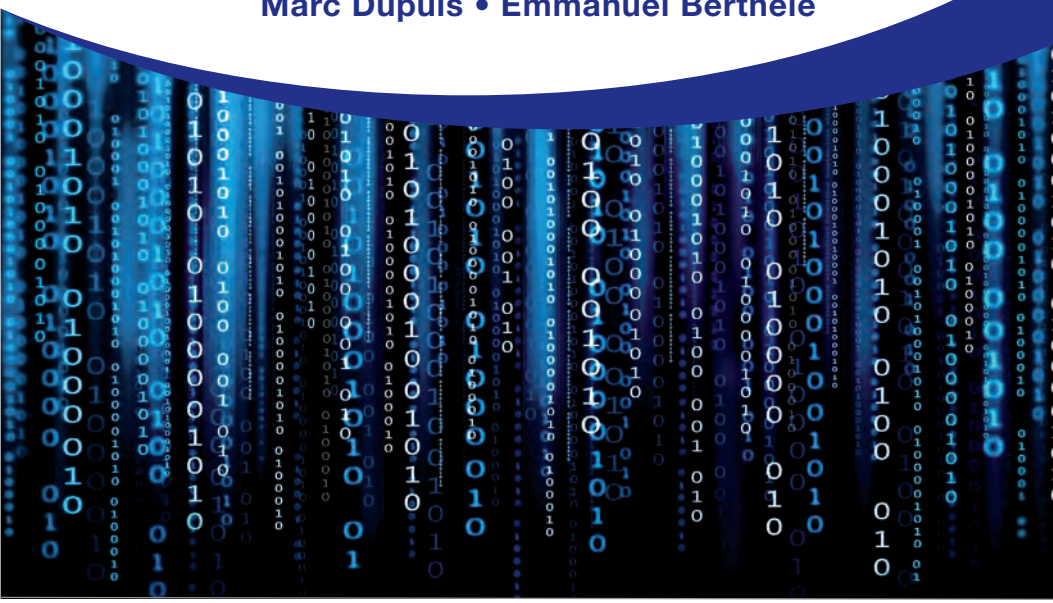
Les Essentiels

Le Big Data dans l'assurance

Marc Dupuis • Emmanuel Berthelé



L'ARGUS EDITIONS de l'assurance

Sommaire

Introduction 5

1. Cerner les fondamentaux et les enjeux du Big Data 9
Bien commencer avec quelques définitions et chiffres 9
L'évolution de l'informatique crée un environnement favorable à l'émergence du Big Data 11
La révolution digitale en marche 16
Écosystème du Big Data : les acteurs du Big Data 20
Quelques enjeux pour le Big Data 21

2. Placer la donnée au cœur de ce nouveau paradigme 23
La donnée comme point de départ 23
Les sources de données se diversifient 26
Données non structurées : un des enjeux majeurs du Big Data 31

3. Identifier l'écosystème du Big Data 35
Les besoins comme moteur de l'innovation 35
Les fondements technologiques des outils Big Data 36
Architectures et solutions Big Data 45
Les algorithmes 49

4. L'impact du Big Data sur le secteur de l'assurance 59
L'environnement spécifique de l'assurance 59
Les solutions Big Data : quels impacts en assurance 66
De nouveaux paramètres s'immiscent dans la gestion des risques 69
Modèles et modélisation 71
Aller au-delà des données des portefeuilles 72

5. Appréhender une nouvelle relation entre assureur et assuré 75
Les impacts du Big Data pour l'assureur 75
Une nouvelle approche de l'assurance pour le client 80
Vers une convergence des intérêts 81

6. Passer de la théorie à la pratique 83
Comment concevoir un projet Big Data ? 84
Une nécessaire transformation de l'entreprise 89
Des clés pour le lancement d'un projet Big Data 93

