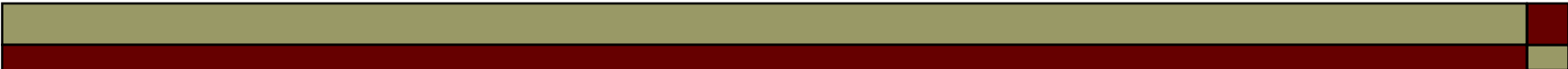




---

# Introduction au BIG DATA

*BAKAYOKO Ibrahima*  
*Enseignant-Chercheur UFR MI*  
*[bakayokosoft@gmail.com](mailto:bakayokosoft@gmail.com)*  
*« La logique sauve de l'ennui. »*



---

*« Sans connaissance de son audience, et sans hypothèses préalables solides, la Big data ne sert à rien. »*

*Benjamin Mercier, Responsable Digital Analytics*

# Définition (1)

---

- La notion de Big Data est un concept s'étant popularisé en 2012 pour traduire le fait que les entreprises sont confrontées à des volumes de données à traiter de plus en plus considérables et présentant un fort enjeux commercial et marketing.
- Ces Grosses Données en deviennent difficiles à travailler avec des outils classiques de gestion de base de données.

## Définition (2)

---

- Il s'agit donc d'un ensemble de technologies, d'architecture, d'outils et de procédures permettant à une organisation de très rapidement capter, traiter et analyser de larges quantités et contenus hétérogènes et changeants, et d'en extraire les informations pertinentes à un coût accessible.

# Les 5 principes du big data (1)

---

Les 5 V du big data font référence aux cinq principes qui servent de base à cette technique de compilation, de stockage et de gestion de données. Ces principes commencent tous par la lettre V.

# Les 5 principes du big data (2)

---

**Variété (Variety)** : Gérer la complexité de plusieurs types de données et de schémas structurés ou non structurés (texte, données de capteurs, son, vidéo, logs, ...). Cette procédure est indispensable pour éviter des dispersions ou des faux positifs.

# Les 5 principes du big data (3)

---

**Volume (Volume)** : Volumes de données croissants de tous types, qui se comptent en téraoctets, pétaoctets ou en exaoctets, voir plus.

Les techniques de big data se caractérisent par le traitement de **quantités considérables** de données, et c'est pourquoi la définition du volume est si importante. Concrètement, le big data permet de rassembler et d'analyser de grands lots de données, pouvant comprendre des millions de registres, et qui pourront être utilisés pour le bien de l'entreprise.

# Les 5 principes du big data (4)

---

1 MégaOctet =  $10^6$  Octets

1 GigaOctet =  $10^9$  Octets

1 TéraOctet =  $10^{12}$  Octets

1 PétaOctet =  $10^{15}$  Octets

1 ExaOctet =  $10^{18}$  Octets

1 ZettaOctet =  $10^{21}$  Octets





# Les 5 principes du big data (5)

---

**Véracité (Veracity)** : L'objectif principal du big data est que les prises de décision de l'entreprise soient basées sur des données réelles, ce qui explique que la véracité soit un des piliers fondamentaux du big data. Aussi, une des tâches les plus ardues pour atteindre cet objectif est celle d'écarter les données qui ne sont pas valables, soit de par leur origine, leur forme, ou bien parce que la manière de les compiler ne correspond pas aux normes.



# Les 5 principes du big data (6)

---

**Vitesse (Velocity)** : tel le volume, la vitesse du flux de données obtenues en ligne est très élevée. Ces changements constants nuisent à la véracité des données, aussi est-il nécessaire qu'elles soient mises à jour pratiquement en temps réel. Les algorithmes de traitement doivent tenir compte de cette instantanéité afin de la traduire en analyse de journaux et d'en tirer les bonnes conclusions.

# Les 5 principes du big data (7)

---

**Valeur (Value)** : Une autre manière de le dire est qu'il s'agit de discerner la véritable utilité des bonnes données pour les adapter à l'entreprise. Cela signifie qu'il faut choisir les registres les plus appropriés pour le traitement, être sélectif et prendre en compte leurs relations véritables.

**L'amélioration des produits, la personnalisation des services ou l'offre de meilleurs prix** ne sont que trois exemples de la valeur que la collecte de données peut apporter à l'entreprise et à son public potentiel.

# Statistiques et chiffres clés (2022) (1)

---

- (1) 1,7 Mo : le nombre de données générées chaque seconde dans le monde.

Source (1) : édition du baromètre « *Data Never Sleeps* » édité par l'entreprise américaine Domo (spécialisée dans le cloud et ses outils)

- (2) Il y a 2,80 milliards d'utilisateurs mensuels de Facebook lorsqu'il s'agit de comptes actifs. Chaque jour, WhatsApp, Messenger, Instagram et Facebook reçoivent la visite de pas moins de 1,8 milliards d'utilisateurs.

