-tiers Vs Architecture Big Data

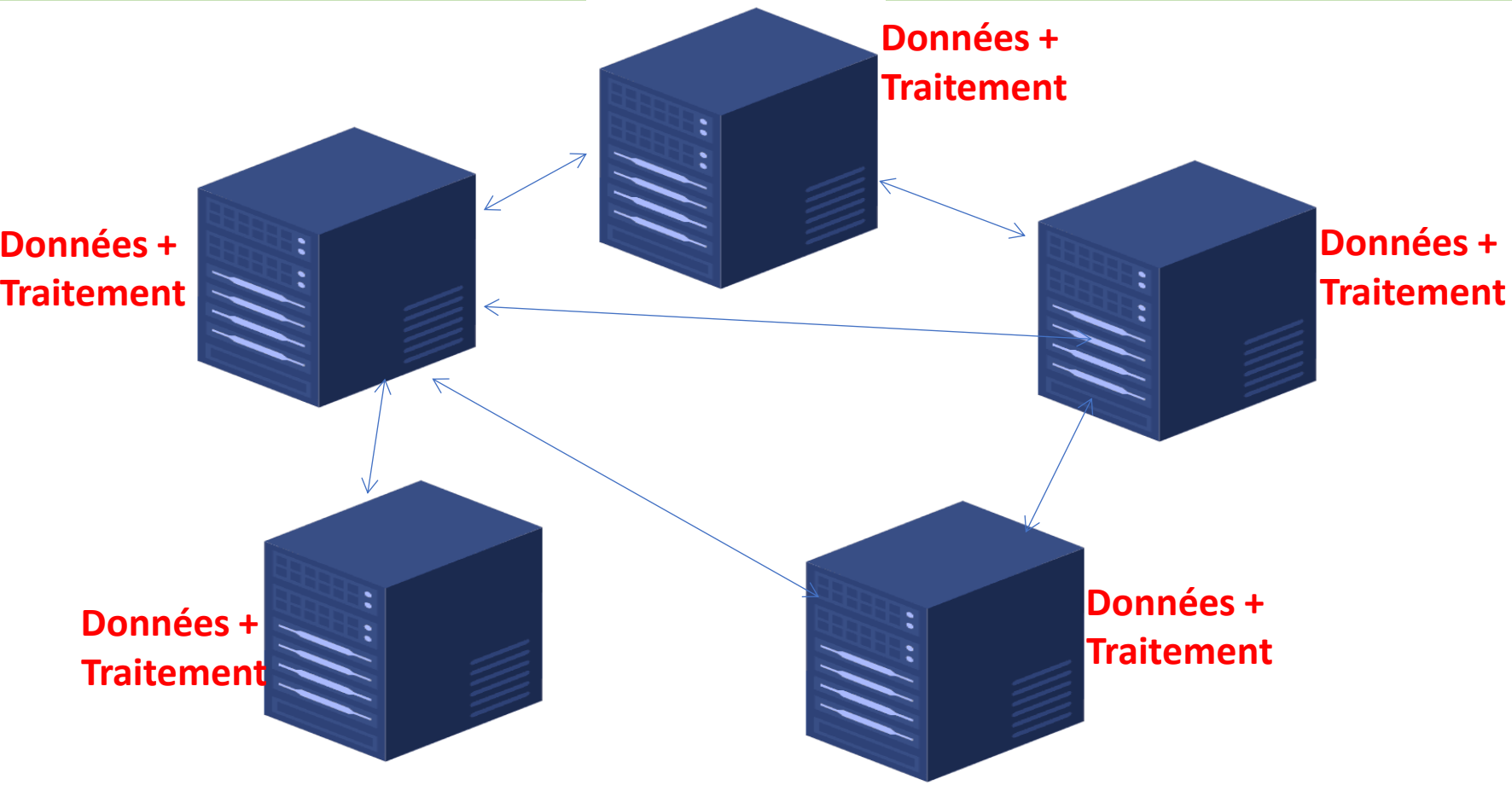
Colocalité

Systèmes Distribués Vs Système Big Data

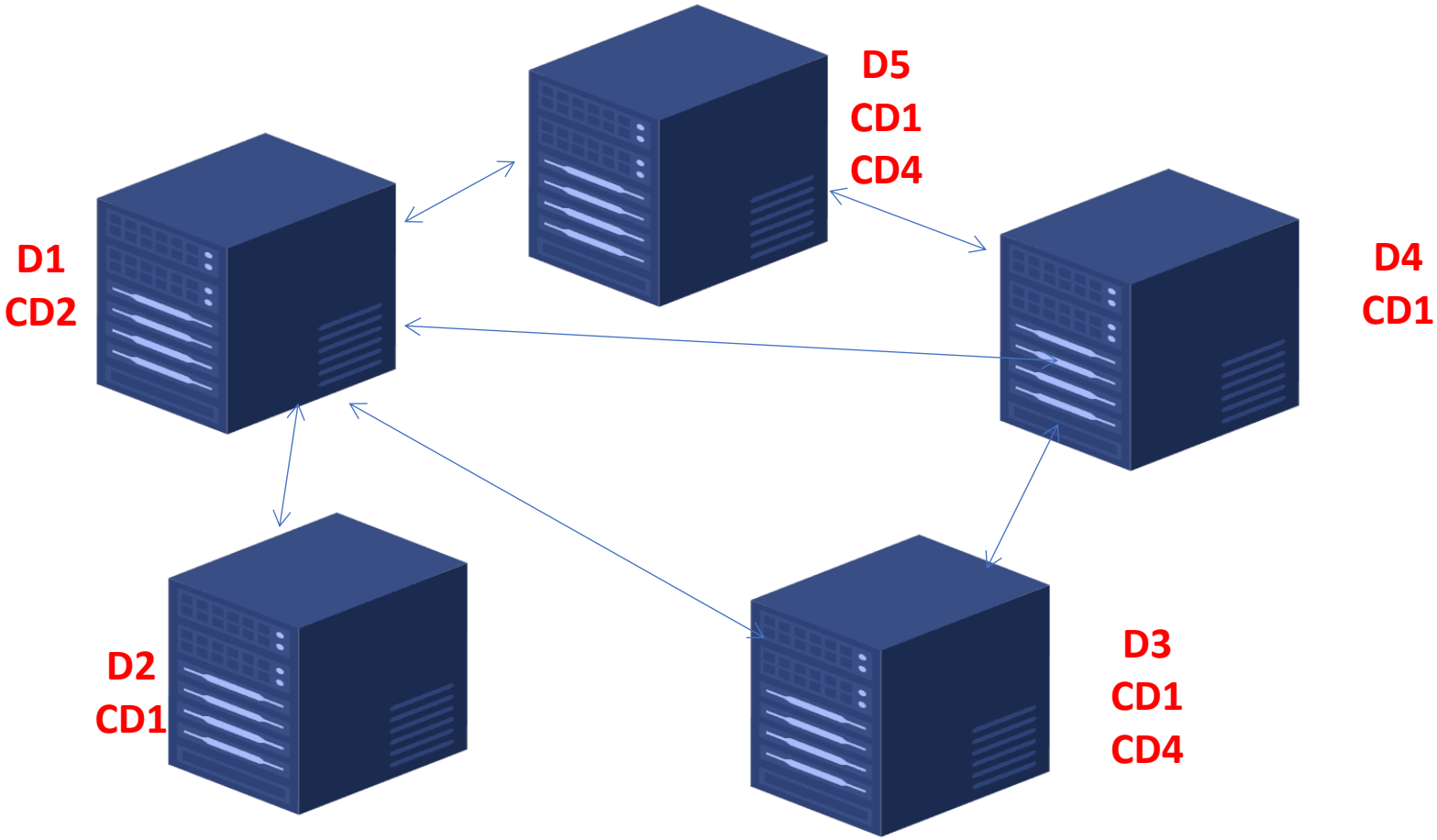
Architecture N-tiers Vs Architecture Big Data

- Un Système Big Data est différent d'un système distribué classique.
- La différence est dans le contenu des machines:
  - Système Distribué: chaque machine est dédiée à une tâche.
  - Système Big Data: Colocalité des données et traitements.

