

Big Data

GL4 (Option Management des Systèmes d'Information) - 2017

Chp1 : Introduction

Approche et Vocabulaire

Faits

Introduction aux Big Data

- 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
- 90% des données générées sont non structurées
- 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

Intérêts

Introduction aux Big Data

- 1 / 3

- 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

Sources

Introduction aux Big Data

- 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 cours, prédire ceux qui vont réussir, d'après les statistiques et modèles réunis au long des années (domaine Big Data in Education)

Challenges

Introduction aux Big Data

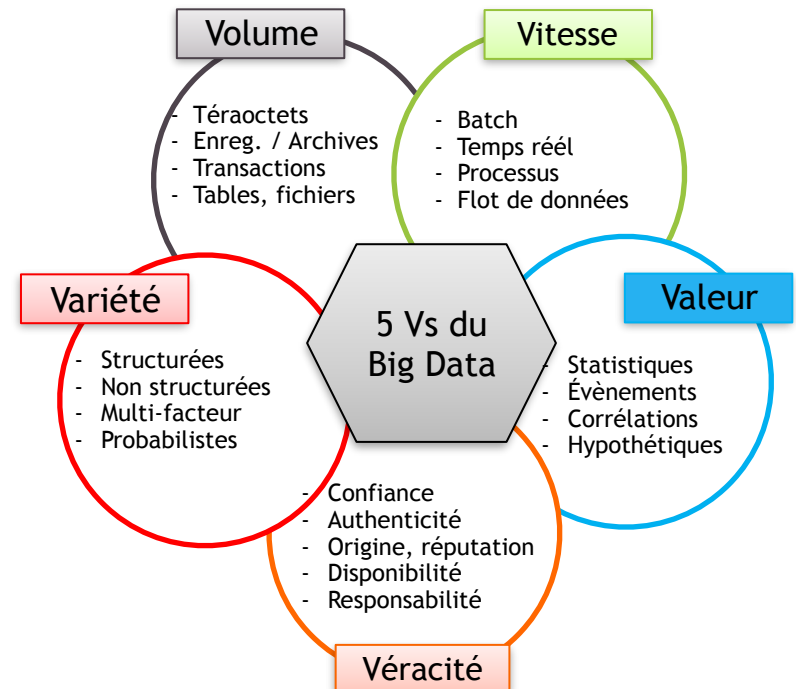
- Réunir un grand volume de données variées pour trouver de nouvelles idées
- Capturer des données créées rapidement
- Sauvegarder toutes ces données
- Traiter ces données et les utiliser

Les 5 V

Introduction aux Big Data

- Extraction d'informations et décisions à partir de données, caractérisées par les 5 V:

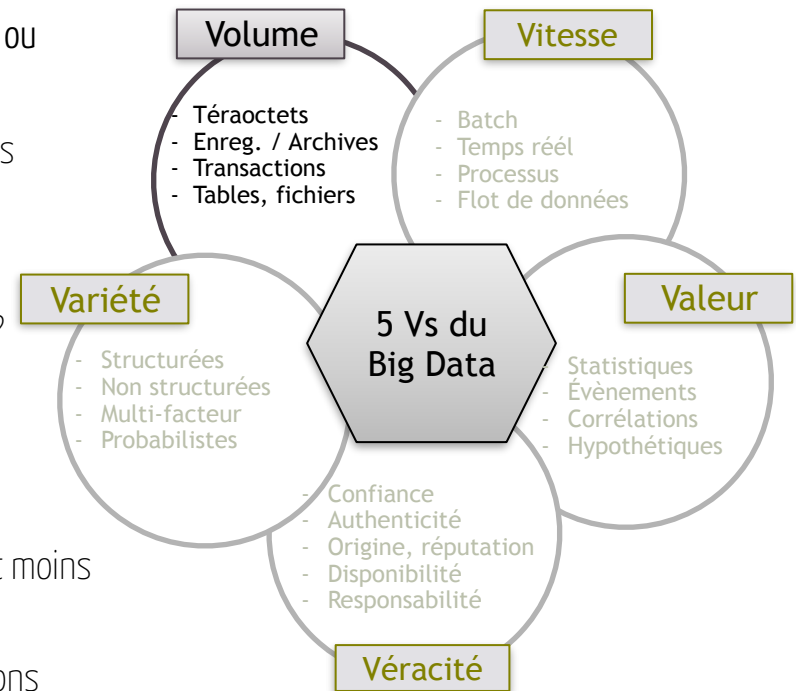
- Volume (Volume)
- Variété (Variety)
- Vitesse (Velocity)
- Véracité (Veracity)
- Valeur (Value)



