



Big Data : Les fondamentaux de l'analyse de données

=>[Lien de la formation en ligne](#)<=

Dr. FIRAS
Docteur.firass@gmail.com
<https://www.dr-firas.com/>

Table des matières

- 01**  **Deux approches pour l'analyse de données**
- 02**  **Processus de collecte de données**
- 03**  **Cartographier La donnée : quoi, pourquoi, comment ?**
- 04**  **L'ère de la data**
- 05**  **Big data, Smart data, Dark data, Fast data, Open data**
- 06**  **Les techniques du Big data**
- 07**  **Big data avec Python**

Ressources

La boîte à outils de la Stratégie big data.

- Auteur(s): Rissoan, Romain Jouin, Romain
- Editeur: Dunod
- Année de Publication: 2018 pages: 192
- ISBN: 978-2-10-077898-0

La boîte à outils de l'Analyse de données.

- Auteur(s): Coron, Clotilde
- Editeur: Dunod
- Année de Publication: 2020 pages: 314
- ISBN: 978-2-10-080855-7

Deux approches pour l'analyse de données

01

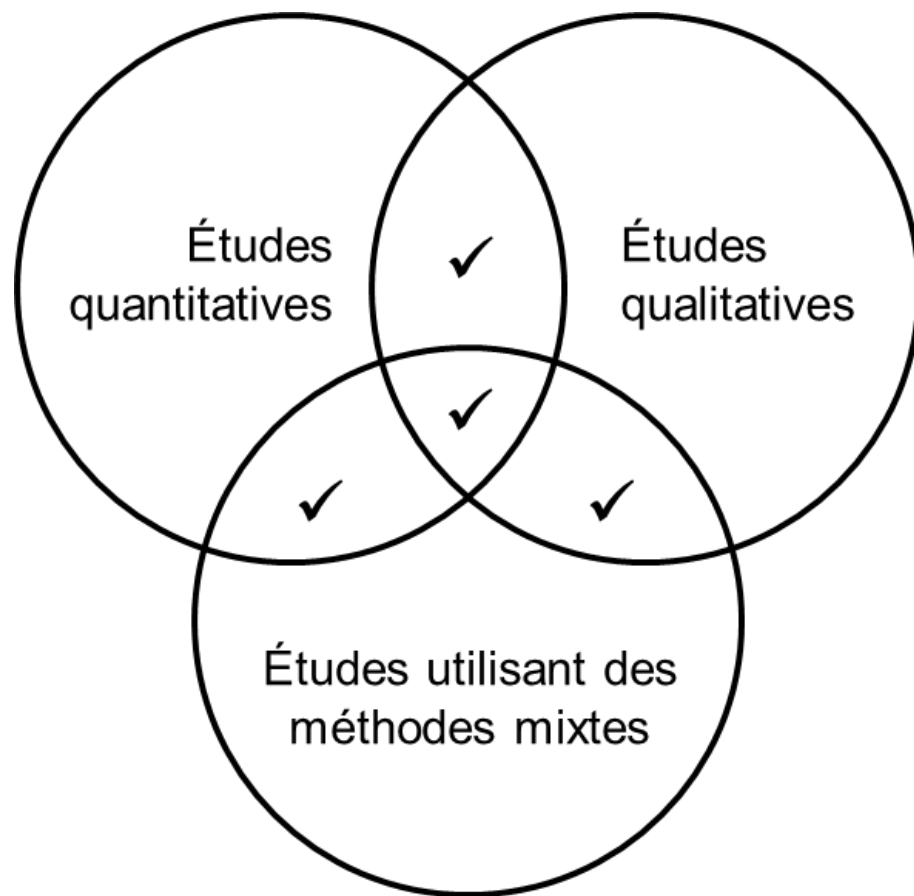


SOURCE :
La boîte à outils de l'Analyse de données.

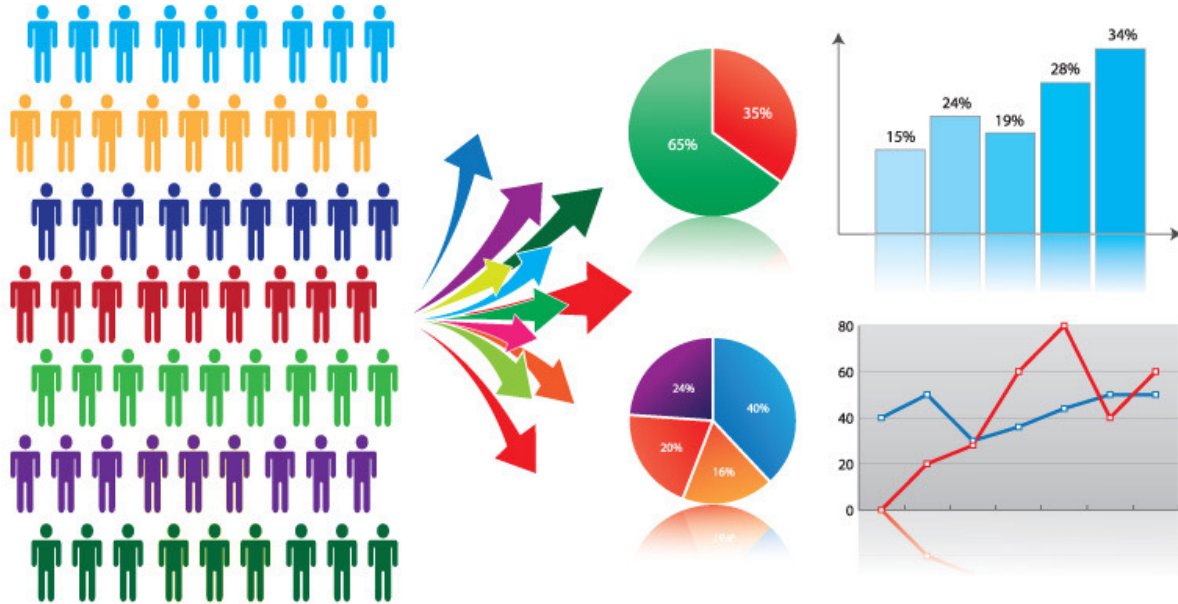
Auteur(s): Coron, Clotilde Editeur: Dunod
Année de Publication: 2020

QUESTIONS

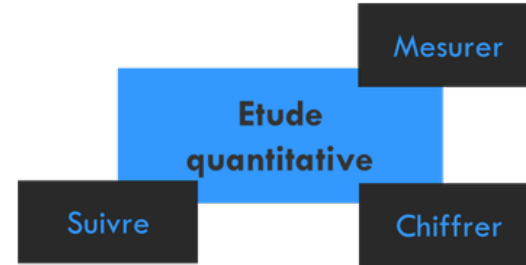
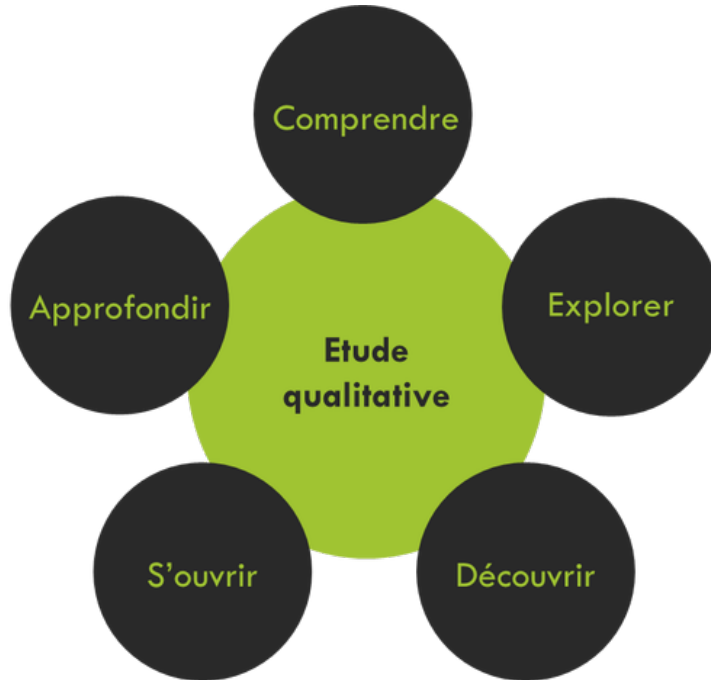
- Une approche quantitative ou une approche qualitative ?
- Quels objectifs cherche-t-on à atteindre avec l'analyse de données ?
- De quels moyens financiers, humains et informatiques ?
- Que cherche-t-on à mesurer ?



