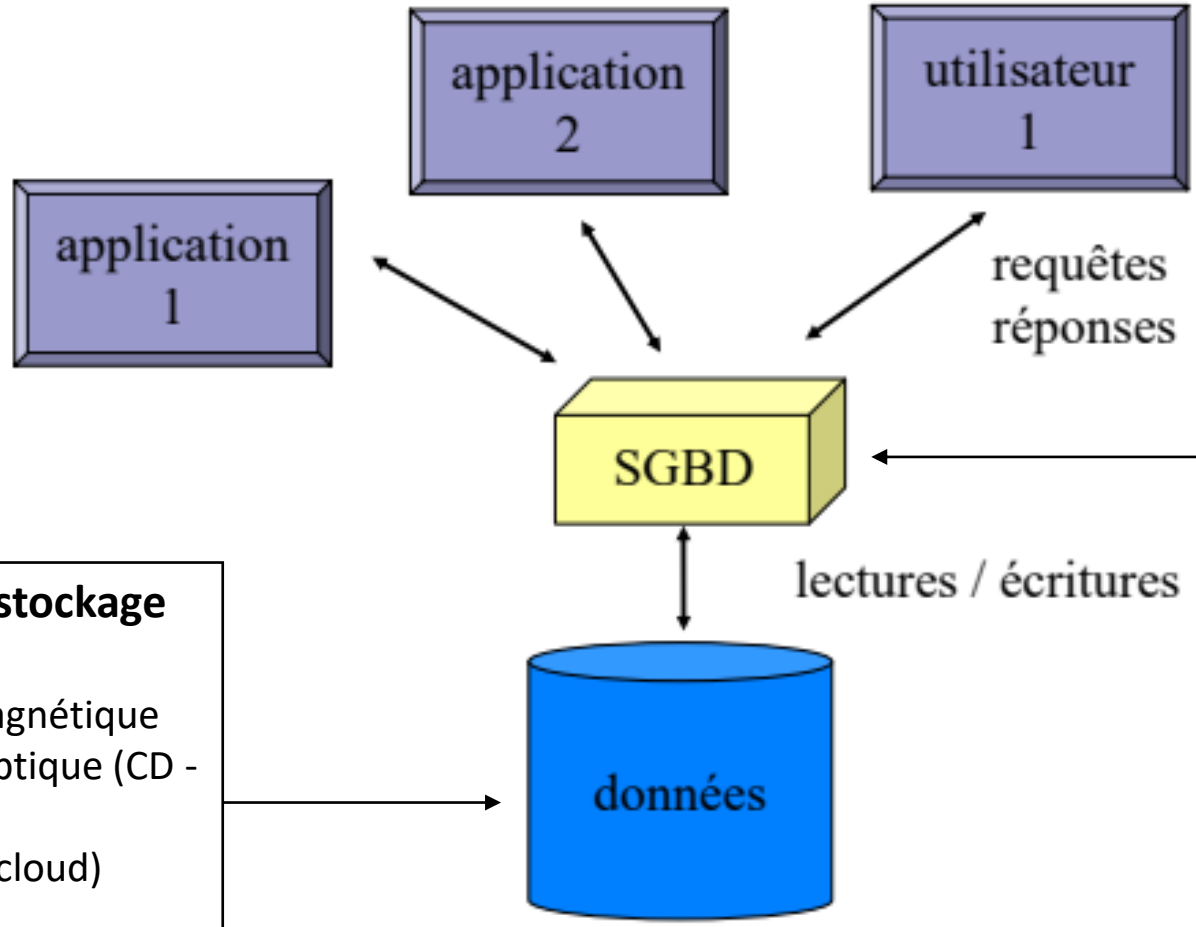
Excelerate SYSTEMS

Big data :
Concepts clés et fonctionnement

Plan de cours:

- Séance 2 :
  - Big Data
    - Concepts clés et fonctionnement.
  - L’analyse de données:
    - Vue d’ensemble.
    - Logiciels et outils.
  - Apprentissage supervisé, non supervisé :
  - Problèmes de classification et de régression.
  - Réseaux de neurones et IA :
    - Vision par ordinateur.
    - NLP.

Rappels:



Système de gestion de base de données:

- **Schémas**
- Structuration de données.
- Consultation, modification de données.
- Gestion d'accès et de rôles.
- Etc.

Support de stockage (centralisé):

- Disque magnétique
- support optique (CD - DVD),
- internet (cloud)
- Etc.

