BIG DATA

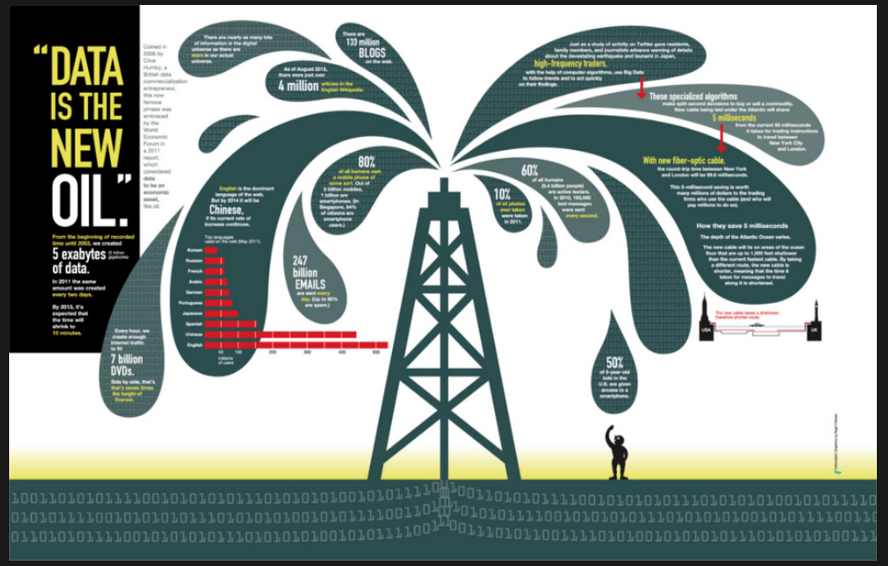
2 - Qu'est-ce que le Big Data ?

POLYTECH NANTES | Pôle Sciences et technologie Nantes Université

SOMMAIRE

LA THÉORIE

L'INFORMATION : UNE NOUVELLE RICHESSE À EXPLOITER



Information is the oil of the 21st century, and analytics is the combustion engine.

Peter Sondergaard - Gartner Senior VP

LA THÉORIE

L'INFORMATION : UNE NOUVELLE RICHESSE À EXPLOITER

There were 5 exabytes (=5000 petabytes) of information created between the dawn of civilisation through 2003, but that much information is new created every 2 days.

Eric Schmidt - PDG de Google - 2010

LA THÉORIE

LE BESOIN DE TRAITEMENT SUR DE FORTES VOLUMÉTRIES

- Internet scale
- Cheap computing
- Massive amount of information sources (social networks, sensors)
- Non-structured data
- Images / Videos

LA THÉORIE

ET POURQUOI PAS LES SGBD/DWH CLASSIQUES ?

- Not horizontally scalable
- Schema on-write
- Not easy to work with semi-structured data
- Normalization makes partitioning harder
- SQL is fantastic but restrictive
- Need for more specialized systems

LA THÉORIE

LES NOUVELLES SOURCES DES DONNÉES

- **Données d'entreprise**: Uniquement 30% des données d'entreprise sont exploitées
- **Internet Of Things**: 22 milliards dans le monde en 2020
- **Open Data**: Valeur estimée de plus de 3 milliards
- **Social Networks**: 3,2 milliards d'utilisateurs actifs soit 42% de la population mondiale

DÉFINITION

Le big data (litt. "grosses données" en anglais), les mégadonnées ou les 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.

Wikipédia

DÉFINITION

> *Le big data (litt. "grosses données" en anglais), les mégadonnées ou les 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.*

Wikipédia

Le Big Data vise à tirer un **avantage concurrentiel** au travers de méthodes de **collecte**, d'**analyse** et d'**exploitation** des données qu'on ne pouvait utiliser jusqu'à présent du fait des contraintes **économiques**, **fonctionnelles** et **techniques** liées aux **volumétries**, à la **vitesse** de traitement et à la **variété** des données à considérer.

DÉFINITION

LES 3V DU BIG DATA

- **Volume**: Données brutes d'origine (Téraoctets à Pétaoctets de données disponibles)
- **Vélocité**: Rapidité avec laquelle les données sont générées et traitées (temps réel)
- **Variété**: Données hétérogènes (structurées, non structurées, vidéo, géolocalisation, logs, etc.)

LA PRATIQUE

RÉVOLUTION OU EXPRESSION EN VOGUE ?

Comme pour toute innovation, il y a des abus ... mais aussi de vrais cas d'usages !!!

Cas d'usages appropriés :

- Exploration web (ex: bots)
- Réseaux sociaux et analyse de graphes (ex: suggestions d'amis sur Facebook)
- Indexage et recherche en temps réels (ex: Google search)
- Internet des objets (ex: localisation flotte de véhicules)
- Machine learning, système de recommandation (ex : recommandation Spotify)

Des dérives au niveau des entreprises et des gouvernements : plus rien n'est supprimé !

