



Institut Supérieur d'Informatique
et de Multimédia de Gabes

Introduction au BigData

Dr: Tasnim ABAR

tasnim,abar@isimg.tn

BigData

- Explosion des volumes des données générées sur le web, web mobile...
- Plus de 90% de toutes les données dans le monde ont été créées au cours des 2 dernières années.
- On prévoit que d'ici 2035, la quantité d'informations numériques existantes sera passée de 45 zettaoctets aujourd'hui à 2000 zettaoctets
- Chaque minute, nous envoyons 204 millions d'e-mails, générons 1,8 million de likes Facebook, envoyons 278 000 tweets et téléchargeons 200 000 photos sur Facebook.
- Google traite en moyenne plus de 40 000 requêtes de recherche par seconde, ce qui en fait plus de 3,5 milliards en une seule journée.
- Environ 100 heures de vidéo sont téléchargées sur YouTube chaque minute et il vous faudrait environ 15 ans pour regarder chaque vidéo téléchargée par les utilisateurs en une journée.
- 570 nouveaux sites Web voient le jour chaque minute de chaque jour.
- La quantité de données transférées sur les réseaux mobiles a augmenté de 81% à 1,5 exaoctets (1,5 milliard de gigaoctets) par mois entre 2012 et 2014. La vidéo représente 53% de ce total.