Rappels:

- **Les systèmes de stockage classiques (relationnels)** :

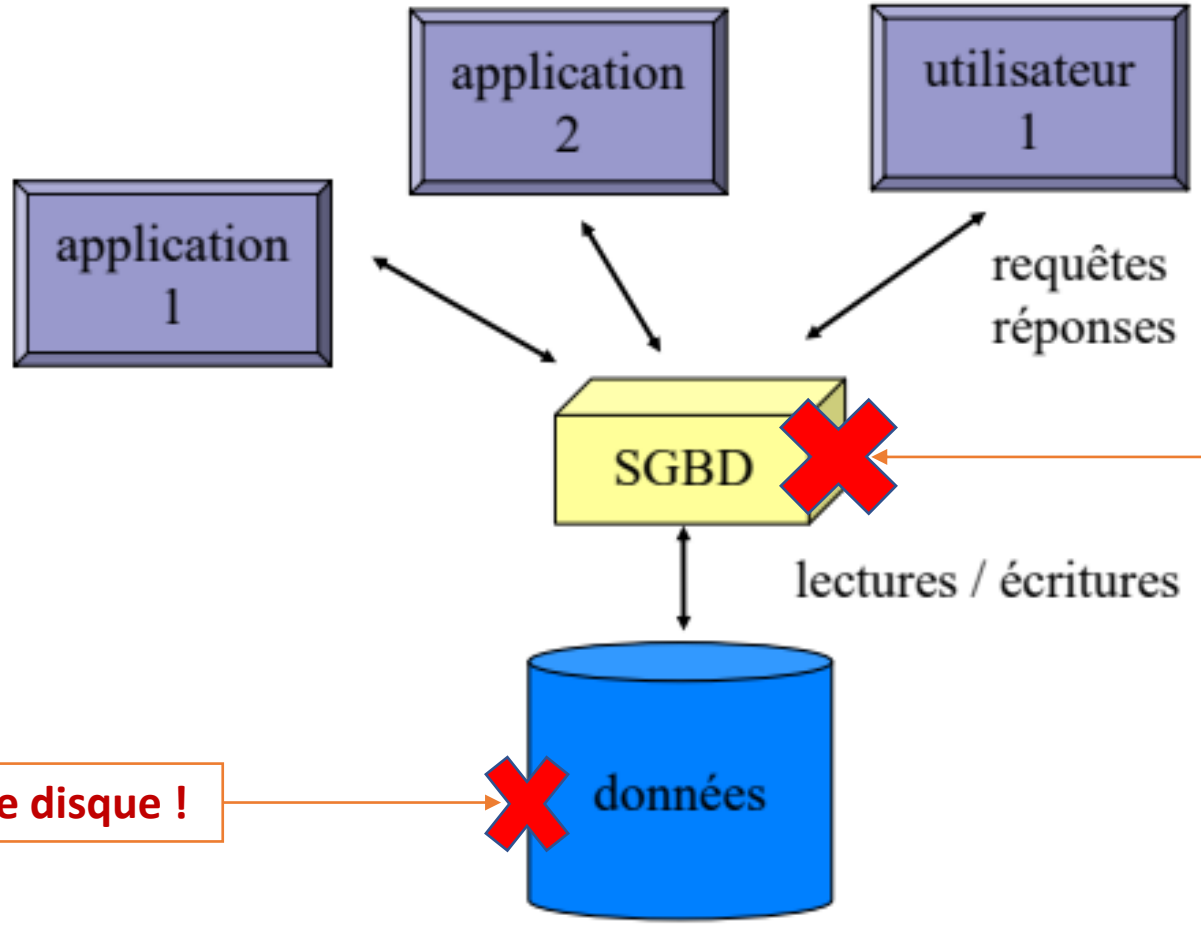
  - ‘Single Node’ = serveurs (architecture horizontale).
  - Centralisés.
  - Langages et outils d’interrogation standardisées et bien définis.
  - Lecture et écriture de données sont gérés par les systèmes de gestion de bases de données (SGBD).
  - **Limités en termes de scalabilité. (1)**
  - **Limités en termes de performances.**
  - **Single Point Of Failure(2).**
  - **Des données structurés uniquement(3).**

1- Scalabilité :

Scale-up (Architecture horizontale):

+ Mémoire
+ puissance de calculs.
+ stockage.

2-Single Point Of Failure: arrêt complet du système!



3- Données structurées uniquement !

