Big Data

Une escroquerie mondiale

29 Avril 2022

claude@lemarson.com
https://www.lemarson.com

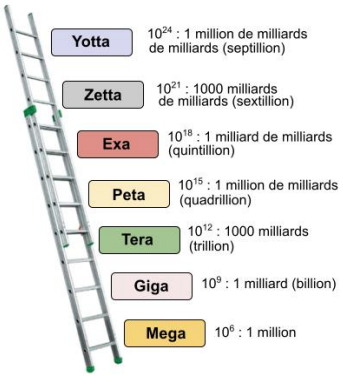
Sommaire

Big Data : une escroquerie mondiale

273,4 G$ en 2026 (+ 11 % par an) : pas en rapport avec les prévisions euphoriques des prestataires (MarketsandMarkets)

Quand le marketing prend le pas sur la réalité

- Depuis 10 ans, la mode est à l'extension
- Plus le volume de données est élevé, plus les décisions doivent être fiables (forcément)
- Si on ne comprend rien aux unités, ça n'en sera que mieux
- On cherche à exploiter au mieux les performances des systèmes : traitement et stockage
- On accommode l'IA à toutes les sauces : sans Intelligence Artificielle il ne saurait y avoir de projet sérieux…
- Idem pour la Blockchain, qui apporte en plus la liberté de choix
- Les prestataires jouent sur la corde sensible : la crainte des entreprises d'être dépassées par leurs concurrents
- Internet est forcément la solution : c'est magique
- Si c'est pour ajouter quelques données issues des réseaux sociaux à la base clients, ce n'est pas du Big Data
- Si on est capable d'influencer les élections américaines, on ne voit pas pourquoi on ne ferait pas la même chose avec le BI en entreprise, pour déstabiliser son concurrent
- Le TI est de plus souvent un objet de pouvoir et le Big Data associé à la transformation numérique est considéré comme le XXI ème siècle, le nirvana en matière d'efficacité : tout ce qui a été fait de bon par le passé n'était que le fruit du hasard…



McKinsey et Gartner pas d'accord

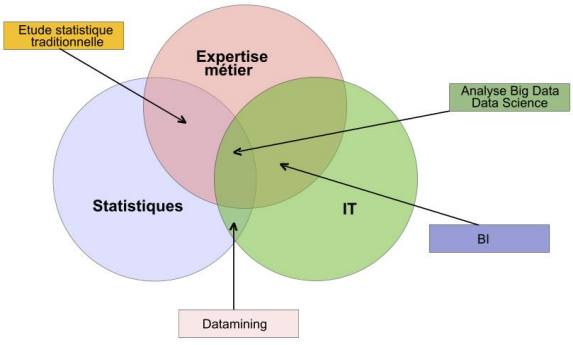
- Il y a quelques années McKinsey :
- Les entreprises qui pratiquent le Big Data ont 23 fois plus de chances d'acquérir des nouveaux clients, 9 fois plus de chances de ne pas les perdre et 19 fois plus de chances d'être profitables
- Gartner : 60 % des projets Big Data n'aboutiront pas
- McKinsey avait tort, Gartner aussi : le vrai taux d'échecs est de 85 % !!!