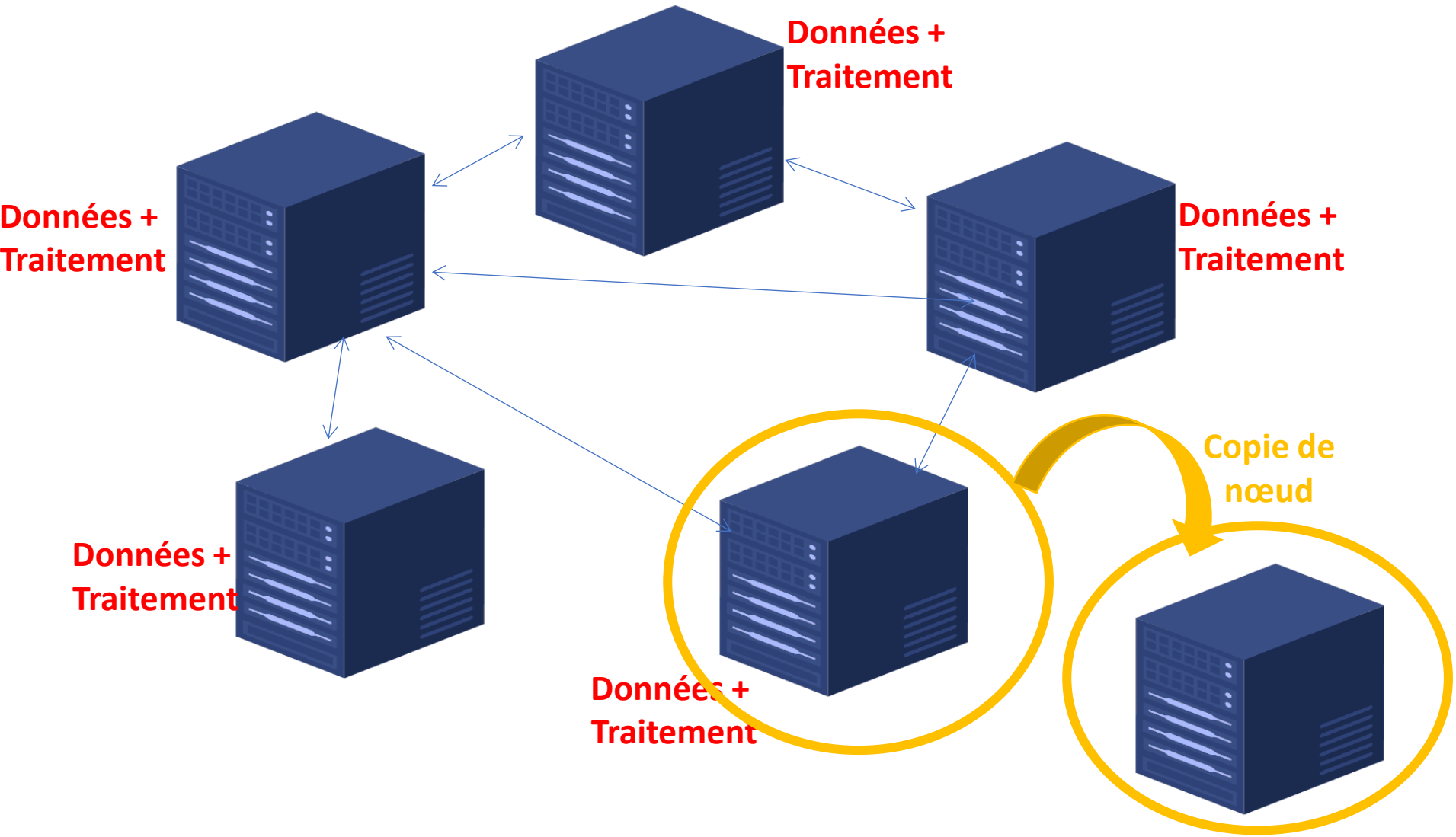
Colocalité



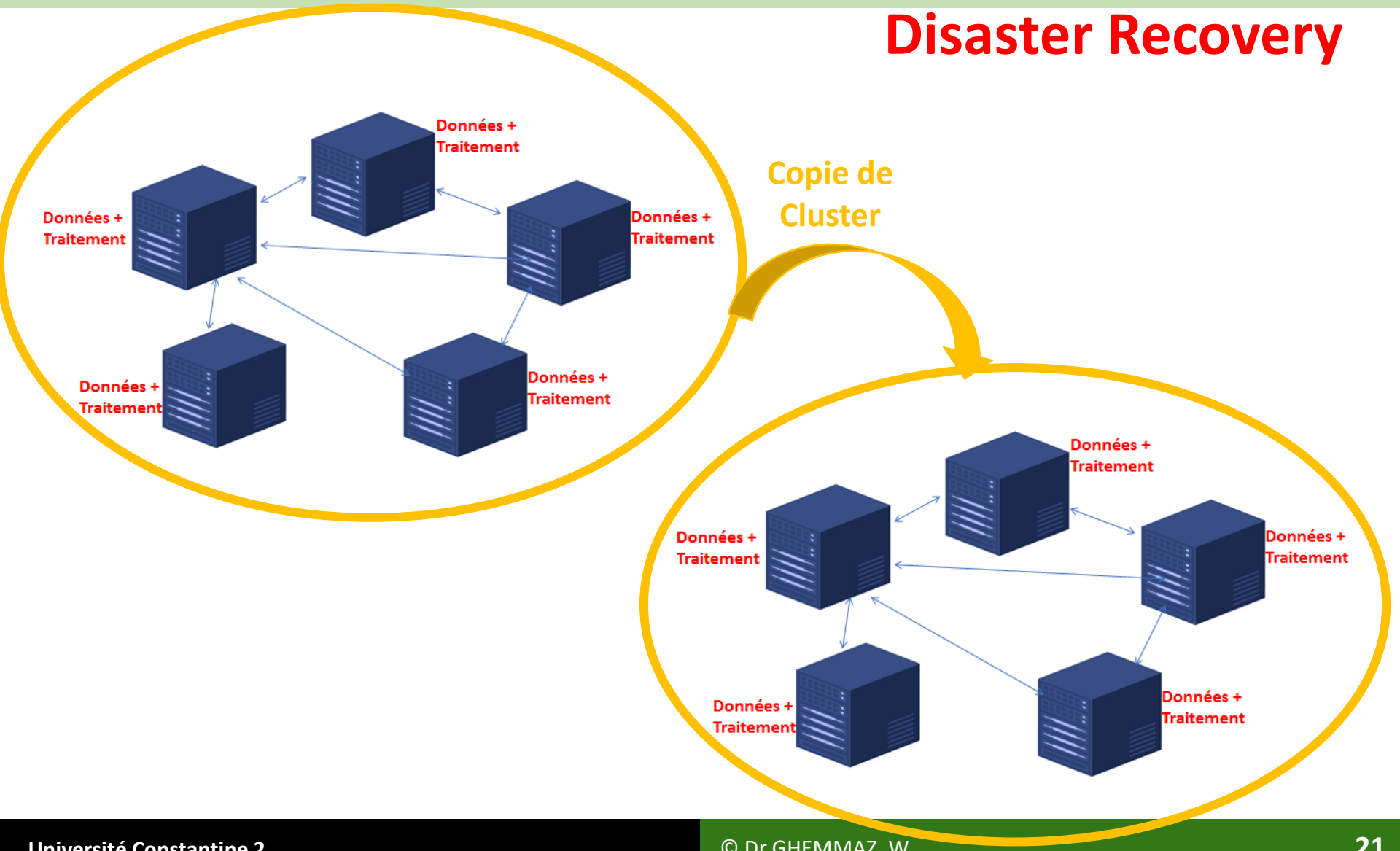
Réplication: 1ère stratégie



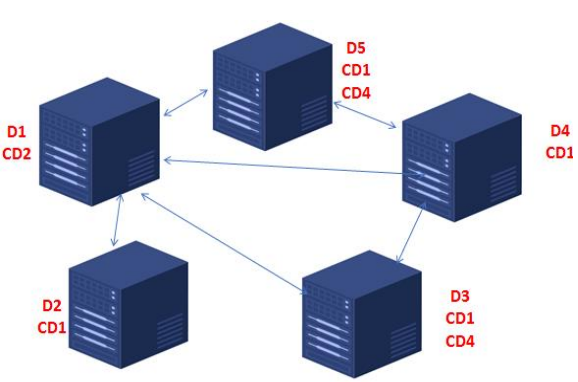
Réplication: 2ème stratégie

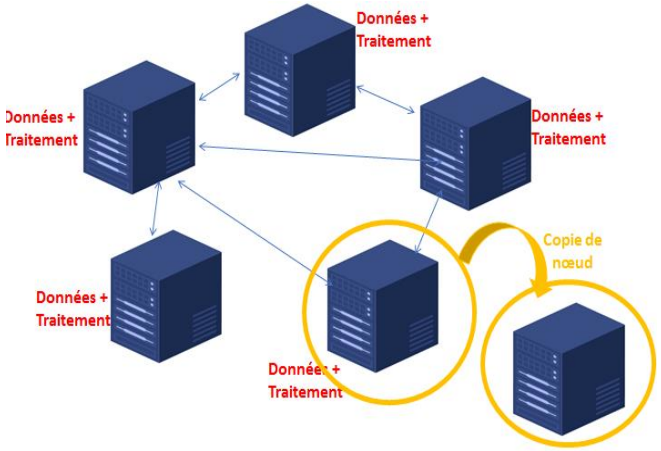


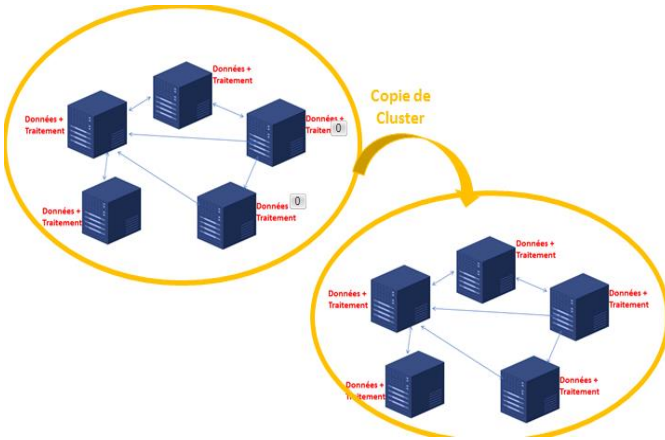
Réplication: 3ème stratégie



Réplication:







Problèmes :

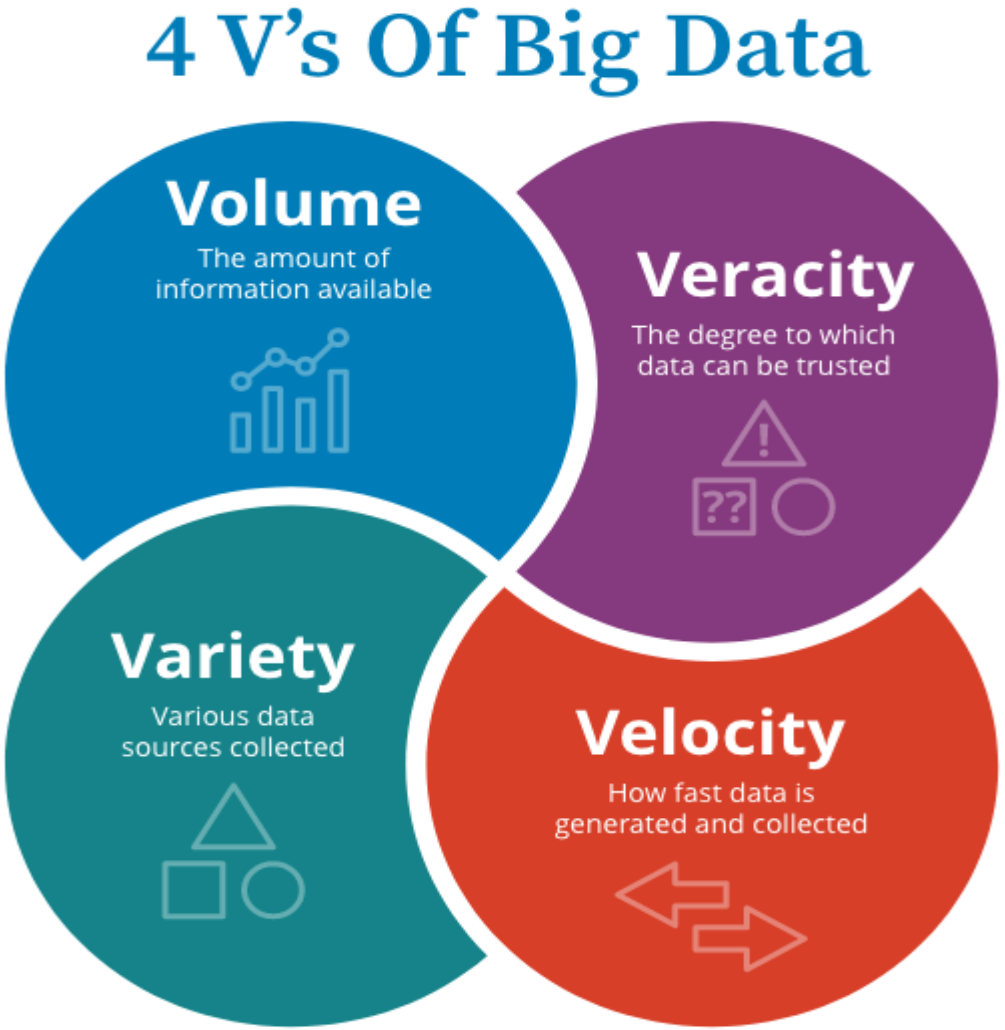
- **Synchronisation des données**
- **Volume multiplié (++)**

Plan

- Big Data
- Les nV des Big Data
- ACID Vs BASE
- Framework Big Data
- BDR Vs NoSQL

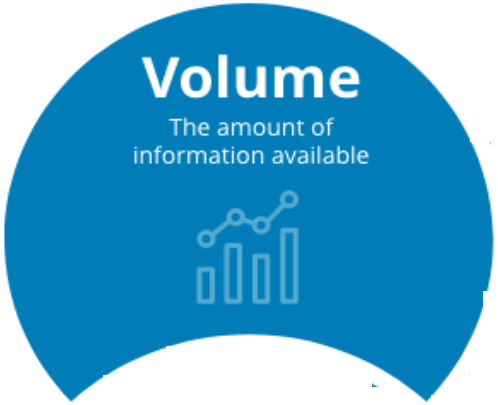
Les V du Big Data

Les 4 Vs (Selon McKinsey & Company, IBM, Gartner...)



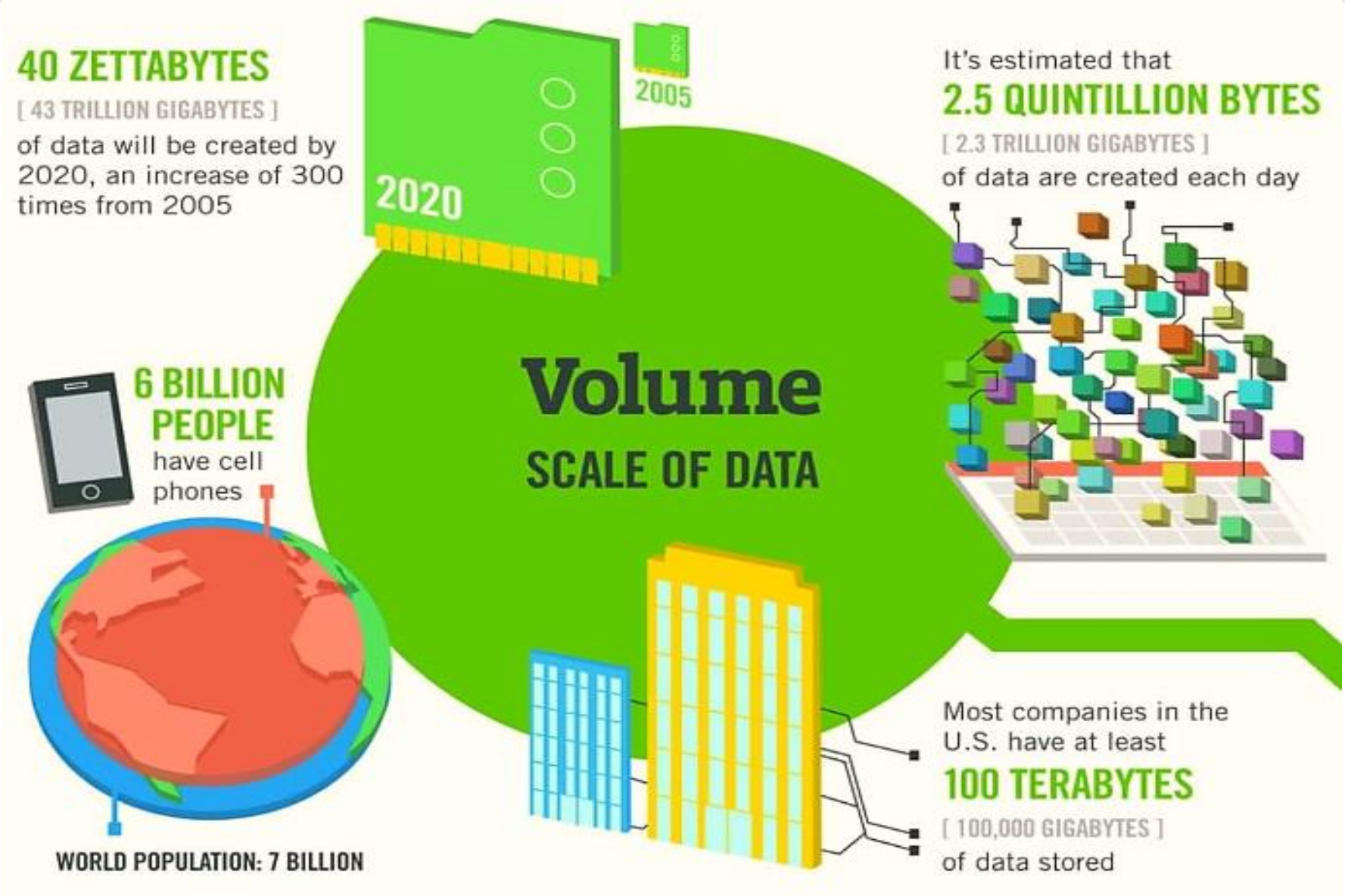
Le n V du Big Data

Les 4 Vs de base (Selon McKinsey & Company, IBM, Gartner...)



