

Cours : Big DATA

Chapitre 1: Introduction au Big Data

MAJ: 05/09/2023

Asma KERKENI asma.kerkeni@gmail.com

Plan

- ☐ Data: c'est quoi?
- ☐ Typologie de données
- ☐ Explosion des données
- ☐ Définition du Big Data
- ☐ Les 3V, 4V, 5V
- ☐ Applications du Big Data
- ☐ Big Data et NOSQL

BigData

2

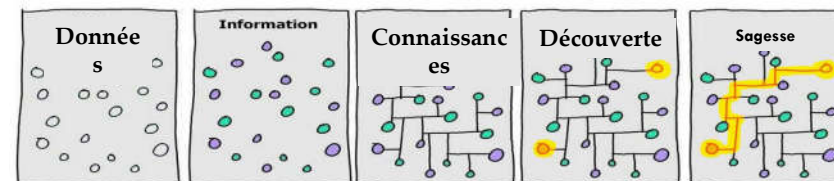
- ☐ Au terme de ce chapitre, vous serez capable de:
 - ☐ Définir le Big Data
 - ☐ Définir les 3V
 - ☐ Donner des exemples d'applications du Big DATA
 - ☐ Donner des chiffres sur le données massives
 - ☐ Comprendre la nature de ces données
 - ☐ Situer les bases de données NOSQL dans le Big Data

BigData

Data : C'est quoi ?

4

- ☐ Les données sont des symboles qui représentent les propriétés des objets et événements.
 - ☐ Elle n'a pas de sens en elle-même. Elle existe tout simplement.
 - ☐ C'est la matière première sur laquelle la plupart des informations sont basées.
 - ☐ **Exemple:**
 - **Données :** { « Ali », « Ben Ahmed », 55}
 - **Information:** Données +sens : {Prénom: « Ali », Nom:« Ben Ahmed », age: 55}



BigData

Typologie des données

6

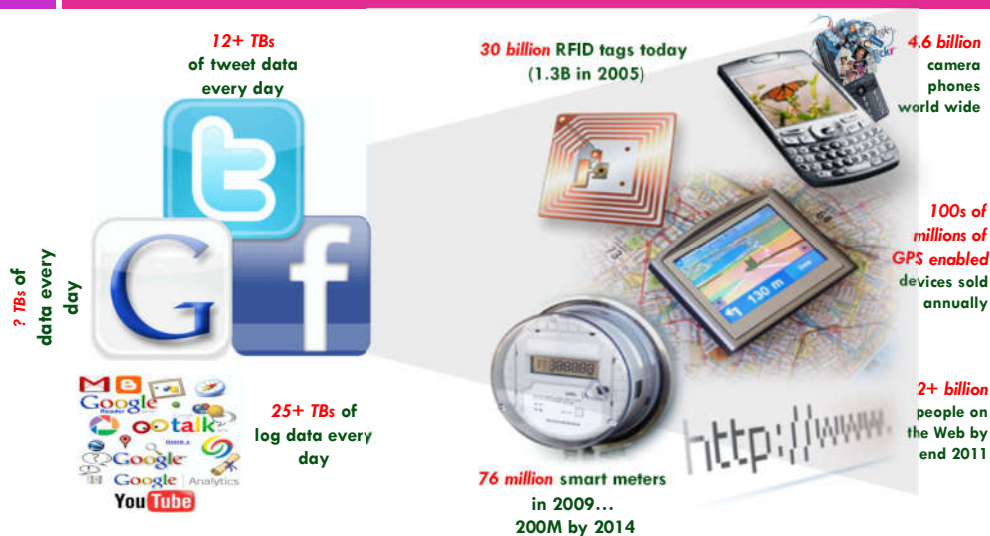
Données qualitatives/ quantitatives:

- Quantitative:** sont des données qui peuvent être mesurées (taille, poids...) ou repérées (température...)
- Qualitative:** sont des données auxquelles on ne peut pas attribuer une valeur ou une caractéristique (la couleur, la texture, le goût, l'odeur).

BigData

D'où viennent les données?