# Statistiques et chiffres clés (2022) (2)

---

- (3) Chaque minute, plus de 500 000 tweets sont publiés sur Twitter
- (4) Chaque jour, Google reçoit plus de 3,5 milliards de recherches. Parmi celles-ci, pas moins de 15 % des nouvelles requêtes n'ont jamais été effectuées auparavant.
- (5) Jusqu'à 65 milliards de messages sont envoyés via WhatsApp chaque jour.

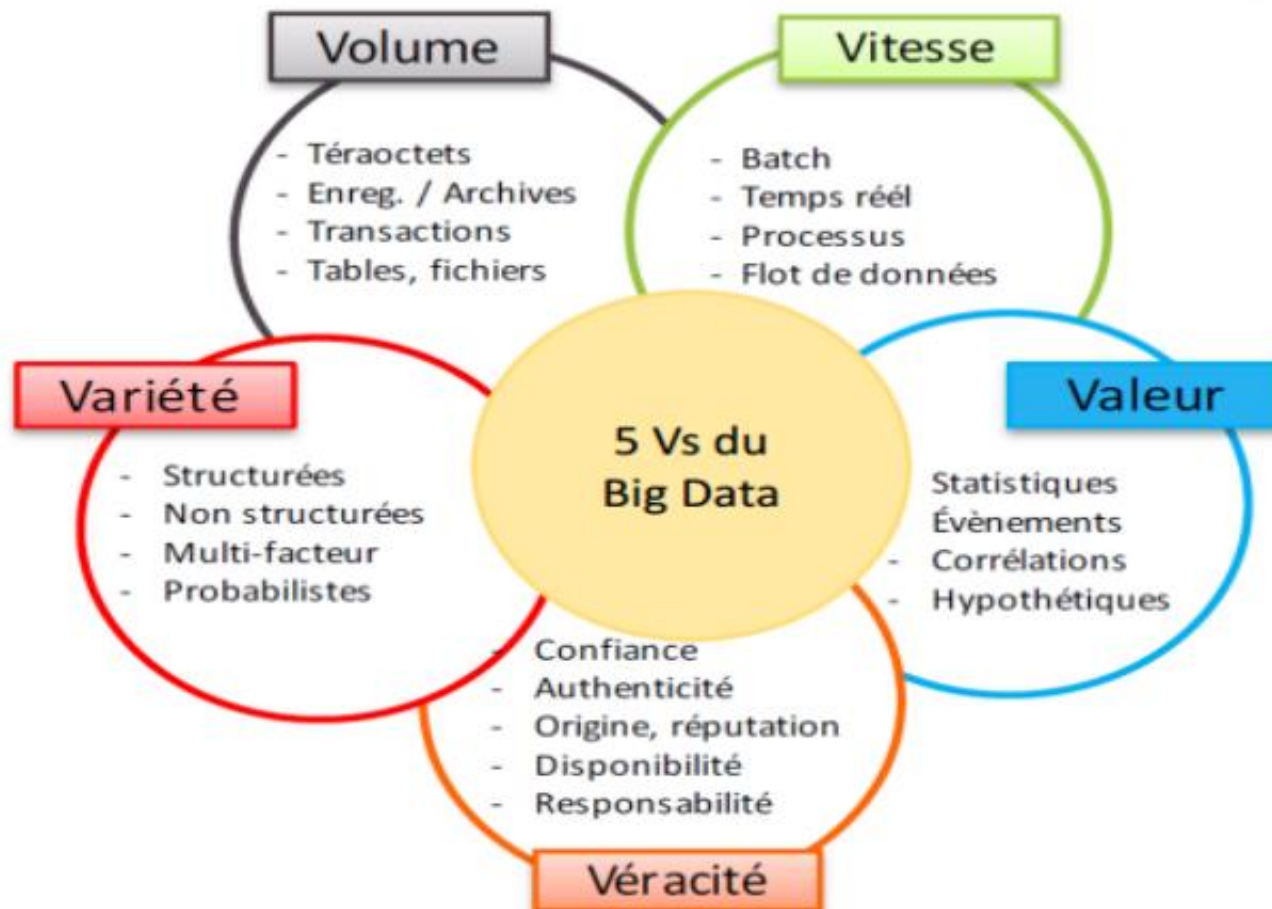
# Statistiques et chiffres clés (2022) (3)

---

- ❑ (6) Netflix économise 1 milliard de dollars chaque année en utilisant le Big Data.
- ❑ (7) Le marché de l'analyse du Big Data dans le secteur bancaire devrait atteindre 62,10 milliards de dollars d'ici 2025.
- ❑ (8) La quantité de données créées devrait atteindre plus de 180 zettaoctets d'ici 2025.