Le BigData



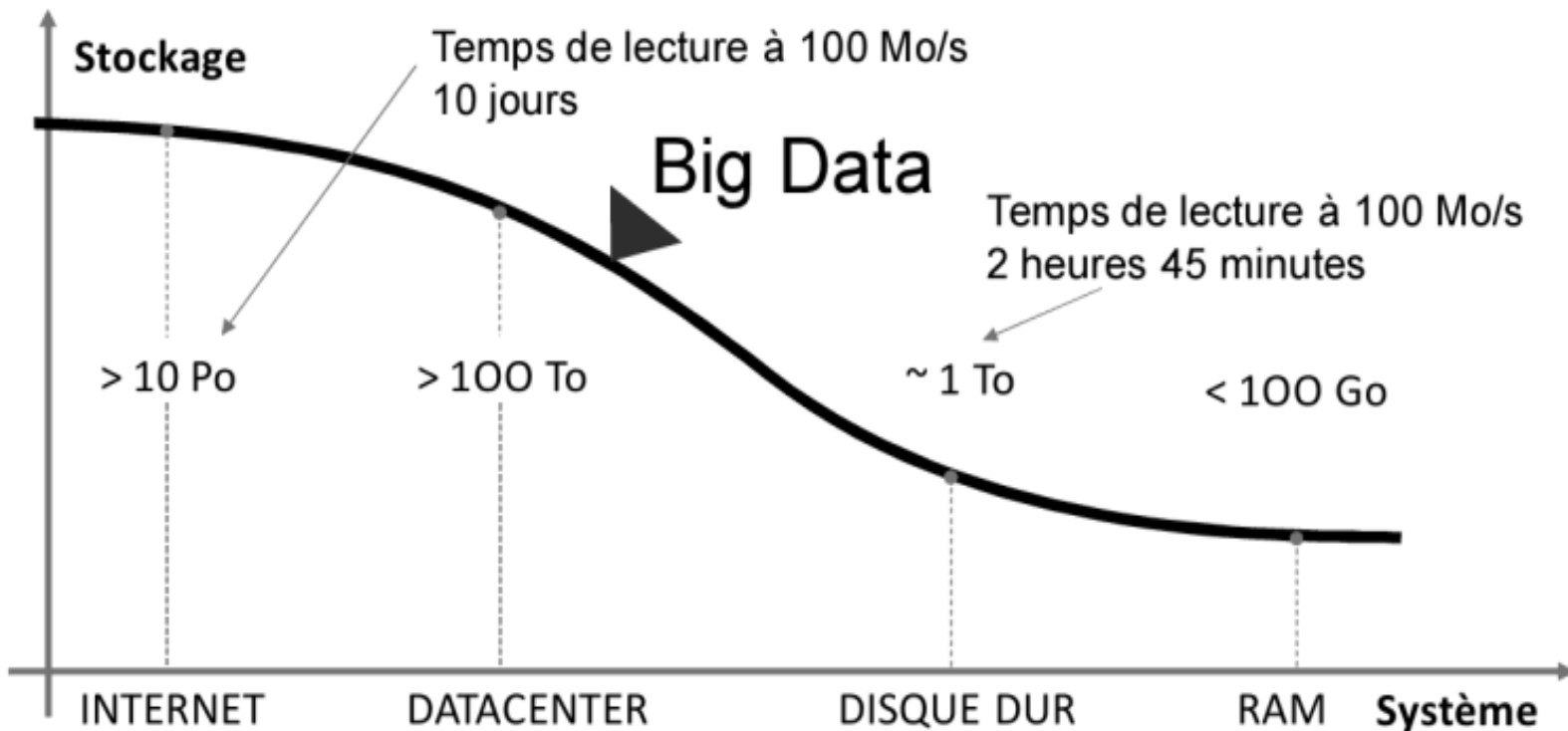
Data Torrent



Computing Anytime, Anywhere

Big Data

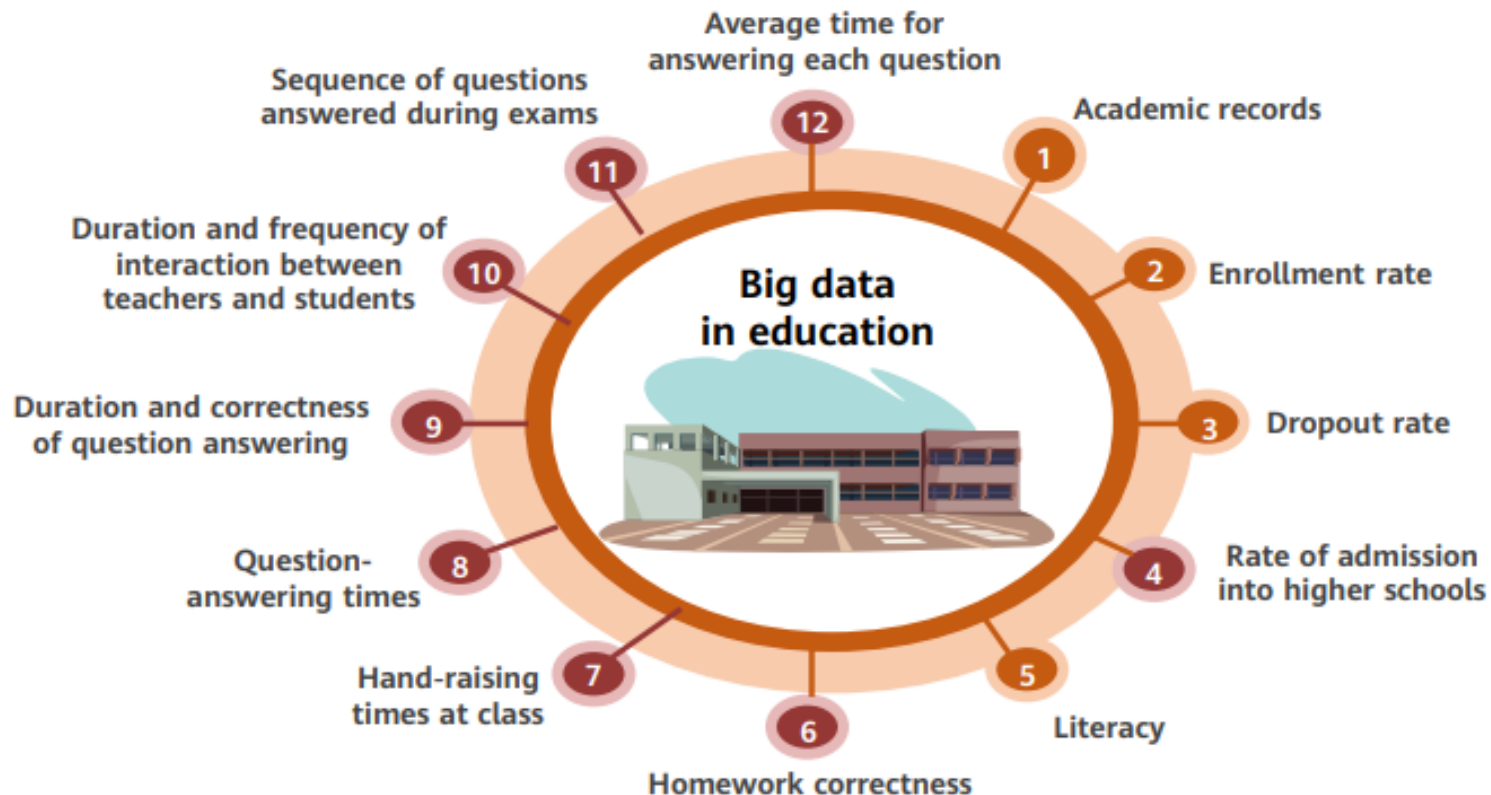
- On parle du BigData quand le traitement devient trop long et trop lourd pour un seul serveur