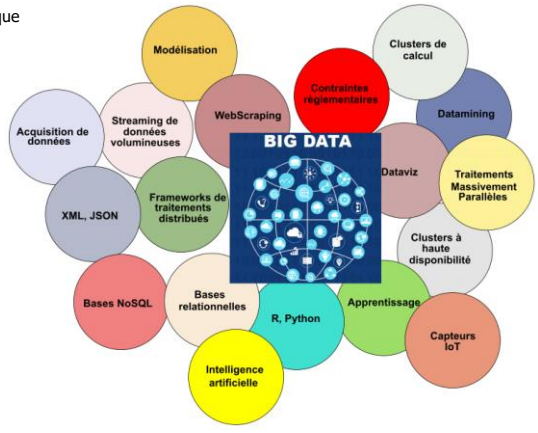
Gartner

McKinsey &Company

C'est quoi le Big Data

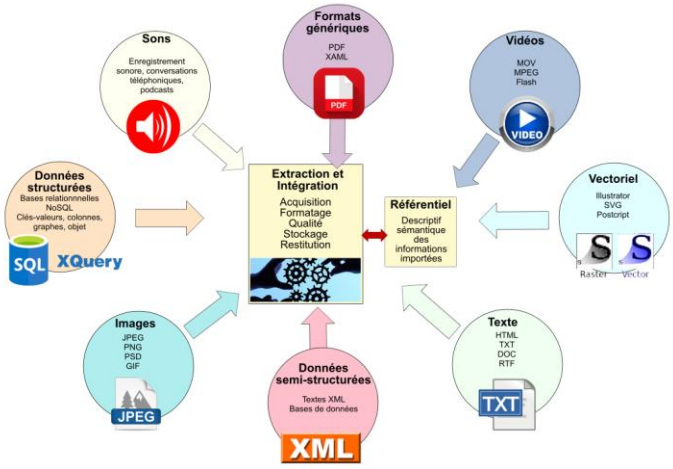
- Ambition de tirer un avantage économique de l'analyse quantitative des données volumineuses internes et externes de l'entreprise (petabytes)
- Capacité à traiter des données "difficiles", du fait de leur complexité. Ce serait restrictif de limiter le Big Data aux seules questions de volumes.
- Ouverture à tous les formats autres que relationnels : vidéos, textes, XML…
- A la croisée de 3 grands domaines : statistiques, expertise métier et informatique

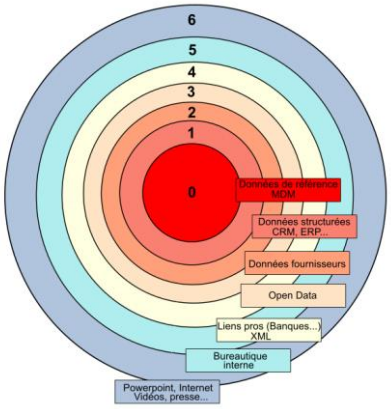


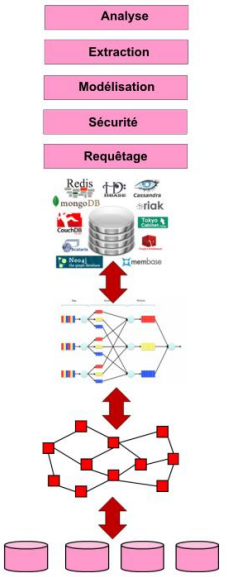


C'est aussi et surtout la capacité à s'étendre dynamiquement dans des domaines de données non prévus au départ
C'est cette extension qui sonnera le glas des installations totalement locales

Les sources de données







Niveau 9 : Applications analytiques

Niveau 8 : Peuplement de l'infrastructure logique et physique de stockage

Niveau 7 : Constitution des référentiels, metadatas

Niveau 6 : Modules de sécurité, authentification forte, Kerberos, SAML, certificats, fédération

Niveau 5 : Choix du langage de requête ou d'un module de consultation lié à une application analytique

Niveau 4 : Organisation logique des données : relationnel, NoSQL, "time series"

Niveau 3 : Distribution des traitements (Map and Reduce), GFS (Google File System)...

Niveau 2 : Gestion physique des fichiers, locale ou distribuée

Niveau 1 : Caractéristiques physiques du stockage et de l'infrastructure de soutien

Le modèle à neuf niveaux de la hiérarchie Big Data

A quoi devrait servir le Big Data

On ne manque pas d'idées … encore faut-il les concrétiser

Le BI est indissociable du Big Data, on change simplement de dimension

Traitement des données en temps réel (streaming) pour faire ressortir des tendances immédiates

Vision dynamique des comportements des clients. Déduction des parcours réels par "sequence mining": historique des achats, regroupement par séquences ("sequence clustering", prédiction des opérations : offres promotions

Liens avec les réseaux sociaux. Mesure de la e-reputation et analyse des tendances

Rapprocher les fournisseurs de services de leurs usagers : meilleure efficacité.