Définir les types d'objets mesurés

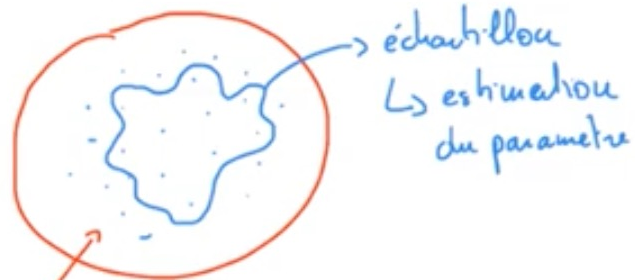


Définir les questions



ÉTUDES STATISTIQUES

Échantillonnage



↳ estimation du paramètre

population

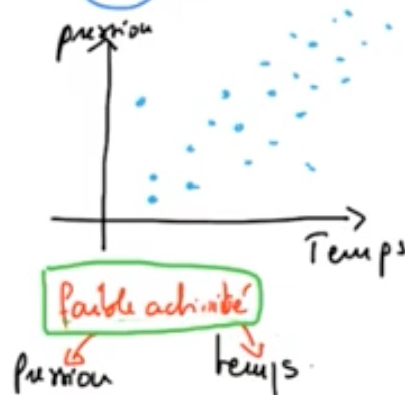
paramètre : temps moyen passé de vant un ordinateur chaque jour.

Étude observationnelle

1000 personnes



Temps ordi
et
pression artérielle



Étude expérimentale

100 personnes

aléatoire

30 min
max

+2h



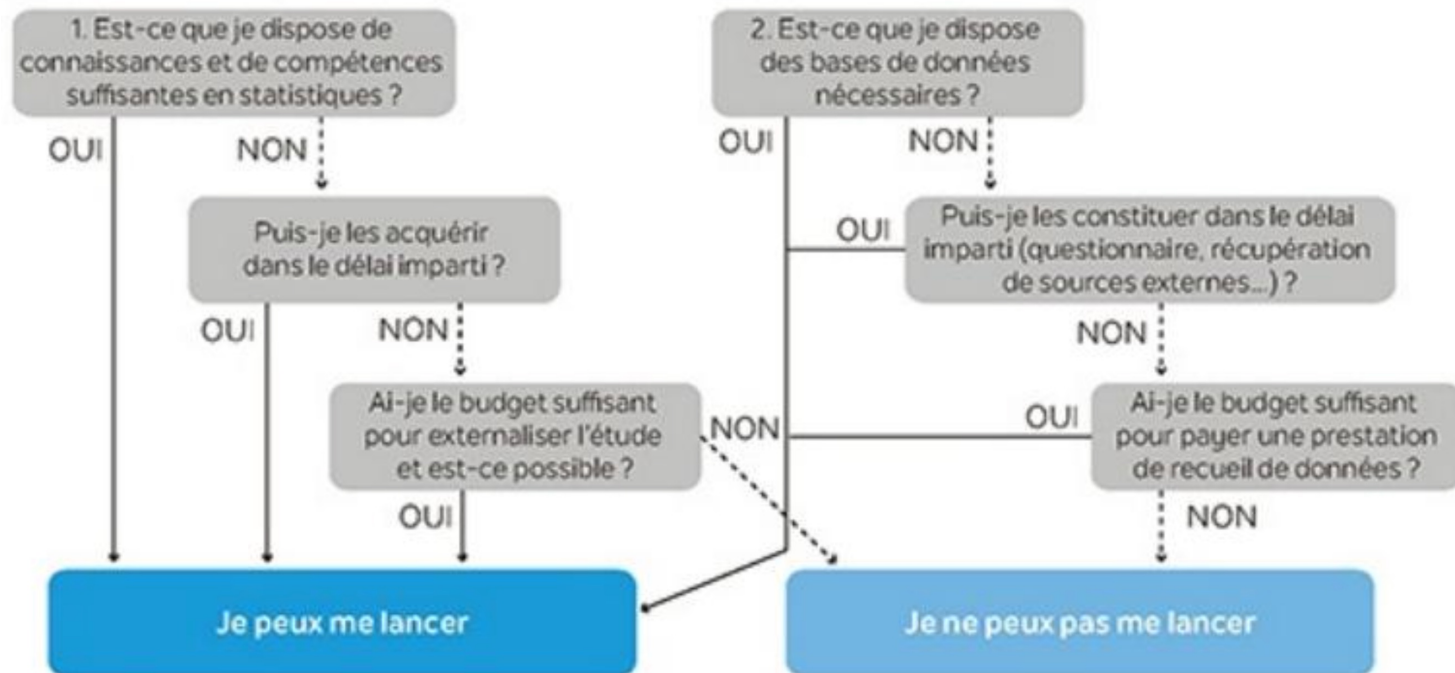
contrôle
(temoin)



Traitement

pression + élève

Quels sont les moyens ?