Big Data

- Les principaux acteurs du web tels que Google, Yahoo, Facebook, Twitter, LinkedIn, etc ont été les premiers à traiter des volumes de données extrêmement importants et ont été à l'origine des premières innovations dans ce domaine, principalement centrées sur deux types de technologies:
 - Les plateformes de développement et de traitement des données (Hadoop, Spark,...)
 - Les bases de données (NoSql)

Exemples d'utilisation de BigData



Exemples d'utilisation de BigData

Traffic planning: multi-dimensional analysis of crowds

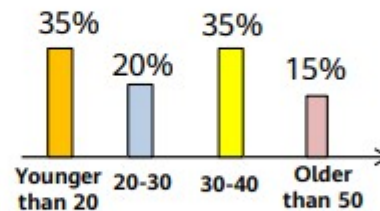


Traffic prediction based on crowd analysis

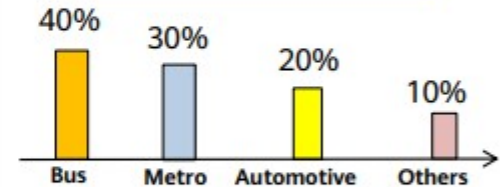
Areas where the people flow ever exceeded the specified threshold

- North gate of the Workers' Stadium: > 500 people/ hour
- Sanlitun: > 800 people/hour
- Beijing Workers' Stadium: > 800 people/hour