Le volume est la principale caractéristique du Big Data. Le terme est en effet directement tiré de l'immense masse de données générées au quotidien.

Les V du Big Data



Les V du Big Data

Les 4 Vs (Selon McKinsey & Company, IBM, Gartner...)

La vitesse et les directions à partir desquelles les données arrivent dans l'entreprise augmentent en raison de l'interconnexion et des progrès de la technologie des réseaux, de sorte qu'elles arrivent parfois plus vite que nous ne pouvons en tirer un sens.

Les V du Big Data

The New York Stock Exchange captures

1 TB OF TRADE INFORMATION

during each trading session

Modern cars have close to

100 SENSORS

that monitor items such as fuel level and tire pressure

Velocity

ANALYSIS OF STREAMING DATA

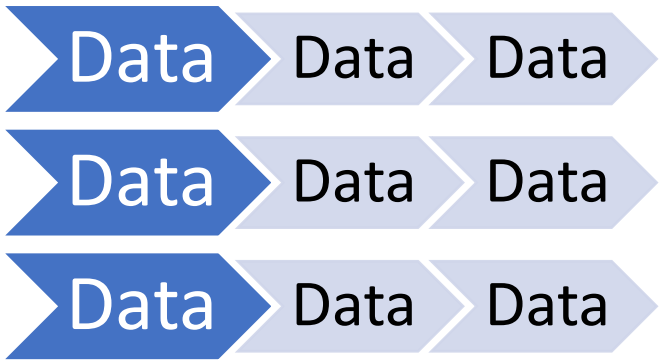
By 2016, it is projected there will be

18.9 BILLION NETWORK CONNECTIONS

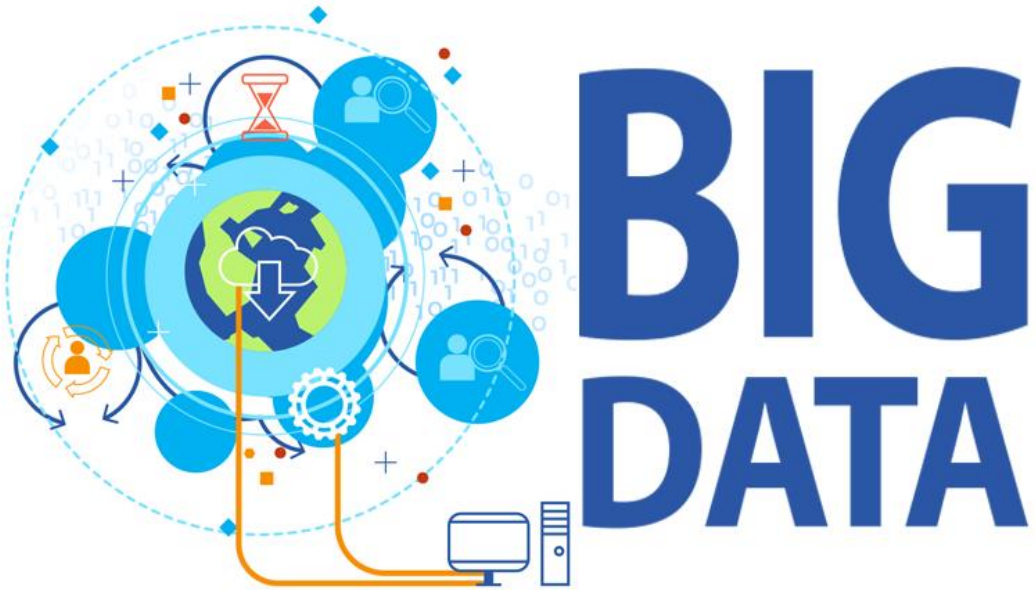
– almost 2.5 connections per person on earth

Le n V du Big Data

Vélocité:



Flux de données

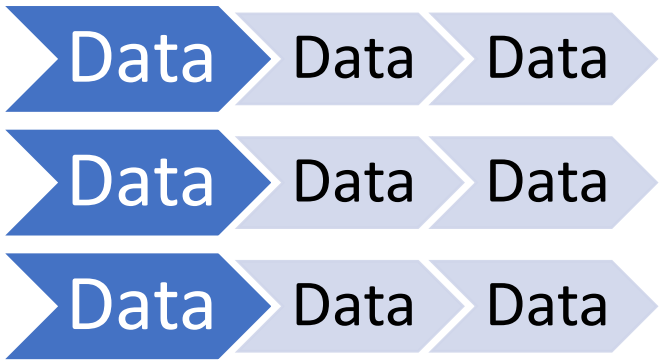


Système Big Data

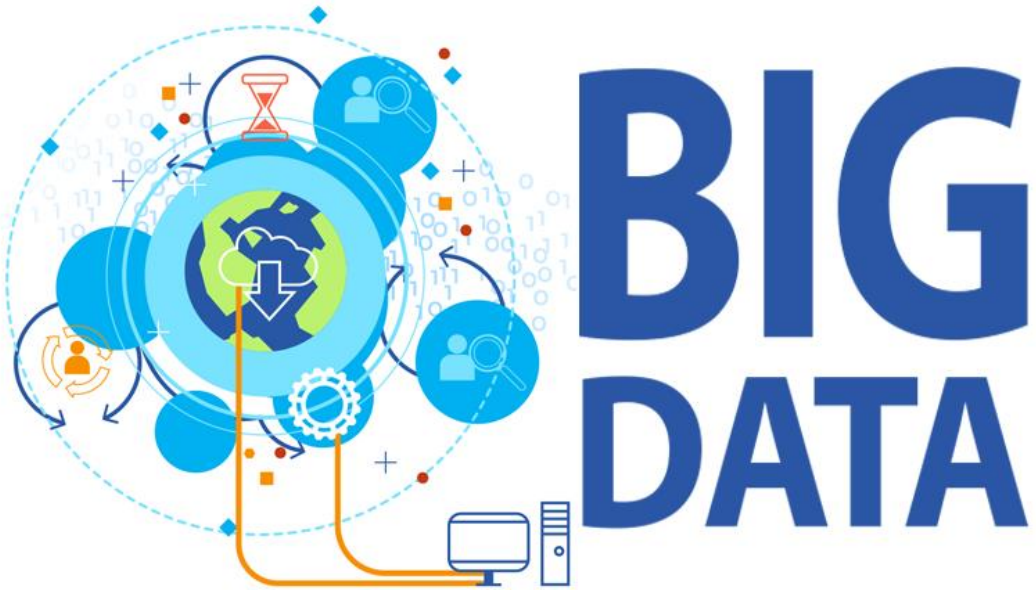
- Le système Big Data doit assurer un traitement à la volée
- Le Problème de traitement à la volée:
  → La vitesse de traitement des données doit être supérieure à la vitesse d'arrivée de données.

Le n V du Big Data

Vélocité:



Flux de données

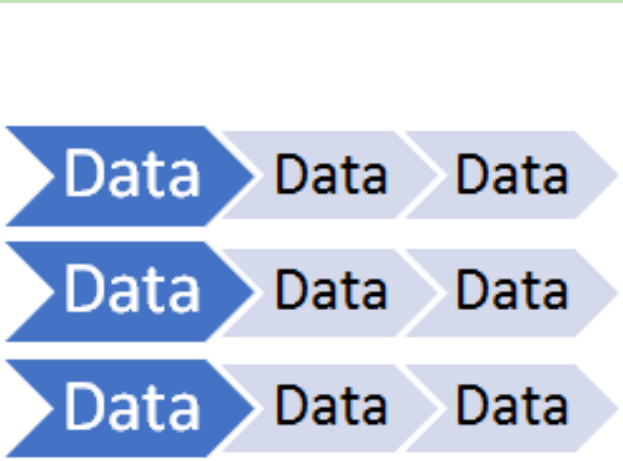


Système Big Data

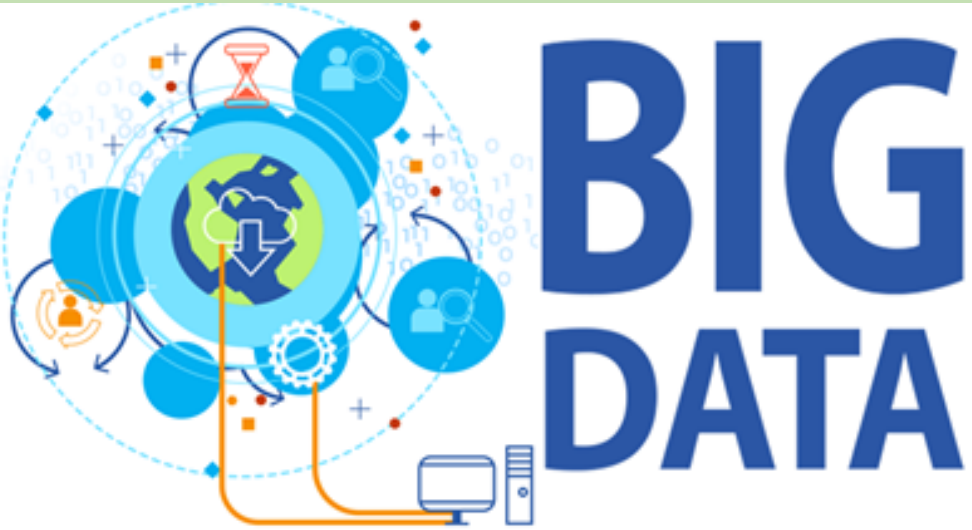
- **Le système Big Data doit être disponible tout le temps**
  (même en cas d'un flux de données massif).

Le n V du Big Data

Vélocité:



Flux de données



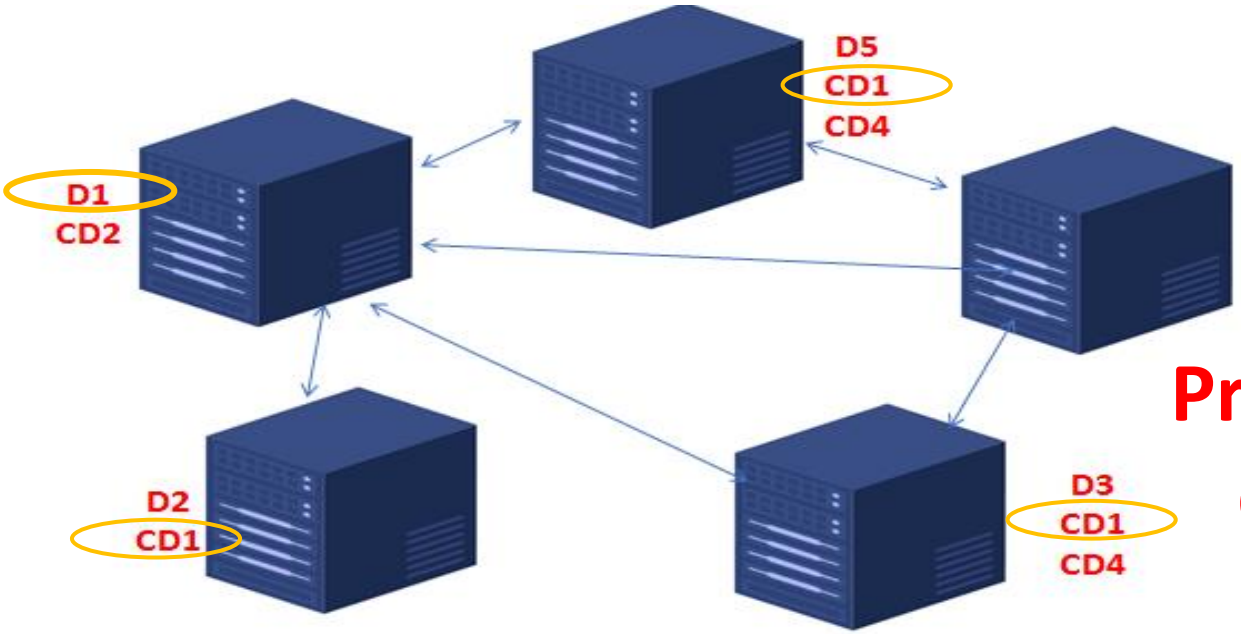
Système Big Data

Disponibilité (Availability)

- **Principe:** **Stocker d'abords, Réfléchir ensuite**
  (contrairement au BDR et BI)

Le n V du Big Data

Vélocité:



Propagation des MAJ

Le n V du Big Data

Vélocité:

Disponibilité (Availability)

Cohérence (Consistency)