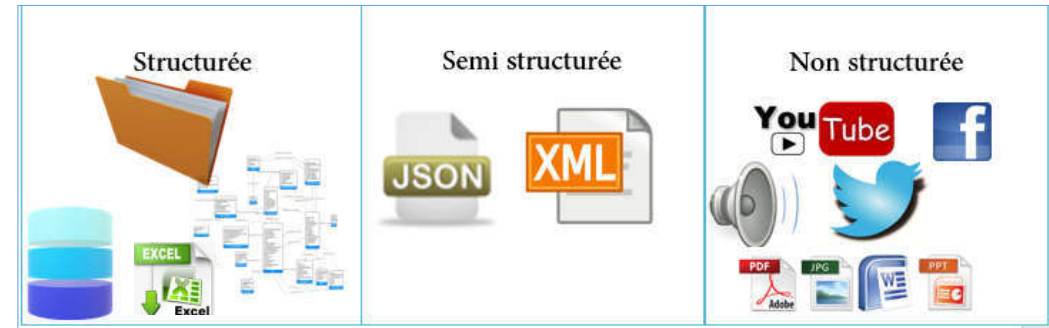
8



Typologie des données

7

Données structurées/Semi-structurées/Non structurées

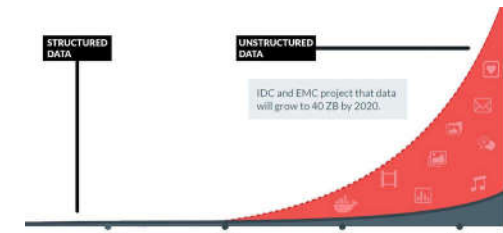


BigData

D'où viennent les données?

10

- Généralement, ce ne sont pas seulement des données structurées (comme dans les BD traditionnels) normalisées.
- Ce sont des données semi ou non structurées:
 - Tweets
 - Logs des serveurs
 - Pages web
 - Objets connectés
 - Fichiers CSV, JSON, etc. volumineux



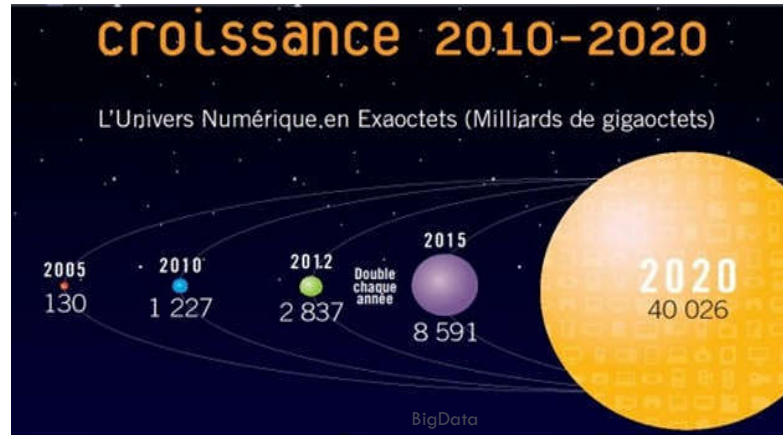
```
{
  "hashtags": {
    "text": "TulsaAirport",
    "indices": [0, 13],
    "text": "Oklahoma",
    "indices": [14, 23]
  },
  "trends": [],
  "urls": {
    "url": "http://t.co/Vt.coVSnC8ST3gQC",
    "expanded_url": "http://bit.ly/188eNcw",
    "display_url": "bit.ly/188eNcw",
    "indices": [93, 115]
  },
  "user_mentions": [],
  "symbols": [],
  "favorited": false,
  "retweeted": false,
  "possibly_sensitive": false,
  "filter_level": "low",
  "lang": "en",
  "timestamp_ms": "1421853664710",
  "created_at": "Wed Jan 21 15:21:04 +0000 2015",
  "id": "557920823877464064",
  "id_str": "557920823877464064",
  "text": "An imeepisode updated: Kyoukaino Kanata: Mini Theater # 6 ( http://t.co/VkEPWveEHM)"
}
```

BigData

Explosion des données

13

- Il n'y avait que **130 exaoctets** de données dans l'univers numérique en 2005.
- Il devrait y en avoir plus de **40 000 exaoctets** à l'horizon 2020.
- En 2020, les données représenteront l'équivalent de plus de **5 000 GO** par personne!