Analysis by crowd



Analysis by transportation method



Road network planning



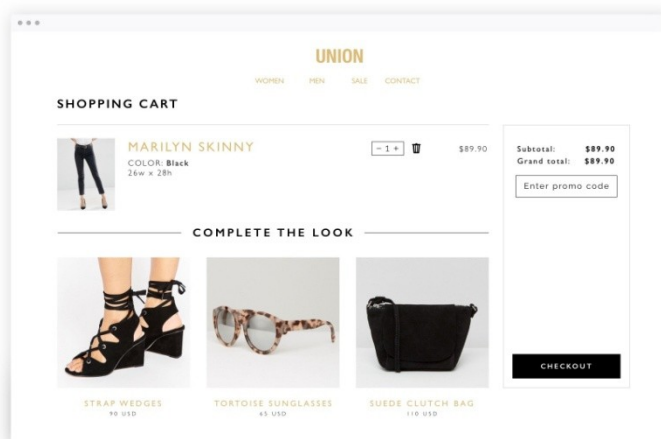
Bus network planning



Exemples d'utilisation de BigData

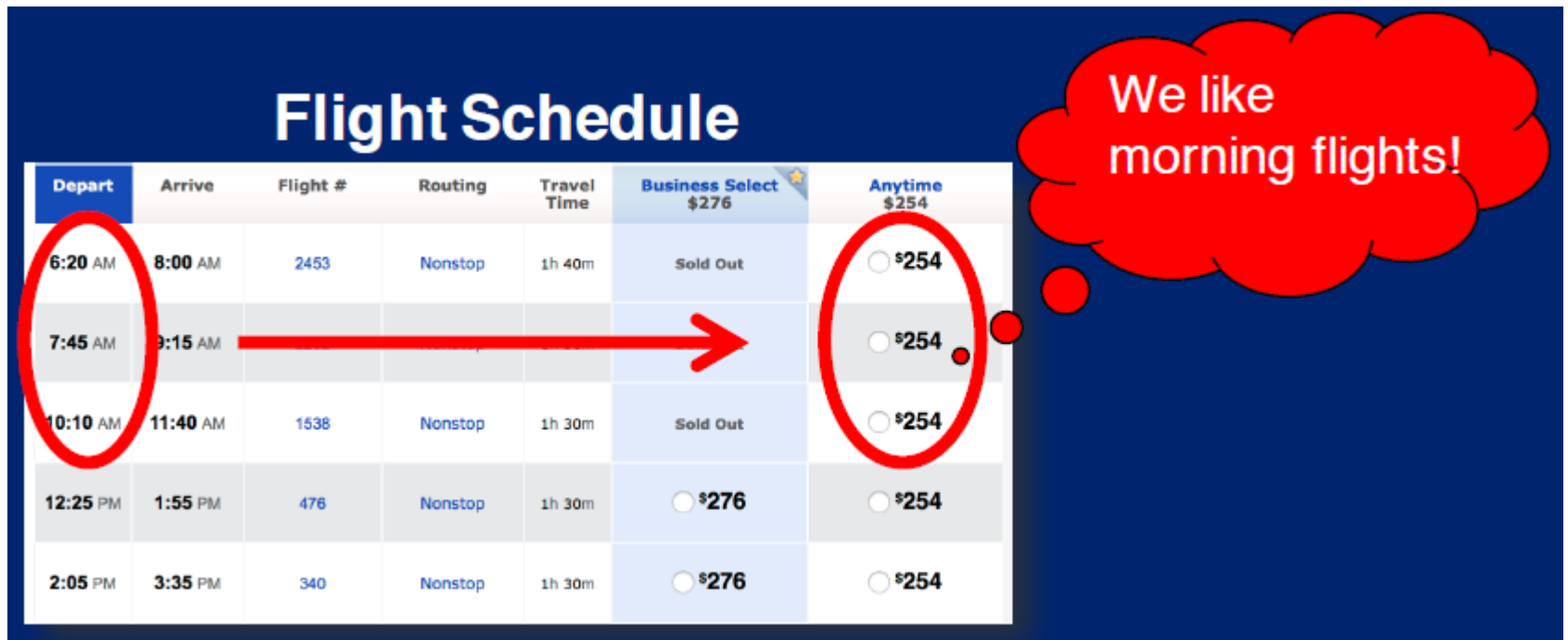
Moteurs de recommandation

Customers Who Bought This Item Also Bought



Exemples d'utilisation de BigData

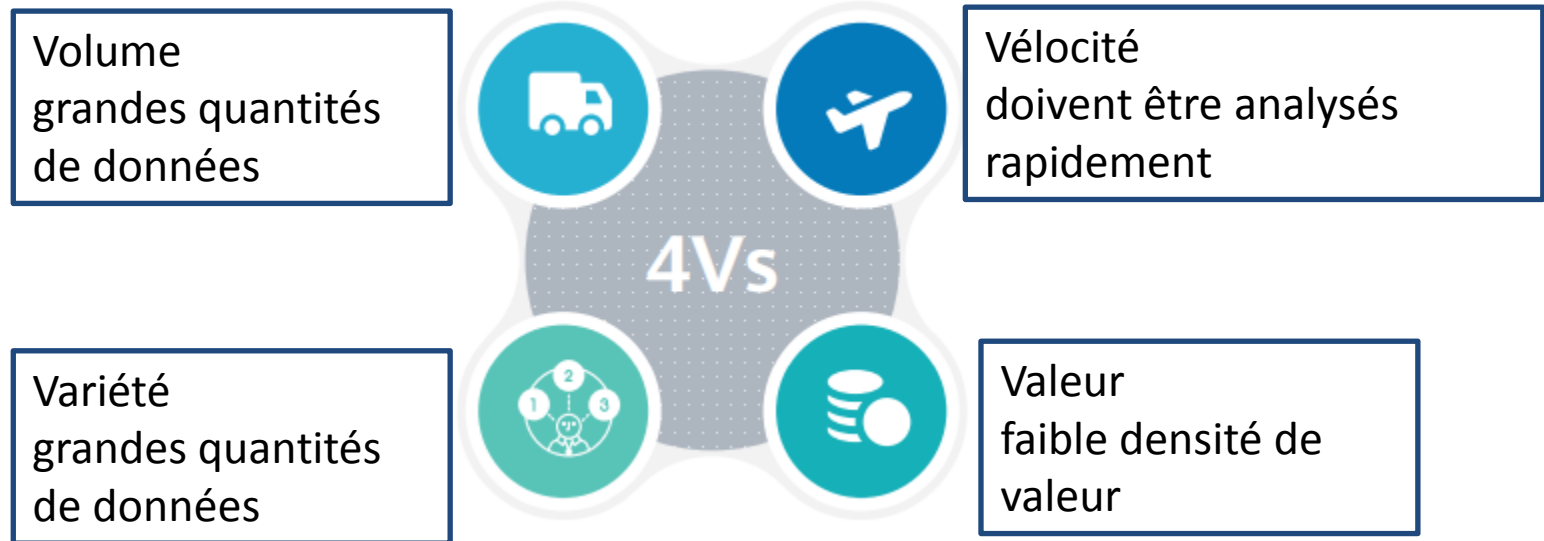
La croissance des consommateurs pour guider la croissance des produits