7. Cerner les enjeux réglementaires, juridiques et sociétaux 103
Réglementation : contraintes et opportunités 103
Comment la certification peut fournir une légitimité 114
Réflexions sociologiques et sociétales 120

Conclusion 123

Index alphabétique 125

Cerner les fondamentaux et les enjeux du Big Data

1. Bien commencer avec quelques définitions et chiffres

1.1 Définitions

Chaque acteur tend à donner sa propre définition du Big Data. Cependant une d'entre elles fait référence, celle du Gartner Group :

« Big Data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. »

Ce qui peut se traduire par :

Les Big Data sont des actifs (informations) de gros volume, de haute vitesse et de grande variété qui exigent des formes de traitements d'un bon rapport coût-performance et innovantes pour une compréhension accrue et la prise de décision.

Parmi toutes les définitions, des constantes reviennent :
- volume : de gigantesques volumes de données ;
- vitesse : besoin d'une grande vitesse de traitement associée à une évolution très rapide des données disponibles ;
- variété : des données non structurées de tout type (données numériques, textes, images, vidéos, sons), stockées sur des médias sociaux (réseaux sociaux, blogs…), échangées entre personnes (emails, sms), organisées par l'entreprise ou sur des supports plus classiques (bases de données internes).

Le principal intérêt du Big Data et du développement des approches correspondantes réside toutefois dans un 4e V : la valeur potentielle attendue de leur exploitation.

Enfin, certains considèrent le Big Data dans une acception plus large et intègrent l'open data voire tout simplement les immenses bases de données internes dont disposent certains acteurs. C'est dans l'esprit de la première définition, plus stricte, que s'inscrit cet ouvrage.

L'open data ou « données ouvertes », un phénomène connexe au Big Data décrit plus en détail dans le chapitre 2, émerge. Il s'agit de données numériques produites par un service public, une collectivité ou même une entreprise, dont l'accès est libre et gratuit. Les exemples sont nombreux : données statistiques de l'INSEE sur la population ou les entreprises fran-

çaises, de Météo France, des principaux ministères, de certaines villes, et montrent que cette notion n'est pas nouvelle. En revanche la tendance à la généralisation l'est. La nouveauté réside dans l'exploitation qui en est faite et les nouveaux outils qui permettent de les visualiser et les mettre en valeur.

IMPORTANT

Les Big Data sont des données massives, complexes et récoltées en temps réel. Tous les acteurs reconnaissent la nécessité de formes innovantes de traitement de ces nouvelles informations. Ces formes restent encore à améliorer voire à inventer et constituent l'un des enjeux fondamentaux du Big Data.

1.2 Chiffres

Les chiffres ci-après mettent en avant le gigantisme des données disponibles - sources 2014 :
- 90 % des données disponibles aujourd'hui sous forme numérique ont été créées au cours de ces deux dernières années ;
- Twitter génère actuellement 7 Teraoctets de données chaque jour et Facebook 10 Teraoctets (1 teraoctet, noté To, représente mille milliards d'octets) ;
- le nouveau radiotélescope « Square Kilometre Array » produit 50 To de données analysées par jour, à partir d'un flux de 7 000 To de données brutes par seconde.

Plus éloquents encore, sont les chiffres de ce qui se passe en 1 seconde dans le monde numérique :
- 2 000 téléchargements d'applications Apple et Android sont effectués ;
- 6 000 tweets sont générés ;
- 11 000 fichiers sont sauvegardés sur Dropbox ;
- 40 000 recherches sont effectuées sur Google ;
- 3 400 000 mails sont envoyés.

Tant et si bien que l'unité de mesure des données numériques est aujourd'hui le Zettaoctet (noté Zo) soit 10^{21} octets ! En effet, selon IDC, les données numériques créées dans le monde seraient passées de 1,2 Zo par an en 2010 à 1,8 Zo en 2011 puis 2,8 Zo en 2012 et s'élèveront à 40 Zo en 2020.

