

Introduction au Big Data

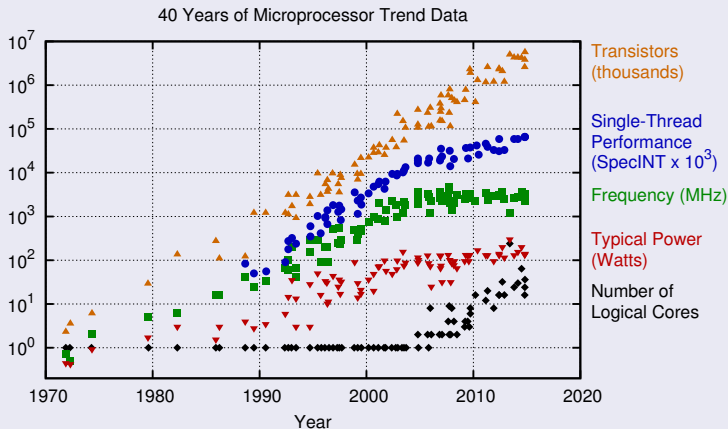
Jonathan Lejeune

Sorbonne Université/LIP6-INRIA

DataCloud – Master 2 SAR 2021/2022

Évolution des ressources informatiques

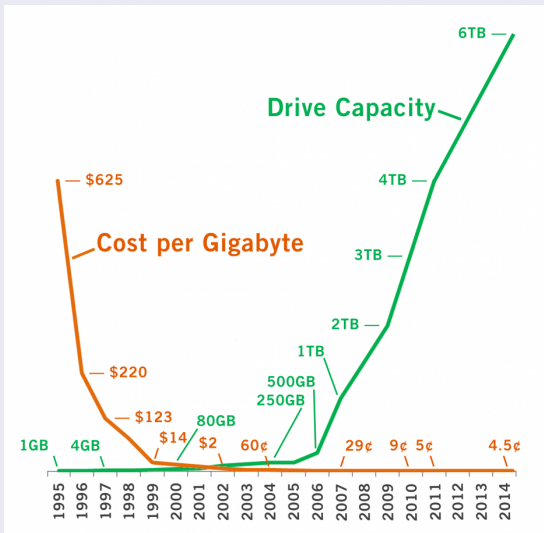
Processeurs



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

Évolution des ressources informatiques

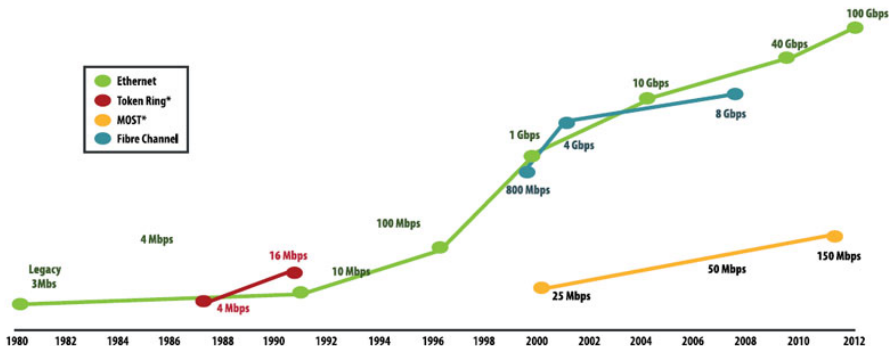
Stockage



Évolution des ressources informatiques

Performances des réseaux

Evolution of Network Bandwidth



*Shared network architecture

© 2013 Broadcom Corporation. All rights reserved

Une évolution exponentielle

En résumé

Des ressources informatiques :

- de + en + performantes
- de - en - chères

Conséquences

- de plus en plus de terminaux :
PC, smartphone, tablettes, objets connectés ...
- de plus en plus d'utilisateurs
- de plus en plus d'applications en ligne :
réseaux sociaux, commerces en ligne, partage de contenu, multimédia ...
De plus en plus de données produites chaque jour

Les ordres de grandeur

Nom	Puissance	Facteur de comparaison
octet (o)	10^0 octets	grain de riz
Kilo-octet (Ko)	10^3 octets	bol de riz
Megaoctet (Mo)	10^6 octets	8 sacs de riz
Gigaoctet (Go)	10^9 octets	3 semi-remorques de riz
Téraoctet (To)	10^{12} octets	2 porte-containers de riz
Pétaoctet (Po)	10^{15} octets	Manhattan couverte de riz
Exaoctet (Eo)	10^{18} octets	Cote ouest des USA couverte de riz
Zéttaoctet (Zo)	10^{21} octets	Océan Pacifique rempli de riz
Yottaoctet (Yo)	10^{24} octets	Remplir le volume de la terre de riz

Tous les jours nous produisons plusieurs Exa-Octets de données

(sources : IBM, <https://www.internetlivestats.com/>)

Définition(s) du Big Data

La définition de Wikipédia

"Désignent 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."

Une deuxième définition

"La notion de big data est un concept s'étant popularisé dès 2012 pour traduire le fait que les entreprises sont confrontées à des volumes de données (data) à traiter de plus en plus considérables et présentant de forts enjeux commerciaux et marketing."

Une troisième définition

"Le Big Data est l'ambition de tirer un avantage économique de l'analyse quantitative des données internes et externes de l'entreprise."

Les caractéristiques du Big Data

Les 3 V fondamentaux

- **Volume** : traiter et stocker une grande masse de données
- **Vélocité** : production massive de données en peu de temps
- **Variété** : données structurées/non structurées sur des formats variés

Un 4ème V éventuel

- **Véracité** : assurer l'intégrité des données (obsolescence, justesse ...)

Voire un 5ème...

- **Valeur** : évaluer les données à leur juste valeur pour qu'elles soient rentables


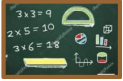



Les données sont précieuses pour les entreprises

- Meilleure prise de décision
- Amélioration des processus opérationnels
- Amélioration de la connaissance client
- Réduction des coûts

Problématique principale

Comment traiter/analyser efficacement des données massives ?

Les métiers du Big data

	Management 	Mathématiques /statistiques 	Informatique IA 	Informatique Infrastructure 	Droit digital 
Data miner (50 k€)	●	●			
Chief data officer (50 k€)	●				
Business Intelligence Manager (42 k€)	●		●		
Ingénieur Big Data (jusqu'à 60k€)		●	●	●	
Data analyst (36 k€)		●	●		
Data Scientist (55 k€, 120k€ aux US)		●		●	
Data Protection Officer (40 k€)				●	●
Architecte Big Data (40 k€)				●	

Vue globale d'un système Big Data