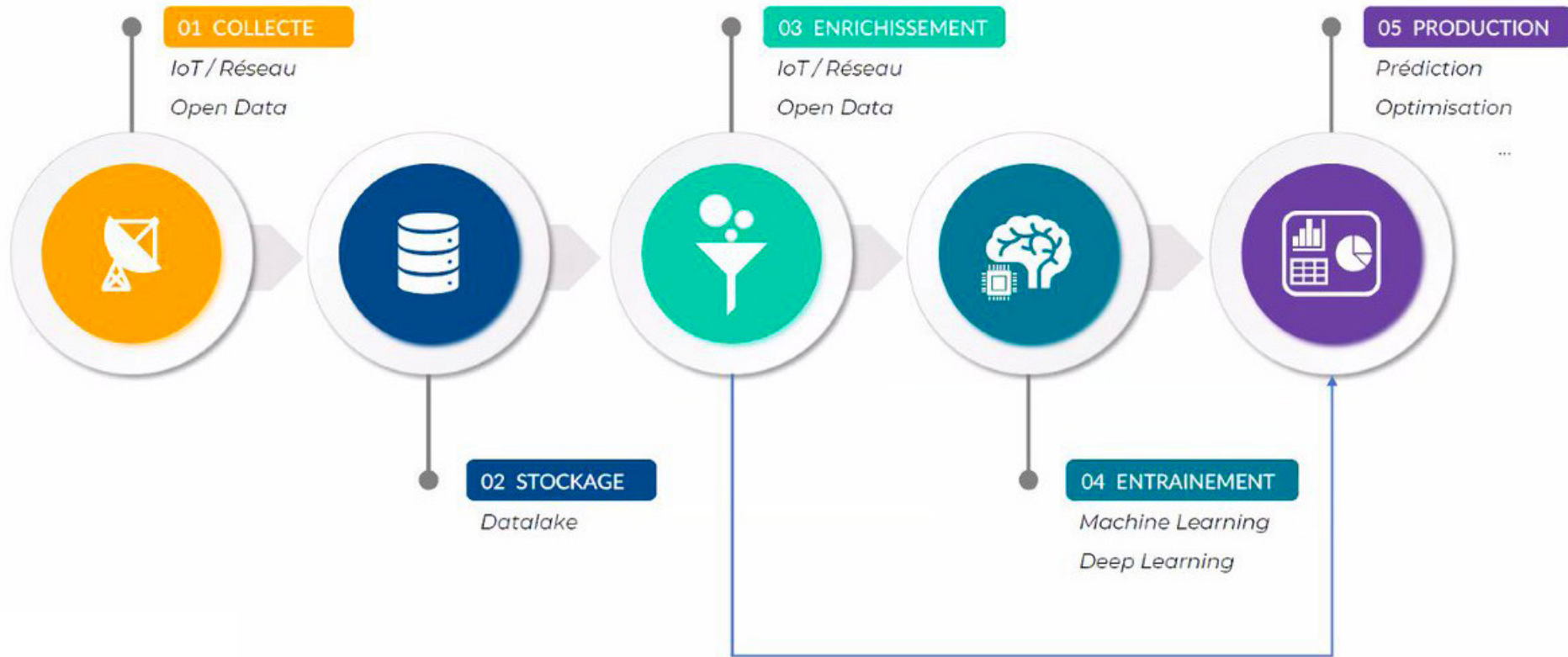
Processus de collecte de données

02

SOURCE :
La boîte à outils de l'Analyse de données.

Auteur(s): Coron, Clotilde Editeur: Dunod
Année de Publication: 2020

Le cycle de vie de la donnée



Les données externes



Données externes non-structurées
(réseaux sociaux + UGC...)



Données internalisées
(ERP, CRM, datawarehouse)



Données externes structurées
(Sondages, panels, partenaires...)



Données utilisables par l'entreprise

Cartographeur

La donnée : quoi, pourquoi, comment ?

03



SOURCE :
Data Galaxy

<https://www.datagalaxy.com/blog/cartographie-des-donnees-quoi-pourquoi/>

Exploitation de la donnée

À mesure que **les données sont massivement collectées** par l'entreprise, elles doivent être exploitées :

- **Analysées,**
- **Triées,**
- **Segmentées,**
- **Qualifiées.**

Qu'est-ce que la cartographie des données ?

la cartographie des données est un processus permettant de recenser, puis de visualiser, les points d'entrée et de traitement de la data.

- • **L'outil sémantique**, consistant à recenser dans un glossaire métier les métadonnées des données et des objets métier propres à l'entreprise, afin d'optimiser la compréhension et le contexte de la data.

Qu'est-ce que la cartographie des données ?

- ➔ • **Les modèles de données**, apportant des indications précises sur la façon dont la data est modélisée et stockée dans les différents systèmes de stockage (structurées, semi structurées ou non structurées). Un autre outil lui est consubstantiel : **le dictionnaire des données techniques**.
- ➔ • **L'outil de design des traitements des flux de données**, apportant des informations sur les méthodes de transformation, de manipulation et de traitement de la data au travers des différents SI de l'entreprise.

Qu'est-ce que la cartographie des données ?

➡ À ces trois outils, il faut ajouter **la formalisation** du format de mise à disposition, d'accès et de conditions d'utilisation des données, pour que le partage de la cartographie ne se heurte pas aux obstacles relatifs **à la confidentialité** et à **l'intégrité de la compréhension des données**.

Pourquoi ce processus est-il capital ?

➡ Obtenir une vision à 360° de la masse **des données en circulation dans une entreprise** : telle est la solution offerte par le processus de **cartographie de la donnée**. Le but étant de l'universaliser (à l'échelle de la structure, du moins), c'est-à-dire de la rendre totalement accessible et compréhensible à l'ensemble des collaborateurs, afin que chacun soit à même d'identifier la provenance d'une donnée, de mesurer son mode de calcul, de repérer d'éventuelles redondances

Pourquoi ce processus est-il capital ?

→ **Cartographier les données** permet donc de tisser des liens entre ***la vision technique*** (stockage et transformation de la data dans les applications des SI) et ***la vision métier*** (l'utilisation qui en est faite au sein des différents services). Ce processus vise à faire tomber les silos historiques, à décloisonner les services – bref, à édifier un espace commun dans lequel tous les collaborateurs parlent le même langage, afin de donner à chacun les outils d'une meilleure ***gouvernance opérationnelle de la donnée***.