Source (2) (3) (4) (5) (6) (7) et (8) : <https://www.sales-hacking.com/post/statistiques-big-data>

# Synthèse de « 5 V »



# Architecture traditionnelle

---

## Applications traditionnelles

- Architecture n-tiers

le cas le plus fréquent est l'architectures 3-tiers,

- Un (01) Frontend;
  - Un (01) Backend;
  - Et la partie qui permet de gérer la base de données.
- Bases de données centralisées
- Transport des données vers les serveurs d'application



# Architecture BIG DATA

---

## Applications BIG DATA

- Pipelines complexes (*nous y reviendront lors d'un cas concret*)
- Bases de données distribuées
- Transport du code des applications vers les serveurs où les données sont stockées



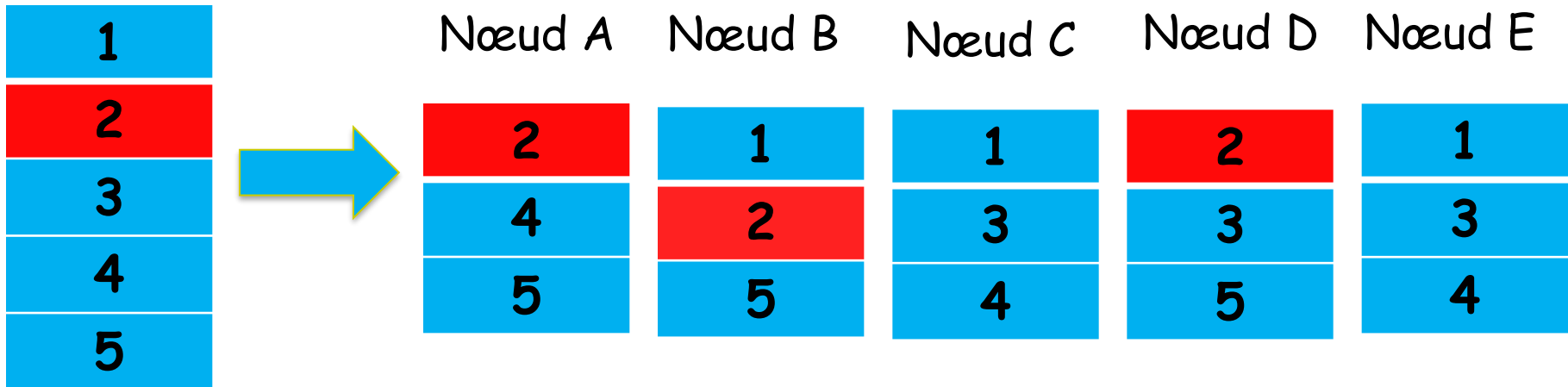
# Gestion des situations complexes? (1)

La plupart des outils et des Frameworks de Big Data sont construits en gardant à l'esprit plusieurs caractéristiques.

## Gestion des situations complexes? (2)

- **La distribution des données** : Le grand ensemble de données est divisé en morceaux ou en petits blocs et réparti sur un nombre  $N$  de nœuds ou de machines. Ainsi les données sont réparties sur plusieurs nœuds et sont prêtes au traitement parallèle. Dans le monde du Big Data, ce type de distribution des données est réalisé à l'aide d'un Système de Fichiers Distribués - DFS (Distributed File System).

# Distribution des données



Fichier d'entrée

Concrètement, cela revient à diviser la charge de stockage et de traitement d'une machine sur plusieurs machines afin de gagner en rapidité, en réactivité et en performance.

## Gestion des situations complexes? (3)

- **Le traitement en parallèle** : Les données distribuées obtiennent la puissance de N nombre de serveurs et de machines dont les données résident. Ces serveurs travaillent en parallèle pour le traitement et l'analyse. Après le traitement, les données sont fusionnées pour le résultat final recherché. (Actuellement ce processus est réalisé par **MapReduce de Google**)

## Gestion des situations complexes? (4)

- **La tolérance aux pannes** : En général, nous gardons la réplique d'un seul bloc (ou chunk) de données plus qu'une fois. Par conséquent, même si l'un des serveurs ou des machines est complètement en panne, nous pouvons obtenir nos données à partir d'une autre machine ou d'un autre « data center ». Encore une fois, nous pouvons penser que la réplication de données pourrait coûter beaucoup d'espace.

