



# Big Data - Hadoop

Sébastien Jardin - DevOps

Oct. 2018

# Définition



“ Le big data, littéralement « grosses données », ou mégadonnées, parfois appelées données massives, désigne des ensembles de données devenus si volumineux qu'ils dépassent l'intuition et les capacités humaines d'analyse et même celles des outils informatiques classiques de gestion de base de données ou de l'information.

-- [BigData - Wikipedia, the free encyclopedia](#)

”

# Le BIG DATA en 3 phrases



**1** Historiquement,  
les données d'une  
entreprise étaient  
saisies par ses employés  
(factures, fichiers...)



**2**

L'informatisation de nos sociétés et l'avènement du web ont permis à des tiers de générer des informations pour l'entreprise (emails, fichiers, réponses à des enquêtes électroniques, données sur le web, tweets et échanges sur les réseaux sociaux...)

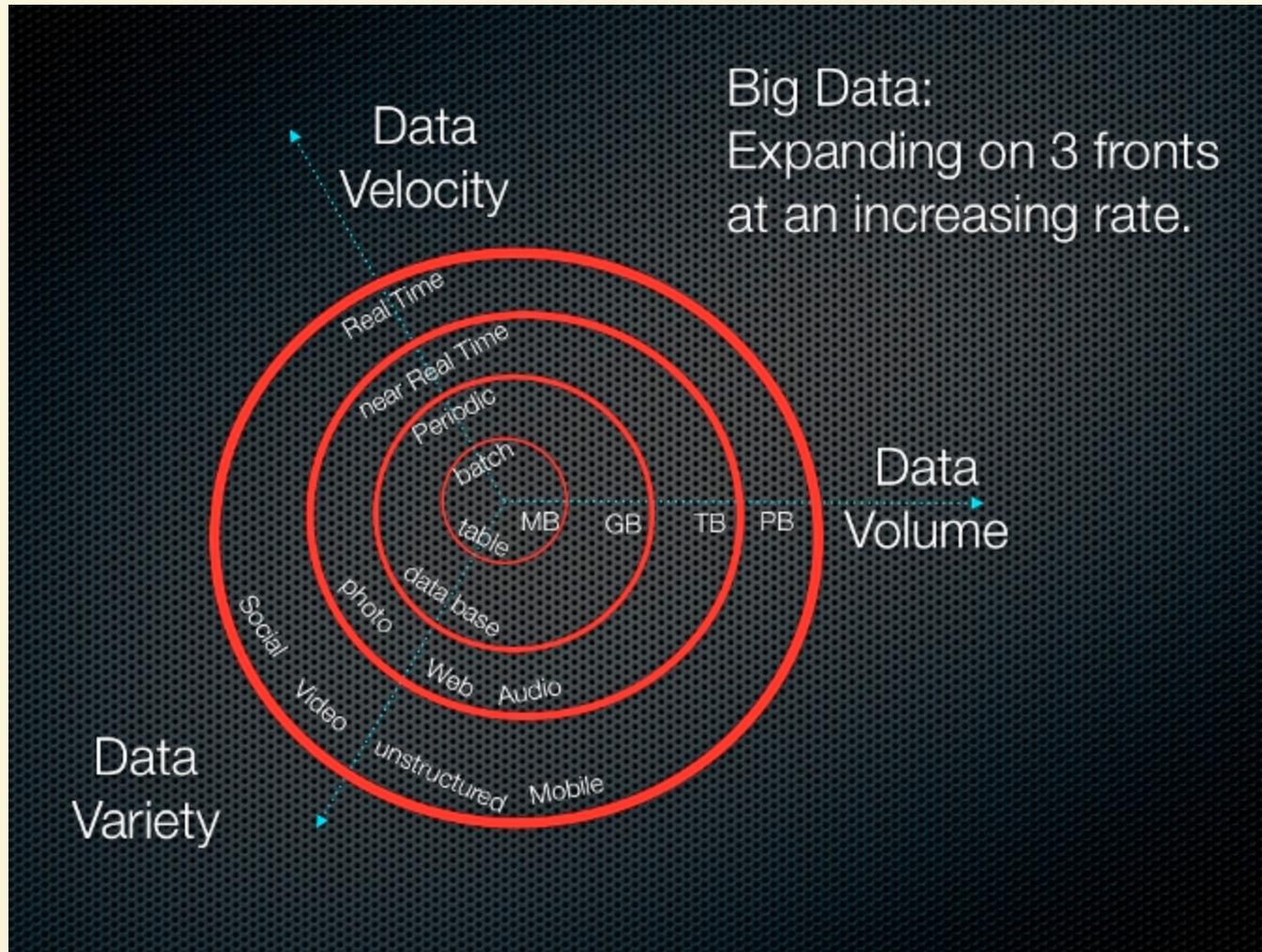


**3**

Aujourd'hui, d'immenses masses de données utiles à l'entreprise sont générées en permanence et de manière automatique par des machines à l'intérieur et à l'extérieur de l'entreprise : sites web, standards téléphoniques, systèmes d'encaissement, serveurs, mobiles, cartes, véhicules, objets connectés...

Le BIG DATA, c'est toutes ces données à la fois, de plus en plus variées, nombreuses et rapides à obtenir, et qui comportent une richesse analytique que chaque entreprise peut (doit) exploiter.