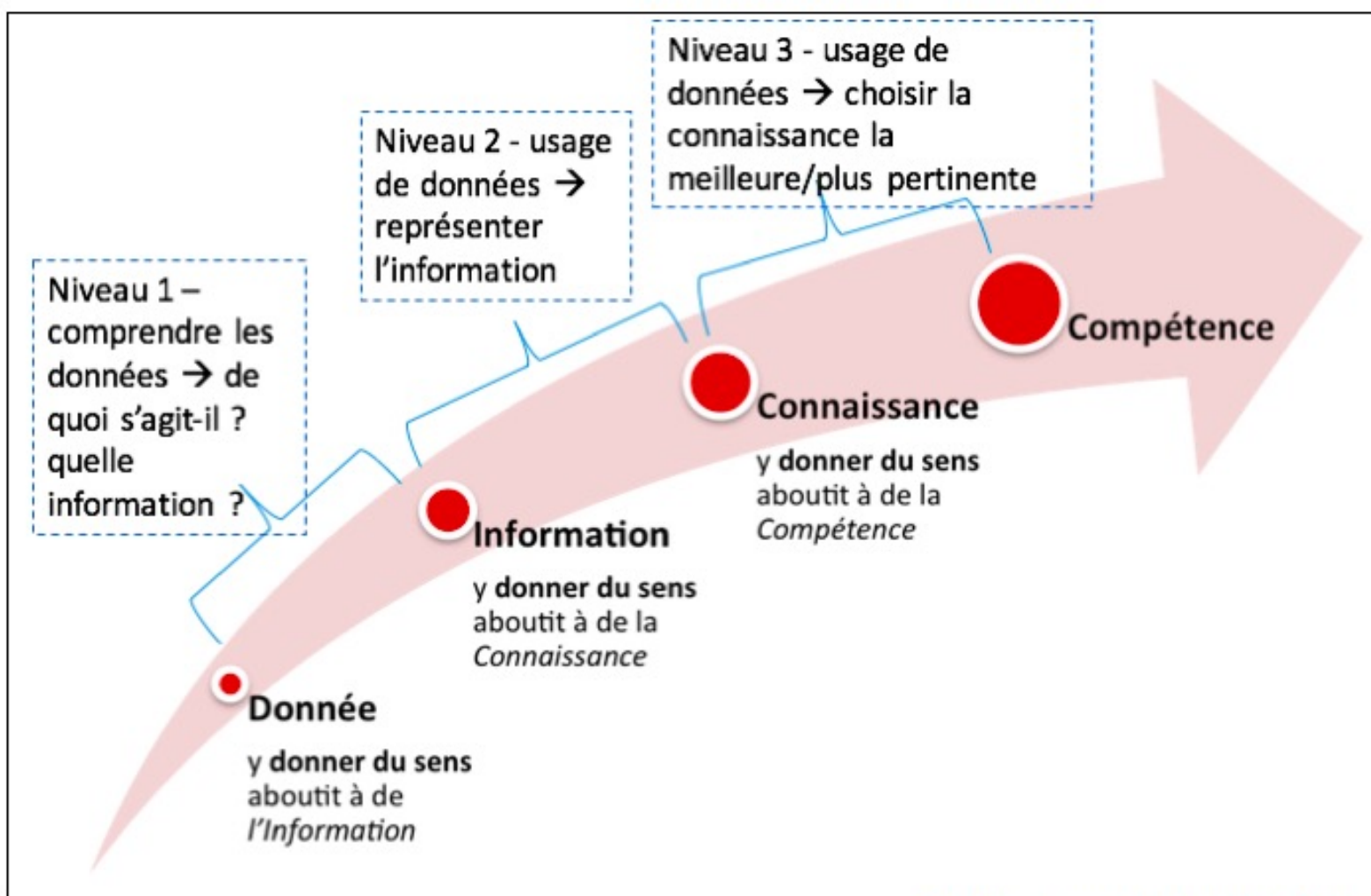
L'ÈRE DE LA DATA



SOURCE :
La boîte à outils de la Stratégie big data.

Auteur(s): Coron, Clotilde Editeur: Dunod
Année de Publication: 2020

04



LE SENS DE LA DONNÉE

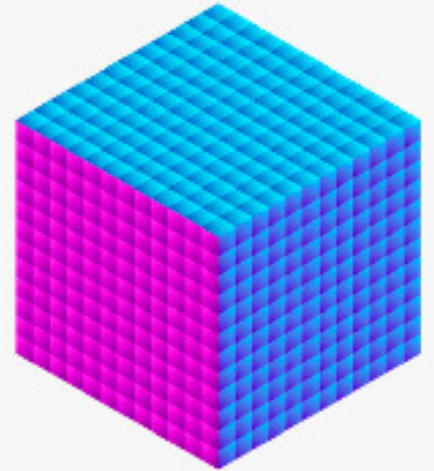
BIG DATA



▶ ANALYTICS



▶ **DECISIONS**



LE V DU SENS DE LA DONNÉE

Valeur

Quel potentiel nous accorde cette donnée?

- Combien peut-on revendre cette donnée ?

Véracité

Quelle précision nous accorde cette donnée?

- Quel outil mesure cette donnée?

Viabilité

Quel avenir nous accorde cette donnée?

- Quel acteur génère cette donnée?

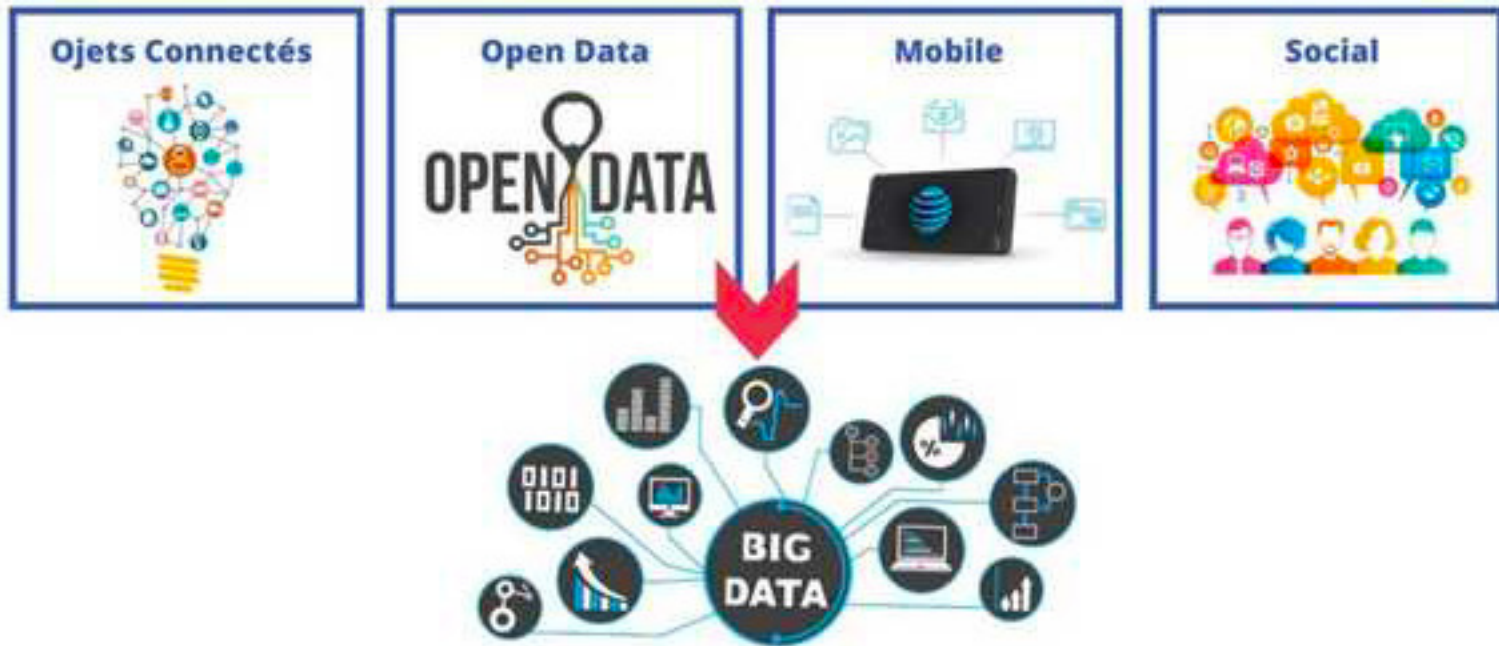
Validité

Quel processus de collecte a subi cette donnée?

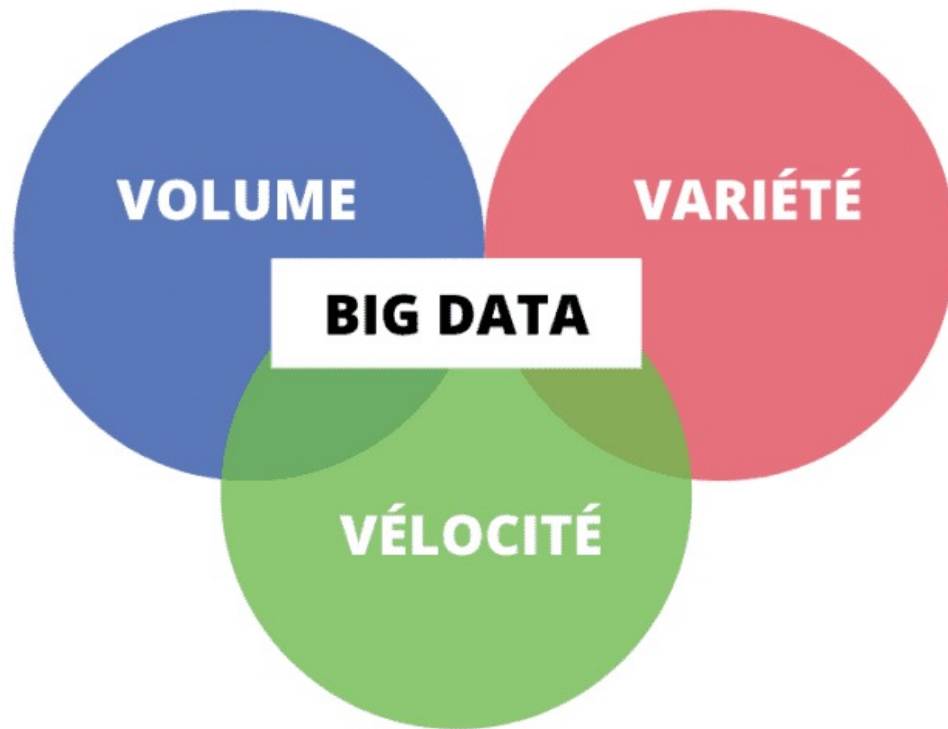
- Comment est générée cette donnée?