Traitement des données volumineuses, issues des capteurs et IoT

Recommandations de services et de films : You Tube, Netflix

Mise sur le marché de nouveaux produits et analyse des réactions en temps réel

Analyse des tweets et messages issus des réseaux sociaux

Comportements, actions et géolocalisation des usagers de cartes bancaires

Analyse des données échangées sur les réseaux de jeux pour déterminer les orientations futures des produits

Analyse des comportements des automobilistes pour les assurances : vers la règlementation temps réel

Optimisation du support technique, basée sur l'analyse des tickets de "help desk", croisée avec les contenus publics disponibles sur les forums techniques.

Suivi des déplacements des personnes à risque ou stratégiques

Un exemple Big Data

La constitution d'un profil client

Hôpitaux santé

Cookies de browsers

Collections, Art

Assurances

Réseaux sociaux

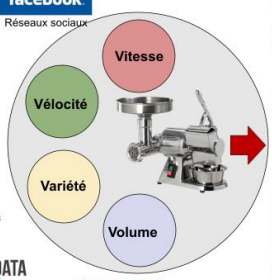
Opérateurs téléphoniques

Agences gouvernementales

Institutions financières

Compteurs électriques, eau

Voitures connectées



Publication articles, contributions

Commentaires hôtels restaurants, spectacles

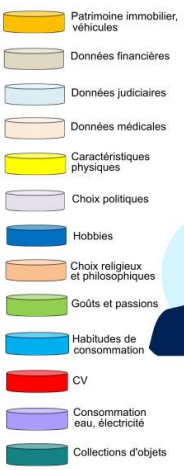
Agences de voyages

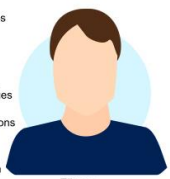
ISP (Internet)

Communautés d'intérêt

Websites de CV

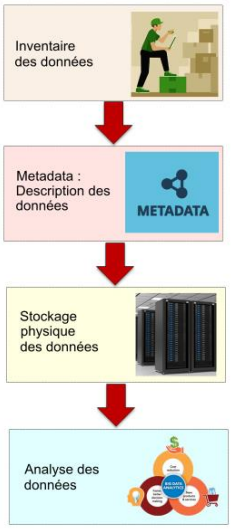
IoT





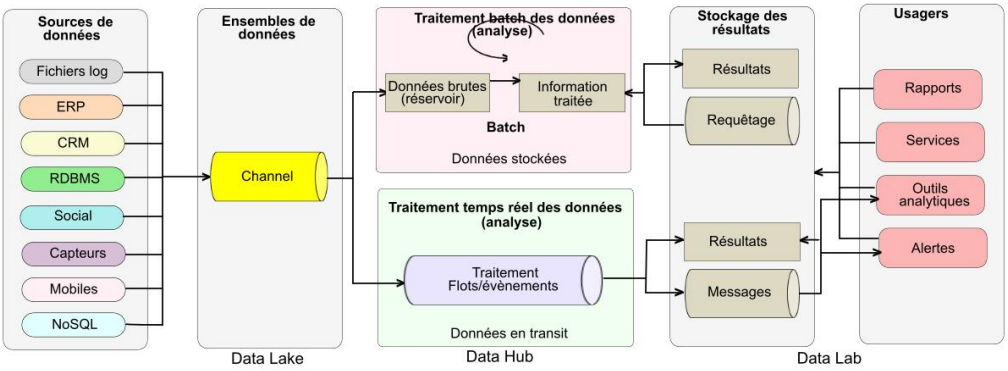
Filtrage
Traitements analytiques
Consolidation