The image shows a flight schedule table with several annotations. A red circle highlights the departure times 6:20 AM and 7:45 AM in the first column. A red arrow points from the 7:45 AM row to the 'Anytime \$254' column. Another red circle highlights the '\$254' prices in the 'Anytime' column for the first three rows. A red thought bubble on the right contains the text 'We like morning flights!'.

Depart	Arrive	Flight #	Routing	Travel Time	Business Select \$276	Anytime \$254
6:20 AM	8:00 AM	2453	Nonstop	1h 40m	Sold Out	<input type="radio"/> \$254
7:45 AM	9:15 AM					<input type="radio"/> \$254
10:10 AM	11:40 AM	1538	Nonstop	1h 30m	Sold Out	<input type="radio"/> \$254
12:25 PM	1:55 PM	476	Nonstop	1h 30m	<input type="radio"/> \$276	<input type="radio"/> \$254
2:05 PM	3:35 PM	340	Nonstop	1h 30m	<input type="radio"/> \$276	<input type="radio"/> \$254

Caractéristiques du Big Data



Caractéristiques du Big Data: Volume

Volume == taille

Chaque minute



204 Million emails



200,000 photos

1.8 Million 

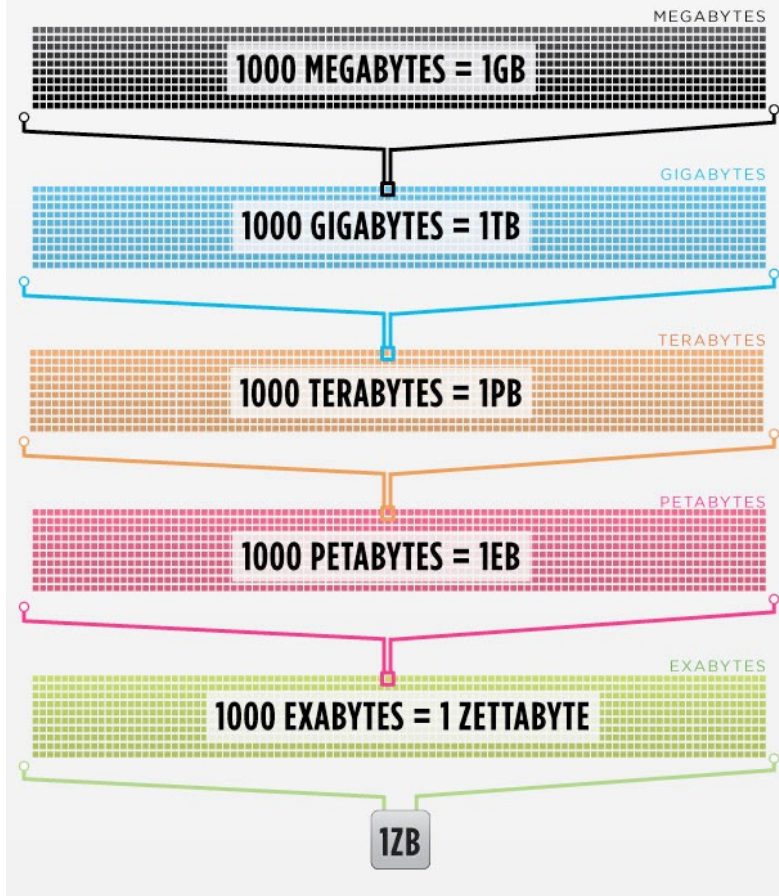


1.3 Million video views

72 hours of video uploads

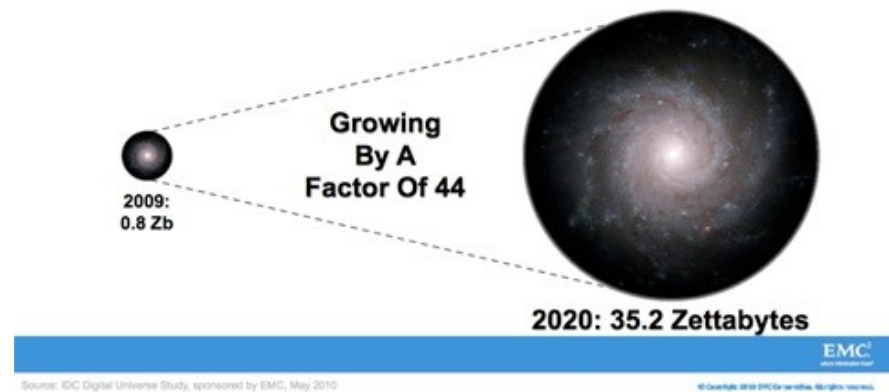
Caractéristiques du Big Data: Volume

But how much data are we talking about?



1 TB \approx 300 heures de vidéo de bonne qualité

The Digital Universe 2009-2020



Caractéristiques du Big Data: Variété

Variété == hétérogénéité & complexité

Avant: Les données ont été limitées