Illustration tirée du livre Boite à Outils de la Stratégie Big Data de Romain Rissoan & Romain Jouin

Les nouvelles sources de données



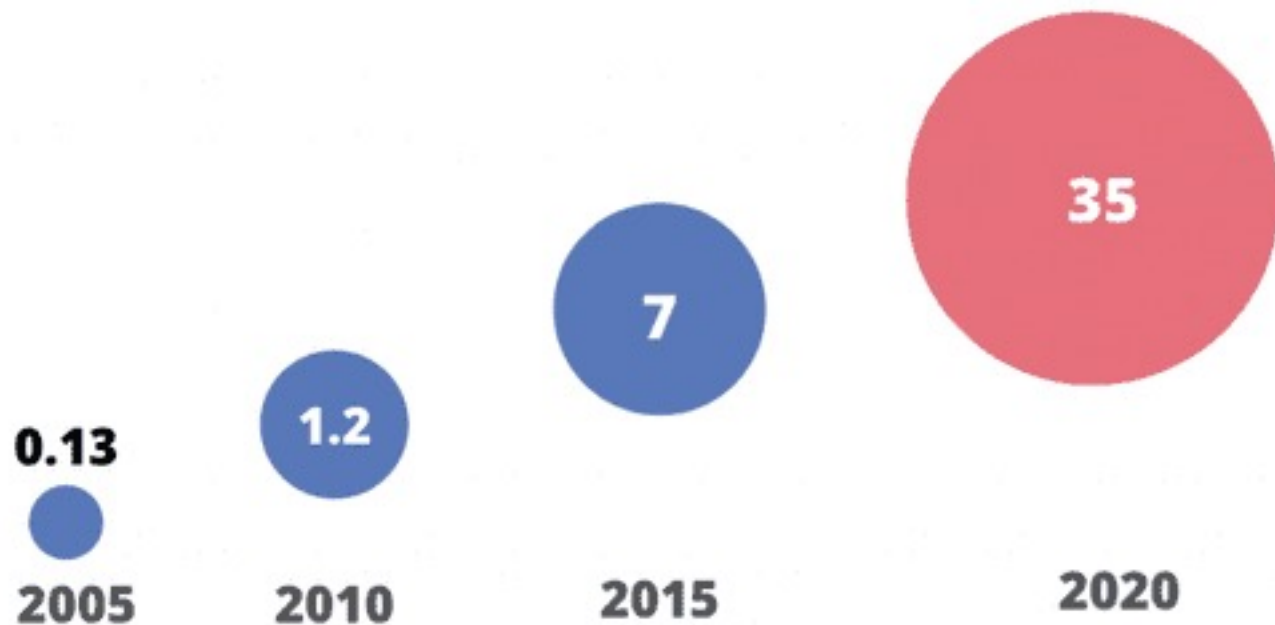
BIG DATA: LES 3Vs



Volume

Le volume mondial de données créées et répliquées

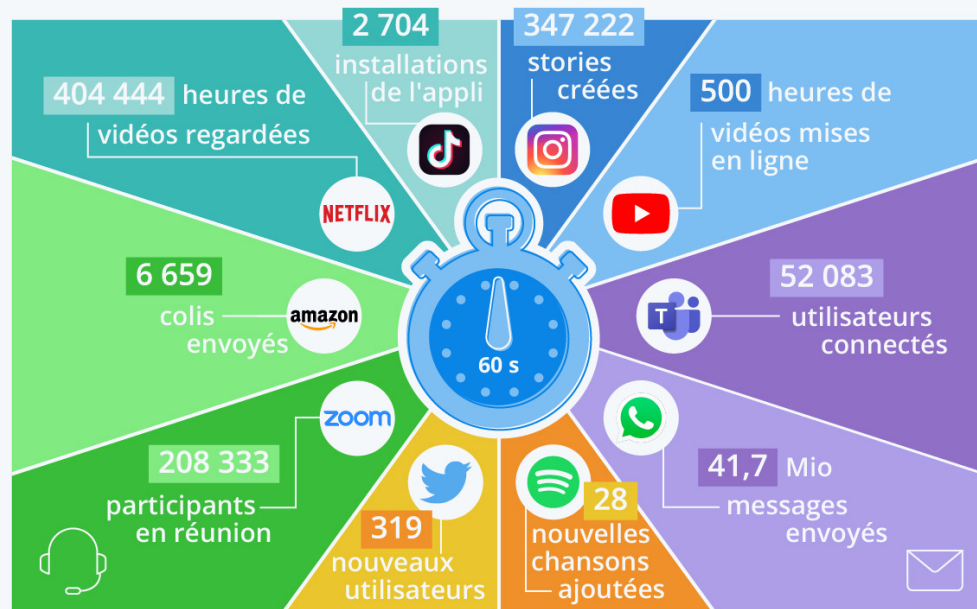
En ZettaOctet (1 ZettaOctet = 1 Trillion de GigaOctets)



Vélocité

Une minute sur Internet en 2020

Estimation de l'activité et des données générées sur Internet en l'espace d'une minute



Source : Visual Capitalist



Variété

BIG DATA - Les Types de données

Web et Rseaux sociaux

Contenu Web
File des réseaux sociaux
(Facebook, Twitter, Instagram...)
Données Clickstream