Les quatre phases du Big data
à des fins analytiques



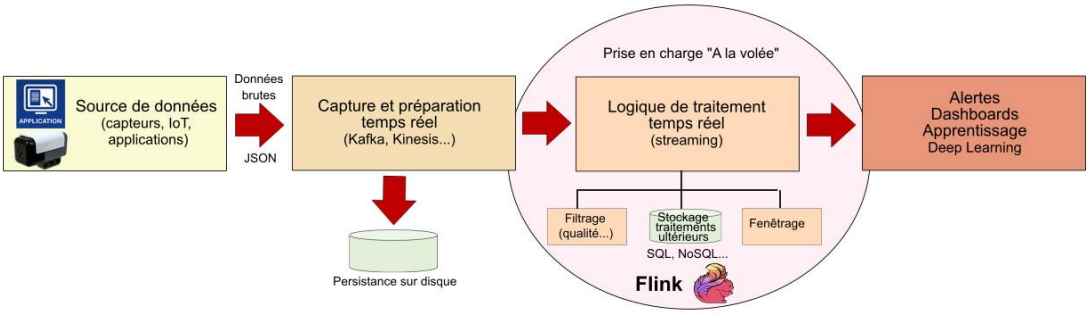
- La mise en œuvre d'une architecture dite de Big Data à des fins analytiques, repose sur 4 éléments :
- D'abord des données 3V, volumineuses et variées.
- La mise en place d'un référentiel (repository) de toutes ces données, celles-ci étant rapatriées ou non, voire perçues à travers le verre grossissant d'un "data lake".
- Eventuellement, le stockage physique des données, s'il faut le prendre en compte. Il peut être crédible de passer par une phase d'importation, comme on le faisait dans les premiers temps du datawarehouse et de tout concentrer dans un même silo. Sauf, si certaines de ces données doivent être traitées en temps réel. Dans ce cas, il faudra les prendre telles quelles, là où elles sont.
- Le recours à des outils d'analyse, relativement classiques, Pentaho, SAP, Map/R, Hortonworks, Cloudera, EMR Amazon, de nombreux utilitaires Open Source, etc, sauf que cette fois, ils vont pointer sur des volumes de données gigantesques.

L'architecture globale du Big Data
(ce que l'on essaie de faire)



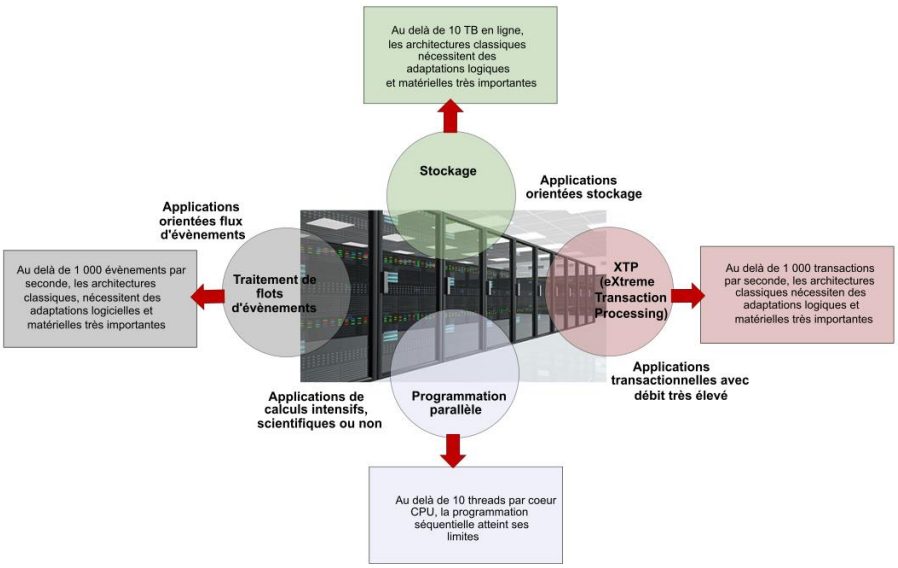
- Data Lake : collecte et stockage de toutes les données de l'entreprise
  - Pas de filtrage, ni purge, sans présupposition des modèles à appliquer : un simple réservoir de données
- Data Hub : transformation récurrente des données (data flow)
  - Batch, flux et temps réel, apprentissage et exploitation de modèles prédictifs sur des volumes de données importants
  - Requêtage, agrégation ad-hoc
- Data Lab : analyse des données de production, apprentissage et exploitation de modèles prédictifs, interrogation, agrégation, requêtes ad-hoc