uniquement aux tableaux

Cars marketplace				
vendor	Model	Price	Mileage	VIN Code
Chevrolet	Corvette	17226	25965.0	ILLAKAWAZDZ
Chevrolet	Corvette	34229	46429.0	RCPNSRYGXOI
Chevrolet	Corvette	27982	50209.0	NWLGCVEHGI
Chevrolet	Corvette	51825	72998.0	NGVZSCIZGSM
Chevrolet	Corvette	52845	34364.0	PSDRUYYOJIG
Chevrolet	Malibu	37874	37273.0	VLFPQPWNEFD
Chevrolet	Malibu	15600	71441.0	EXLJGDWOZSA
Chevrolet	Malibu	52447	46700.0	NLMGJZAKBRD
Chevrolet	Malibu	27129	36254.0	OIPFUENLEHSX
Chevrolet	Malibu	28846	77162.0	WRCOOFREZLI
Chevrolet	Malibu	46165	60590.0	HUFTTHQHSFJF
Chevrolet	Malibu	18263	37790.0	JLMHNAFESHVD

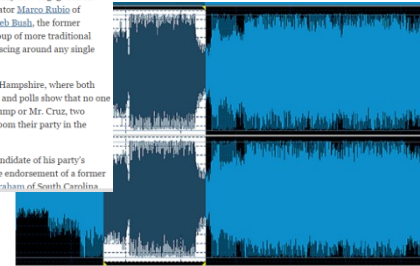
Aujourd'hui: Aujourd'hui, les données
sont plus hétérogènes

The battle for the Republican nomination appeared more splintered than ever between two halves of a bitterly divided party as several candidates scrambled Friday to consolidate the support of more moderate conservatives a day after a raucous debate.

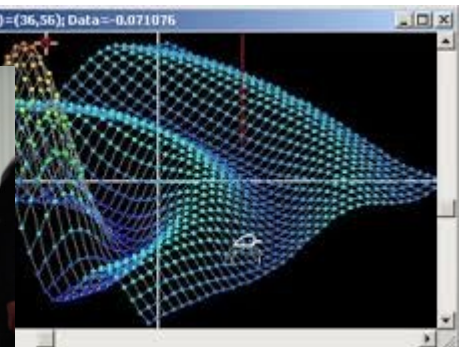
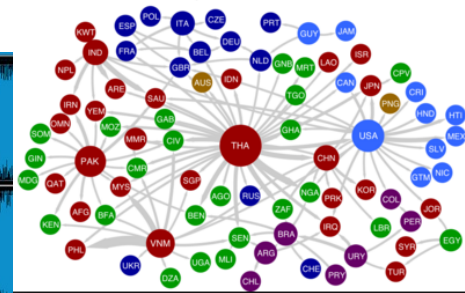
With Donald J. Trump and Senator Ted Cruz finally now engaged in an open feud for the most disillusioned voters, Senator Marco Rubio of Florida, Gov. Chris Christie of New Jersey and Jeb Bush, the former Florida governor, were battling to win over a group of more traditional Republicans who are showing little sign of coalescing around any single candidate.