Biométrie

Reconnaissance Faciale
Génétique

Machine à Machine

RFID
Signaux GPS
Météo

IOT

Transaction de grandes données

Données géospatiales
Télécommunications
Données Bancaires

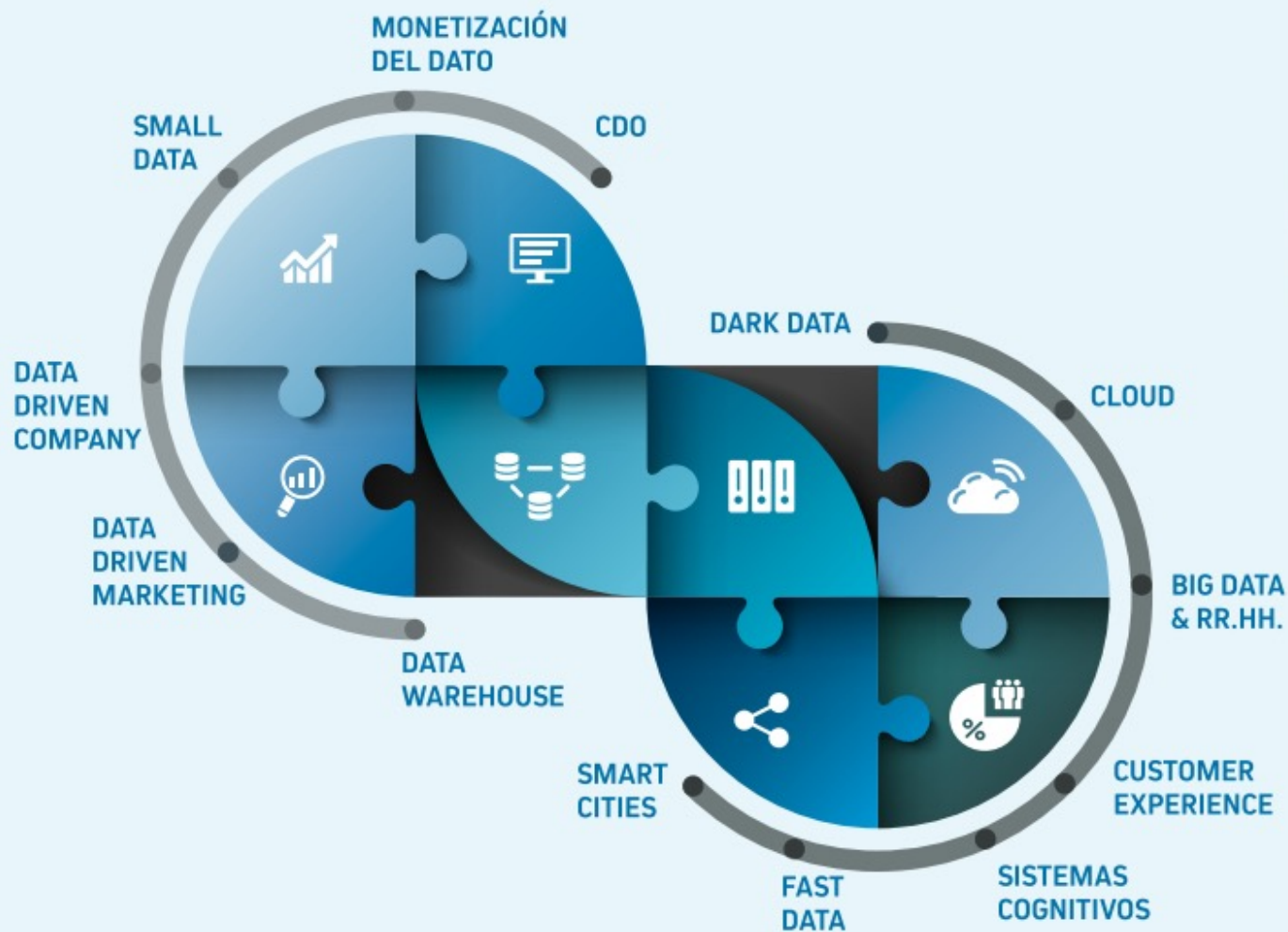
Humain

Centre d'appel
Email
Enregistrements médicaux
Electronique

05



Auteur(s): Coron, Clotilde Editeur: Dunod
Année de Publication: 2020



INSCRÍBETE
GRATIS!



Plazas
limitadas!

Les métiers du big data

BIG DATA - Les Types de données

Web et Rseaux sociaux

Contenu Web
File des réseaux sociaux
(Facebook, Twitter, Instagram...)
Données Clickstream

Biométrie

Reconnaissance
Faciale
Génétique

Machine à Machine

RFID
Signaux GPS
Météo

IOT

Transaction de grandes données

Données géospatiales
Télécommunications
Données Bancaires

Humain

Centre d'appel
Email
Enregistrements médicaux
Electronique

LES MÉTIERS DU BIG DATA



Les métiers techs

- Data Manager
- Data Visualizer
- Architecte BDD / Infrastructures
- Ingénieur Réseaux
- Expert HADOOP,...
- Développeur BI / Big Data
- ...



Les métiers business :

- Chief Data Officer
- Data marketing
- Data Management Platform
- Data Scientist
- Data Analytics
- Consultant Data
- ...

**Tendance forte : multiplication des métiers
liés à l'exploitation business**

Big data vs Business Intelligence

Périmètre	Business Intelligence	Big data
Périmètre d'analyse des données Quelles données analysons-nous ?	Sous-ensemble de données Ex. : un échantillon de la base clients	Toutes les données et métadonnées Ex. : l'intégralité de la base client croisée avec la base Google Maps
Fraîcheur des données Quand ont-elles été produites ?	Données passées Ex. : un export de la base clients de la semaine dernière	Proche du temps réel Ex. : les données en direct sur le système
Structure des données Comment sont-elles organisées dans le système ?	Données structurées Ex. : nom, prénom, e-mail	Données structurées et non structurées Ex. : nom, prénom, e-mail, description de son profil client, commentaires du client

Origine et utilisation de la donnée	Une source ou une seule application Ex. : un système informatique pour les commerciaux	Multi source et multi application Ex. : un système informatique croisé avec la base Google Map pour servir la base de données des recruteurs
Mouvement de la donnée	Statique Ex. : une base clients	Dynamique Ex. : les coordonnées GPS en temps réel des clients dans un magasin
Propriété de la donnée	Données privées Ex. : une base clients	Données privées et publiques Ex. : une base clients couplée à la base de données open data des entreprises
Perception de la croissance de la donnée	Croissance de la donnée gênante Ex. : les coordonnées GPS client en temps réel ne sont pas gérables sur Excel	Donnée comme nouvelle source d'avantage concurrentiel Ex. : les coordonnées GPS client en temps réel permettent de savoir qu'à cet instant si on lance une promotion sur ce produit, cela devrait marcher



LES TECHNIQUES DU BIG DATA

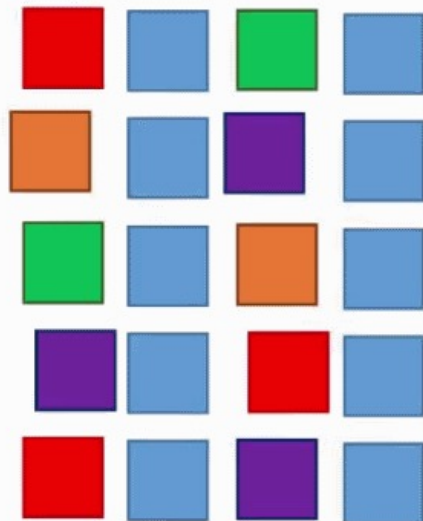
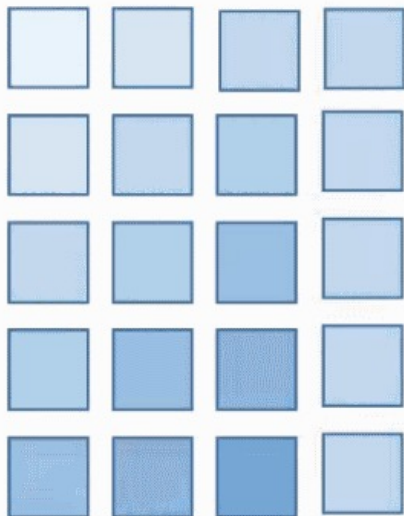
06

SOURCE :
La boîte à outils de la Stratégie big data.