Le traitement à la volée des données massives



- Trois phases : capture, traitements à la volée, traitements en aval
- Il faut :
  - Une couche de récupération des données, interfaçage avec les fournisseurs, les données étant générées en continu, sans limites
  - Une couche de préparation des données, l'intermédiaire entre les flux de données et leur mise à disposition applicative : le broker Kafka ou Kinesis
  - Une logique de traitement temps réel sur des données individuelles ou regroupées en mini-batch
- Flink est le projet étendard, hébergé chez Apache

Big Data : on atteint les limites



Des obstacles techniques ignorés
(ce que l'on ne sait pas faire)

- Big data ne veut pas dire Good Data
- Avec des volumes proches et au-delà des Petabytes, les bases relationnelles ne sont plus adaptées, qui entraînent des "big" temps de réponse
- Le mode distribué, type "data mesh" n'est pas adapté aux gros volumes et pose des problèmes de sécurité et d'usage pour le transactionnel
- Complexité : les bases relationnelles explosent dès que l'on dépasse un certain nombre d'enregistrements distincts dans le schéma, sans qu'il soit possible de modéliser le comportement du SGBD
- La solution est de disposer de SGBD hyper puissants, comme ceux d'Oracle, mais cela ne règle pas le problème de fond : reculer pour mieux sauter
- Hétérogénéité : le mélange de données structurées, non structurées et semi-structurées (XML) n'est pas maîtrisé en termes de performances, on est très loin du compte
- On ne sait pas unifier les données à des fins transactionnelles et analytiques : le modèle HTAP n'a convaincu personne
- Il faut des ressources importantes : traitement et stockage
- Le problème du stockage est mal traité, d'où l'arrivée de solutions DaaS, type Snowflake, des Clouds fédérateurs de données, qui prennent tout en charge et libèrent d'autant les utilisateurs
- Il n'existe pas de solution de requêtage simple et satisfaisante, avec formalisme proche de SQL, qui soit adaptée aux données hétérogènes
- Les outils de BI analytique sont déjà nombreux mais décevants quand on les pousse aux limites

Ce que l'on ne sait pas faire

Un management qui attend des résultats trop vite via les outils, alors qu'il s'agit d'un travail de fond : on ne se comprend pas.

Le transactionnel est encore hors de portée du Big Data. Sauf cas spécifiques avec de lourds développements propriétaires. Ne pas confondre avec le BI.

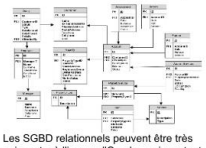
La qualité des données relationnelles est insuffisamment maîtrisée : "Garbage In, Garbage Out". C'est encore pire avec les autres données, surtout sur Internet.

Les réalisations orientées Big Data sont fortement teintées de solutions propriétaires et d'un fort investissement des usagers en termes de développement

La gestion de fichiers physiques dans un cluster, très orientée multimédia, est inadaptée aux besoins de gestion et au transactionnel. D'où le recul d'HDFS d'Apache Hadoop.

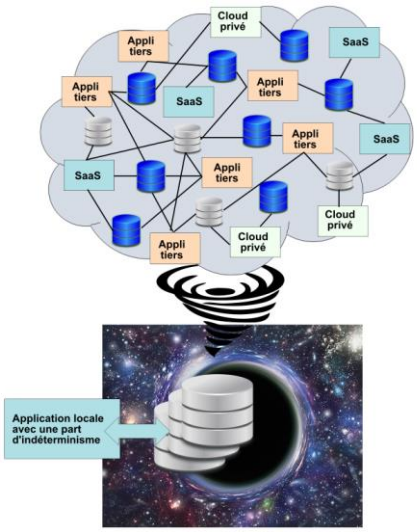
Le Big Data nécessite de lourds investissements en termes d'infrastructures. Pour les réseaux, entre autres, dans un contexte de déport dans le Cloud.