Volume

Introduction aux Big Data

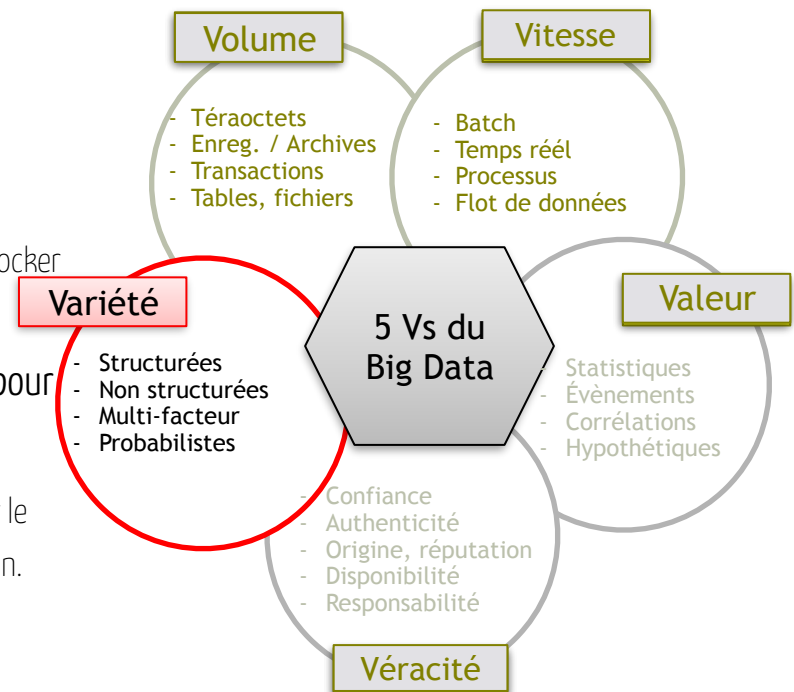
- Le prix de stockage des données a beaucoup diminué ces 30 dernières années:
 - De \$100,000 / Go (1980)
 - À \$0.10 / Go (2013)
 - Les lieux de stockage fiables (comme des SAN: Storage Area Network) ou réseaux de stockage peuvent être très coûteux
 - Choisir de ne stocker que certaines données, jugées 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?



Variété

Introduction aux Big Data

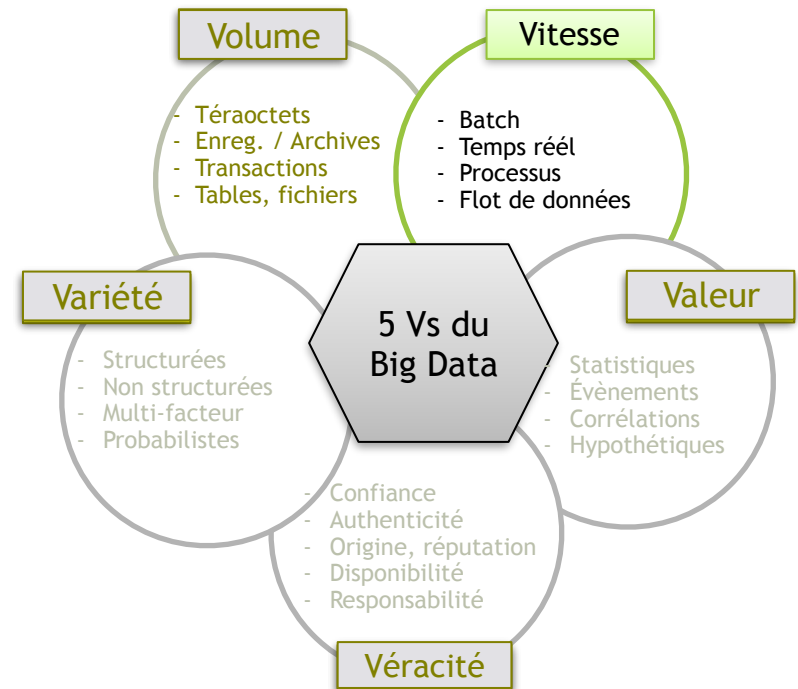
- Pour un stockage dans des bases de données ou dans des entrepôts de données, les données doivent respecter un format prédéfini.
- La plupart des données existantes sont non-structurées ou semi-structurées
- Données sous plusieurs formats et types
- On veut tout stocker:
 - Exemple: pour une discussion 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 problèmes peuvent intervenir.
 - Besoin de : Données GPS, Plan de livraison du camion, Circulation, Chargement du camion, Niveau d'essence...