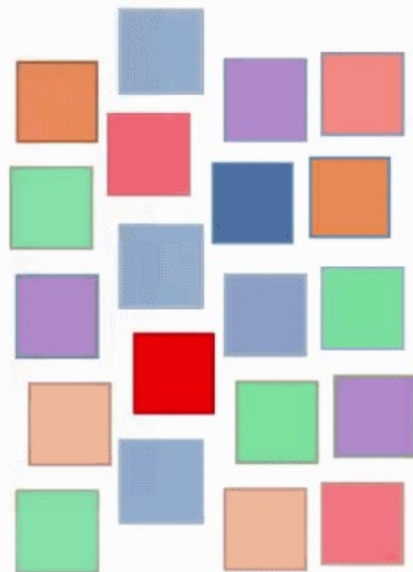
Auteur(s): Coron, Clotilde Editeur: Dunod
Année de Publication: 2020

Semi-Structured Data

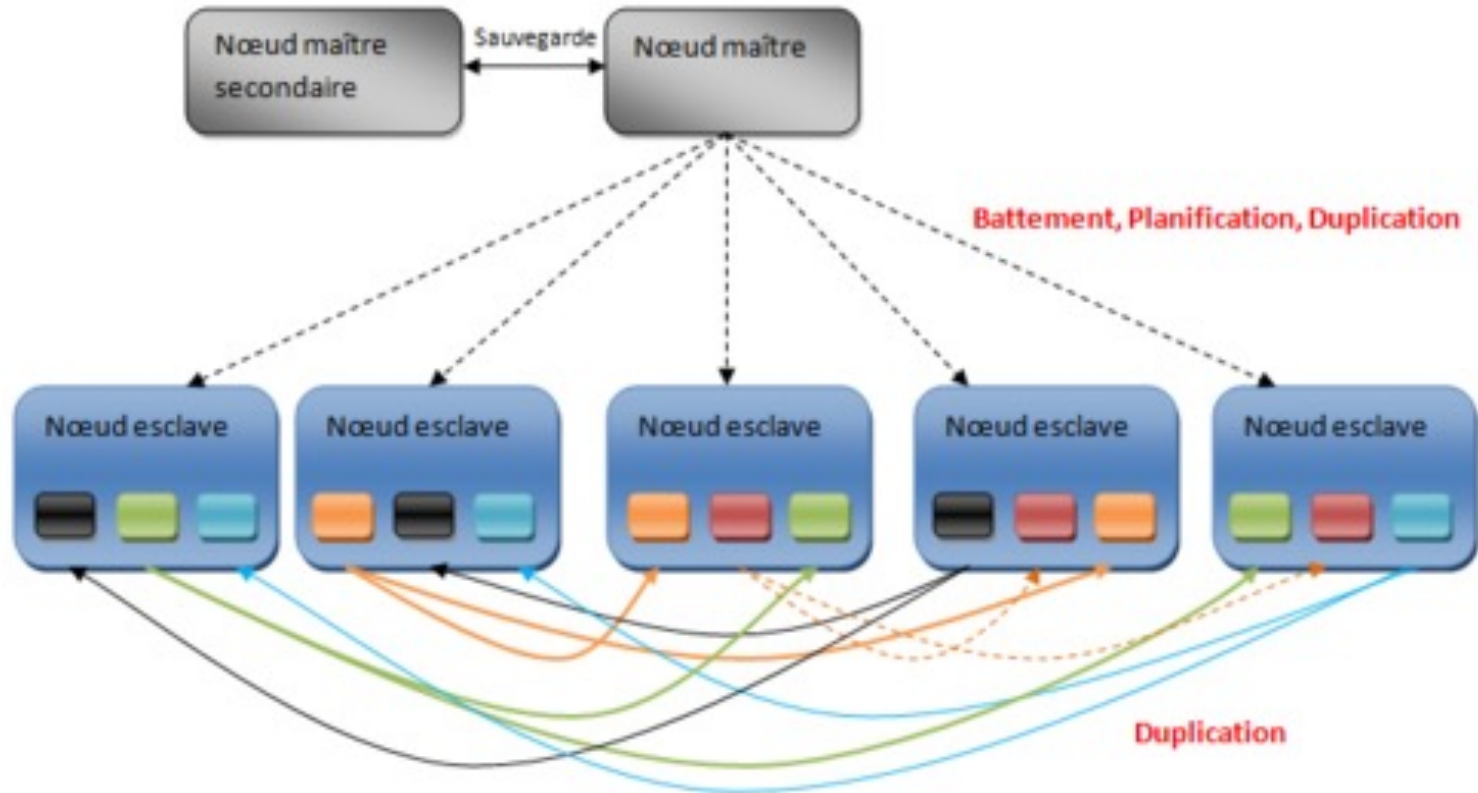
Structured Data



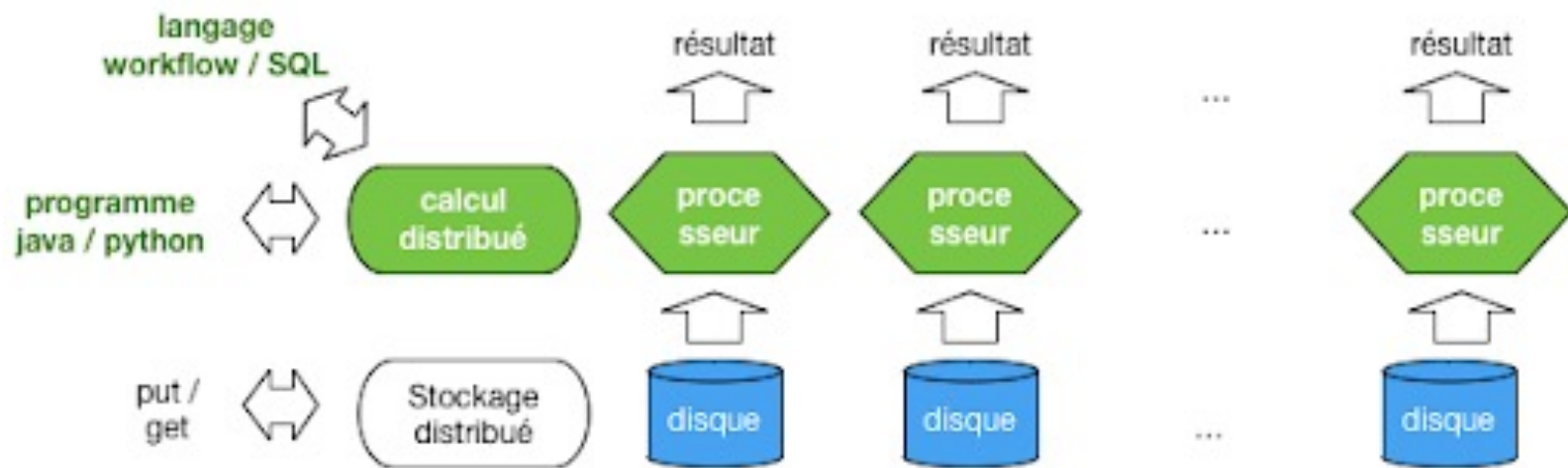
Unstructured Data



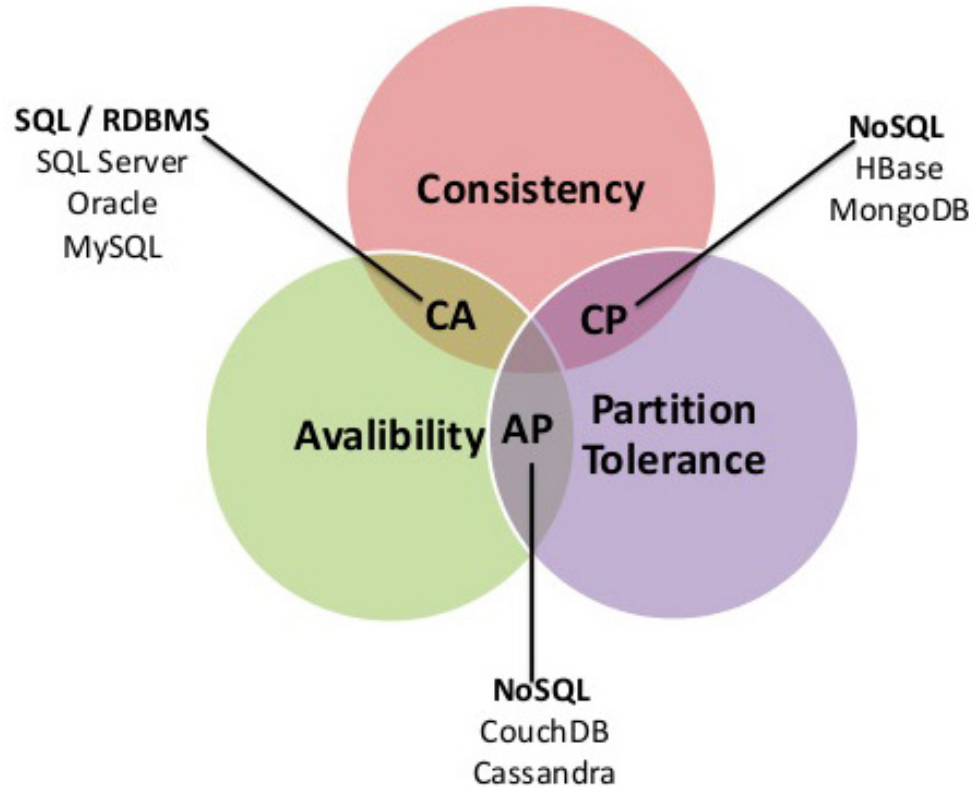
Stockage distribué



Calcul distribué



Théorème de CAP



Python

07



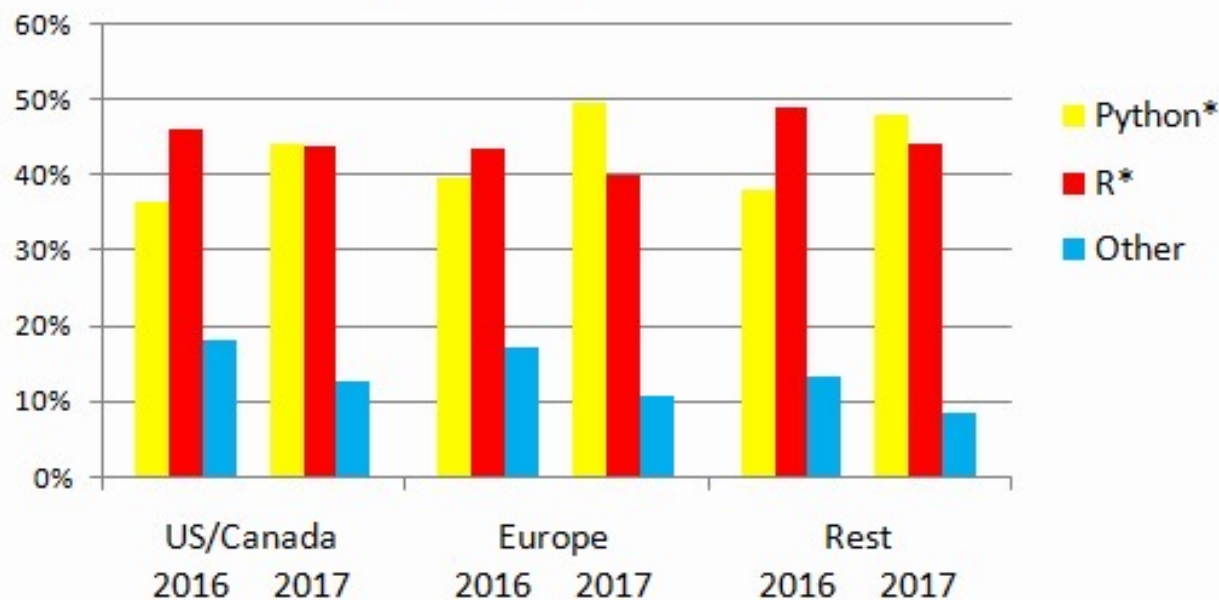
SOURCE :
La boîte à outils de la Stratégie big data.

Auteur(s): Coron, Clotilde Editeur: Dunod
Année de Publication: 2020

Grilles de Calcul



R vs Python vs Rest by Region, 2016 vs 2017



Numpy

Package de base pour les calculs matriciels

Scipy

Calculs pré-codés de routines mathématiques

Stats models

Statistique descriptives, tests, séries temporelles

