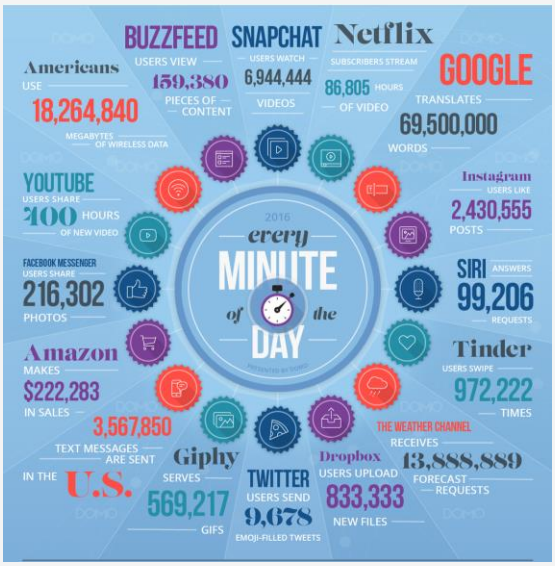
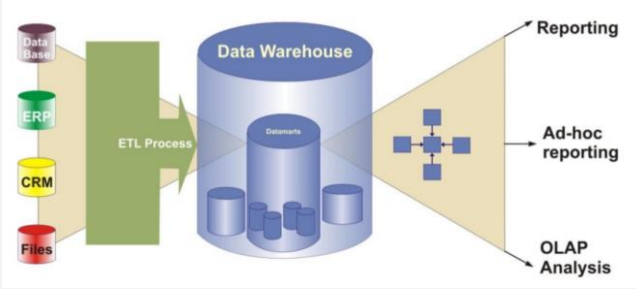
Big Data
Introduction

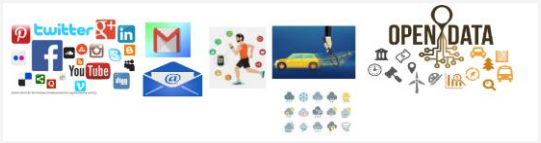
Les données chez les géants du Web

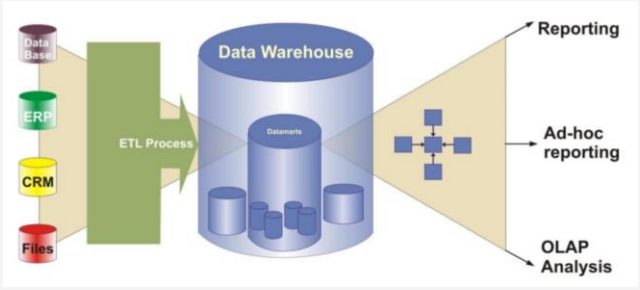


Mais dans une entreprise ordinaire, ça gère vraiment autant ?



Elles en gérent de plus en plus





C’est quoi le Big Data ?

- Un **concept** popularisé en 2012 Traduit le fait que les entreprises sont **confrontées** à des **volumes de données** à **traiter** de plus en plus **considérables** et présentant un fort **enjeu** commercial et **maketing**
- Ces données devenant de plus en plus importantes qu’il est devenu (presque) impossible à travailler avec les outils classiques de gestion de bases de données

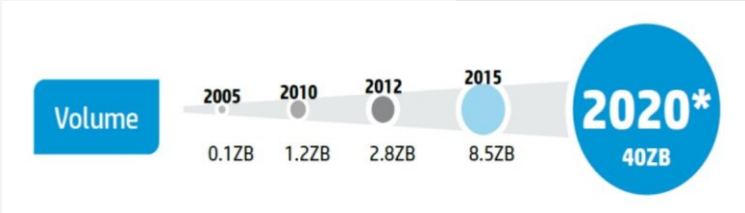
C’est quoi le Big Data ?