## Gestion des situations complexes? (5)

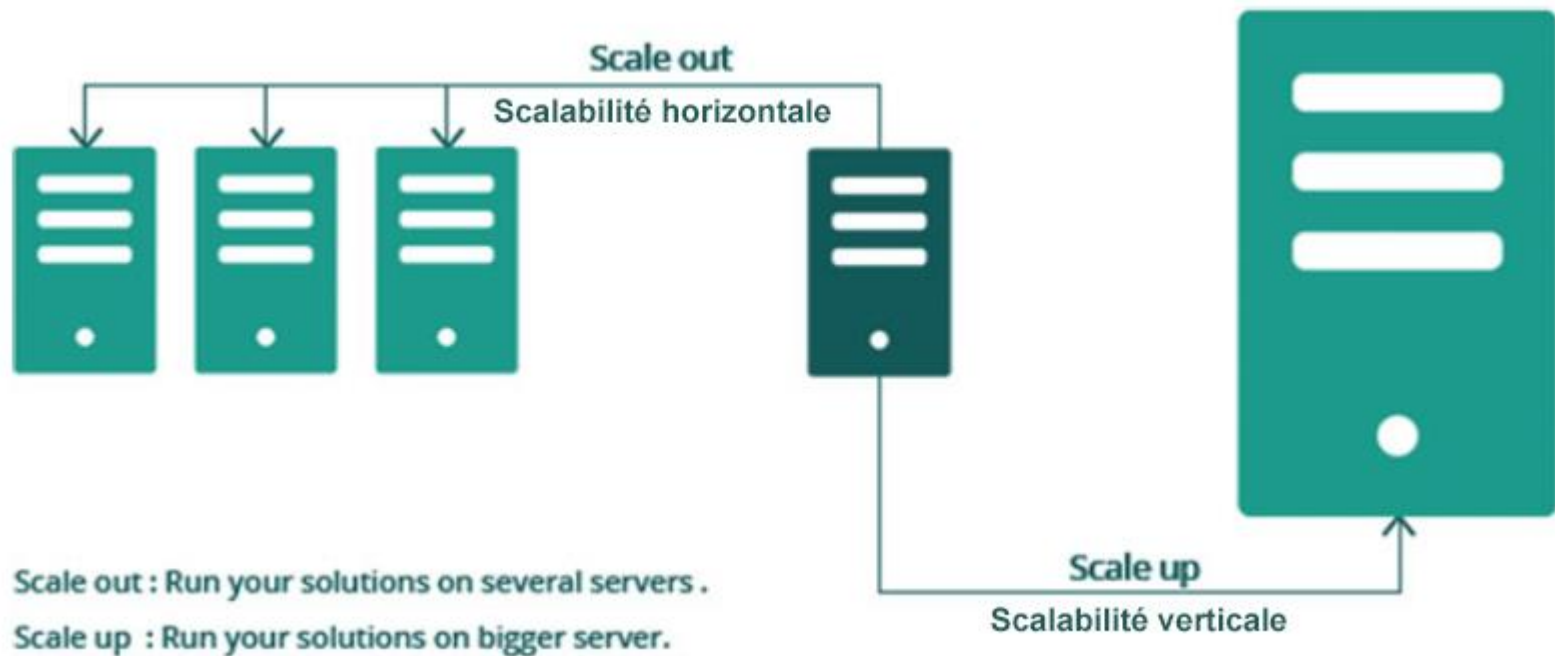
- **L'utilisation de matériel standard** : La plupart des outils et des frameworks Big Data ont besoin du matériel standard pour leur travail. Donc nous n'avons pas besoin de matériel spécialisé avec un conteneur spécial des données « RAID ». Cela réduit le coût de l'infrastructure totale.

## Gestion des situations complexes? (6)

- **Flexibilité, évolutivité et scalabilité** : Il est assez facile d'ajouter de plus en plus de nœuds dans le cluster quand la demande pour l'espace augmente. De plus, la façon dont les architectures de ces Frameworks sont faites, convient très bien le scénario de Big Data.



# Gestion des situations complexes? (7)





## Les applications concrètes du big data (1)

Dans le domaine de la santé, par exemple, le Big Data favorise une médecine préventive et personnalisée. Ainsi, l'analyse des recherches des internautes sur un moteur de recherche a déjà permis de détecter plus rapidement l'arrivée d'une épidémie de grippe. Dans un futur proche, les appareils connectés devraient permettre l'analyse en continu des données biométriques des patients.



## Les applications concrètes du big data (2)

Dans le **domaine des transports**, l'analyse des données du Big Data (données provenant des pass de transport en commun, géolocalisation des personnes et des voitures, etc.) permet de modéliser les déplacements des populations afin d'adapter les infrastructures et les services (horaires et fréquence des trains, par exemple).



## Les applications concrètes du big data (3)

Dans le domaine de la gestion énergétique, l'analyse des données issues du Big Data intervient dans la gestion de réseaux énergétiques complexes via les réseaux électriques intelligents (smartgrids) qui utilisent des technologies informatiques pour optimiser la production, la distribution et la consommation de l'électricité.



## Les applications concrètes du big data (4)

De la même manière, l'analyse des données provenant de capteurs sur les avions (données de vol) associées à des données météo permet de modifier les couloirs aériens afin de réaliser des économies de carburant et d'améliorer la conception et la maintenance des avions.