Vitesse

Introduction aux Big Data

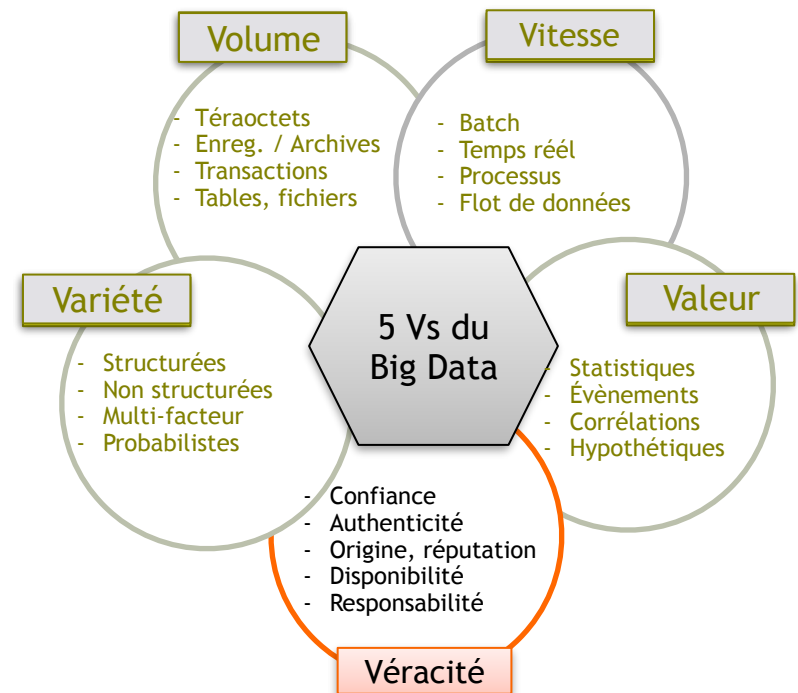
- Rapidité d'arrivée des données
- Vitesse de traitement
- Les données doivent être stockées à l'arrivée, parfois même des Teraoctets par jour
 - Sinon, risque de perte d'informations
- Exemple
 - Il ne suffit pas de savoir quel article un client a acheté ou réservé
 - Si on sait que vous avez passé 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.



Véracité

Introduction aux Big Data

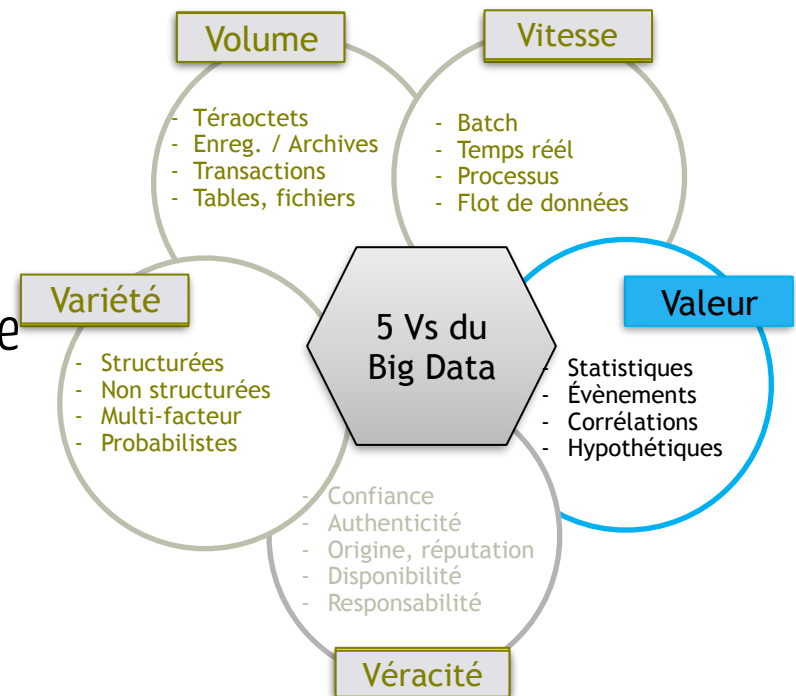
- Cela fait référence au désordre ou la fiabilité des données. Avec l'augmentation de la quantité, la qualité et précision se perdent (abréviations, typos, déformations, source peu fiable...)
- Les solutions Big Data doivent remédier à cela en se référant au volume des données existantes
- Nécessité d'une (très) grande rigueur dans l'organisation de la collecte et le recoupement, croisement, enrichissement des données pour lever l'incertitude et la nature imprévisible des données introduites dans les modèles mais aussi pour respecter le cadre légal pour créer la confiance et garantir la sécurité et l'intégrité des données.