- Il s’agit donc d’un ensemble de **concepts, paradigmes, technologies, architectures et procédures** permettant à une organisation de très rapidement et efficacement capter, stocker, traiter et analyser de gros volumes de données **hétérogènes et changeantes** et d’en extraire des connaissances pouvant améliorer leurs positions sur le marché

Caractéristiques du Big Data: les 5 V

Caractéristiques du Big Data

- **Volume:**
  - Croissance sans cesse des données à gérer de tout type
  - Chaque jour, 2.5 trillions d'octets sont générés
  - 90% des données créées dans le monde l'ont été au cours des 2 dernières années (2014)
  - Prévision d'une croissance de 800% des quantités de données générées



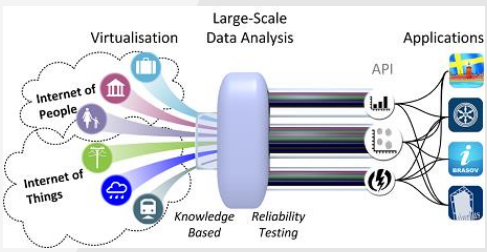
Caractéristiques du Big Data

- **Variété:**
  - Nature des données:
    - Données structurées: bases de données relationnelles, données tabulaires, …
    - Données non-structurées: textes, sons, images, vidéos, fichiers journaux, réseaux sociaux, email, …

  - Diversité des sources:
    - les entreprises consomment de plus en plus de données générées à l'extérieure de son SI

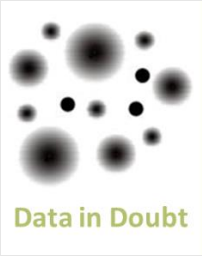
Caractéristiques du Big Data

- **Vélocité (Vitesse):**
  - Elle fait référence à la vitesse à laquelle de nouvelles données sont générées et la vitesse à laquelle elles sont traitées par les systèmes
  - Les technologies d'aujourd'hui permettent d'analyser les données dés qu'elles sont générées sans même les stocker parfois
  - Exemple d'usage:
    - Détection de fraude dans les transactions bancaires
    - Monitoring temps réel d'infrastructures sensibles



Caractéristiques du Big Data

- **Véracité:**
  - Elle fait référence à la qualité et à la fiabilité des données qui sont analysées ainsi qu'à la confiance attachées aux résultats de leur analyse
  - Les données Big Data peuvent être bruitées, imprécises, incomplètes, inconsistantes
  - Quelques faits
    - 1/3 des chefs d'entreprise ne font pas confiance à l'information qu'ils utilisent au quotidien
    - La pauvreté de la qualité des données coûte 3.1 billions de $ annuellement à l'économie américaine



Caractéristiques du Big Data

- **Valeur:**
  - La démarche Big Data doit s'inscrire dans une perspective de création de valeurs pour les entreprises et leurs clients:
    - Meilleure connaissance des clients
    - Optimisation des processus
    - Meilleure maitrise des marchés
    - Meilleure compréhension d'un phénomène
  - Les projets Big Data doivent générer des analyses/études actionnables permettant aux organisations d'atteindre leurs objectifs

Histoire – Google: indexer le web

- **PageRank: l'algorithme d'indexation du Web de Google**
  - **Google annonce que 130 mille milliards de page sont indexées par leur moteur de recherche en Novembre 2016**
  - Google annonce réponde à 5.48 milliards de requêtes par **jour** !
- Problème:
  - Comment stocker et analyser 260 PetaBytes ?
  - Il aurait fallu ~668 jours pour les charger d'un disque avec une capacité de lecture de 150 MB/s

- L'accroissement des besoins ne suit pas l'accroissement des puissances de calcul et de stockage
- Il y'aura une limite physique qu'il sera difficile à contourner
- Ces machines sont extrêmement chers et les délais de construction sont longs

C'est du scale up

L'idée de base: diviser pour mieux régner

Utiliser plusieurs machines avec des performances moyennes au lieu d'une seule grosse machine