Il existe des utilisations concrètes du Big Data dans de nombreux autres domaines : recherche scientifique, marketing, développement durable, commerce, éducation, loisirs, sécurité, etc.



# BIG DATA : Acteurs et Solutions

---

Les grands acteurs du web tel que Google, Yahoo, Facebook, Twitter, LinkedIn ... ont été les premiers à être confrontés à des volumétries de données extrêmement importantes et ont été à l'origine des premières innovations en la matière portées principalement sur deux types de technologies :

- ❑ Les plateformes de développement et de traitement des données (GFS, Hadoop, Spark, ...)
- ❑ Les bases de données (NoSql)

# Historique : Big Data, Google : Le système de fichier GFS (1)

---

Pour stocker son index grandissant, quelle solution pour Google ?

1. Utilisation d'un SGBDR ?
  - Problème de distribution des données
  - Problème du nombre d'utilisateurs
  - Problème de Vitesse du moteur de recherche
2. Invention d'un nouveau système propriétaire : GFS ( Google File Système) en 2003

# Historique : Big Data, Google : Le système de fichier GFS (2)

---

- La notion de Big Data est intimement lié à la capacité de traitements de gros volumes.
- MapReduce un algorithme inventé par Google afin de distribuer des traitements sur un ensemble de machines avec le système GFS.
- Google possède aujourd'hui plus de 1 000 000 de serveurs interconnectés dans le monde.



# BIG DATA: Plateforme - Technologies - Outils

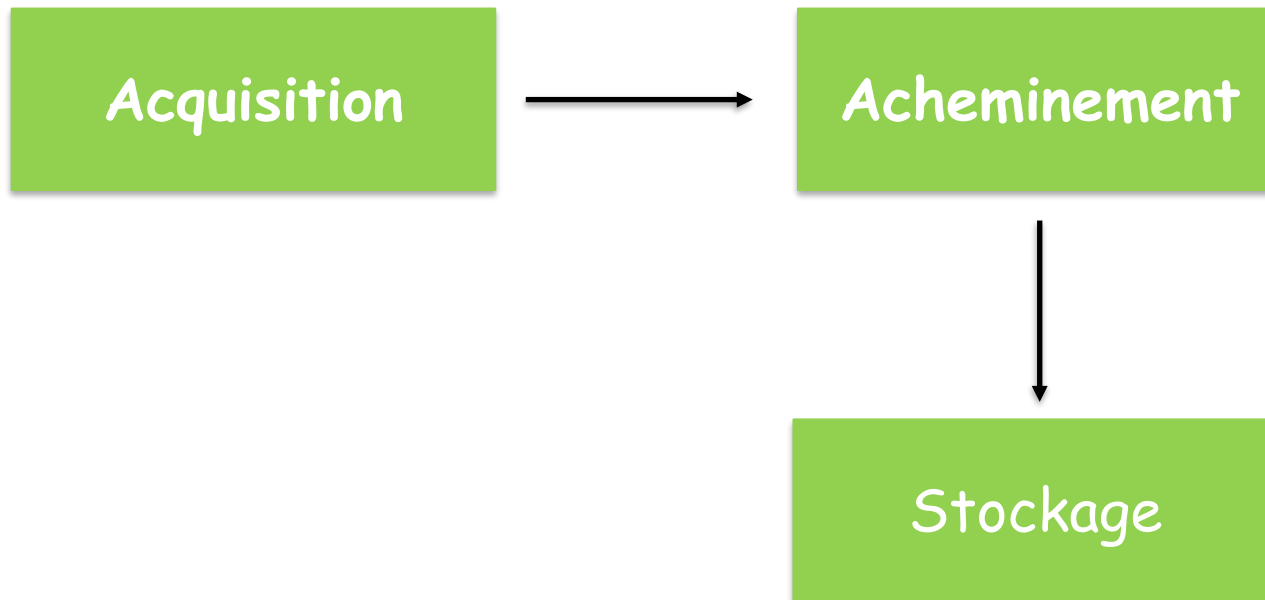
Société	Technologie développée	Type de technologie
Google	Big Table	Système de base de données distribuée propriétaire reposant sur GFS (Google File System). Technologie non Open Source, mais qui a inspiré Hbase qui est Open Source.
	MapReduce	Plate-forme de développement pour traitements distribués
Yahoo	Hadoop	Plateforme Java destinée aux applications distribuées et à la gestion intensive des données. Issue à l'origine de GFS et MapReduce.
	S4	Plateforme de développement dédiée aux applications de traitement continu de flux de données.
Facebook	Cassandra	Base de données de type NoSQL et distribuée.
	Hive	Logiciel d'analyse de données utilisant Hadoop.
Twitter	Storm	Plateforme de traitement de données massives.
	FockDB	Base de données distribuée de type graphe.
LinkedIn	SenseiDB	Base de données temps réel distribuée et semi-structurée.
	Voldemort	Base de données distribuée destinée aux très grosses volumétries.