Le n V du Big Data

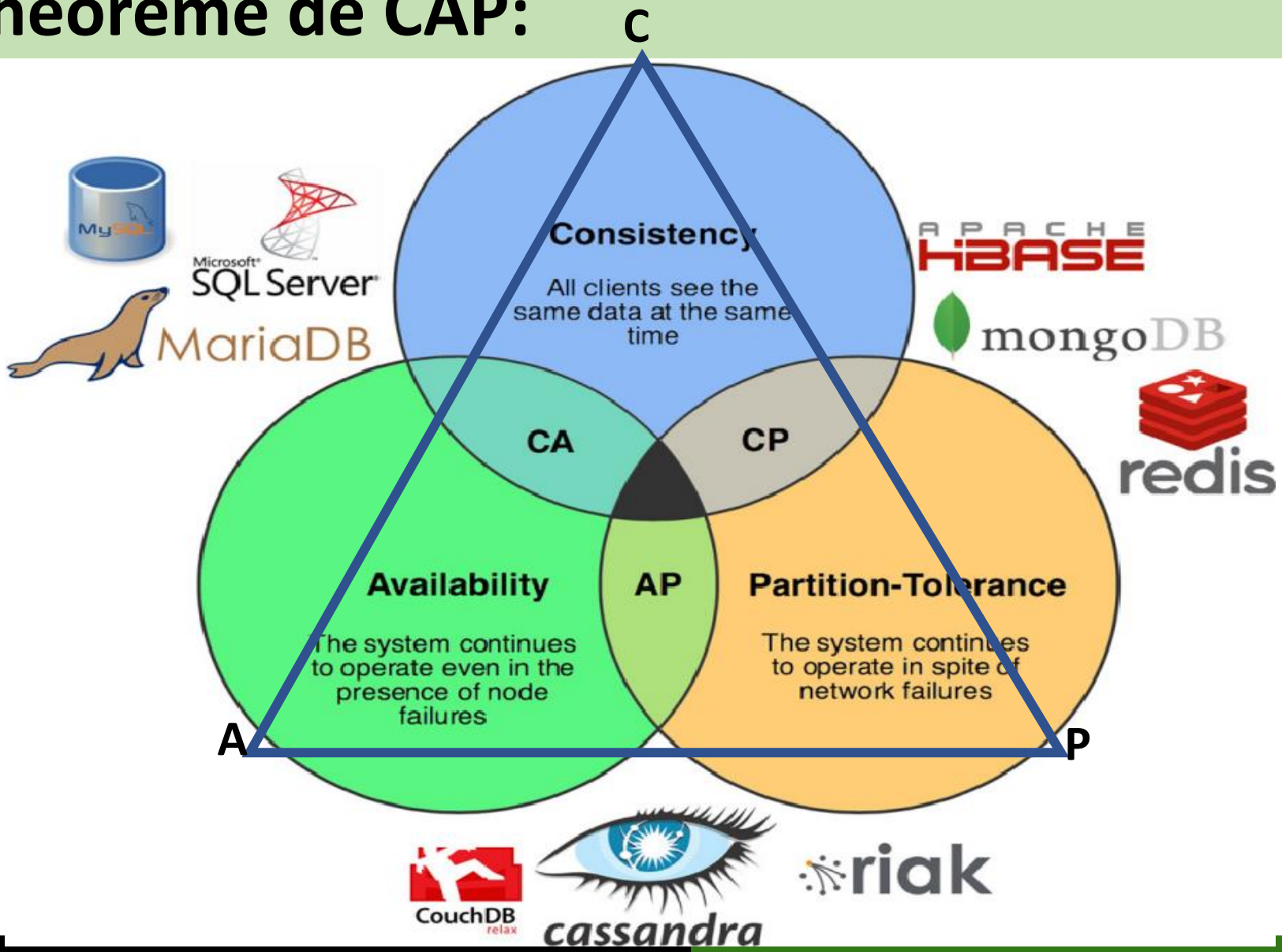
Vélocité:

Cohérence (Consistency)

Disponibilité (Availability)

Le n V du Big Data

Théorème de CAP:



Le n V du Big Data

Théorème de CAP:

Vélocité

Cohérence (Consistency)

Disponibilité (Availability)

Volume

Distribution (Partition-tolerance)

Le n V du Big Data

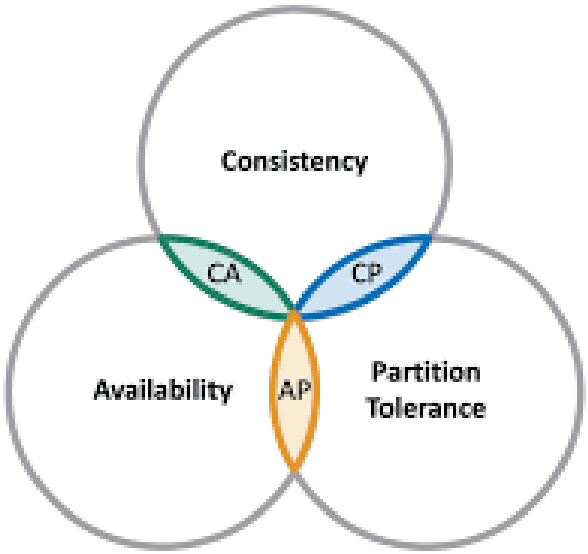
Théorème de CAP:

Exemple:

BDR→ Assure CA

BDD→ Assure CP

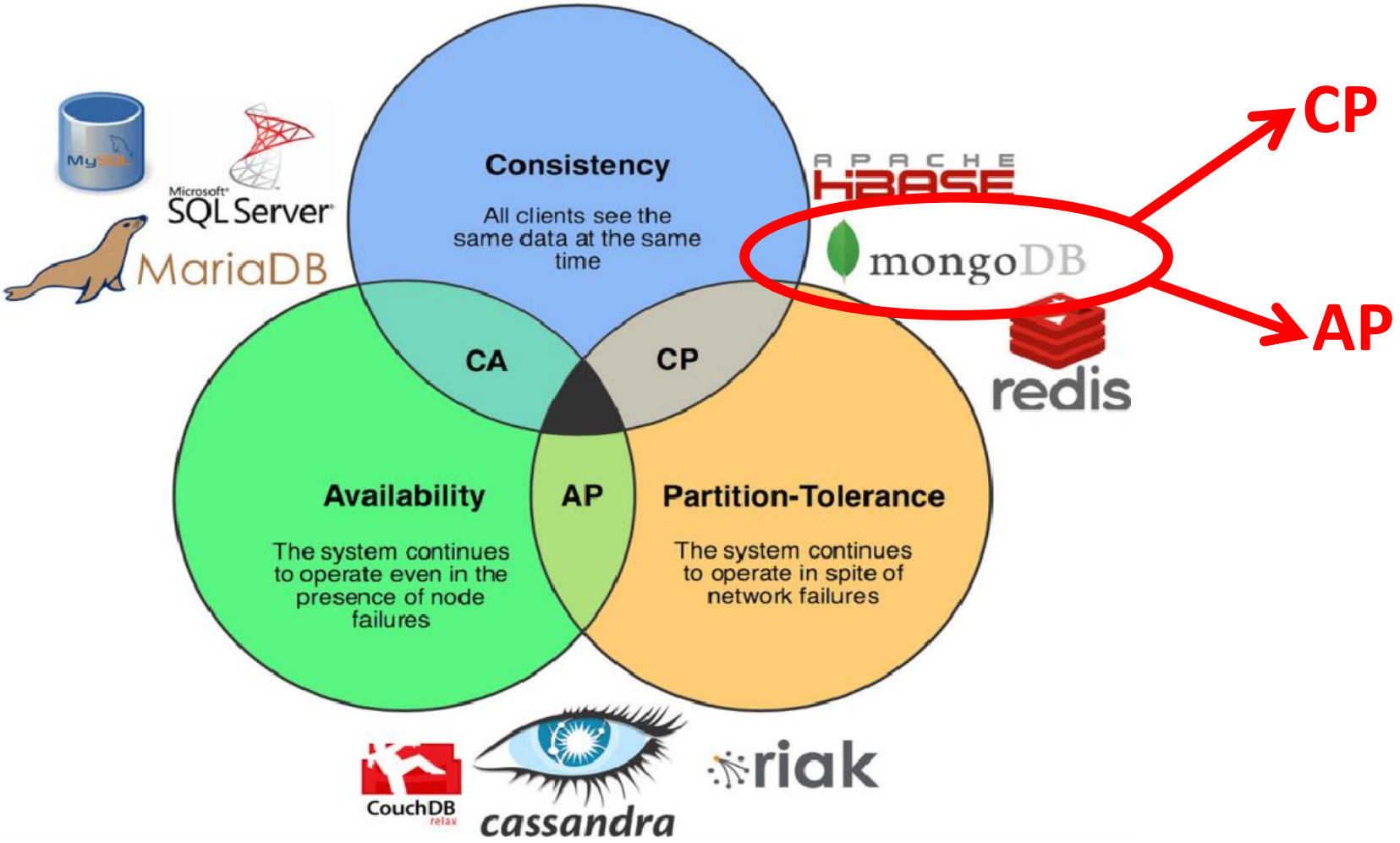
BBigData→ Assure PA



Attention:

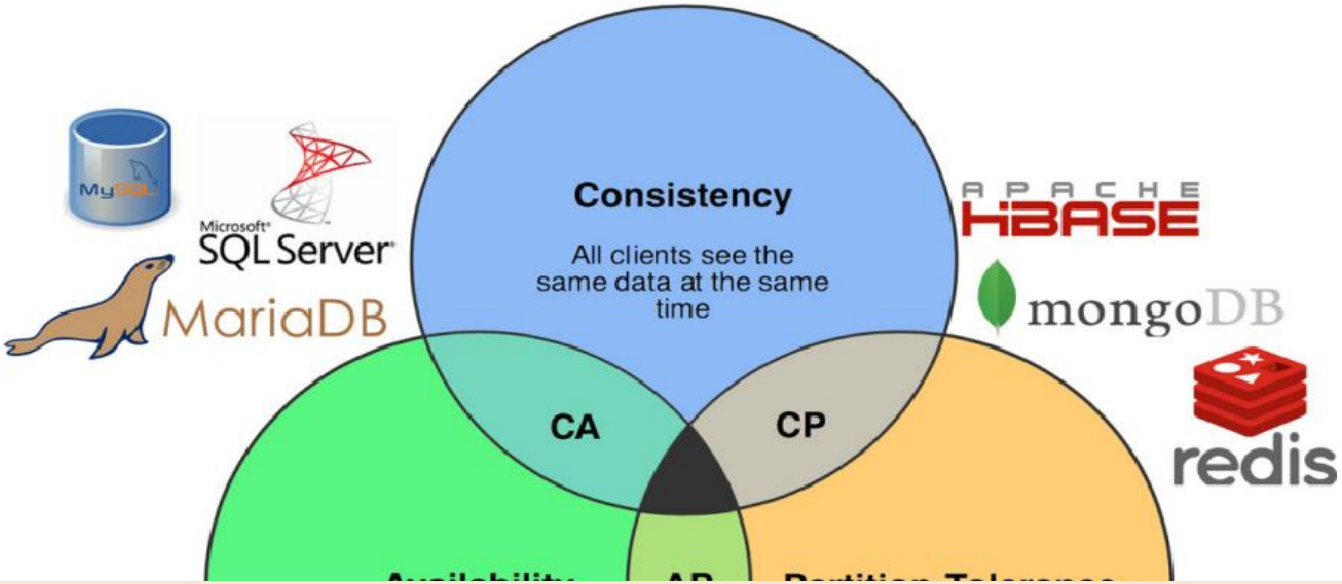
Les Bases Big Data par défaut favorisent la disponibilité que la consistance mais proposent également des mécanismes pour assurer la cohérence.