Encore des chiffres

15

- Volumes de données estimées:

- Google: 15 000PB (=15 Exabytes)
- Facebook: 300PB
- Ebay: 90PB

- Volumes de données par jour:

- Google: 100 PB (5 milliards de requêtes par jour)
- Twitter: 100 TB (>200 millions de tweets par jour)
- Facebook: 600 TB

"90% of the data in the world today has been created in the **last two years alone.**"
- IBM

BigData

Explosion des données

14

1 KB	1024 B	B = byte
1 MB	1024 KB	KB = Kilobyte
1 GB	1024 MB	MB = Megabyte
1 TB	1024 GB	GB = Gigabyte
1 PB	1024 TB	TB = Terabyte
1 EB	1024 PB	PB = Petabyte
1 ZB	1024 EB	EB = Exabyte
1 YB	1024 ZB	ZB = Zettabyte
		YB = Yottabyte

BigData

Big Data: C'est quoi donc ?



- En français: Mégadonnées, Données massives

- C'est donc le volume de données??

Réponse: Non.

- Il n'y a pas que la dimension volume

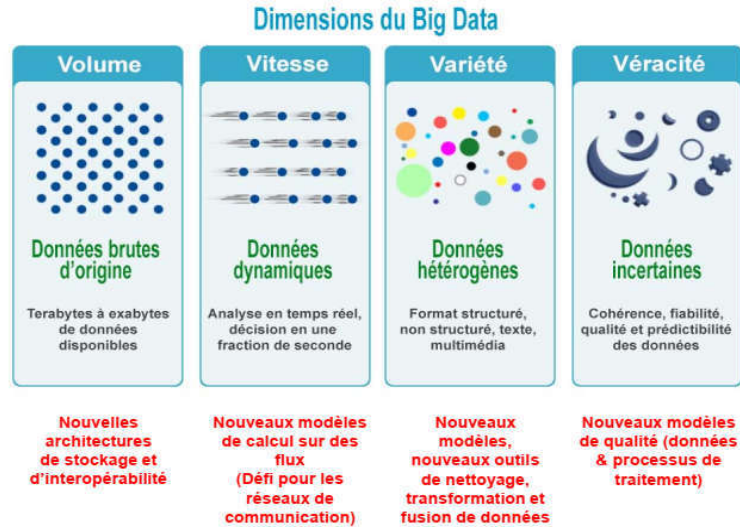
- Le Big Data représente les collections de données caractérisées par un **volume**, une **vélocité** et une **variété** si **grands** que leur transformation en **valeur** utilisable requiert l'utilisation de **technologies et de méthodes analytiques spécifiques**. (d'après Gartner)

BigData

Big data : 3V, 4V, 5V, ?V

18

- **3V:** Volume, Variety, Velocity.
- **4V:** Volume, Variety, Velocity, Value.
- **5V:** Volume, Variety, Velocity, Value, Veracity.



<http://www.astrosurf.com/luxorion/big-data-mining.htm>

BigData

Entreprise et Big Data

-Data Lake-

Forget data warehousing, it's 'data lakes' now

Digital News Asia Mar 31, 2015

- **Data Lakes becoming corporate priority because they fill a critical gap**
- **Federation Business Data Lake simplifies complex task of building a data lake: EMC**