Tableau : Quelques technologies Open Source du Big Data

# Pipeline big data (1)

---

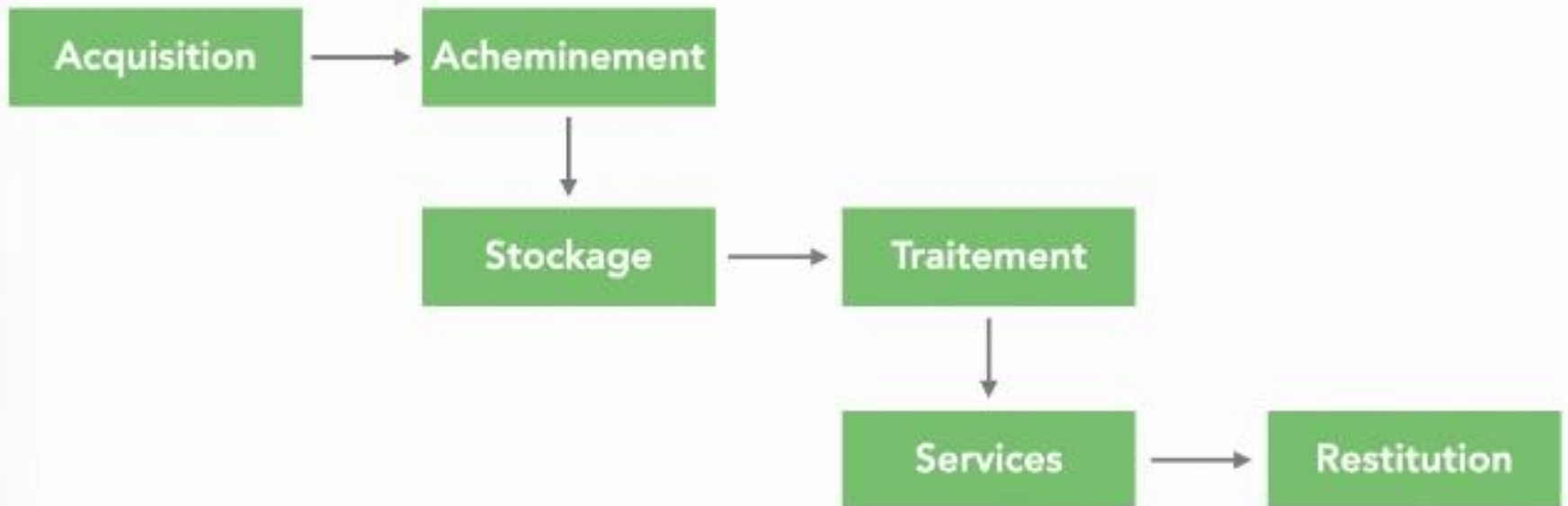
- Décrit la façon dont les données vont être reçues, traitées et exploitées
- Le pipeline générique se présente comme suit :