Théorème de CAP:



Le n V du Big Data

Théorème de CAP:



Attention:

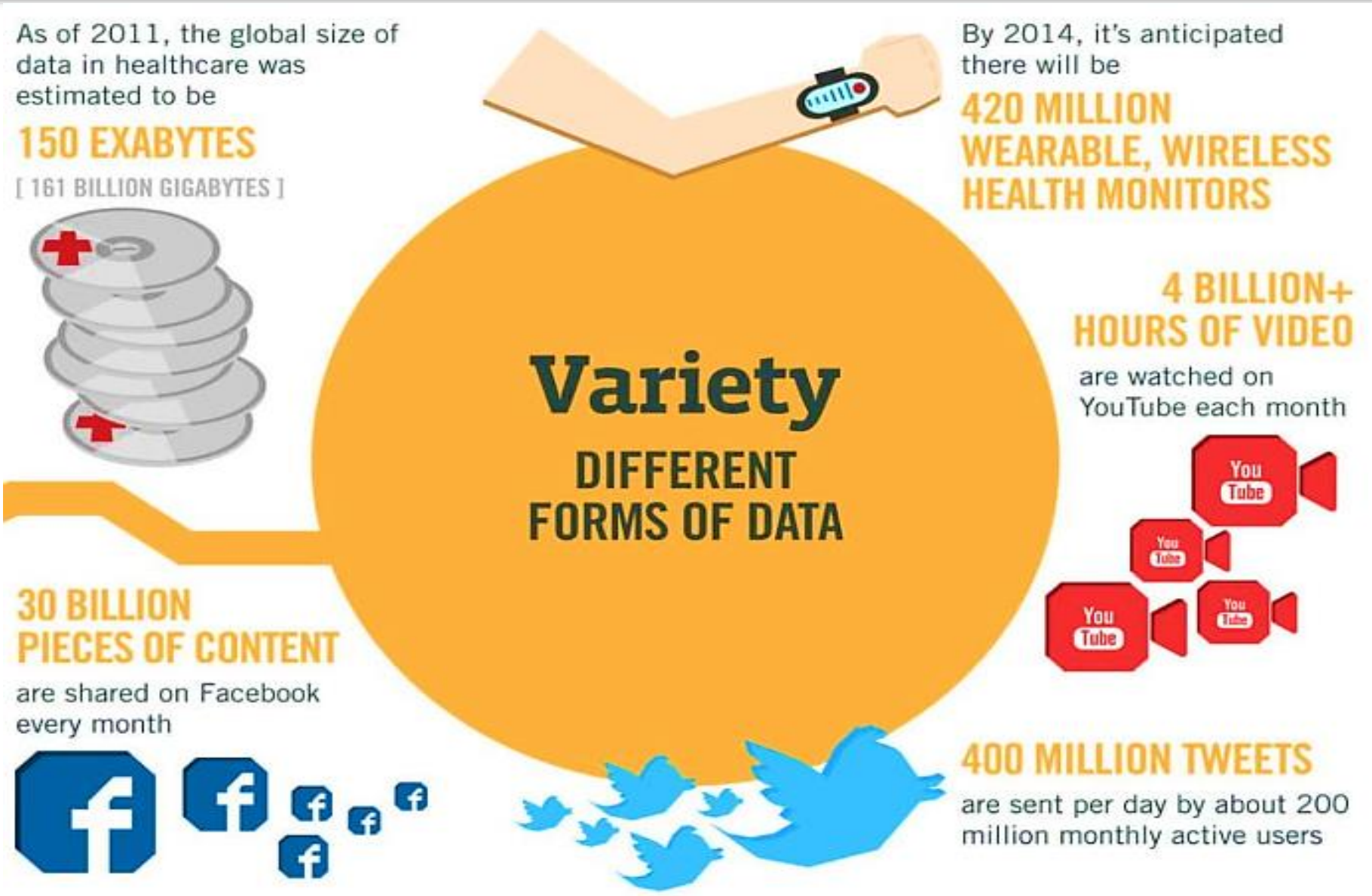
- Tous les systèmes Big Data peuvent stocker des données mais la façon de gérer les données et de traiter les requêtes se diffère d'un système à un autre,
- Le choix d'un Système Big Data doit être bien étudié selon nos besoins.

Les V du Big Data

Les 4 Vs (Selon McKinsey & Company, IBM, Gartner...)

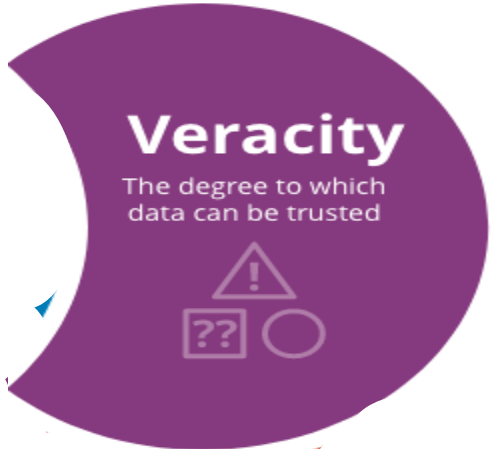
Les données sont également plus diversifiées que jamais. Ce phénomène est lié à la diversification des usages d'internet et du numérique. La provenance des données, leur format, mais également le domaine auquel elles sont liées connaissent une variété sans précédent.

Les V du Big Data



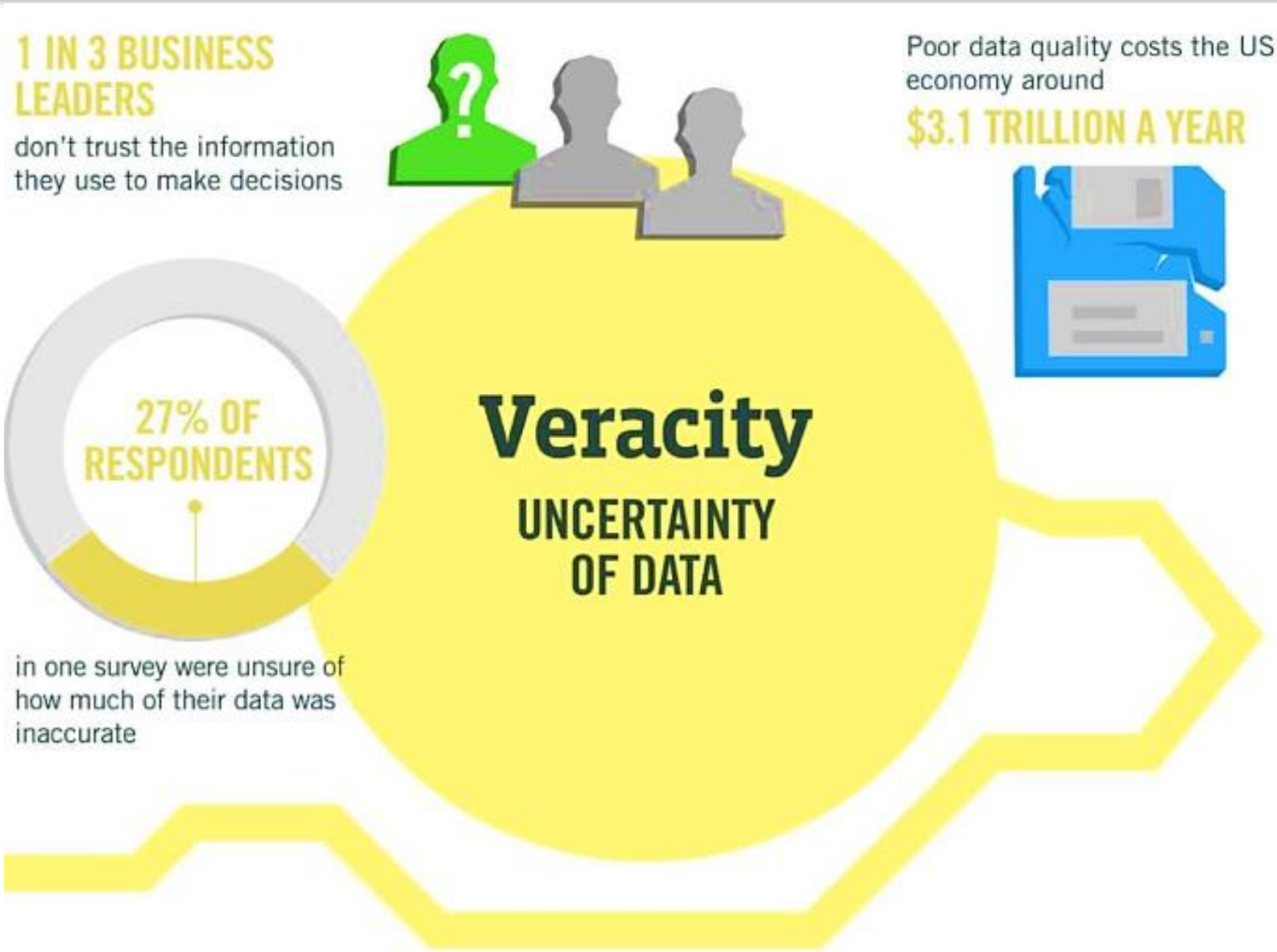
Les V du Big Data

Les 4 Vs (Selon McKinsey & Company, IBM, Gartner...)



La véracité des données ou la quantité de données fiables lorsque des décisions clés doivent être prises sur des volumes aussi importants et collectés aussi rapidement est indispensable.

Les V du Big Data



Le n V du Big Data

Résumé:

Volume → **Scalabilité**

Vélocité → **Disponibilité**

Variété → **Flexibilité**

Autres V liés au Big Data

Autres V liés au Big Data

Valeur :

La valeur fait référence à la valeur que le Big Data apporte et est directement liée à ce que les entreprises peuvent faire avec les données qu'elles collectent. Être capable de tirer de la valeur des Big Data est une condition préalable, car la valeur des Big Data augmente considérablement en fonction des informations qui peuvent en être tirées.

Autres V liés au Big Data

Variabilité :

La notion de variabilité qui traduit le changement des formats de données qui impose aux entreprises de disposer d'un système d'information en perpétuelle évolution

Autres Vs…41…

Plan

- Big Data
- Les nV des Big Data
- ACID Vs BASE
- Framework Big Data
- BDR Vs NoSQL

ACID vs BASE

ACID

- ✓ **Atomicité:** une transaction s'effectue entièrement ou pas du tout
- ✓ **Cohérence**: le contenu d'une base doit être cohérent au début et à la fin d'une transaction (mais pas forcément durant son exécution)
- ✓ **Isolation**: les modifications d'une transaction ne sont visibles/modifiables que quand celle-ci a validé
- ✓ **Durabilité**: une fois la transaction validée, l'état de la base est permanent (non affecté par les pannes ou autre)

ACID vs BASE

BASE

- ✓ **Basically Available** : garantie minimale pour taux de disponibilité face au grande quantité de requêtes
- ✓ **Soft-state** : l'état du système peut changer au cours du temps même sans nouveaux inputs (cela est du au modèle de consistance).
- ✓ **Eventually Consistent** : tous les réplicas atteignent le même état, et le système devient à un moment consistant, si on stoppe les inputs

ACID vs BASE

ACIDE

Atomicité: Lock , Rollback, Log

Cohérence: Contraintes d'intégrité (PK, FK)

Isolation: Lock

Durabilité: Log (Write Ahead Log (WAL))

BASE

Basically Available } **Disponibilité**

Soft-state : Etat mou (Non rigide)

Eventually consistent

Plan

- Big Data
- Les nV des Big Data
- ACID Vs BASE
- Framework Big Data
- BDR Vs NoSQL

Big Data

- **Big Data sert à designer des ensembles de données tellement importants qu'il est difficile de les traiter avec les moyens habituels.**
- **La volumétrie de ces données oblige à concevoir des outils de stockage et de manipulation spécifique.**
- **Le développement de ces outils suscite un intérêt grandissant auprès des acteurs scientifiques et économiques en leur offrant la possibilité d'extraire de nouvelles informations à partir de la masse de données.**

Big Data

La problématique du Big Data s'inscrit dans un contexte complexe à la croisée du 3 préoccupations majeurs:

1. **La mise en œuvre de nouvelles solutions de stockage de masse.**
2. **La capture d'informations à grande vitesse et si possible en temps réel.**
3. **La connexion du SI avec de nouvelles sources de données liées au Web et son adaptation à la notion de l'Open Data (Données Ouvertes)**

Big Data

Pour répondre à ces 3 préoccupations, il faut développer des solutions technologiques comprenant:

1. **Des outils innovants de restitution.**
2. **Des outils innovants de visualisation de données adaptés aux volumétries.**
3. **Des nouvelles solutions de stockage.**

Big Data

Pour répondre à ces 3 préoccupations, il faut développer des solutions technologiques comprenant:

1. Des outils innovants de restitution.
2. Des outils innovants de visualisation de données adaptés aux volumétries.
3. **Des nouvelles solutions de stockage.**