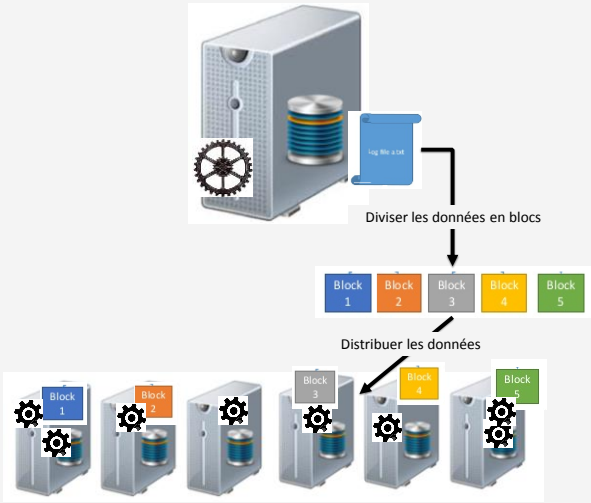
- **Très consommatrice**
- **Très cher**
- **Très fiable**

- **Petites machine**
- **Consomment peu**
- **Pas fiables**
- **Pas chers**

C'est du scale out

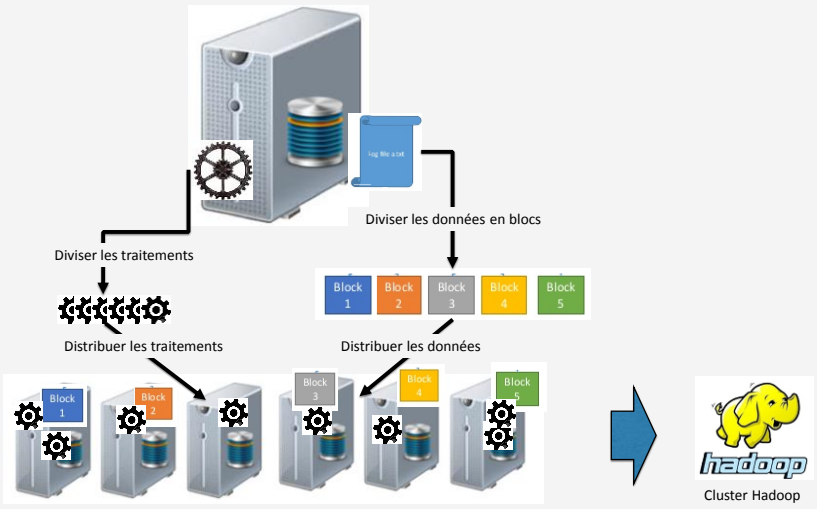
Comment faire ?

Distribuer les données et les calculs sur plusieurs machines



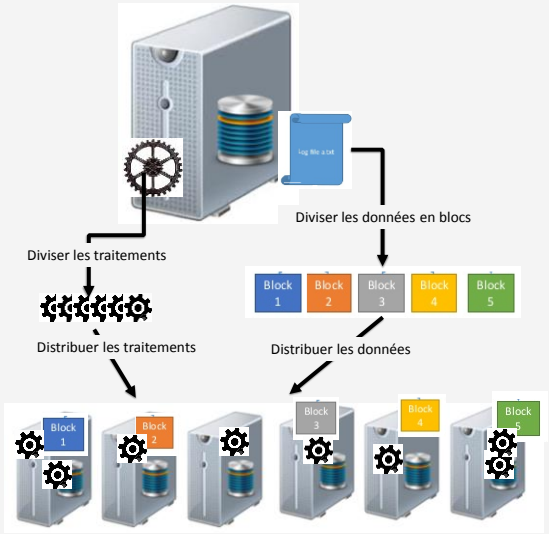
Comment faire ?

Distribuer les données et les calculs sur plusieurs machines



Comment faire ?