Unstructured data

The university has 5600 students.
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

Semi-structured data

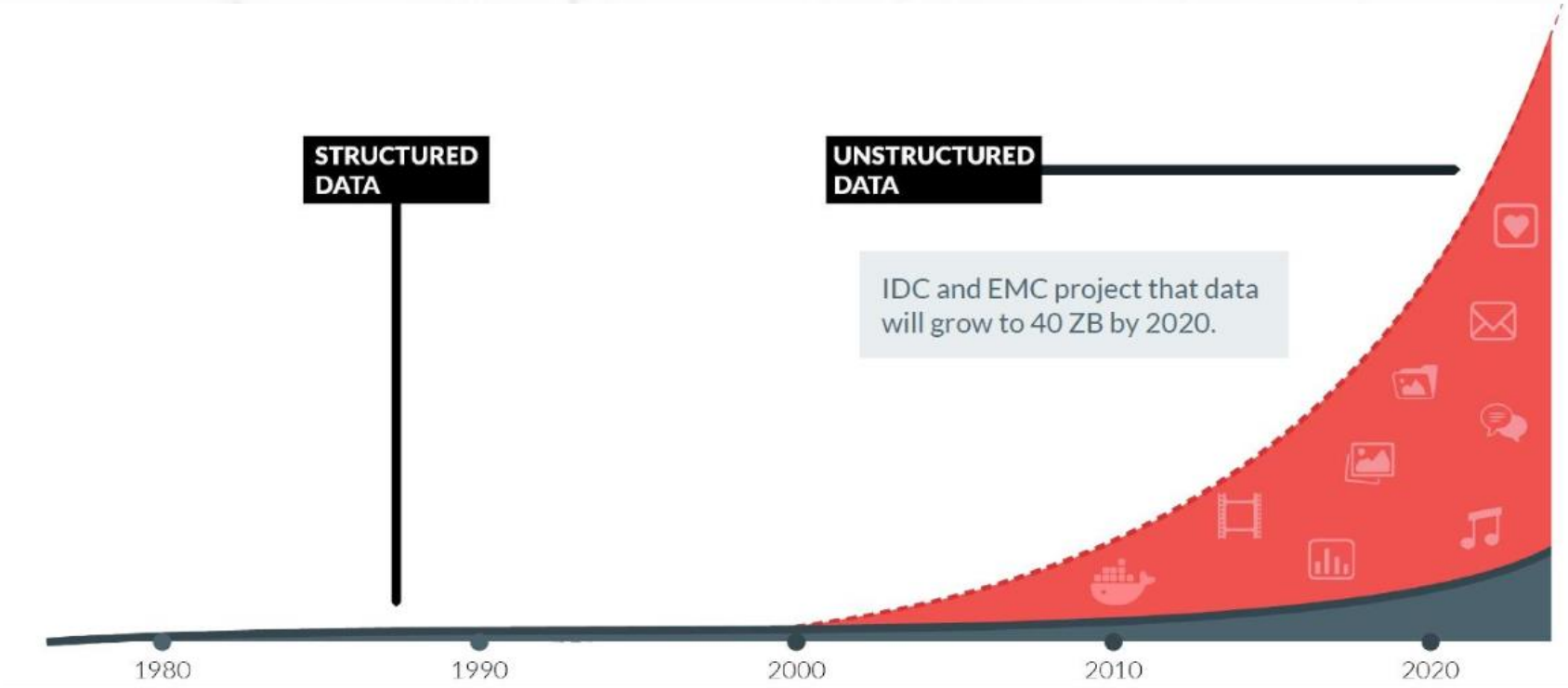
```
<University>
  <Student ID="1">
    <Name>John</Name>
    <Age>18</Age>
    <Degree>B.Sc.</Degree>
  </Student>
  <Student ID="2">
    <Name>David</Name>
    <Age>31</Age>
    <Degree>Ph.D. </Degree>
 </Student>
....
</University>
```

Structured data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

On ne peut pas stocker ses types directement dans une base de de données SQL !

3- Données structurées uniquement !



On produit plus de données non structurées
(fichiers audio, images, vidéos,text,sms,etc)

Alors ?

- Les technologies existantes ne sont pas conçues pour répondre à ces problématiques !
- De “nouvelles” technologies de stockage et gestion de données sont nécessaires :
  - Google file system – google 2003
  - MapReduce – google 2004
  - Hadoop – Yahoo 2006
  - …