LA PRATIQUE

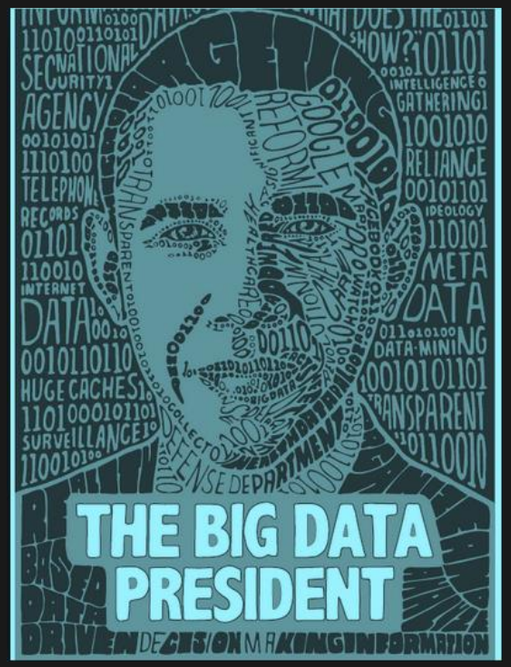
FACEBOOK

- En 2010, Facebook prétend avoir le plus grand cluster Hadoop dans le monde avec 21 PB (21000TB) de stockage.
- En juin 2012, ils annoncent avoir 100PB de données.
- Plus tard dans l'année 2012, ils annoncent que 0,5PB de données sont générés chaque jour.
- En 2014, Facebook Hive générait 4 000 To de data par jour.

Wikipédia

LA PRATIQUE

ÉLECTION DE BARACK OBAMA EN 2012



- Les moyens :
  - Maillage géographique très précis des électeurs
  - Utilisation de données sociologiques et des réseaux sociaux
  - Données récoltées en amont par le porte-à-porte et croisées avec des données internet
- Actions sur la Campagne :
  - Les bénévoles de la campagne Obama savent quels arguments mettre en avant vis-à-vis des femmes célibataires, des jeunes, ou des différentes communautés
  - Ciblage précis des publicités, coups de fil et appels à donations
- Actions le jour du scrutin :
  - Suivi en temps réel et à un niveau de détail important du niveau de participation

LA PRATIQUE

PANAMA PAPERS

Les Panama Papers désignent la fuite massive de documents confidentiels issus du cabinet d'avocats panaméen Mossack Fonseca:

- Début de la fuite en 2015
- Données envoyées vers un quotidien allemand
- Avril 2016 l'affaire des Panama Papers met en évidence un dispositif de sociétés offshore
- Nouveau défi en termes de journalisme de données

- Avec l'aide de l'ICIJ, mise en place d'une plateforme Big Data:
  - Provision de 30 à 40 serveurs
  - Plusieurs mois pour tout scanner
  - 99% des données scannées
- Données sources :
  - 40 années d'historique
  - 2,6 To
  - 11 millions de documents non structurés (pdf, xls, doc, email, images, etc.)
- Enrichissement avec de l'Open Data :
  - Registres de commerce de chaque pays
  - Scraping d'informations d'entreprise

LA PRATIQUE

IA - LES NOUVEAUX CAS D'USAGE

- Médecine:
  - Médecine préventive (prédiction d'une maladie et/ou son évolution)
  - Médecine de précision (recommandation de traitement personnalisé)
  - Aide à la décision (diagnostique et thérapeutique)
  - Robots compagnons
  - Etc.
- Automobile:
  - Maintenance prédictive
  - Sécurité
  - Bien-être
  - Etc.
- Industrie:
  - Maintenance prédictive
  - Sécurité
  - Optimisation
  - Etc.
- Agriculture:
  - Suivi d'exploitation
  - Traitement ciblé
  - Robots
  - Etc.

LA PRATIQUE

LES NOUVELLES PROBLÉMATIQUES

A distributed system is one in which the failure of a computer you didn't even know existed can render your own computer unusable

Leslie LAMPORT - spécialiste de l'algorithmique répartie (et concepteur de LaTeX)

LA PRATIQUE

LES NOUVELLES PROBLÉMATIQUES

- Tolérance aux pannes
- Redondance
- Scalabilité
- Lantence
- Localisation des données
- Théorème CAP

LA PRATIQUE

LES NOUVELLES PROBLÉMATIQUES

Le théorème CAP, aussi connu sous le nom de théorème de Brewer dit qu'il est impossible sur un système informatique de calcul distribué de garantir en même temps les trois contraintes suivantes :

- Cohérence (Consistency en anglais) : tous les nœuds du système voient exactement les mêmes données au même moment ;
- Disponibilité (Availability en anglais) : garantie que toutes les requêtes reçoivent une réponse ;
- Tolérance au partitionnement (Partition Tolerance en anglais) : aucune panne moins importante qu'une coupure totale du réseau ne doit empêcher le système de répondre correctement.

QUESTIONS ?