Distribuer les données et les calculs sur plusieurs machines

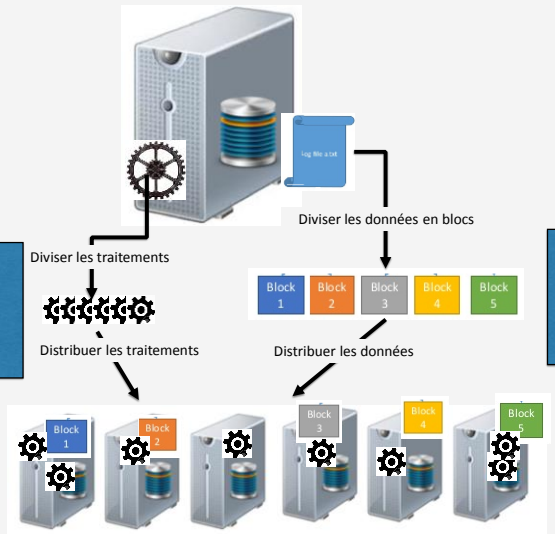


Comment faire ?

Distribuer les données et les calculs sur plusieurs machines

- Framework de développement distribué
- Facilité le développement d'applications distribuées en s'abstrayant de tous les détails réseaux/sérialisation….
- Conçus pour résister aux pannes

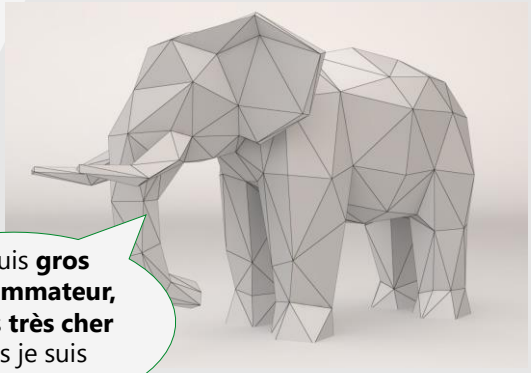
- Yet Another Resource Negociator
- Alloue les ressources (CPU/RAM) pour les différents traitements qui s'exécutent
- Il gère les traitements du cluster
- Il gère les pannes des noeuds



- Hadoop Distributed File System
- Système de gestion de fichiers distribué
- Les fichiers sont divisés en blocks
- Les blocks sont gérés par plusieurs machines
- Tolérant aux pannes

La scalabilité – illustration

Scale-Up



Je suis **gros consommateur,** je suis **très cher** mais je suis **Fiable**

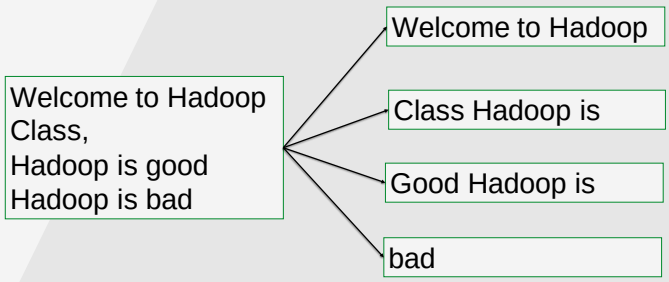
Nous sommes **petits consommateurs,** moins fiables, **pas chers** mais ensemble nous sommes plus **forts**

Scale-Out

Tolérance aux pannes

Un meilleur protocole :

- Découper les données en plusieurs parties
- Lancer des petits traitements sur les petites parties
- Si une panne survient sur une machine, relancer le traitement qui a subit la panne



Un gros traitement (job) doit être découpé en plusieurs job simples

Besoin d'un orchestrateur pour gérer les traitements

Tolérance aux pannes

Dans les systèmes massivement parallélisés, les pannes sont très courantes

Les pannes surviennent car :

Le **Matériel** qui ne fonctionne pas correctement :
- Panne de disque
- Panne de mémoire
- Refroidissement inadéquat

Indisponibilité de **ressources**
- Surcharge du système

Le système doit permettre une tolérance aux pannes dans le cluster

Tolérance aux pannes

Protocole simple : Si une panne survient on redémarre le traitement qui tournait sur la machine

Si un job qui requiert une semaine de traitement, si on a une panne toute les semaines, le traitement ne se terminera jamais !!!