Valeur

Introduction aux Big Data

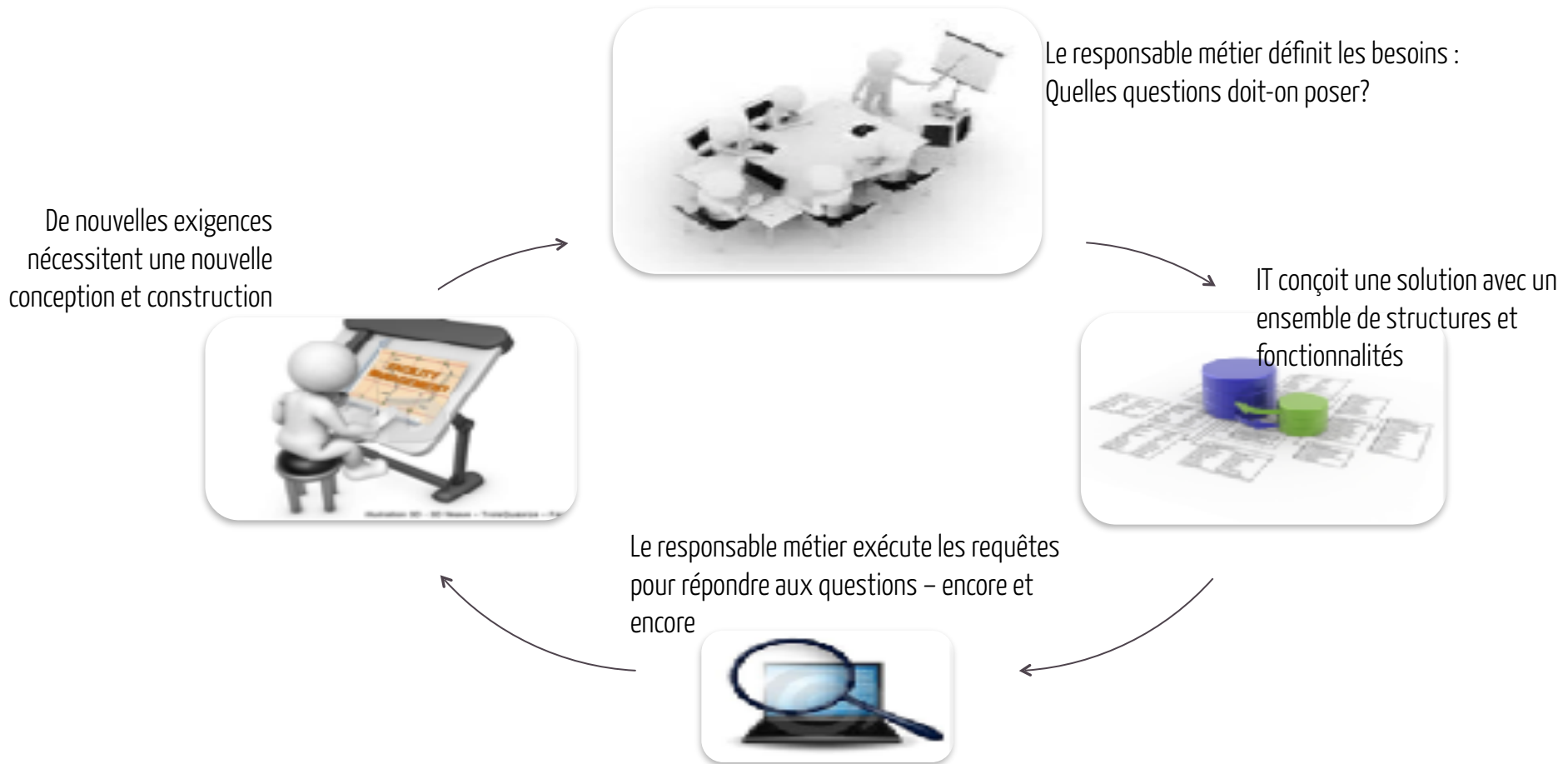
- Le V le plus important
- Il faut transformer toutes les données en valeurs exploitables: les données sans valeur sont inutiles
- Atteindre des objectifs stratégiques de création de valeur pour les clients et pour l'entreprise dans tous les domaines d'activité