Big Data: Définition

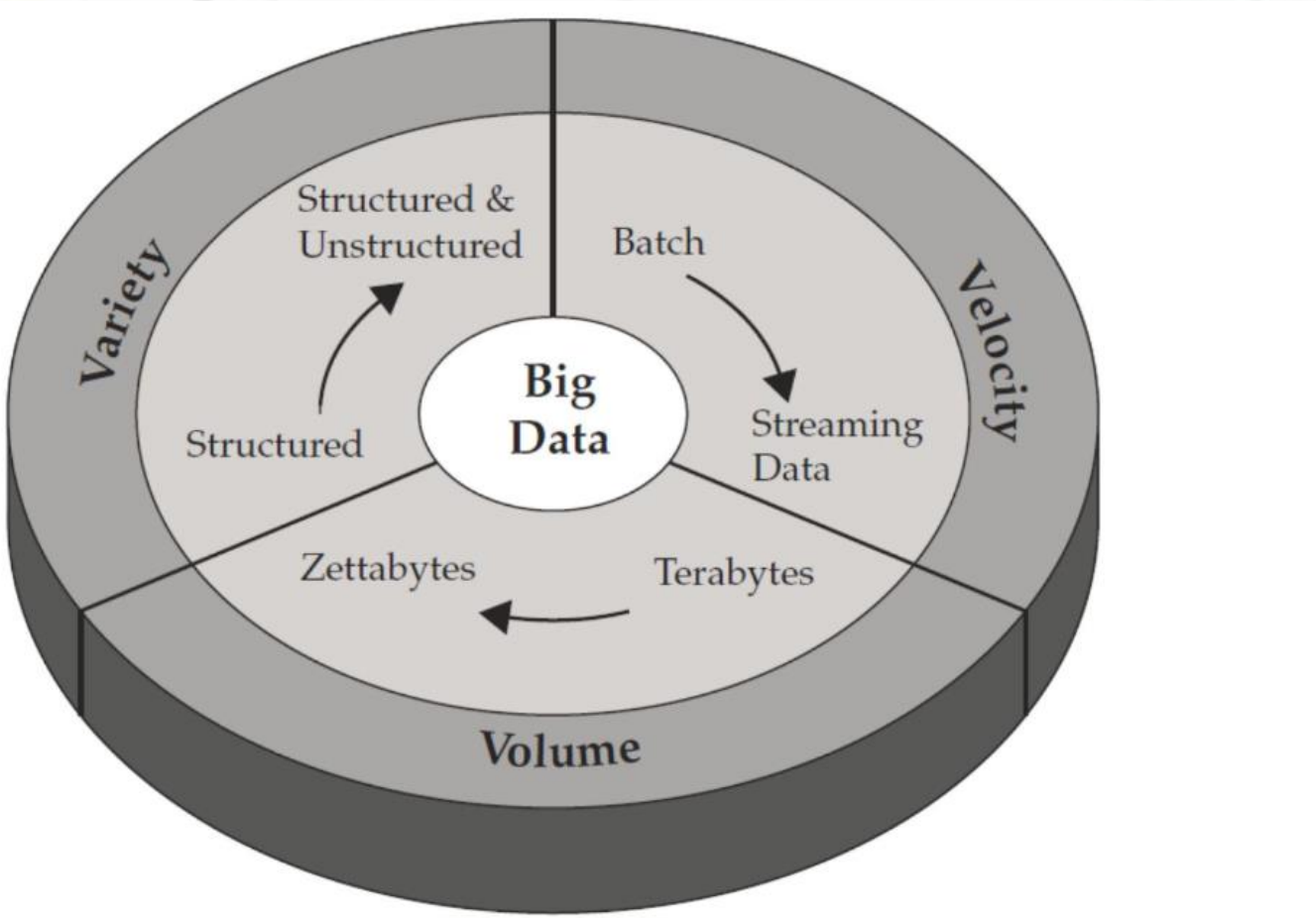
« Le Big Data (ou mégadonnées) représente les collections de données caractérisées par un ***volume****, une* ***vélocité*** *et une* ***variété*** *si grands que leur transformation en* ***valeur*** *utilisable requiert l'utilisation de technologies et de méthodes analytiques spécifiques. »* ***3V***

Big Data: Définition

Rappels :

- On a plus des données bien définies et connus. On doit gérer des différentes types et structures de données (**Variété**).
- La rapidité avec laquelle les données sont générées et traitées(**Vélocité/Vitesse**).
- On doit être capable de traiter des volumes immenses de données (**Volume**).