Résumons ce que nous offre une plateforme Big Data

1. Scalabilité horizontale
2. Tolérance aux pannes
3. Facilité d'usage
4. Capacité à traiter des données variées
5. Schéma à la lecture
5. Pas de mise à jour

Use cases: Marketing

- **Cas d'utilisation : Marketing**
  - **Analyse prédictive :** En analysant l'historique des achats du client ou les fichiers Logs qui contiennent les pages visitées, l'entreprise peut prévoir ce que le client cherche et les mettre dans les zones des offres et publicités afin d'augmenter les achats.
  - **Analyse des sentiments :** De Nombreuses sociétés utilisent les échanges sur les réseaux sociaux comme le reflet de l'opinion publique. Celle-ci devient une nouvelle source d'informations en temps réel directement fournie par le consommateur. Les questions d'e-réputation « à quoi est associée mon image ? » ou « comment est accueilli le nouveau produit que je viens de lancer ? » peuvent être analysées avec ces données. Le Big Data permet de prendre le pouls quasiment en direct, mesurer l'impact de sa marque, savoir comment est perçue la société par le public et anticiper les mauvaises critiques.
  - **Analyse des comportements :** L'analyse du comportement des clients en magasin permet d'améliorer l'aménagement du magasin, le mix produit et la disposition des produits dans les rayons et sur les étagères. Les dernières innovations ont également permis de suivre les habitudes d'achat (compter le nombre de pas effectués et le temps passé dans chaque rayon du magasin), géolocaliser en temps réel les clients,.... Les données issues des tickets de caisse, captées depuis longtemps, peuvent maintenant être analysées et révèlent les habitudes d'achat des clients.

Use cases: analyse de Tweet

- **Cas d'utilisation : Analyse de tweets en temps réel**

Use cases: sport

- La première source de données recueillie s'appuie sur des capteurs intégrés aux protège-tibias ou aux chaussures. Ces minuscules composants remontent des informations biométriques sur les joueurs :
  - la distance parcourue
  - les vitesses en sprint
  - les accélérations
  - le nombre de ballons touchés
  - le rythme cardiaque, etc.

→ A terme et quand l'analyse en temps réel sera réellement possible, on peut très bien imaginer qu'une alerte remonte lorsqu'un joueur fatigue afin que l'entraîneur le remplace.

- Une deuxième source de récolte de données provient de caméras installées en hauteur autour du terrain. Tous les déplacements des joueurs et leurs positions les uns par rapport aux autres sont ainsi filmés et enregistrés. Lors de son débriefing, le tacticien peut ainsi comparer plusieurs fois par match la position géométrique de son équipe au moment des temps forts, quand l'équipe se montre offensive, s'ouvre des occasions et marque des buts.

- Le tacticien a également la capacité d'analyser le comportement de son équipe en fonction de la réaction de l'équipe concurrente. Ces données peuvent ensuite être agrégées avec d'autres sources telles que l'historique des matchs joués ou les données recueillies pendant les entraînements.

Use cases: sécurité publique

- Aujourd'hui, avec le Big Data, la vidéosurveillance va beaucoup plus loin : elle permet d'analyser automatiquement les images et les situations, de croiser les informations, et d'envoyer des alertes.
- Cette analyse de vidéo avancée est utilisée en particulier pour :
  - la sécurité du trafic (routier, ferroviaire, maritime et aérien)
  - la protection des espaces et des bâtiments publics
  - la sécurité personnelle.
- Il est aujourd'hui possible à travers l'analyse des images vidéo de faire de :
  - la reconnaissance d'objets et de mouvements
  - la lecture de plaques minéralogiques
  - la détection de véhicule non autorisé
  - la reconnaissance faciale
  - l'auto-surveillance avec possibilité de déclenchement d'alertes ou autres actions automatisées.
- A titre d'exemple la ville de Londres avait, quant à elle, mis en place un système de reconnaissance faciale lors des jeux olympiques de 2012 organisés dans la capitale, afin de lutter contre le terrorisme pour lequel l'alerte était à son maximum.