L'analyse des informations issues d'Internet est entâchée d'un manque évident de confiance. Ne sortir que les données crédibles nécessite des efforts encore nors de portée de la plupart des entreprises (modules d'IA).



Les SGBD relationnels peuvent être très puissants, à l'image d'Oracle, mais restent insuffisants pour ce qu'induit le Big Data.

Le "trou noir" des données



- Au fur et à mesure que les données deviennent plus volumineuses, elles attirent de plus en plus d'autres données, qu'elles engloutissent comme un trou noir.
- C'est l'opposition entre un modèle statique où les données sont mises à jour mais restent fixes quant à leur schéma et un modèle dynamique où les données sont connectées à d'autres données, avec une maillage de plus en plus touffu.
- C'est ce modèle qui s'instaure progressivement, avec des applications locales qui « attaquent » des données privées ou Cloud, mais sont également connectées à d'autres sources et applications, publiques ou privées, qui elles-mêmes accèdent à des données tiers. Pour constituer « in fine » un entrelacs d'applications et de fichiers, dont on a de plus en plus de mal à démêler les liens. C'est le trou noir…

Gravité des données

$$\frac{\left(\text{Data} \times \text{Application}\right) \times \text{Number of Requests}}{\left(\text{Latency} + \left(\text{Average Size of Requests} / \text{Bandwidth}\right)\right)^2}$$

- McCrory nous propose une formule censée exprimer l'importance des critères qui vont contribuer à la gravité des données et à l'incertitude des résultats.
- La gravité est proportionnelle au volume de données, à celle des applications qui les manipulent, de même qu'au volume de requêtes effectuées sur ces données par seconde. Mais elle est aussi inversement proportionnelle au carré de tout ce qui peut la gêner : temps d'accès, taille moyenne des requêtes divisée par la bande passante réseau disponible.

Big Data et Intelligence Artificielle

Intéressant… pour demain

- Le principe est d'utiliser les technologies d'IA pour extraire du sens et prendre de meilleures décisions à partir de sources de connées massives.
- Il ne faut pas tomber dans le délire "prémonitoire" qui voudrait que le Big Data pourrait permettre de "surmonter des défis comme le chômage, l'environnement, l'économie, la sécurité ou la santé". Rien de moins…
- L'IA sera peut-être au cœur de certaines avancées de compréhension du Big Data, avec une forte spécialisation des algorithmes, mais cela prendra du temps
- Ne connaissant pas les spécificités des contenus, en plus de leur volume, nous ne sommes plus capables d'exprimer nos requêtes de manière explicite : il faut passer par des algorithmes d'apprentissage qui exécuteront les introspections à notre place
- A terme, l'IA peut jouer un rôle utile en deux phases :
  - Apprentissage sur des volumes de données élevés, pour "comprendre" la situation et faire ressortir les principales caractéristiques des données manipulées
  - Algorithmes Big Data, éventuellement locaux, sur des échantillons réduits de données, qui fourniront par leur adéquation des résultats équivalents

Le Big Data est un Hymalaya dont les usagers n'ont pas conscience, jusqu'au moment où il faut payer l'addition

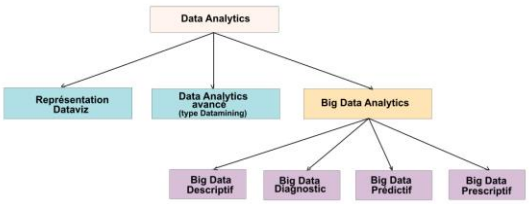
Le Big Data et la Blockchain
Une fausse bonne idée