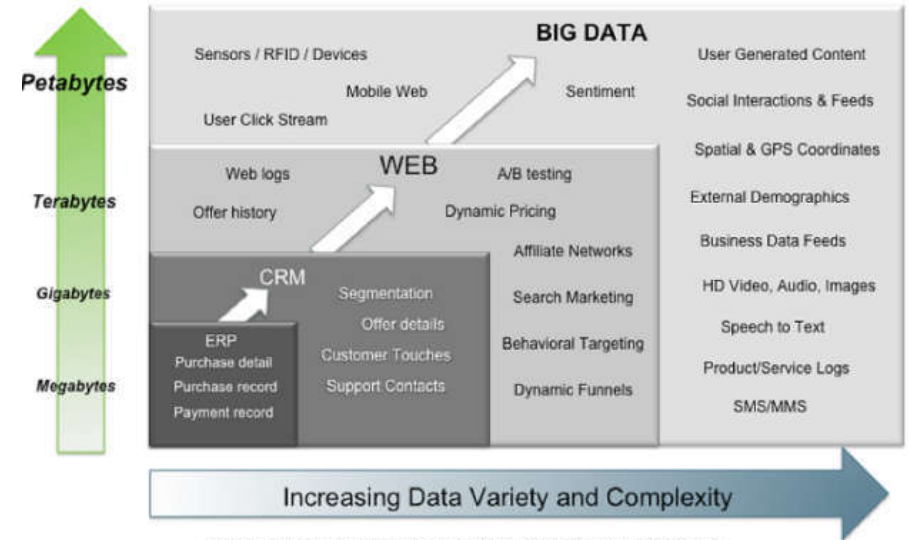


BigData

Entreprise et Big Data

28

Big Data = Transactions + Interactions + Observations



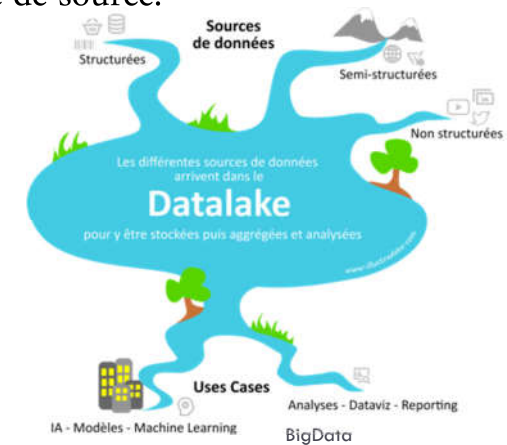
Source: Contents of above graphic created in partnership with Teradata, Inc.

Entreprise et Big Data

-Data Lake-

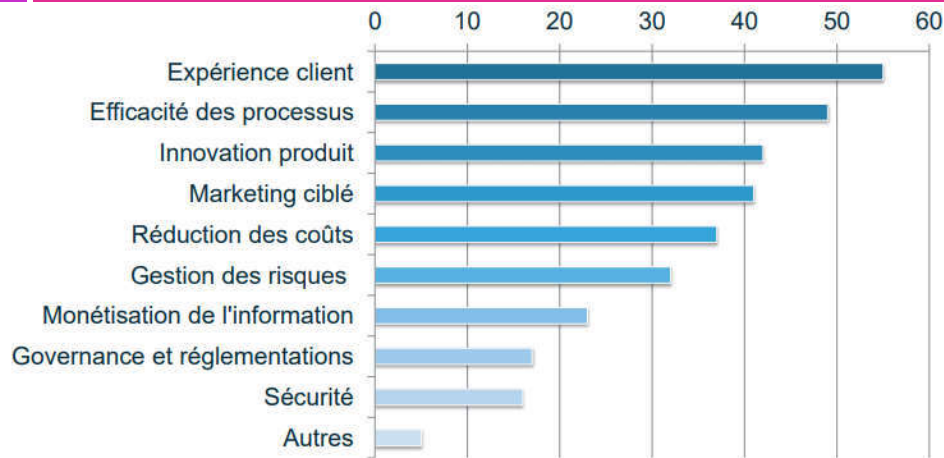
30

- Un data lake est un emplacement de stockage centralisé qui contient des big data sous un format brut et granulaire provenant d'un grand nombre de source.



Applications du Big Data: Pour qui, pour quoi?

36



□ La relation client avant tout, puis l'efficacité des processus et l'innovation

BigData

Exemples d'applications

40

□ Recommandations basées sur les données utilisateur:

□ « *There are 33 million different versions of Netflix.* » (Joris Evers)

□ 75% des vidéos regardées viennent des suggestions de Netflix à ses utilisateurs.

□ Le Big Data a motivé l'achat des droits (pour 100 million US\$) et le tournage de la série House of Cards (popularité de séries similaires, choix du producteur, du réalisateur, des acteurs).

□ Bande annonce personnalisée en fonction du type d'utilisateur.