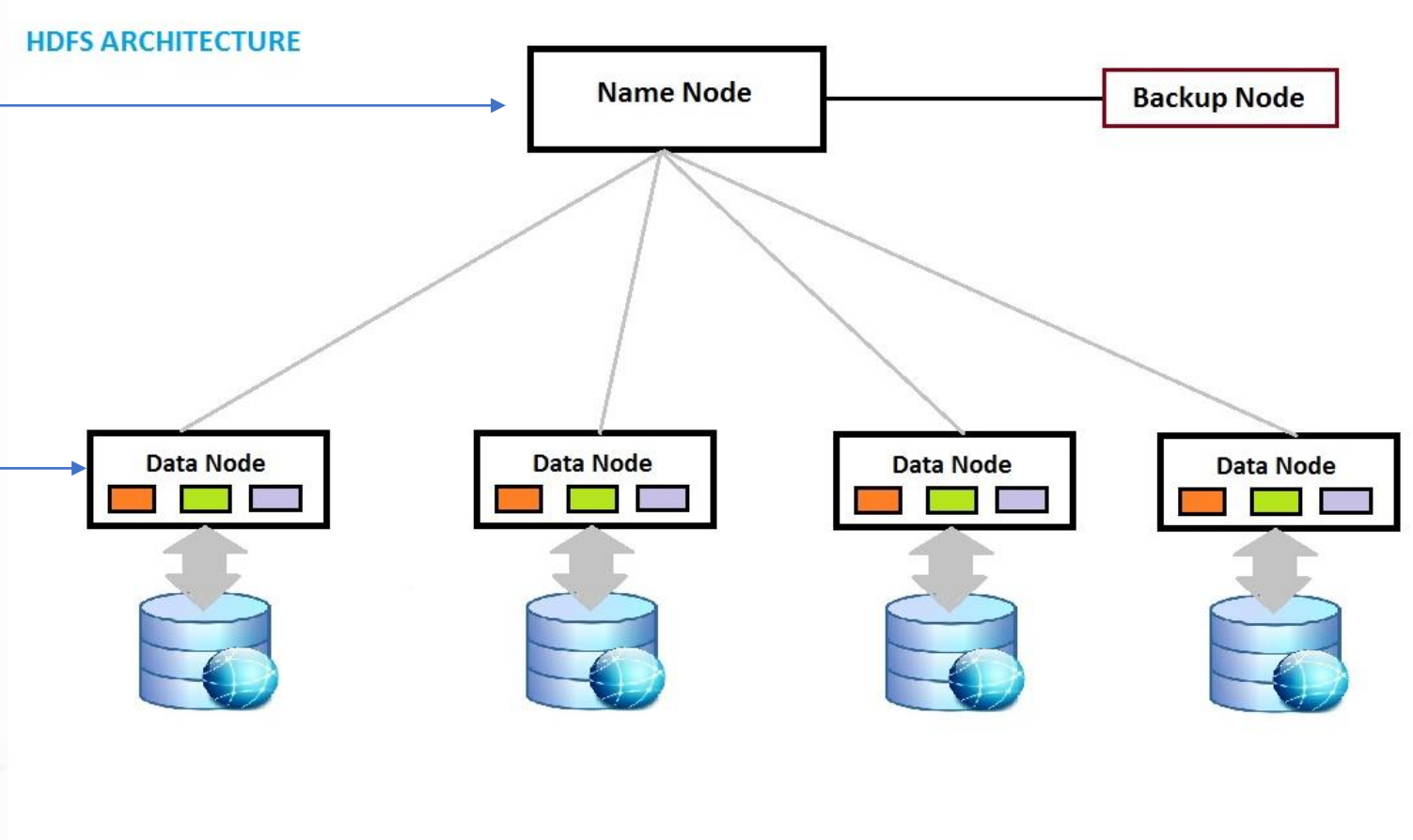
Big Data: Définition



Big Data: Architecture et stockage de données (HDFS)

Master Node : service de métadonnées
Il sait ce que chaque nœud de données contient.

