# Big Data

*Introduction*

# Big Data : Faits



- Chaque jour, nous générons 2,5 trillions d'octets de données
- 90% des données dans le monde ont été créées au cours des deux dernières années
- Source:
  - Capteurs utilisés pour collecter les informations climatiques
  - Messages sur les médias sociaux
  - Images numériques et vidéos publiées en ligne
  - Enregistrements transactionnels d'achat en ligne
  - Signaux GPS de téléphones mobiles
  - …
- Données appelées **Big Data** ou **Données Massives**

# Big Data : Intérêts



| | |
|---|---|
| 3/ 1 | •Chefs d'entreprise prennent fréquemment des décisions basées sur des informations en lesquelles ils n'ont pas confiance, ou qu'ils n'ont pas |
| 1 / 2 | •Chefs d'entreprise disent qu'ils n'ont pas accès aux informations dont ils ont besoin pour faire leur travail |
| % 83 | •Des DSI (Directeurs des SI) citent : « L'informatique décisionnelle et analytique » comme faisant partie de leurs plans pour améliorer leur compétitivité |
| % 60 | •Des PDG ont besoin d'améliorer la capture et la compréhension des informations pour prendre des décisions plus rapidement |

# Big Data : Sources



- Sources multiples: sites, bases de données, téléphones, serveurs:
  - Détecter les sentiments et réactions des clients
  - Détecter les conditions critiques ou potentiellement mortelles dans les hôpitaux , et à temps pour intervenir
  - Prédire des modèles météorologiques pour planifier l'usage optimal des éoliennes
  - Prendre des décisions risquées basées sur des données transactionnelles en temps réel
  - Identifier les criminels et les menaces à partir de vidéos, sons et flux de données
  - Étudier les réactions des étudiants pendant un cour, prédire ceux qui vont réussir, d'après les statistiques et modèles réunis au long des années (domaine *Big Data in Education*(

# Big Data : Challenges



- Réunir un grand volume de données variées pour trouver de nouvelles idées
- Difficulté pour sauvegarder toutes ces données
- Difficulté pour traiter ces données et les utiliser
- Les données sont créées rapidement

# Big Data : Termes Clefs



- Extraction d'informations et décisions à partir de données caractérisées par les 3 V:
  - **Volume (*Volume*(**
    - ✓✓ L'entreprise est submergée de volumes de données croissants de tous types, qui se comptent en *téraoctets*, ou même en *pétaoctet*
  - **Variété (*Variety*(**
    - ✓✓ Gérer la complexité de plusieurs types de données et de schémas structurés ou non structurés )texte, données de capteurs, son, vidéo, logs(…
  - **Vitesse (*Velocity*(**
    - ✓✓ Parfois, les données doivent être saisies et traitées au fil de l'eau, au fur et à mesure de leur collection par l'entreprise, pour la détection des fraudes par exemple
- *Objectif* : relever ce qui est important et ce qui l'est moins

# Big Data : **Volume**



- Le prix de stockage des données a beaucoup diminué ces 30 demières années:
  - De $100,000 / Go ((1980
  - À $0.10 / Go ((2013
- Les lieux de stockage fiables (comme des SAN: Storage Area Network) ou réseaux de stockage peuvent être très coûteux
  - Choisir de ne stocker que certaines données, jugées sensibles
  - Perte de données, pouvant être très utiles, comme les logs
- Comment déterminer les données qui méritent d'être stockées?
  - Transactions? Logs? Métier? Utilisateur? Capteurs? Médicales? Sociales?

→ **Aucune donnée n'est inutile**. Certaines n'ont juste pas encore servi.

- Problèmes:
  - Comment stocker les données dans un endroit fiable, qui soit moins cher
  - Comment parcourir ces données et en extraire des informations facilement et rapidement?

# Big Data : **Variété**



- Pour un stockage dans des bases de données ou dans des entrepôts de données ,les données doivent respecter un format prédéfini.
- La plupart des données existantes sont non-structurées ou semi-structurées
- Données sous plusieurs formats et types
- On veut tout stocker:
  - *Exemple*: pour une discussion dans un centre d'appel, on peut la stocker sous forme textuelle pour son contenu, comme on peut stocker l'enregistrement en entier, pour interpréter le ton de voix du client
- Certaines données peuvent paraître obsolètes, mais sont utiles pour certaines décisions:
  - *Exemple*: Pour le transport de marchandise, on a tendance à choisir le camion le plus proche. Mais parfois, ce n'est pas la meilleure solution. D'autres pb peuvent intervenir.
  - Besoin de : Données GPS, Plan de livraison du camion, Circulation, Chargement du camion, Niveau d'essence…

# Big Data : **Vitesse**



- Rapidité d'arrivée des données
- Vitesse de traitement
- Les données doivent être stockées à l'arrivée, parfois même des Teraoctets par jour
  - Sinon, risque de perte d'informations
- Exemple
  - Il ne suffit pas de savoir quel article un client a acheté ou réservé
  - Si on sait que vous avez passé plus de 5mn à consulter un article dans une boutique d'achat en ligne, il est possible de vous envoyer un email dès que cet article est soldé.