Use cases

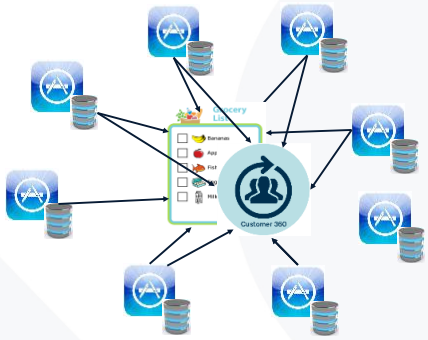
DataLake: le data warehouse du Big Data (1)

Performances opérationnelles

- Optimisation du staffing
- Optimisation des approvisionnements des rayons
- Réduction des déchets
- Reporting en temps réel
- ….

Logistique

- Optimisation des approvisionnements
- Optimisation des stocks
- Organisation des livraisons
- …



- Multiplicité des flux de données
- Consommateur en temps
- Redondance des flux
- Source d'erreurs

Marketing

- Vision 360 des clients
- Offres promotionnelles personnalisées
- Profiling des Clients
- Conception de catalogues
- ….

Expérience Clients

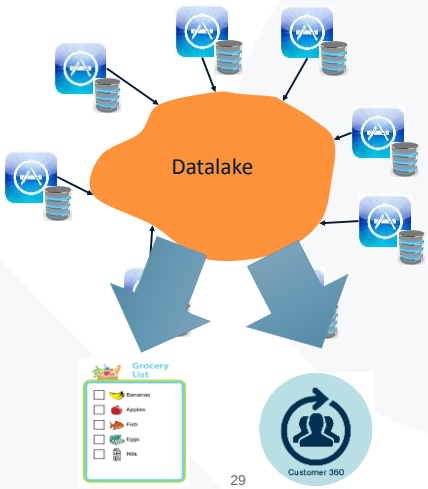
- Construction de listes de courses
- Recommandation de produits
- Analyse comportementales des clients
- …

Use cases

DataLake: le data warehouse du Big Data (2)

Opérations

- Optimisation du staffing
- Optimisation des approvisionnements des rayons
- Réduction des déchets
- Reporting en temps réel
- ….



Marketing

- Vision 360 des clients
- Offres promotionnelles personnalisées
- Profiling des clients
- Conception de catalogues
- ….

Logistiques

- Optimisation des approvisionnements
- Optimisation des stocks
- Organisation des livraisons
- …

Expérience clients

- Construction de listes de courses
- Recommandation de produits
- Analyse comportementales des clients
- …

Use cases

DataLake: le data warehouse du Big Data (3)

Capacité de stockage et puissance de calcul permettant de mener à bien tous les projets data d'une entreprise

- Données aux formats natifs:
  - Définition du schéma à l'exploitation de la donnée
- Scalabilité:
  - capacité de stockage et puissance de calcul extensibles
- Versalité:
  - Stocker n'importe quel type de données (relationnel, XML, Logs, message, données réseaux sociaux, ….)
- Flexibilité:
  - Facilité d'intégrer de nouvelles sources de données
- Accessibilité:
  - Plusieurs langages/outils pour requêter/analyser les données
- Evolutivité:
  - Capacité à changer le schéma des sources sans impacter les traitements en aval

- Démocratisation de l'accès aux données de toute l'entreprise
- Favoriser l'émergence d'une entreprise data-driven
- Evolution plus facile du datalake
- Favorisation de l'esprit d'innovation
- Accélération du time to market