Paandas

Manipulation de DataFrame

Scikit Learn

Machine Learning: classification / Regression, SVM, Random Forest, Knn, etc.

Tensor Flow

Réseau de neurones, réseaux convolutionnels

Matplotlib

Librairie de visualisation de données (tracé de courbes)

Seaborn

Librairie de visualisation de donnée plus jolie et simple que matplotlib

Plotly

Librairie de visualisation de donnée interactive sur le Web

Bokeh JS

Librairie de visualisation de donnée interactive sur le Web

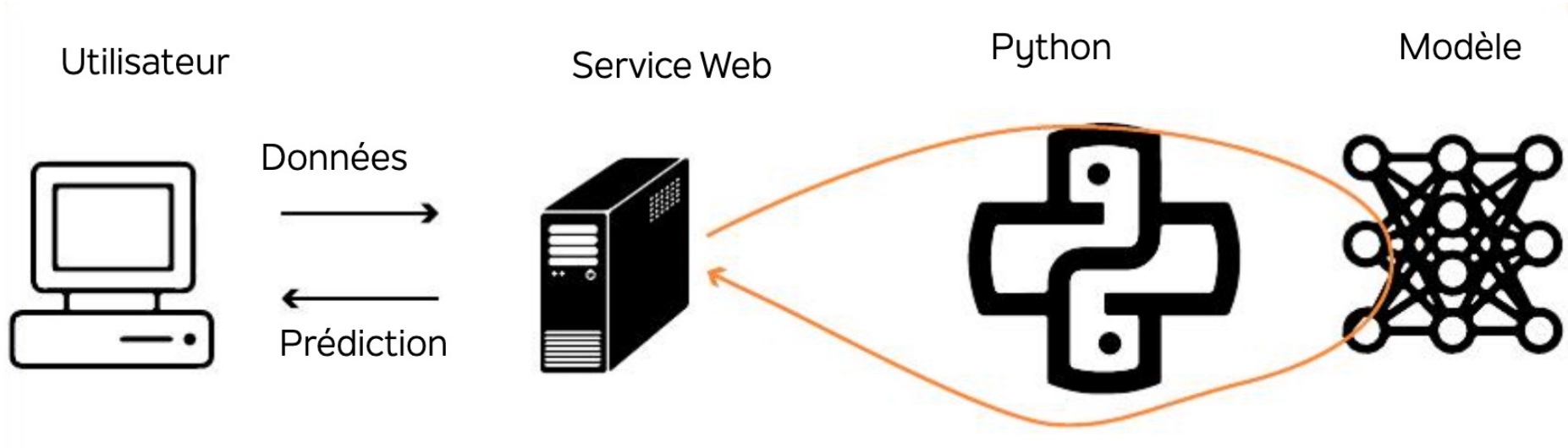
Django

Module pour créer des sites Web

Dask

Librairie de parallélisations de calculs

Schéma d'une API prédictive avec Python



MERCI !

Dr. FIRAS
Docteur.firass@gmail.com
<https://www.dr-firas.com/>
=>Lien de la formation en ligne<=