This fracture was most vividly apparent in New Hampshire, where both Mr. Bush and Mr. Rubio campaigned on Friday, and polls show that no one is emerging as the obvious alternative to Mr. Trump or Mr. Cruz, two candidates that many Republicans fear would doom their party in the general election.

Mr. Bush sought to highlight his image as the candidate of his party's seasoned, sober-minded wing on Friday with the endorsement of a former rival in the presidential race, Senator Lindsey Graham of South Carolina.



Rice Trade Network, 2009



Caractéristiques du Big Data: Variété

Variété au sein d'un type

Penser à une collection de email

Pièce semblable à une table

from: Banikumar Maiti (GMAIL) <banikumar.maiti@gmail.com>
to: Reghu Rajan <reghurajan@gmail.com>
cc: Amarnath Gupta <aguptasd@gmail.com>
date: Tue, Feb 2, 2016 at 2:29 PM
subject: Re: Connecting
mailed-by: gmail.com
signed-by: gmail.com

Expéditeur, destinataire, date... Bien structuré

Texte non structuré Dear All,

I would like to congratulate you for putting together a wonderful show.
It was only possible by your hard work.

Dreaming of an UNIQUE show. This credit goes to Zubair. You dreamed about it and made it happen.

Caractéristiques du Big Data: Variété

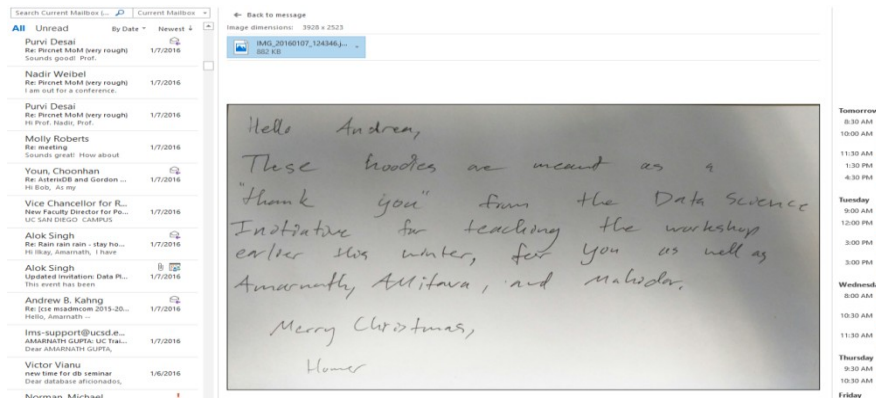
Variété au sein d'un type

Penser à une collection de email

Expéditeur, destinataire, date... Bien structuré

Corps de l'e-mail Texte

Media



Qui envoie à qui Network

Temps réel? Disponibilité

Caractéristiques du Big Data: Variété

Impact de la variété des données

- Plus difficile à ingérer
- Difficile de créer un stockage commun
- Difficile de comparer et de faire correspondre les données d'une variété
- Défis de gestion

Caractéristiques du Big Data: Vitesse

Vélocité == Vitesse

$$V = \frac{\Delta x}{\Delta t}$$

- Rapidité de création des données
- Rapidité de stockage des données
- Rapidité d'analyse des données

Big Data



Real-time action



Traitement BigData VS. Traitement des données traditionnelles

	Traitement BigData	Traitement des données traditionnelles
Echelle de données	Large (GB, TB, PB)	Petit (MB)
Type de données	Plusieurs type de données (structuré, semi-structuré, non structuré)	Un seul type de données (structuré)
Objet à traiter	poisson dans l'océan	poisson dans l'étang
outil de traitement	aucune taille ne convient à tous	taille unique

Principaux modes de calcul de BigData

- Batch Computing (par lot)
 - permet de traiter une grande quantité de données par lots: MapReduce et Spark



- Stream Computing
 - Vous permet de calculer et de traiter les données de flux en temps réel: Spark, Storm, Flink, Flume et Dstream