Bases de données NoSQL

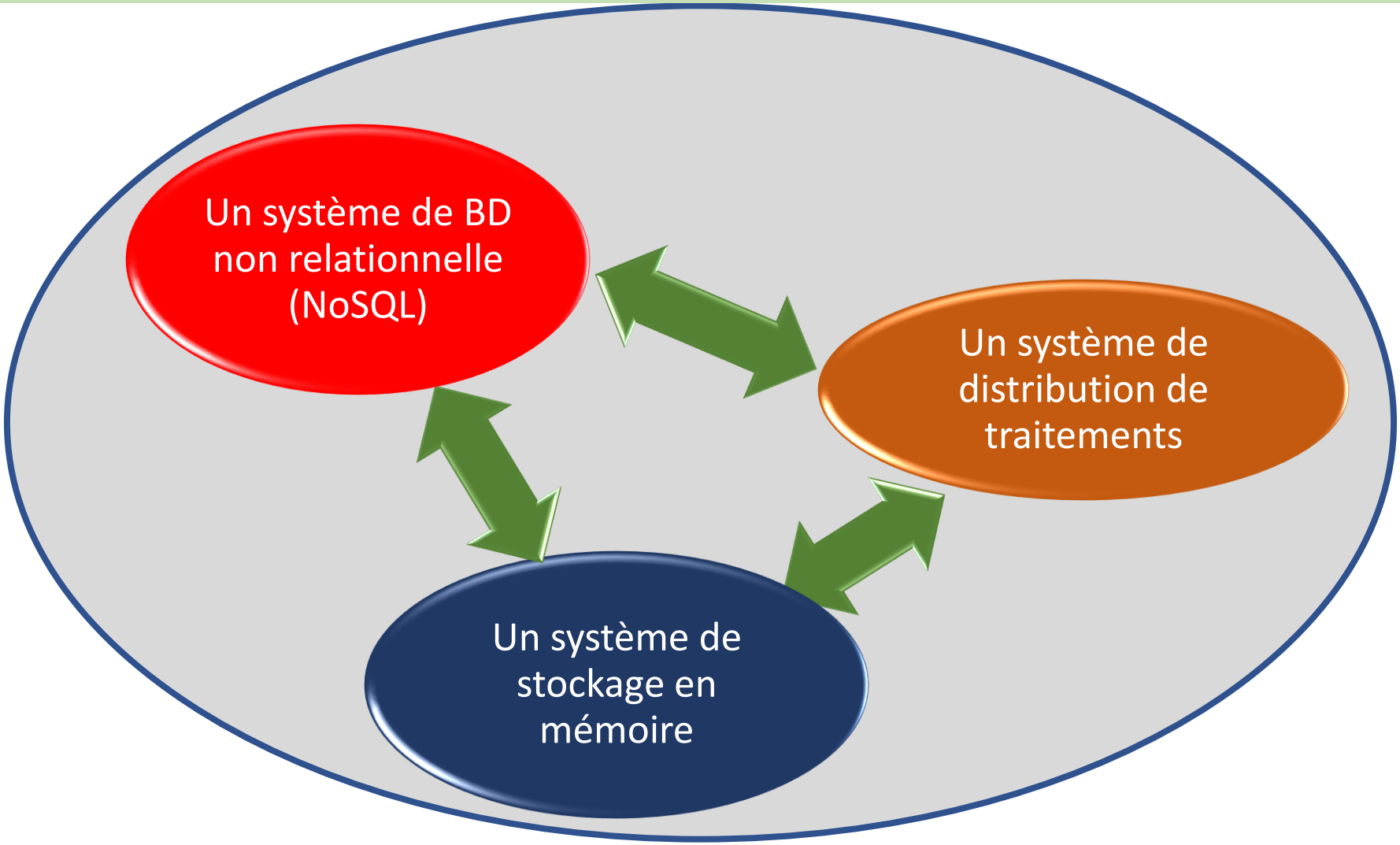
La problématique du Big Data

Le terme Base de données « **NoSQL** » Regroupe des solutions récentes qui se différencient du modèle SQL par une logique de représentation des données différentes.

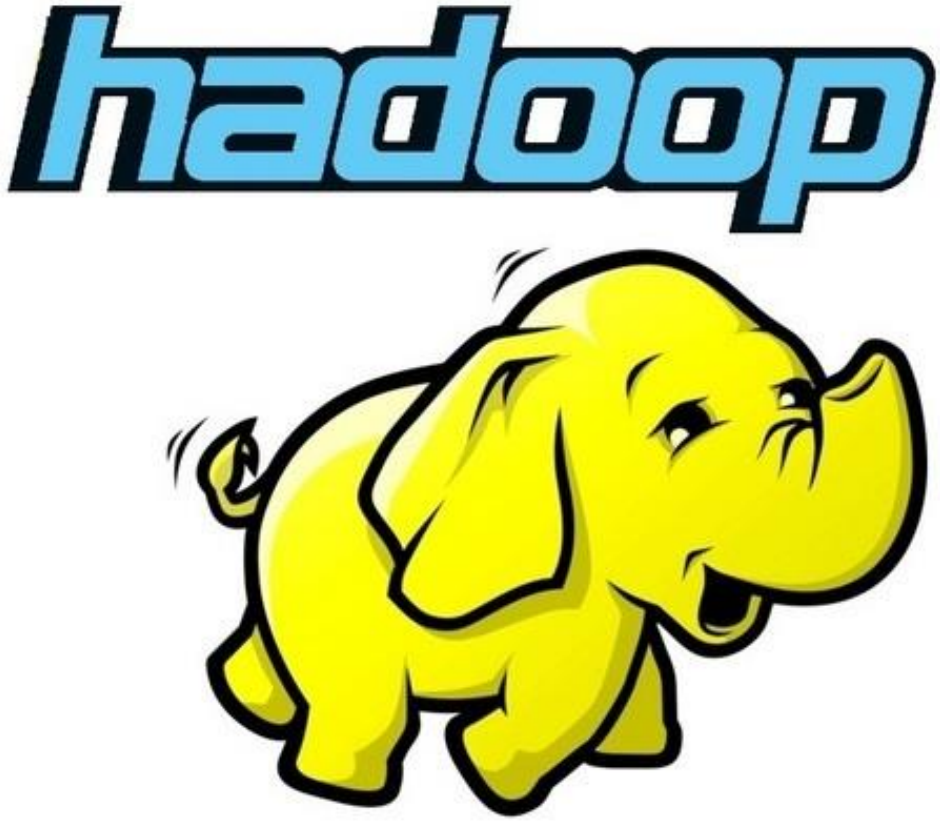
Les principaux avantages de ces solutions sont leurs performances et leur capacité à traiter de très grands volume de données; en particulier lorsque cela concerne le stockage de données dont la structure varie.

La problématique du Big Data

Définition d'un Framework pour le Big Data



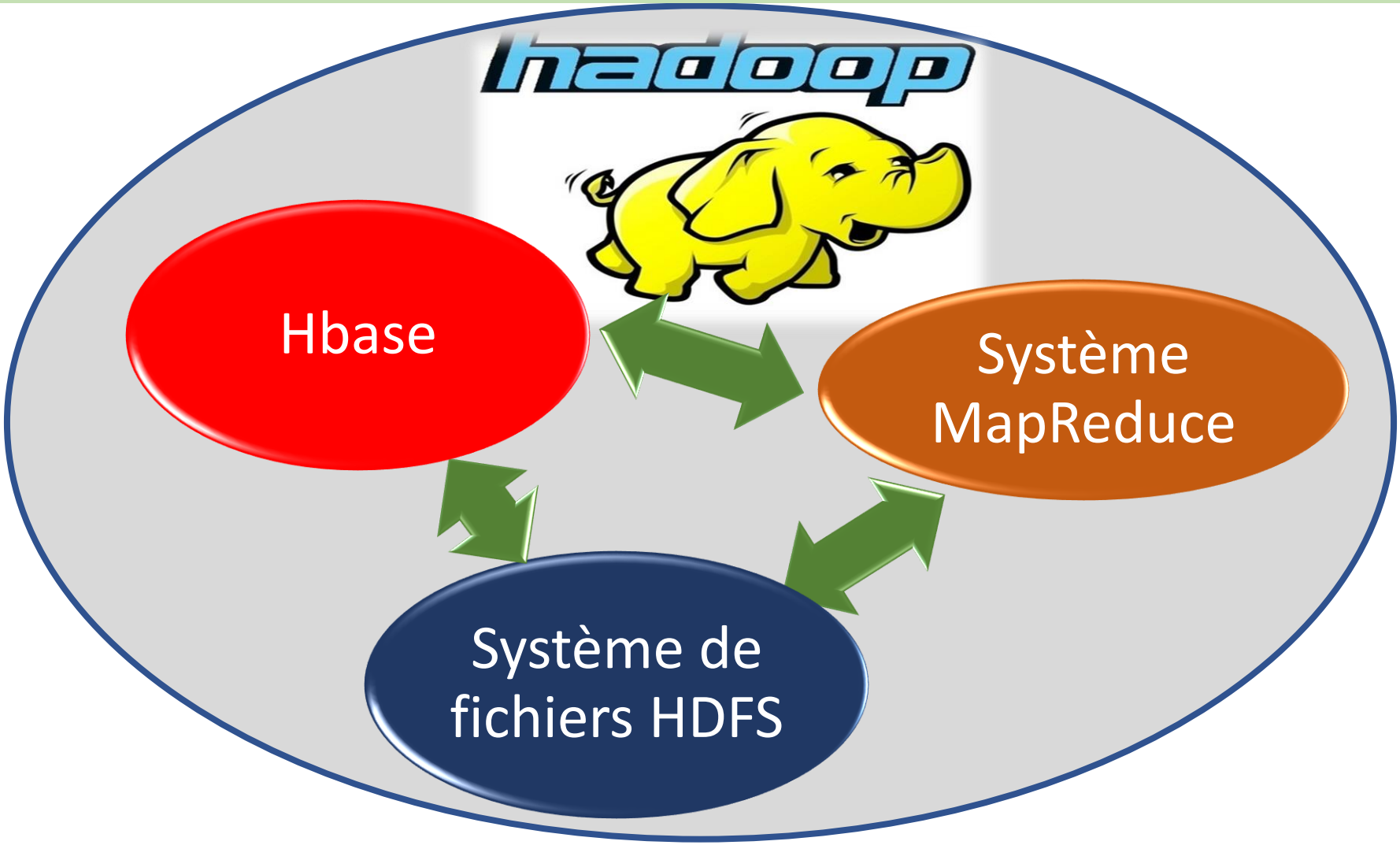
Définition d'un Framework pour le Big Data



Le plus connu des frameworks de Big Data et sans conteste: Hadoop

La problématique du Big Data

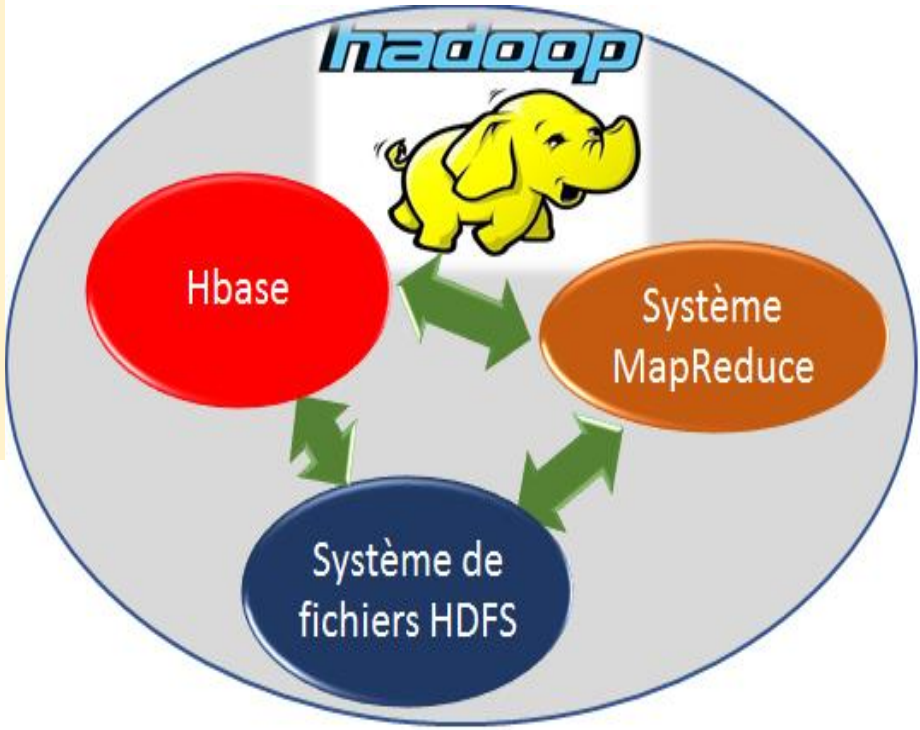
Définition d'un Framework pour le Big Data



Définition d'un Framework pour le Big Data

Attention:

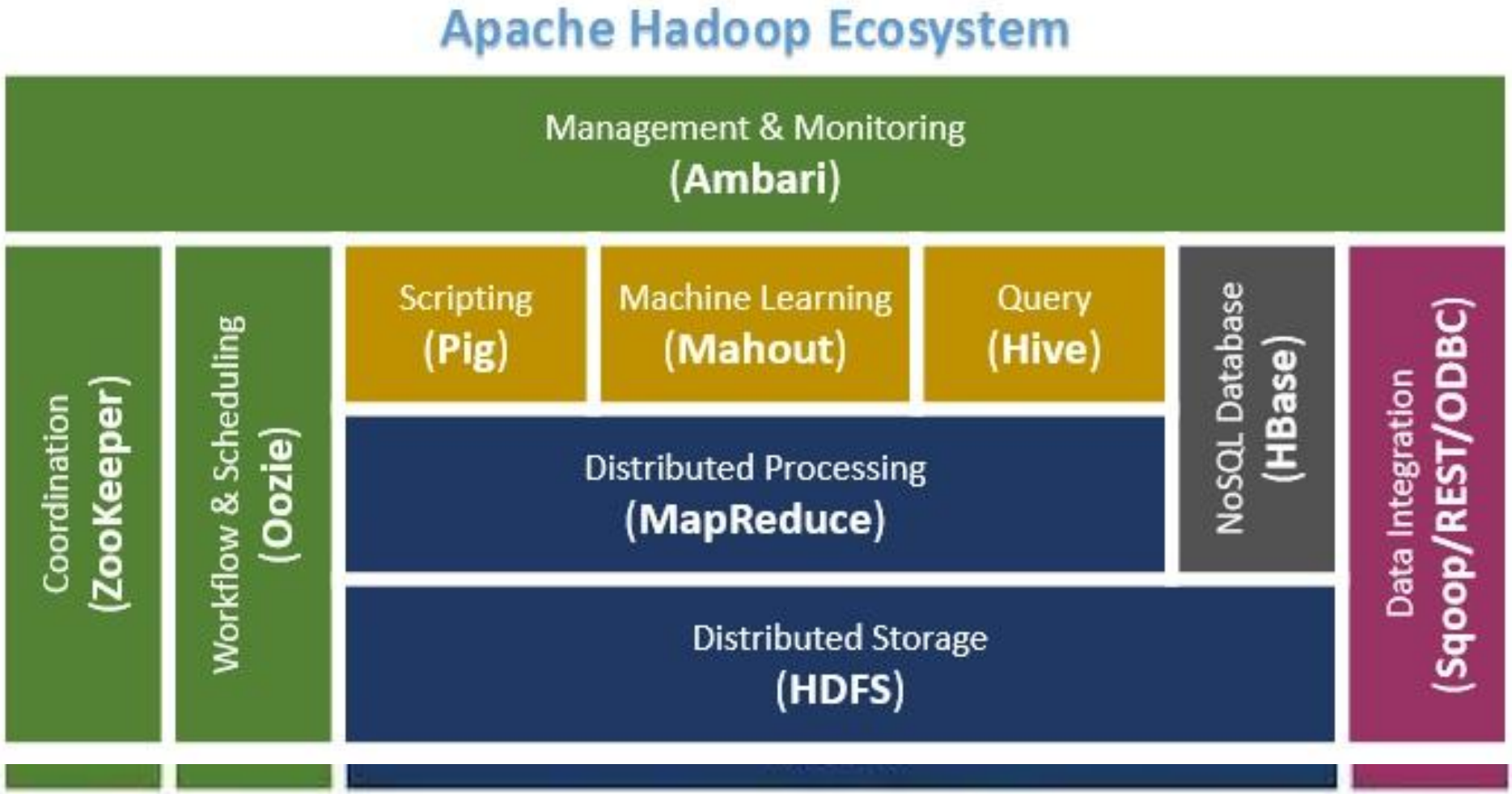
Cette représentation ne peut être que schématique. Seuls les principaux composants y sont mentionnés.



Pour Chaque framework dédié au Big Data, il existe un **écosystème**

Un **écosystème** est un ensemble d'applications prévues pour fonctionner ensemble. Il couvre à la fois les problématiques de stockages de données mais aussi de restitution des résultats (Tableaux de bord, fouille de données, etc.)

Définition d'un Framework pour le Big Data

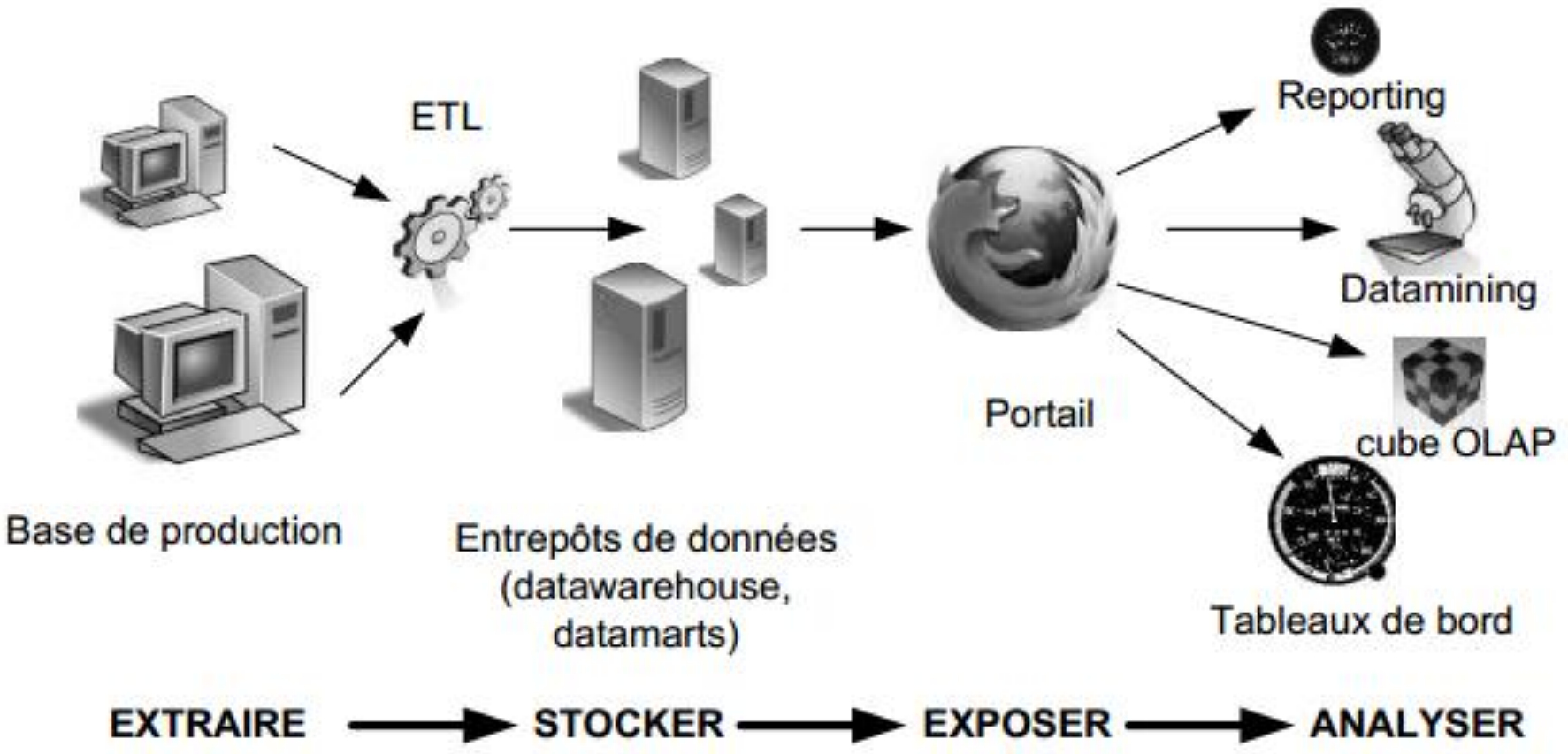


Big Data & BI

BIG DATA

BUSINESS INTELLIGENCE

Big Data & BI


ETL
Reporting
Datamining
Portail
cube OLAP
Base de production
Entrepôts de données
(datawarehouse,
datamarts)
Tableaux de bord
EXTRAIRE
STOCKER
EXPOSER
ANALYSER


Talend ETL
ProjectPro

Big Data & Cloud Computing

Cloud Computing est une solution logicielle permettant de proposer des solutions de Big Data à des acteurs économiques dont les ressources financières et matérielles ne leur permettent pas de mettre en place eux-mêmes les moyens nécessaires pour rentrer dans le traitement des données volumineuses en temps réels comme promettent de le faire les solutions de Big Data.

Big Data & Cloud Computing

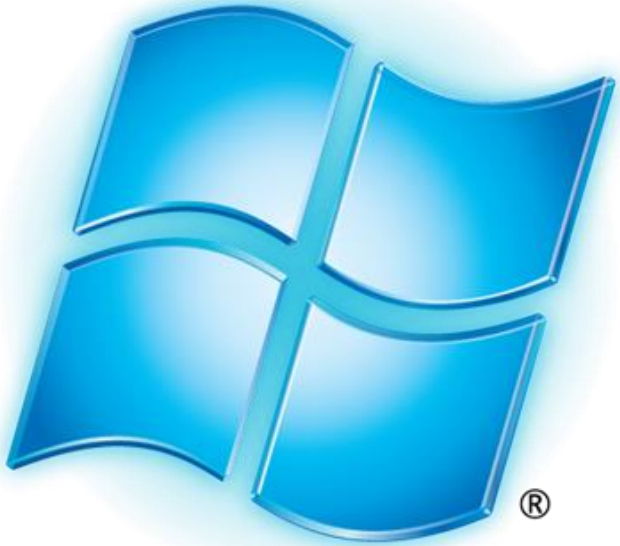
Cloud Computing consiste à proposer des ressources de calcul via Internet, ces ressources de calcul étant fournies par un tiers et donnant lieu à une facturation en fonction de volume de données ou des consommations des ressources de calculs.

Big Data & Cloud Computing & BI

Cloud Computing une solution technique qui permet de mettre en œuvre des outils d'analyse de Big Data (le plus souvent des BDs NoSQL) et de developper des nouvelles approches de traitement des données orientées BI

Les principaux fournisseurs des solutions Cloud proposent des services Big Data, la plupart basés sur des distribtions Hadoop


Google Cloud Platform



Windows Azure™


amazon
web services

IBM Cloud

Plan

- Big Data
- Les nV des Big Data
- ACID Vs BASE
- Framework Big Data
- BDR Vs NoSQL

Spécificités des BDRs

Lorsqu'on réalise l'analyse des données on passe généralement par la conception d'un MCD et MLD qui permettent à terme d'identifier:

- Les entités (Classes)
- Les attributs de ces entités
- Les relations qui existent entre ces entités

Cela suppose évidement que les données analysées sont représentatives des données à stocker et il est souhaitable que le format de données en entrée n'évolue pas en cours du temps.

Spécificités des BDRs

Exemple: Gestion d'une bibliothèque (Gestion des Livres)

On considère que:

- Un auteur peut créer plusieurs livres.
- Un livre est créé par un seul auteur.

Tableau 1-1. Liste des livres

Numéro	Titre	Prix	Auteur
10101	aaaaa	10	Emilie Castafiore
11111	ee	54	Emilie Chambord
80808	cccc	45	Pierre Dupont
90909	ddddd	35	Roland Momo
202022	bb	25	Sylvie Fabière

Tableau 1-2. Liste des auteurs

Nom	Prénom	Domicile	Numéro
Castafiore	Emilie	Paris	85478
Chambord	Emilie	Nice	3547
Dupont	Pierre	Avignon	542563
Fabière	Sylvie	Bordeaux	52136
Momo	Roland	Toulouse	8547585
Tintin	Thiery	Clermont	78545

Spécificités des BDRs

Exemple : Gestion d'une bibliothèque (Gestion des Livres)

On considère que:

- Un auteur peut créer plusieurs livres.
- Un livre est créé par un seul auteur.

Tableau 1-1. Liste des livres

Numéro	Titre	Prix	Auteur
10101	aaaaa	10	Emilie Castafiore
11111	ee	54	Emilie Chambord
80808	cccc	45	Pierre Dupont
90909	ddddd	35	Roland Momo
202022	bb	25	Sylvie Fabière

Tableau 1-2. Liste des auteurs

Nom	Prénom	Domicile	Numéro
Castafiore	Emilie	Paris	85478
Chambord	Emilie	Nice	3547
Dupont	Pierre	Avignon	542563
Fabière	Sylvie	Bordeaux	52136
Momo	Roland	Toulouse	8547585
Tintin	Thiery	Clermont	78545

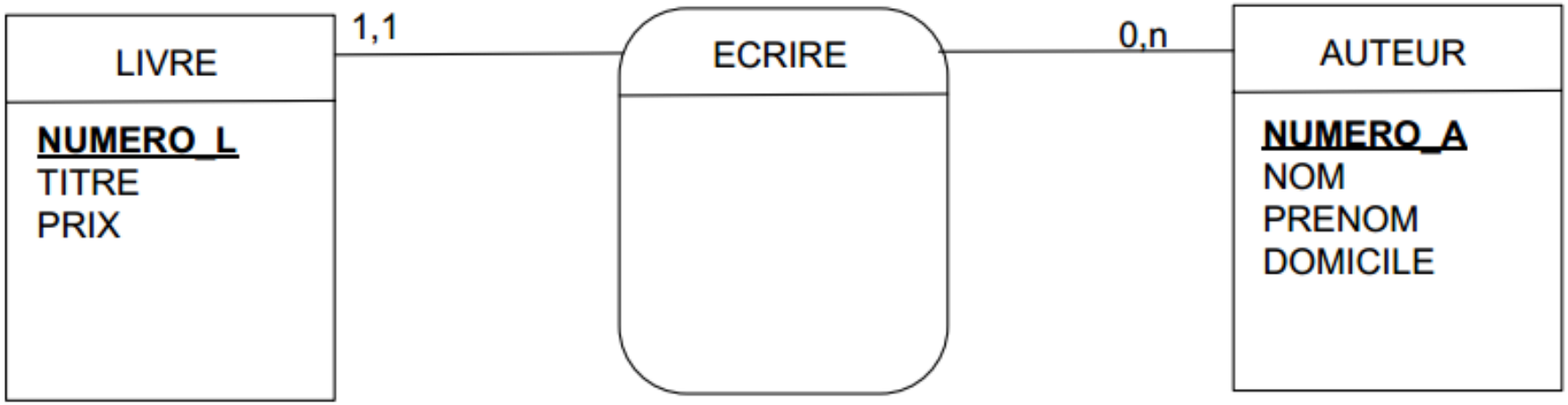
Un auteur est stocké dans la base de données alors qu'aucun de ses livres ne figure dans cette bibliothèque.