## Pipeline big data (2)

---

Traiter et gérer les données dans un projet big data



# Choisir une solution big data (1)

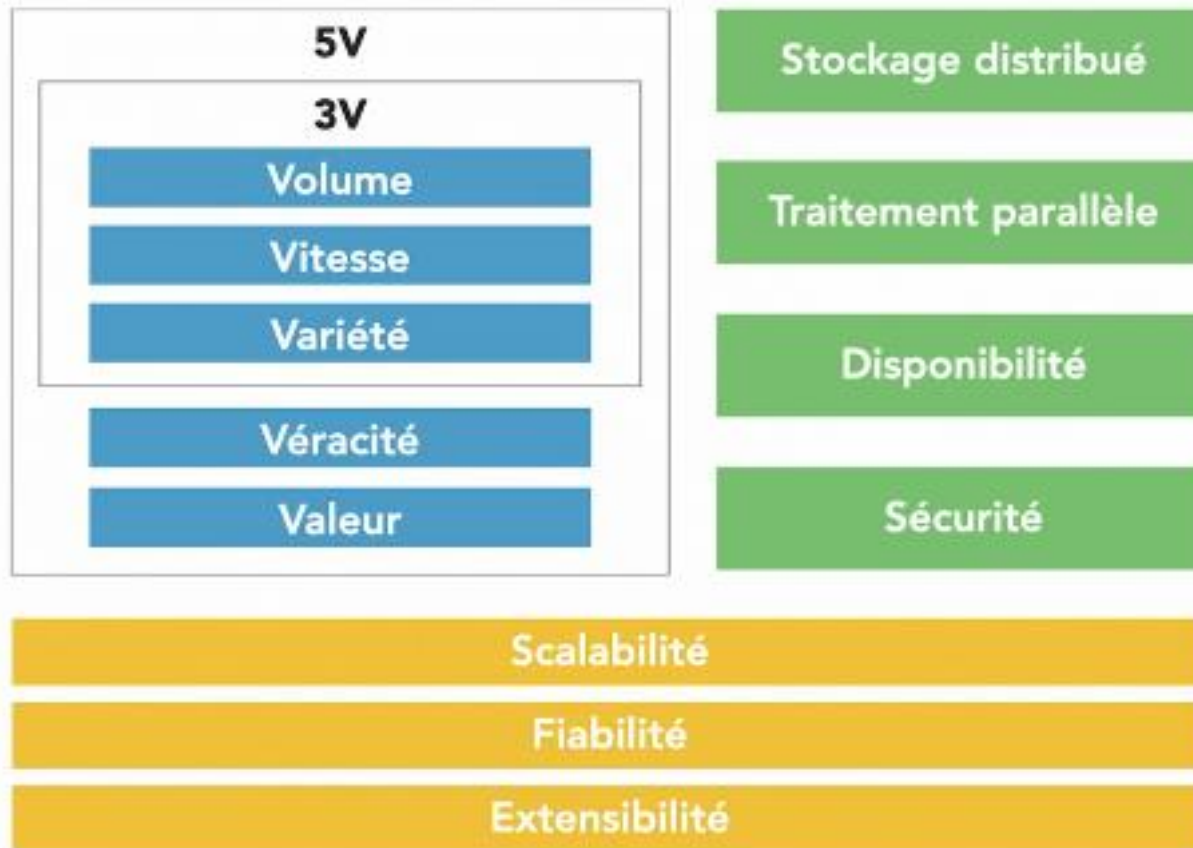
---

Pour choisir les technologies à utiliser pour implémenter un projet big data, il est important de s'intéresser à des technologie Capable de :

- ☐ Gérer des **stockage distribué**
- ☐ D'effectuer des **traitements parallèles** (permettre optimiser le temps de calcul et d'analyse réalisé, obligatoire dans un projet big data ou les données sont distribuées)
- ☐ Garantir un niveau de **disponibilité**
- ☐ **Assurer la sécurité**