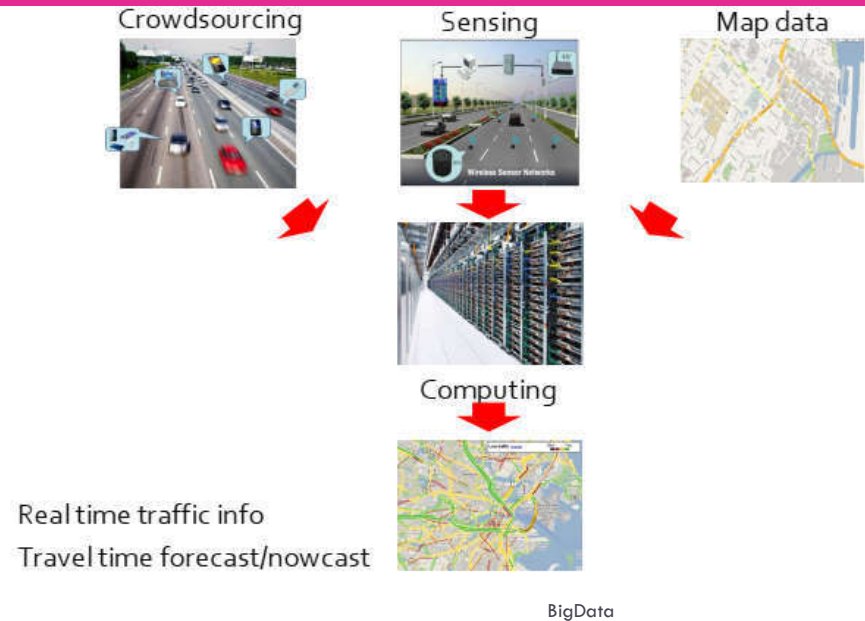
□ House of Cards a amené 2 millions de nouveaux utilisateurs aux USA et 1 million en dehors des USA.

NETFLIX

BigData

Exemples d'applications

37



BigData

Exemples d'applications

41

□ PredPol=« Predictive Policing »

□ prédire la probabilité des crimes et délits (nature, localisation à quelques centaines de mètres près, timing à 12h près)

□ base d'apprentissage de 13 millions d'événements

□ Los Angeles: diminution de 33% des agressions

□ Santa Cruz : diminution de 27% des cambriolages en moins d'un an.

□ D'autres applications:

□ Fraude (détection / prévention),

□ Education, santé,

□ Génomique,

□ etc...

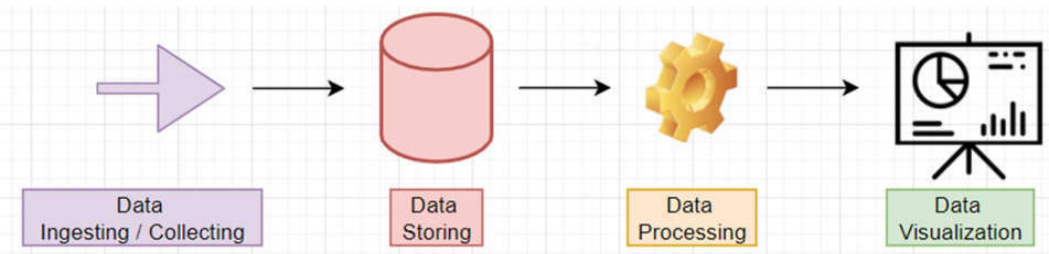
PREDPOL

BigData

Pipeline des données

42

- Système d'acheminement des données d'un point A vers un point B.
- Mis en place par un Data-Engineer.



BigData

Challenges du Big Data

43

- On s'intéresse dans ce cours à deux challenges: **traitement** des données et **stockage**.
- **Traitement:**
 - ▣ Paralléliser le calcul sur les machines
 - ▣ Framework de traitement distribuée (Hadoop /Spark)
- **Stockage:**
 - ▣ Système de fichiers distribué
 - ▣ Bases de données (relationnel/NoSQL) distribués

BigData

NOSQL: C'est quoi?

45

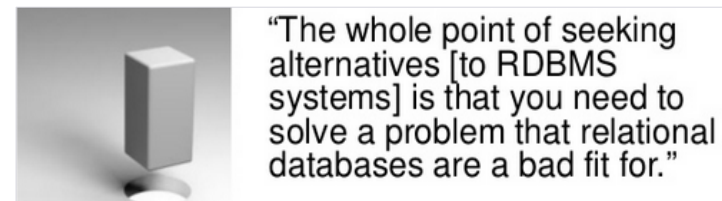
- Nom exact: **Bases de données non relationnelles**.
- Privilégier NOSQL plutôt que NoSQL.
- Ce n'est pas du relationnel, et le contexte d'utilisation n'est donc pas ce lundis SGBDR.
- Ce n'est pas seulement l'opposition à SQL :
 - ▣ il s'agit de compléments aux SGBDR pour des besoins spécifiques et non de solutions de remplacement.
 - ▣ expérimentations, autres modèles de données très simples, nouveaux besoins, nouveaux outils !
 - ▣ logiciels de stockage de données plutôt que SGBD



BigData

NOSQL: C'est quoi?