# Les 3V



# Puis les 5v

## THE 5V'S: TURNING BIG DATA INTO VALUE

With the datafication comes big data, which is often described using the four Vs:

**THE DATAFICATION OF OUR WORLD:**

**ACTIVITY DATA**

Music players, eReaders and smart phones collect data on how we use them; web browsers collect information on what we search for; credit card companies collect data on where we shop; and shops collect data on what we buy.

**CONVERSATION DATA**

Our conversations are being captured - From emails to all the conversations we have on social media sites like Facebook or Twitter as well as our phone conversations are now digitally recorded.

**PHOTO AND VIDEO IMAGE DATA**

All the pictures and videos we take on our smart phones and digital cameras - we upload and share millions of them on social media sites every second.

**SENSOR DATA**

We are surrounded by sensors that collect and share data - devices like our smart phones use sensors to track our location, the speed and direction at which we are travelling, read our fingerprints, detect how light it is outside, etc.

**VELOCITY**

**VOLUME**

...refers to the vast amounts of data generated every second - Today, we create the same amount of data in a single minute, that was created from the beginning of time until the year 2000.

**VERACITY**

**VARIETY**

...refers to the different types of data we can now use - Today, we don't have to rely on nicely structured data, we can now collect and analyse text, images, video, voice, location data, and much more.

**VERACITY**

**VALUE**

...refers to the messiness or trustworthiness of the data - Today, quality and accuracy of data are less controllable (hash tags, abbreviations, typos and colloquial speech) but technology now allows us to deal with it.

**VALUE**

...the final V refers to the need to turn our data into value - Today, big data is used to better understand and target customers, understand and optimize business processes, and improve health care, security and law enforcement. But the possible applications of big data are endless!

**ANALYZING BIG DATA:**

**TEXT ANALYTICS**

**SENTIMENT ANALYTICS**

**FACE RECOGNITION**

**VOICE ANALYTICS**

**MOVEMENT ANALYTICS**

# Tour d'horizon



# Landscape 2018



BIG DATA & AI LANDSCAPE 2018



Final 2018 version, updated 07/15/2018

© Matt Turck (@mattturck), Demi Obavomi (@demi\_ obavomi), & FirstMark (@firstmarkcap)

mattturck.com/bigdata2018

FIRSTMARK  
EARLY STAGE VENTURE CAPITAL

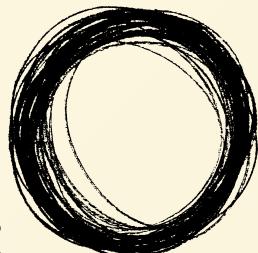
# The solutions

3 leaders : Mapr/Cloudera/HortonWorks (The past years ago)

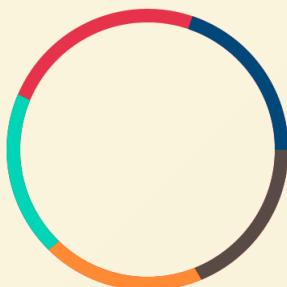
2 leaders: Mapr/Cloudera-HortonWorks (2018 S4)



# Darwin Theory or Not



B.I : Données froides, données structurées/formalisées, indicateurs



DataLake: Data Warehouse V2, données brutes



Big-Data: Révolution, Temps-réel, illimité

# Débouchés



- Chief Data Officer (CDO)
- Business Intelligence Manager
- Data Scientist
- Data Analyst
- Le Data Miner
- Master Data Manager
- Data Protection Officer

# Chiffres clés

- Cdiscount
  - 4 Clusters
  - 200 Tio compressés
- Uber
  - Ubber 100+ Petabytes

(1peta = 1000Tio) donc 102 400 TB

# 10

## CHIFFRES DU BIG DATA

20

ANS

Que l'expression "Big Data" est apparue. Selon la bibliothèque ACM, elle serait apparue en octobre 1997.

6

MILLIONS

De développeurs travaillent à la création d'outils de Big Data, d'IA ou de Machine Learning, dans le monde.

90

POURCENT

Des données créées, l'ont été au cours de ces deux dernières années.

57

MILLIARDS

De recettes réalisées par les vendeurs de services de Big Data, en 2017.

50

MILLE GO

De données sont créées, par seconde, en 2018.

58

POURCENT

Des entreprises françaises n'arrivent à quantifier le ROI de leurs investissement Big Data.

43

POURCENT

Des entreprises se restructurent, ou l'ont déjà fait, pour exploiter le Big Data.

50

POURCENT

Des requêtes sur internet se feront depuis la reconnaissance vocale, d'ici 2020.

1

NUMÉRO 1

Selon le Glassdoor 2017, le métier de Data Scientist est le job n°1 au monde (taux de satisfaction : 4,4/5).

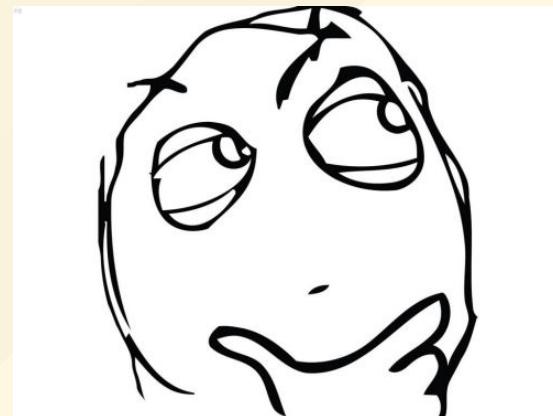
75

POURCENT

Des entreprises vont utiliser le Big Data et l'IA d'ici 2020.

Merci 🙏

Des questions ?



Créer avec [Marp](#) en Markdown, Libre et gratuit !