Data Node : service de bloc de données
Is contient les bloc de données

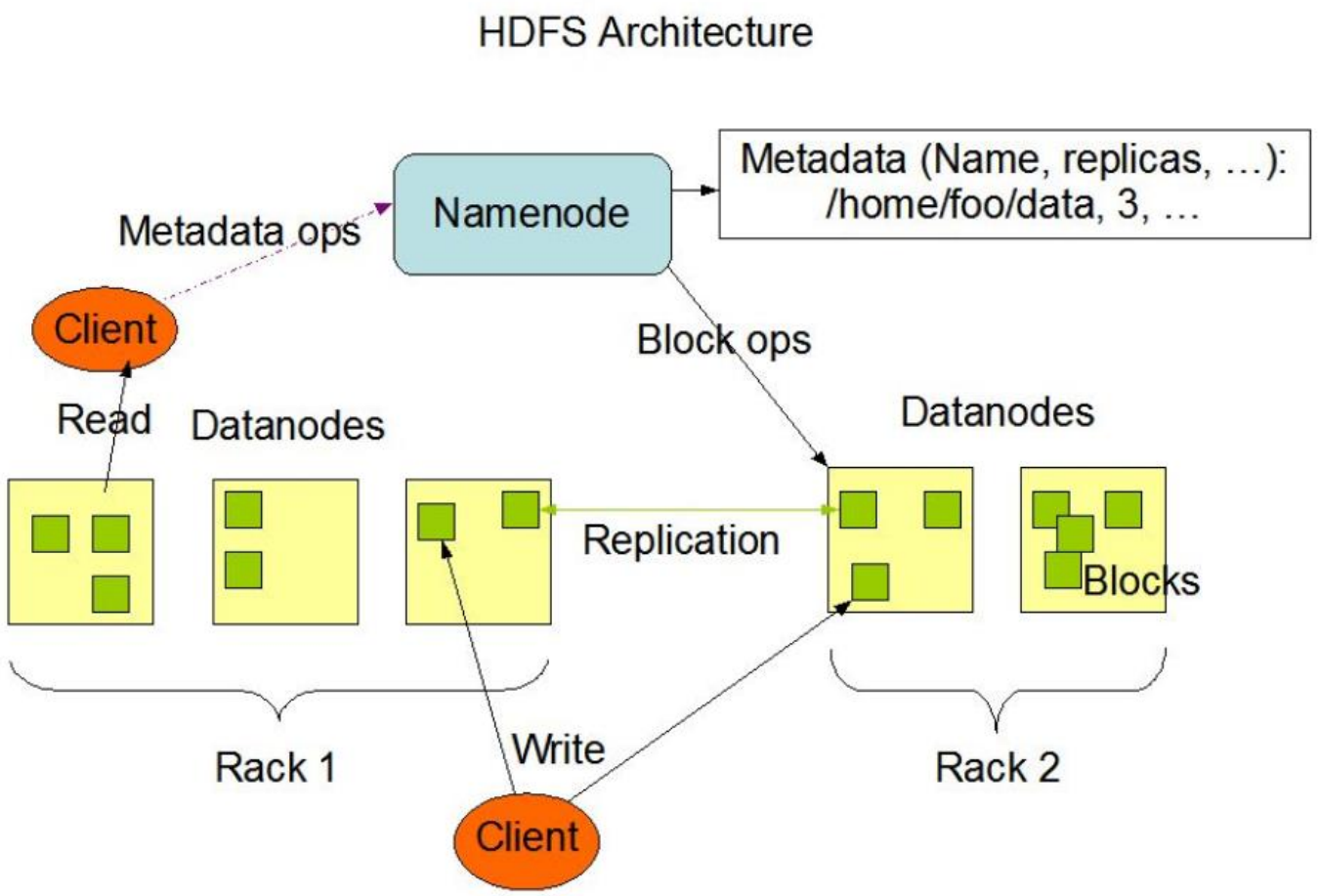


HDFS : Hadoop Distributed File System

Big Data: Architecture et stockage de données

- Chaque fichiers enregistré dans HDFS sera divisé en bloc (fichiers de tailles fixe et généralement de 64 MB chacun).
- Chaque bloc sera répliqué (généralement 3x) et distribués sur plusieurs Data Nodes.
- Le Master/Name Node connait ce que chaque Data Node contient en termes de bloc de données (réplicas ou primaire)
- Le Master Node enregistre tous ses métadonnées dans des journaux avec des copies dans le Secondary Node (Backup).