46



Eric Evans
Rackspace



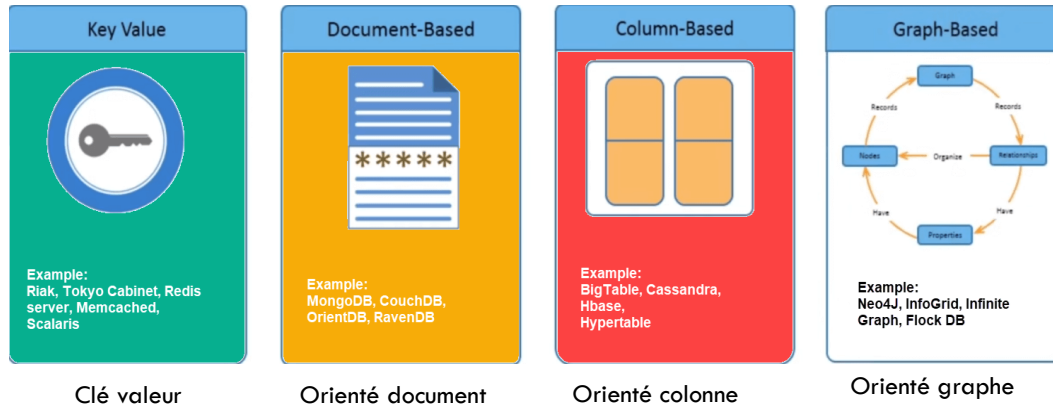
- Un couple de concepteurs de bases de données entre dans un restaurant NoSQL. Une heure après, ils ressortent sans avoir réussi à manger.
- ▣ Pourquoi?

BigData

Types des bases NOSQL

47

4 types de bases NOSQL



BigData

Références

49

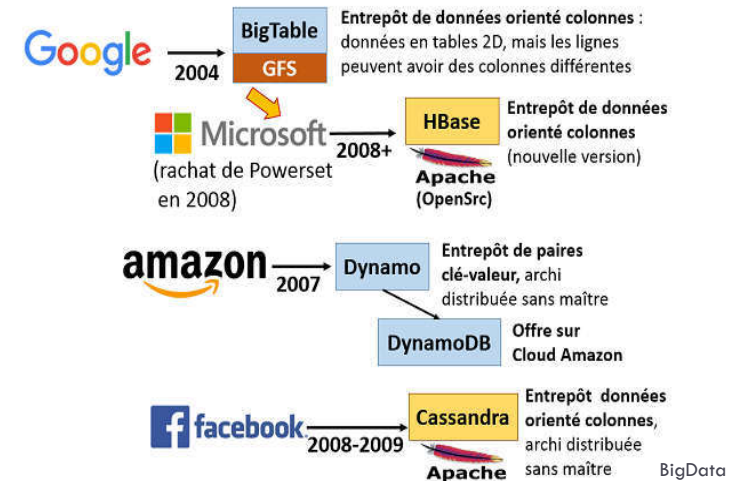
- Laurent Jolia-Ferrier, **Big Data - Concepts et mise en œuvre de Hadoop**, Ed Eni, 2014.
- Pirmin Lemberger et al., **Big Data et Machine Learning -Les concepts et les outils de la data science**, Ed, Dunod, 2016.
- P. Lacomme, S. Aridhi, R. Phan, **Bases de données NoSQL et Big Data: Concevoir des bases de données pour le Big Data**, Cours et travaux pratiques Ellipses ~ Technosup, 2/12/2014, 336 p., 9782340002616.
- Rudi Bruchez, L. **Les bases de données NoSQL et le Big Data: Comprendre et mettre en œuvre**, Ed. Eyrolles, 2015.
- Cours Bases de données NOSQL de Raja CHIKY, Institut Mines Télécom: <http://perso.isep.fr/rchiky/>
- Cours Big Data et NoSQL de Lilia Sfaxi, INSAT, Tunisie: <http://liliasfaxi.wixsite.com/liliasfaxi>
- <https://chewbii.com/>
- Chaine Youtube **Tech Wall**, playlist **Hadoop&cie**.

BigData

Big Data et NOSQL

48

- Les bases de données NoSQL sont des outils Big Data qui permettent de **stocker** et **récupérer** de très gros volumes de données.



BigData

50

Fin