IMPORTANT

Unité	Nombre d'octets	Équivalence
1 yottaoctet (Yo)	10^{24}	
1 zettaoctet (Zo)	10^{21}	Toutes les données créées en 2012 (2,8 Zo)
1 exaoctet (Eo)	10^{18}	Toutes les informations générées jusqu'en 2003 (5 Eo)
1 pétaoctet (Po)	10^{15}	Bases de données de Google, Facebook, etc.
1 téraoctet (To)	10^{12}	6 millions de livres
1 gigaoctet (Go)	10^{9}	1 film de deux heures
1 mégaoctet (Mo)	10^{6}	1 livre de poche (1 Mo), 1 morceau de musique (5 Mo)
1 kilooctet (Ko)	10^{3}	

2. L'évolution de l'informatique crée un environnement favorable à l'émergence du Big Data

L'évolution de l'informatique a permis de générer un écosystème favorable à l'éclosion du Big Data. Dans un environnement économique où la notion de fonction d'utilité devient questionnable, l'innovation et le timing sont plus que jamais essentiels.

IMPORTANT

En 2012, le Gartner Group a publié les résultats d'une enquête sur les 10 technologies les plus stratégiques pour les 3 années à venir. Sans surprise, le Big Data mais aussi des technologies appartenant à l'écosystème du Big Data, y figurent en bonne place :
- tablettes ;
- applications et interfaces mobiles ;
- expérience utilisateur contextuelle et sociale ;
- internet des objets ;
- places de marché et « app stores » ;
- décisionnel de nouvelle génération ;
- Big Data ;
- calculs en mémoire (« in memory ») ;
- serveurs à très basse consommation d'énergie ;
- infonuagique (« Cloud Computing »).

2.1 Systèmes & architectures

Les bonds évolutifs des systèmes et architectures informatiques ont majoritairement été le fruit de ruptures technologiques. La vision centralisée des années 60 a ainsi laissé progressivement place à une vision totalement éclatée depuis une dizaine d'années.

▶ Années 1960-1970

À cette époque, les systèmes centralisés où toutes les applications et les données étaient gérées sur des grands systèmes dans l'entreprise, accessibles depuis des terminaux passifs, constituaient le modèle de référence. Les directions informatiques contrôlaient l'ensemble des applications et leurs évolutions. La sécurité des données (accès, sauvegardes…) demeurait facile à réaliser. Les temps d'accès dépendaient presque exclusivement de la puissance de calcul des systèmes centraux puisqu'une faible quantité de données circulait entre ces systèmes et les terminaux de travail.

▶ Années 1980-1990

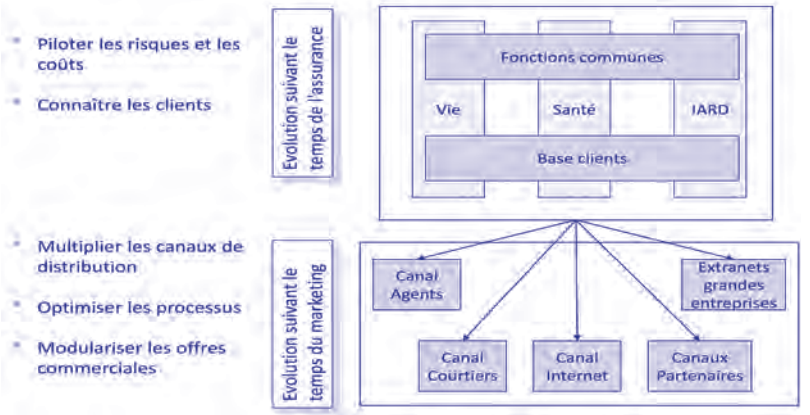
La micro-informatique est arrivée en entreprise ainsi que le développement de terminaux « intelligents » avec la mise en œuvre d'architectures distribuées de type client/serveur à 2 ou 3 couches (couche de gestion et stockage des données, couche de traitements, couche d'interface homme-machine). Grâce à leurs micro-ordinateurs, les utilisateurs ont acquis une certaine autonomie qu'ils ont d'ailleurs rapidement revendiquées. Le développement d'applications qui s'exécutaient entièrement en local (sur les micro-ordinateurs) nécessitait le développement d'architectures plus complexes et en constante évolution. Il était essentiel d'assurer la sauvegarde, la sécurité d'accès et la cohérence des données stockées en central avec celles stockées en local. Les temps d'accès dépendaient des débits des réseaux locaux avec un partage des données entre les serveurs centraux et les terminaux intelligents.