Big Data: Architecture et stockage de données

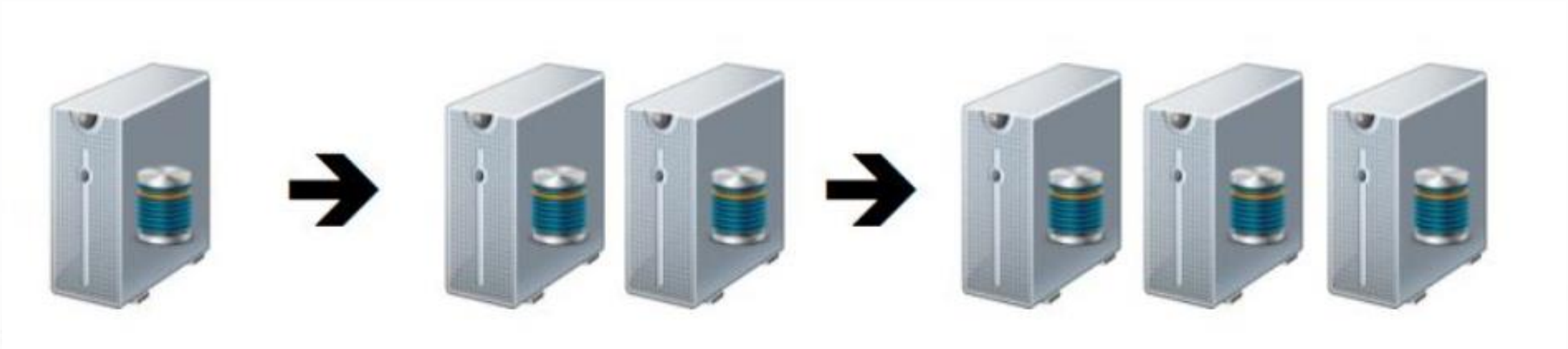


Big Data: Architecture et stockage de données (**HDFS**)

- Les fichiers sont divisés en blocs: la perte d'un seul de ces blocs causerait une corruption du fichier, c'est pour quoi **ils sont tous répliqués et distribués sur plusieurs data nodes (No single point of failure).**
- la perte d'un nœud de données/ bloc de données n'entraînera aucune interruption de service .
- Relativement lent.

Big Data: Architecture et stockage de données (**HDFS**)

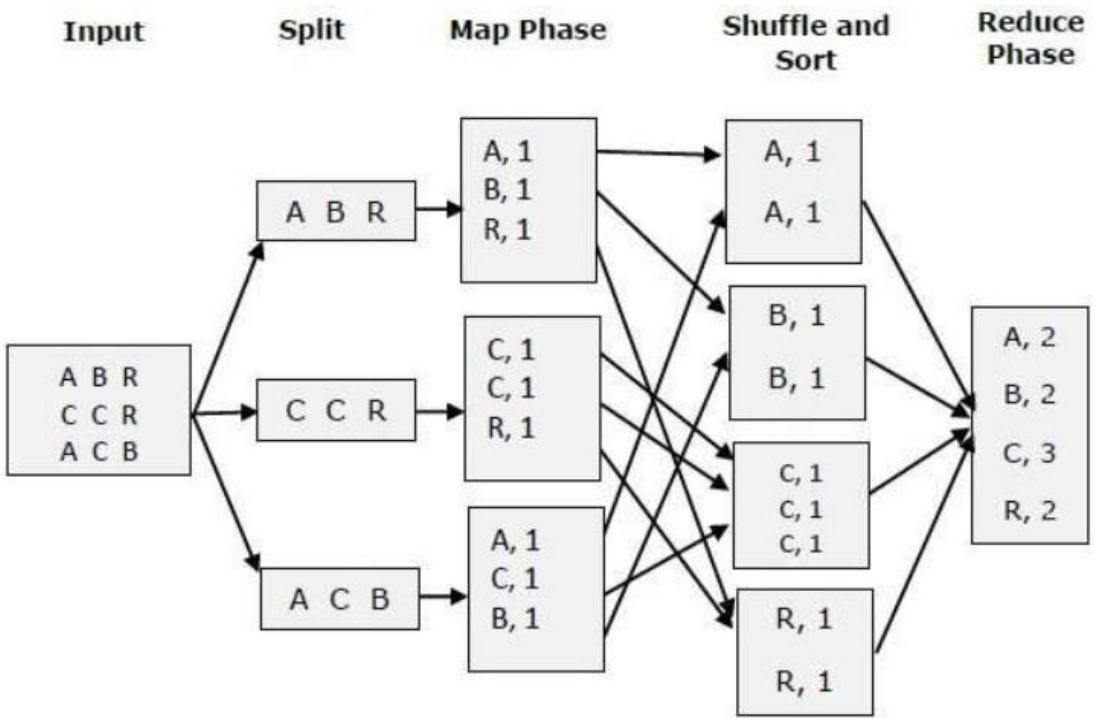
Scale-out (Architecture verticale) :
+ Machines.