Approche Traditionnelle

Introduction aux Big Data

Les besoins métier guident la conception de la solution



Approche Traditionnelle

Introduction aux Big Data

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

Approche Big Data

Introduction aux Big Data

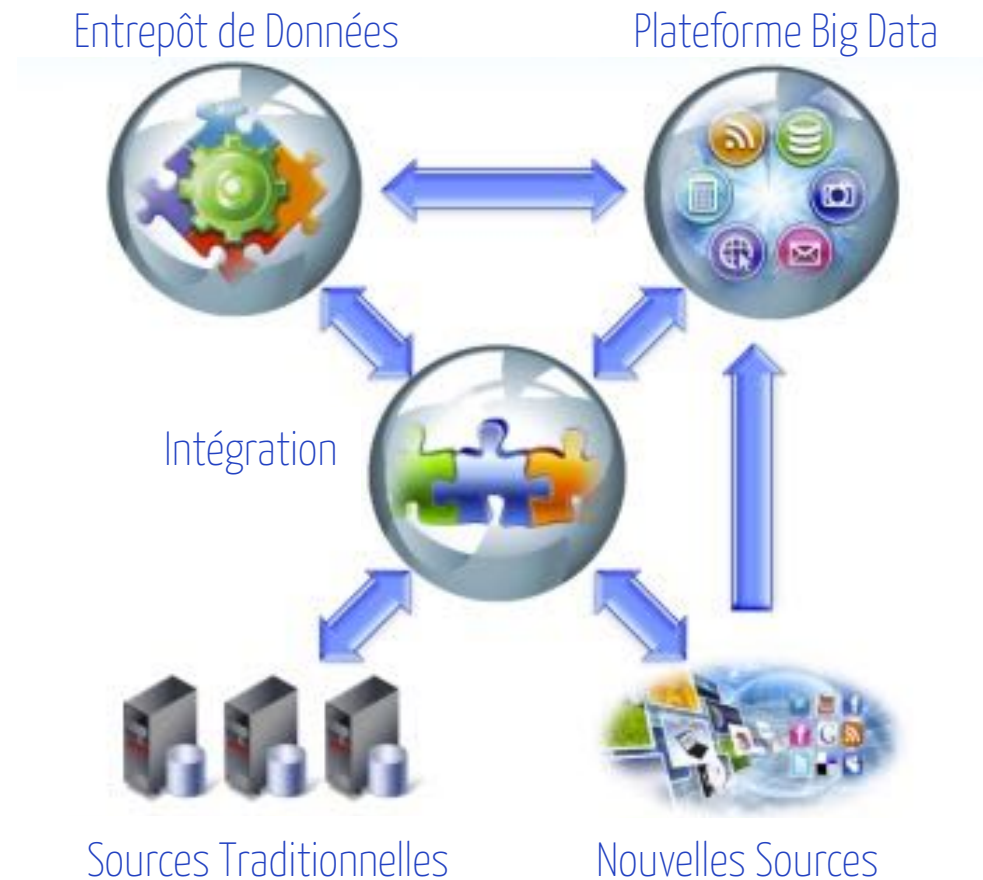
Les sources d'information guident la découverte créative



Approche Big Data vs Approche Traditionnelle

Introduction aux 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?



Approche Big Data vs Approche Traditionnelle

Introduction aux Big Data

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

Introduction aux Big Data

- Cours
 - Big Data Analytics – Lesson 1: What is Big Data, IBM, Big Data University
 - Intro to Hadoop and MapReduce, Coursera, Udacity
- Articles
 - Bernard Marr, “Big Data: The 5 Vs Everyone Must Know”, LinkedIn