# Approche Traditionnelle

## Les besoins métier guident la conception de la solution

Le responsable métier définit les besoins:
Quelles questions doit-on poser?

IT conçoit une solution avec un ensemble de structures et fonctionnalités

Le responsable métier exécute les requêtes pour répondre aux questions – encore et encore

De nouvelles exigences nécessitent une nouvelle conception et construction

# Approche Traditionnelle



- Appropriée pour:
  - Des données structurées
  - Opérations et processus répétitifs
  - Sources relativement stables
  - Besoins bien compris et bien cadrés

# Approche Big Data

## Les sources d'information guident la découverte créative

Le responsable métier et IT identifient les sources de données disponibles

IT fournit une plateforme qui permet une exploration créative de toutes les données disponibles

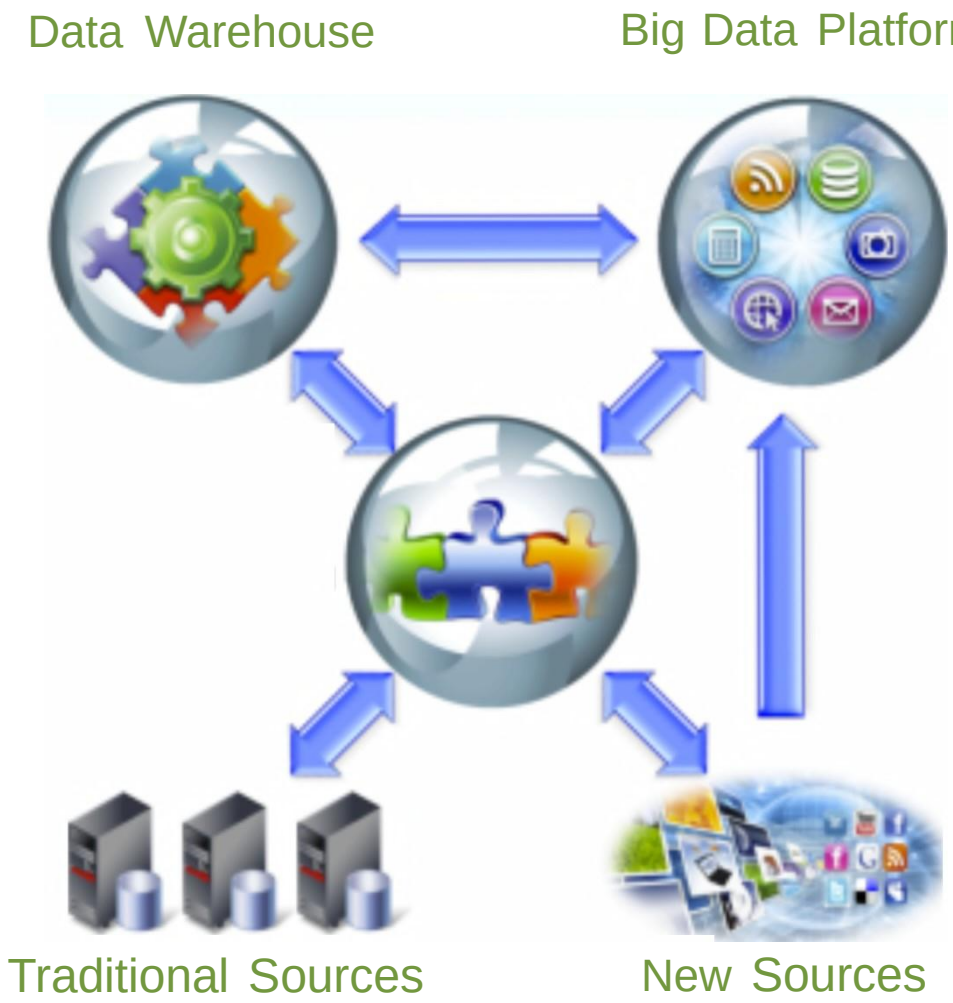
Le responsable métier détermine les questions à poser en explorant les données et relations entre elles

De nouvelles idées conduisent à l'intégration de technologies traditionnelles

# Approche Big Data

- La question **n'est pas**:
  - Dois-je choisir entre l'approche classique et l'approche big data?

- Mais plutôt:
  - Comment les faire fonctionner Ensemble ?



# Fusionner l'approche Big Data avec l'approche Traditionnelle



## Approche Traditionnelle
Analyse Structurée et Répétée

**Responsables Métier**

Déterminent quelles questions poser

**Responsables IT**

Structurent les données pour répondre à ces questions

## Approche Big Data
Analyse Itérative et Exploratoire

**Responsables IT**

Fournissent une plateforme pour permettre la découverte créative

**Responsables Métier**

Explorent la plateforme pour déterminer quelles questions poser

# Sources



- **Cours**
  - *Big Data Analytics – Lesson 1: What is Big Data*, IBM, Big Data University
  - *Intro to Hadoop and MapReduce*, Coursera, Udacity