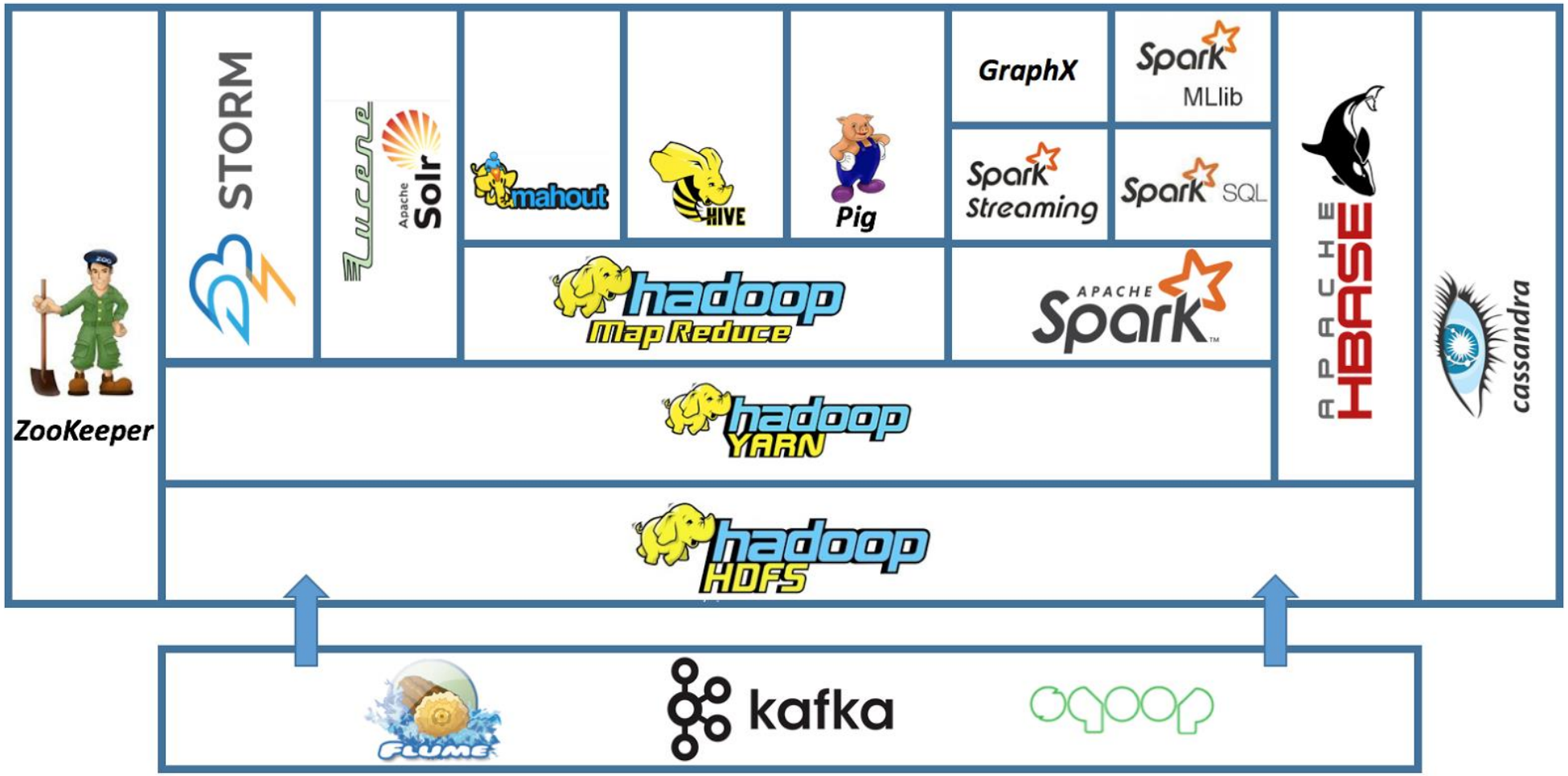
Big Data: Traitement et Analytics (**Map-Reduce**)

- Hadoop Map-Reduce :
  - **Map** : transformation des données en paires de clé-valeur.
  - **Reduce** : opération d'agrégation (somme, moyenne, etc.) par clé.
- Algorithme résilient et s'exécute en parallèle sur les différents data nodes.

Big Data: Traitement et Analytics (**Map-Reduce**)



Big Data:Hadoop ecosystem



Excelerate SYSTEMS

Big Data: uses cases and business opportunites

Recommendation Engine	Sentiment Analysis	Risk Modeling	Fraud Detection	Marketing Campaign Analysis
Customer Churn Analysis	Social Graph Analysis	Customer Experience Analytics	Network Monitoring	Research And Development
Data Analytics				

Travail à faire:

- 7 binômes, chaque binôme traite l'un des uses case suivants.
- **Un rapport de 2 à 3 pages à rendre** en présentant le cas d'usage et l'utilisation des technologie de big data sans des explications techniques ou codes.

Recommendation Engine	Sentiment Analysis	Risk Modeling	Marketing Campaign Analysis
Customer Churn Analysis	Social Graph Analysis	Customer Experience Analytics	Social media analysis