

# Big data

MOC : [\\$INFORMATIQUE\\$DATA ENGINEERING](#)

Projets :

Date : 2024-10-21

---

## Définition

Big data :: Le big data est définie sur 5v

- Volume: grande quantité de données
- Vélocité: vitesse de génération et de traitements des données
- Variété: diversité des types de données
- Véracité: qualité des données
- Valeur: tirer des infos utiles a partir des données

## Quand avons nous besoins de big data ?

- Le volume de données **dépasse la capacité** de la RAM et du stockage traditionnel.
- Le traitement des données doit être effectué en **temps réel** ou dans des délais stricts.
- La variété, la complexité et la vitesse à laquelle les données arrivent exigent des architectures et des outils spécialisés (**distribués, scalables**).

## Comment faire du big data ?

### 1. Scaling pour aller plus vite

- **Vertical scaling** : plus grosse machine ( plus cher )
- **Horizontal scaling** : plus de machines ( moins cher mais moins connecté )

### 2. Composantes distribués connectées mutuellement

[Storage](#)

[Stream](#)

[Processing](#)

### 3. Employés du metier

[Métiers du Big\\_data](#)

### 4. Gestions d'erreurs (les pannes par exemple)

- **At most once** ::: ensemble de donné traité que une fois (panne => risque pas traité)

- **At least once** ::: ensemble de données traité plusieurs fois (panne => tkt car traité plusieurs fois)
- **Exactly once** : a dream, on peut tricher (j'ai pas mon âge + 1 j'ai 22)

## 5. Exemple d'architecture

[Lambda Architecture](#)

[Kappa Architecture](#)

## Solution pour du big data

[5 Garden/Notes/cloud](#)

## Annex

[File Format](#)