Spécificités des BDRs

Exemple : Gestion d'une bibliothèque (Gestion des Livres)

Le schéma fait apparaitre que :

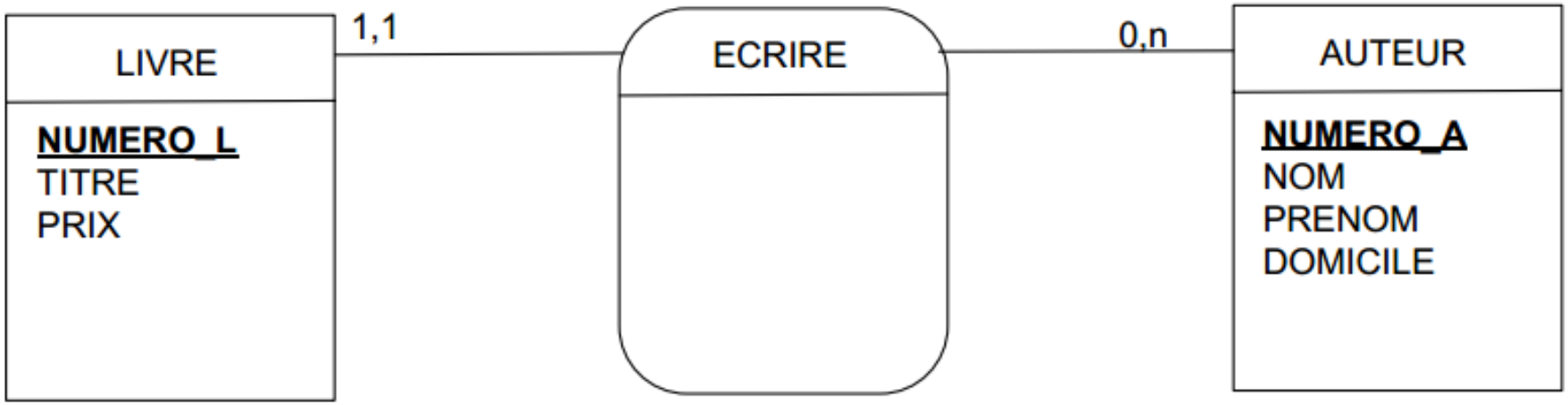
- Un livre est créé par **1et 1** auteur.
- Un auteur peut créer de **0 à n** livres.



Le schéma fait apparaitre qu'un livre est écrit par **1 et 1** seul auteur alors qu'un auteur est à l'origine de **0 à n**.

Spécificités des BDRs

Exemple : Gestion d'une bibliothèque (Gestion des Livres)

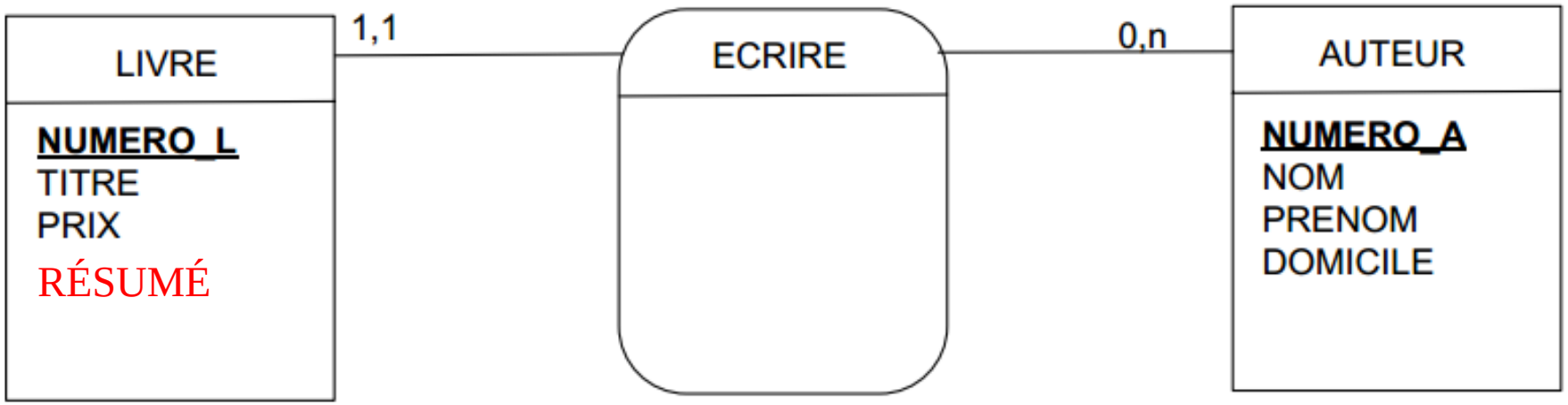


Supposons que le schéma précédent donne lieu à la création d'une BD MySQL et qu'au bout de 3 semaines d'utilisation, la base contienne: 100 000 livres et 105 000 auteurs.

Spécificités des BDRs

Exemple : Gestion d'une bibliothèque (Gestion des Livres)

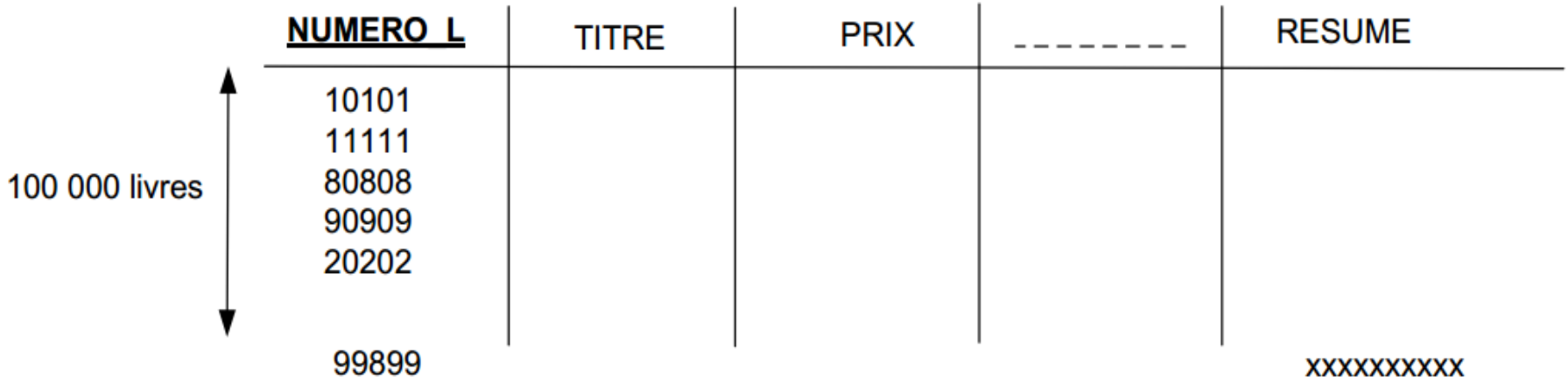
Ajout d'un nouveau champ « **Résumé** » à la table « Livre » après par exemple 3 semaines de mise en œuvre de



Spécificités des BDRs

Exemple : Gestion d'une bibliothèque (Gestion des Livres)

Ajout d'un nouveau champ « **Résumé** » à la table « Livre » après par exemple 3 semaines de mise en œuvre de



On peut constaté facilement que la présence d'une nouvelle information oblige à modifier les 100 000 tuples par l'ajout d'un nouvel attribut (résumé) alors même que le résumé n'existe pas pour ces tuples.

→ **On voit ici les limites du modèle relationnel**

Spécificités des BDRs

Exemple : Gestion d'une bibliothèque (Gestion des Livres)

Une nouvelle condition :

La législation sur les livres avait changé et si pour les livres édités par la suite, 2 taux de TVA s'appliquent:

- 5,50% aux romans
- 10,50% aux livres historiques.

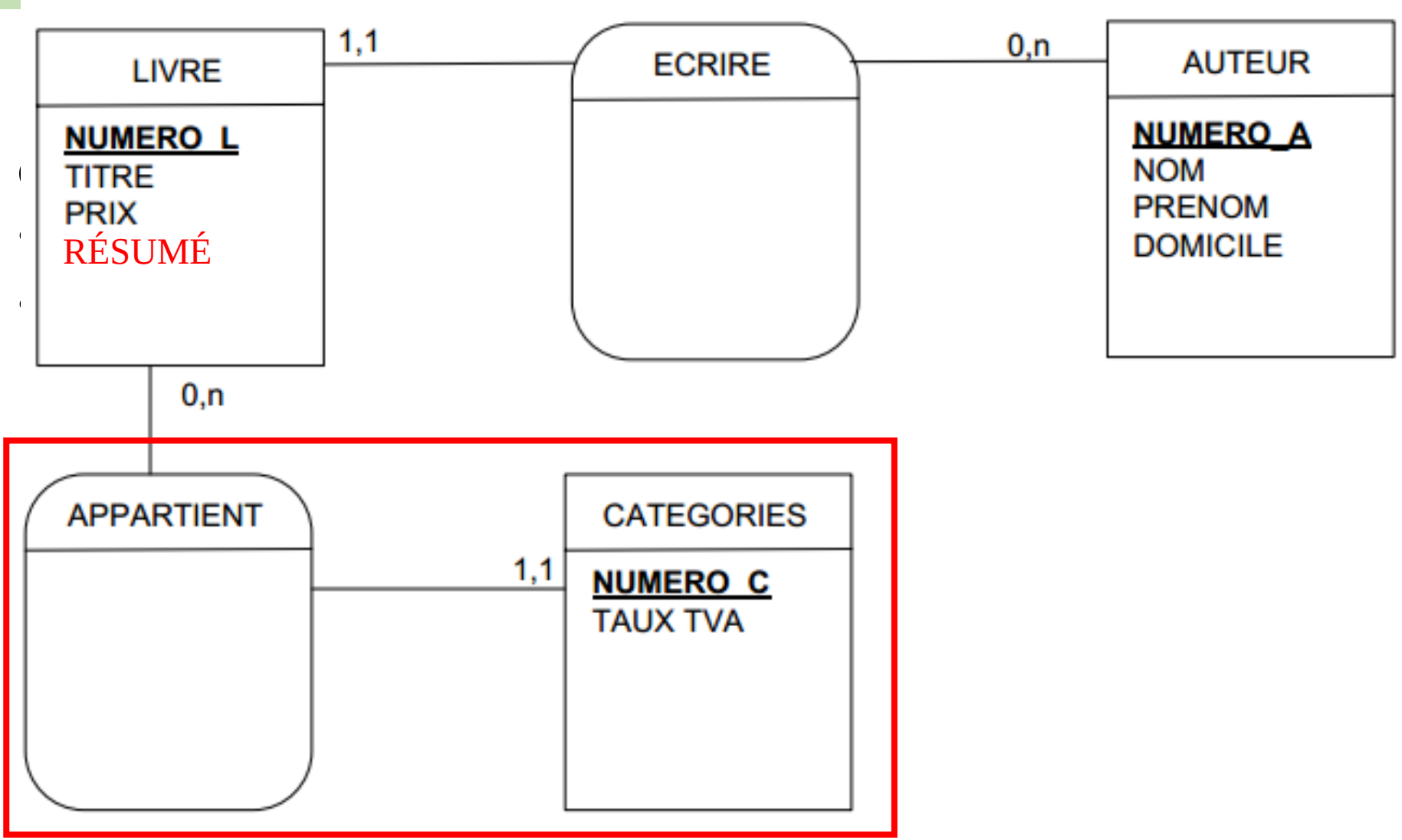
Pour les livres édités avant cette date, le taux unique et usuel = 7,70%.

Dans ce cas et avec cette nouvelle conjoncture:

→ **Il faut modifier profondément le MCD/MLD et introduire une classe CATEGORIE et une relation d'appartenance à une catégorie.**

Spécificités des BDRs

Exemple : Gestion d'une bibliothèque (Gestion des Livres)



Spécificités des BDRs

Exemple : Gestion d'une bibliothèque (Gestion des Livres)

Les BDRs ne sot pas adaptées aux traitements des données peu structurées et/ou ayant une structure changeante.

Avantage des BDs NoSQL

Les BD NoSQL sont parfaitement adaptées aux traitements des données peu structurées et/ou ayant une structure changeante.

Avantage des BDs NoSQL

Le premier acteur économique d'importance ayant fait la promotion des BDS NoSQL (Non relationnelles).

Les BDS NoSQL constituent une nouvelle manière de représenter l'information.
Elles s'affranchissent des contraintes dites ACID et fournissent une architecture technique où il suffit de rajouter des serveurs pour gagner en performance.

Avantage des BDs NoSQL

Attention!

Il ne faut pas opposer les deux approches mais bien souvent les faire **cohabiter** et les BD NoSQL ne visent pas à remplacer les SGBD relationnels mais plutôt à les **compléter**.

Références

- Rudi Bruchez. (2015), « Les bases de données NoSQL et le Big Data », Editeur : Eyrolles, ISBN : 978-2-212-14155-9
- Rudi Bruchez (2021), « Les bases de données NoSQL », Editeur : Eyrolles ISBN : 978-2-212-67866-6
- Juvénal Chokogoue. (2017). « Hadoop - Devenez opérationnel dans le monde du Big Data », Editeur: ENI, ISBN: 978-2409007613