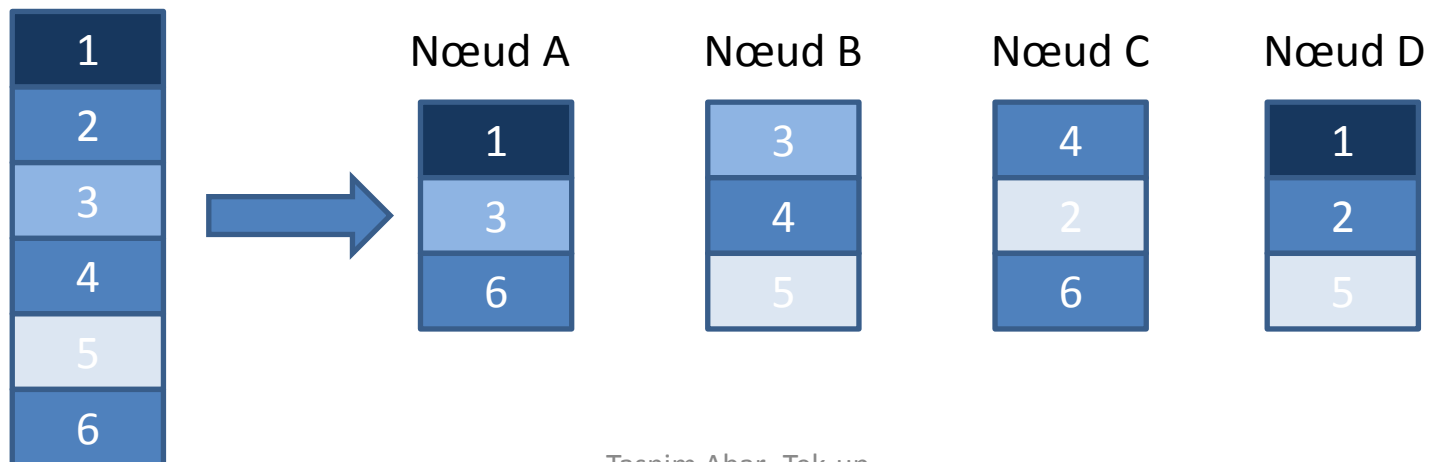


Principaux modes de calcul de BigData

- Informatique graphique
 - permet de traiter de gros volumes de données de structure graphique: GraphX, Gelly, Giraph, et PowerGraph, Neo4J
- Requête et calcul analytique
 - permet de gérer, d'interroger et d'analyser une grande quantité de données stockées: Hive, Impala, et Cassandra

BIG DATA: Généralités

- La plupart des outils et des frameworks de BigData sont construits en gardant à l'esprit les caractéristiques suivantes:
 - La distribution des données: Les données sont distribuées sous forme de bloc (chunks) sur plusieurs noeuds pour un traitement en parallèle.
=> Système de Fichiers Distribués - DFS (DistributedFileSystem).



BIG DATA: Généralités

- La tolérance aux pannes: En général, on fait la réplique d'un seul bloc (ou chunk) de données plusieurs fois sur des serveurs distants. Même si l'un des serveurs tombe en panne, on peut récupérer les données à partir d'un autre serveur ou d'un autre «datacenter».
L'inconvénient ici est que la réplication de données pourrait coûter beaucoup d'espace.
- Le traitement en parallèle: Puisque les données sont distribuées sur plusieurs serveurs. Alors ces serveurs travaillent en parallèle pour l'analyse et le traitement. Les données sont combinées pour obtenir le résultat final souhaité.. (Fameux exemple : Map Reduce de Google).

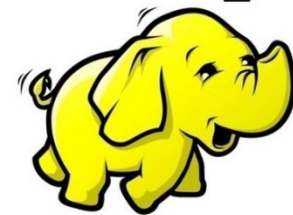
BIG DATA: Plateforme – Technologies -Outils

Société	Technologie développée	Type de technologie
Google	Map Reduce	Patron de traitement distribué et en parallèle
	Big Table	Système de base de données distribuée NoSQL propriétaire reposant sur GFS (Google File System). Technologie non Open Source, mais qui a inspiré Hbase qui est Open Source.
Facebook	Cassandra	Base de données de type NoSQL et distribuée.
	Hive	Outil d'analyse de données utilisant Hadoop.
Yahoo	Hadoop	Plateforme Java destinée aux applications distribuées et à la gestion intensive des données. Issue à l'origine de GFS et MapReduce.
	S4	Outil de développement dédié aux applications de traitement continu de flux de données.

BIG DATA: Plateforme – Technologies -Outils



hadoop



ØMQ