- Un faux problème, purement marketing
- La décentralisation des architectures Blockchain ne justifie pas de l'intégrer dans un process Big Data
- Certaines analystes estiment que la Blockchain pourrait représenter 20 % de l'activité Big Data en 2030 (100 milliards $) : impensable
- Quelles pourraient être les applications :
  - Sécurité et qualité des données (peu envisageable), sauf pour certaines données, mais on sort du périmètre Big Data, pour entrer dans le domaine MDM
  - Analyse des données : les performances liées à l'algorithmique Blockchain, est un handicap à horizon prévisible
  - Le respect des contraintes réglementaires sur certaines données du périmètre Big Data (RGPD, Cloud Act) pourraient être traitées par un processus de Blockchain, mais pas sur des volumes élevés
- Il n'y a aucune raison de vouloir rapprocher artificiellement les deux concepts … même si ça fait si bien dans le paysage…

Les motivations du "data analytics"

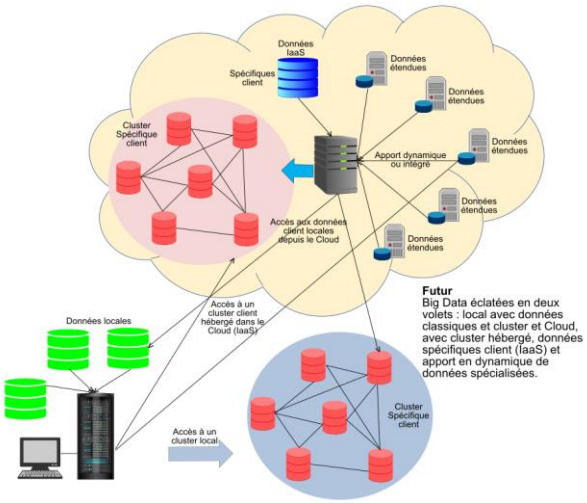
Comme pour les données « normales », les outils d'analyse Big Data peuvent être classés en quatre familles, selon le niveau de prédictivité souhaité

	Descriptive	Diagnostic	Prédictive	Prescriptive
Réponse à la question	Ce qui est arrivé ?	Pourquoi c'est arrivé ?	Ce qui va arriver ?	Ce qu'il faut faire ?
Incorporation des techniques d'IA et d'apprentissage	Non	Parfois	Couramment	En permanence
Popularité	Toutes les entreprises	Nombreuses entreprises	Faible nombre d'entreprises, mais en progression	Pas encore déployé de manière significative



De quoi demain sera-t-il fait ?

- Tout ira dans le Cloud, car les usagers n'ont pas vocation à gérer les infrastructures de données
- Ils n'ont pas les outils pour prendre en charge la transitivité des données (trous noirs) : extension obligatoire si on veut être "résilient"
- La quasi-totalité des projets sérieux seront pris en charge par des consultants très spécialisés, liés aux acteurs du Cloud et les clients auront le même problème de choix qu'avec les ERP et les entrepôts de données : porte ouverte au n'importe quoi
- Le Big Data et le BI qui lui est associé vont sortir du TI et seront directement gérés par les usagers : le TI se contentera de fournir les compétences techniques nécessaires, complémentaires de celles des fournisseurs
- Il y aura de gros "clash" sur des projets non aboutis et dispendieux, avec des procès spectaculaires à la clé… (situation bien connue avec les ERP)
- Le Big Data ne concernera que les gros, voire les très gros joueurs, pas les petites entreprises
- On mettra 10 ans à maîtriser l'hétérogénéité des données