# Choisir une solution big data (2)

---





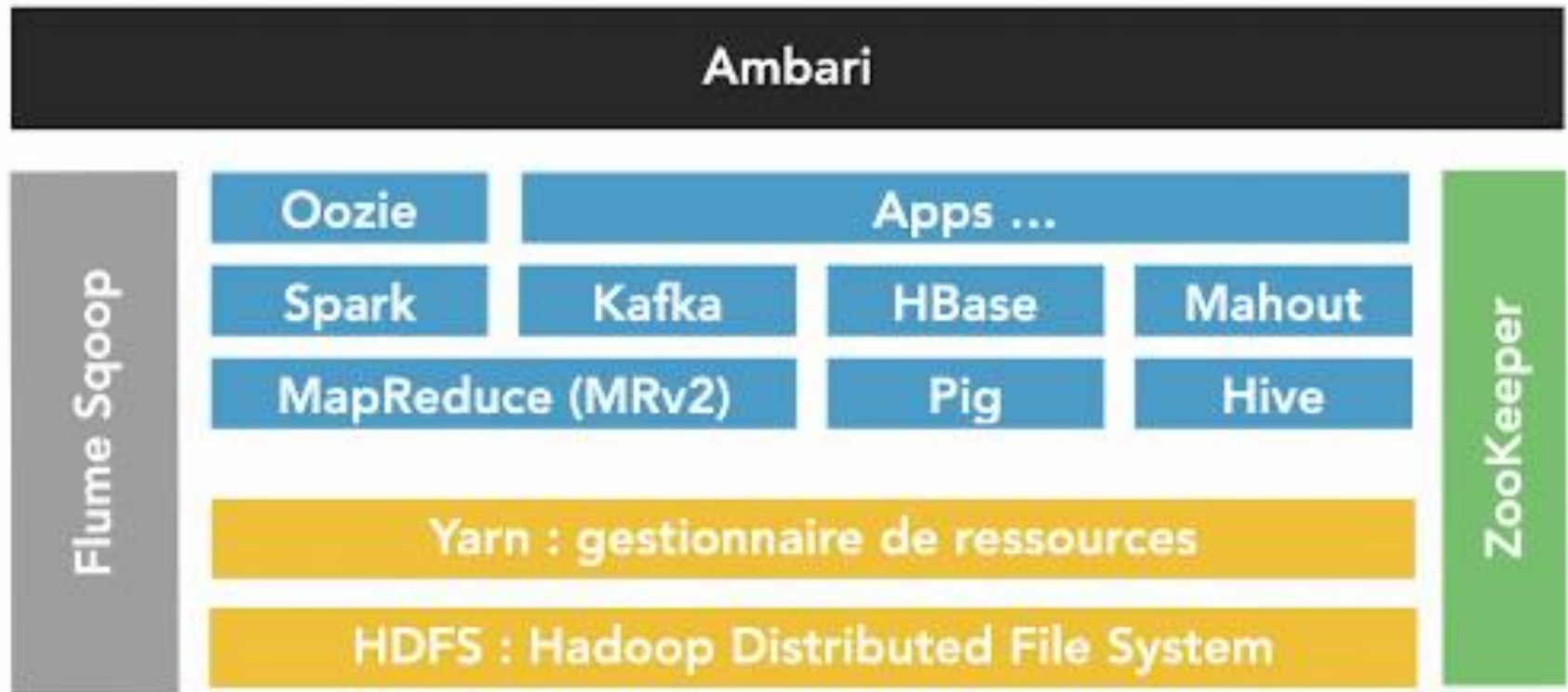
# Introduction à Hadoop

---

- ❑ Framework pour le big data open source
- ❑ Stockage de gros volumes de données
- ❑ Traitement parallèle des données

Il existe un bon nombre de solutions logicielles qui permettent de collaborer avec Hadoop afin d'implémenter des plateformes big data (cf. slide suivant)

# Écosystème Hadoop (1)



# Écosystème Hadoop (2)

---

- ❑ **HDFS** : Système de gestion des fichiers propre à Hadoop et permet de gérer les données de manière distribuée
- ❑ **Yarn** : le moteur de gestion de ressources de Hadoop
- ❑ **MapReduce** : un Framework qui permet de développer des programmes qu'on va pouvoir appliquer sur les données stockées.

HDFS, Yarn et MapReduce sont des modules natifs à Hadoop. C'est-à-dire lorsque nous installons Hadoop sur un cluster d'ordinateur, nous installons par définition sur un HDFS, Yarn et MapReduce



# Écosystème Hadoop (3)

---

- ❑ **Pig** : Une solution logicielle qui permet d'écrire des programmes déployés sur un cluster Hadoop de façon plus simple qu'avec MapReduce
- ❑ **Hive** : un système de requêtage de données issues d'un cluster Hadoop (simple tel les requêtage SQL)
- ❑ **Spark** : un écosystème à part entière dédié au traitement parallèle sur des données massives
- ❑ **Kafka** : est une solution logicielle qui permet principalement de réaliser des applications de STREAM, son utilisation avec Hadoop permet réaliser des applications en temps réel

# Écosystème Hadoop (4)

---

- ❑ **Hbase** : Une solution qui permet de gérer non relationnelle, en effet une base de données NoSql
- ❑ **Mahout** : Une plateforme qui permet de développer des programmes de type Machine Learning appliqués sur les données des clusters Hadoop
- ❑ **Oozie** : un planificateur de Workflow qui permet de planifier l'exécution d'un ensemble de job MapReduce
- ❑ **ZooKeeper** : est un service centralisé qui permet de gérer un certain nombre de configuration relative à notre cluster Hadoop

# Écosystème Hadoop (5)

---

- ❑ **Ambari** : Une solution qui permet de superviser un cluster Hadoop. Il est généralement utilisé lorsque l'on greffe un cluster Hadoop dans un système déjà existant
- ❑ **Flume Sqoop** : **Flume** est une solution qui permet de gérer un gros volume de données distribuées et sur généralement des logs des applications déployées dans un cluster Hadoop. **Sqoop**, quand à lui permet de déplacer les données de Hadoop vers des bases de données relationnelles externes



# S'initier au cloud computing

---

Le cloud computing et le big data ont des **liens forts**, mais ne sont **pas interdépendants**.

Alors c'est quoi le cloud computing ?

- Externalisation de la gestion de toute ou partie et/ou des systèmes d'information



# Acteurs majeurs du cloud computing

---

- Microsoft
- IBM
- Google
- Amazon

# Différents types de cloud

---

## 1 Cloud privé

- Ressources réservées à une entreprise
- Serveurs dédiés
- Plus cher

## 2 Cloud public

- Ressources utilisées en fonction des besoins
- Données de plusieurs entreprises
- Moins cher

## 3 Cloud hybride

- Achat par une entreprise d'une partie du cloud privé et d'une partie du cloud public

# Modèles de services cloud

---

## 1 IaaS

- Déléguer la gestion de son infrastructure
- Destiné aux administrateurs système

## 2 PaaS

- Fournit des VM et les environnements qui vont avec
- Destiné aux développeurs d'applications

## 3 SaaS

- Fournit des logiciels publics prêts à l'usage



# Avantages d'une solution cloud

---

- Coût en fonction de la consommation (pay as you go)
- Accessibilité
- Flexibilité
- Moins de maintenance