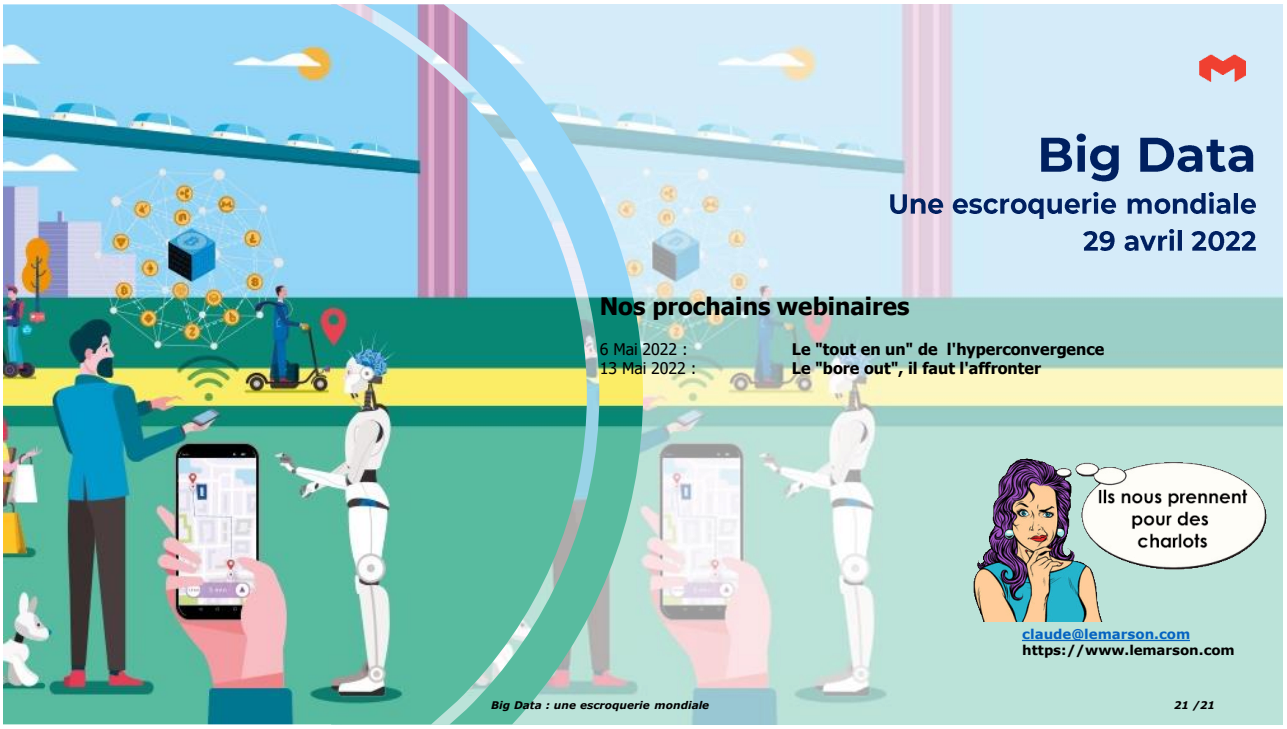


Morale : il ne faut pas écouter les prestataires

- Le Big Data est l'exemple même des projets incontrôlables… mais qui plaisent au management, car il les fait rêver…
- Les fournisseurs ont trouvé là un énorme fromage dont ils comptent bien profiter : profits immédiats avant que la situation se stabilise
- Impossible à contrer car l'entreprise ne peut qu'être gagnante : plus réactive, plus de clients, meilleur chiffre d'affaires, plus efficace…
- Il faut se préparer … mais ne rien précipiter (à faire comprendre par le management)

- C'est un puits sans fin d'interventions de tous ordres : architectes, spécialistes DaaS, gourous de la distribution des ressources, conseils pour le développement, prestataires du Cloud qui vont prendre le pouvoir sur une partie des projets
- La complexité des projets Big Data est minimisée
- Les prestataires mettent en avant des réalisations spectaculaires… développées par les grands de ce monde, elles ne sont pas significatives (même problème que pour l'IA)
- Les objectifs des clients contredisent ceux des prestataires
- Les équipes internes ne sont pas prêtes
  - Schisme de la transformation numérique
  - Big Data ce n'est pas un métier : c'est un assemblage de spécialités
  - Méfiance native du management et des usagers vis-à-vis du TI
- La solution : proposer des contrats d'assistance indexés sur les gains obtenus (après état des lieux)

- On ne sera pas étonné de savoir que les prestataires ne sont pas d'accord…



Big Data

Une escroquerie mondiale
29 avril 2022

Nos prochains webinaires

6 Mai 2022 : **Le "tout en un" de l'hyperconvergence**
13 Mai 2022 : **Le "bore out", il faut l'affronter**

claude@lemarson.com
https://www.lemarson.com