▶ Années 2000-2010

L'émergence rapide de l'internet a entrainé une dispersion des données et corrélativement des applications dans le web donc une urbanisation éclatée des systèmes d'information.

Dans ce contexte, les assureurs se sont vus confrontés à de nouveaux enjeux avec leurs systèmes d'information. Ils les ont amenés à devoir gérer le temps de l'assurance sur le long terme et le temps du marketing à très court terme pour lancer de nouvelles offres, intégrer de nouveaux canaux de distribution et partenaires, le tout avec une articulation délicate à réaliser. Les architectures orientées services y ont largement contribué en rendant les composants applicatifs plus indépendants et autonomes.

Les enjeux des systèmes d'information dans l'assurance



Dans la même décennie, le « grid computing » (grilles de calcul) a fait son apparition, utilisé depuis dans presque tous les secteurs économiques pour réduire considérablement les temps de calcul, sous réserve que les calculs réalisés soient parallélisables. L'une des réponses apportées provient du cloud computing ou « infonuagique ». Ici le temps d'accès devient crucial et les capacités des réseaux augmentent considérablement pour répondre à des besoins de débits de plus en plus exigeants.

▶ Années 2010 et suivantes

Les terminaux mobiles prennent progressivement le dessus : Selon CISCO, il y avait 7 milliards de terminaux mobiles dans le monde en 2013 et 2 milliards d'ordinateurs en service dans le monde en 2014 selon le Gartner Group.

Les directions informatiques se trouvent confrontées à de nouveaux challenges pour renforcer la protection du capital « données » avec une modification des modes de consommation (immédiateté, simplicité, transparence) et en situation de mobilité avec l'exigence de se connecter à un réseau sans contrainte de temps, de localisation ou de terminal. Ainsi se généralisent les notions de BYOD (bring your own device) littéralement « apportez vos appareils personnels » et de mobiquité parfaitement décrite par une terminologie anglo-saxonne : ATAWADAC (any time, any where, any device, any content) tout type de contenu sur tout type d'appareil, n'importe où et n'importe quand. Ces phénomènes concernent potentiellement

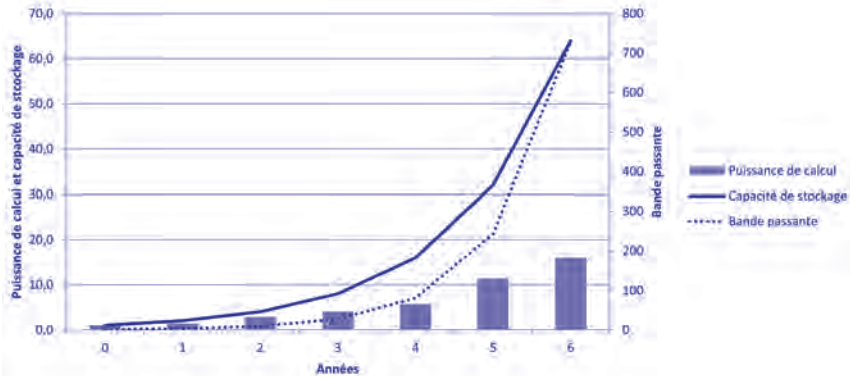
toutes les personnes, que ce soit dans leur univers personnel ou professionnel. Les frontières ne sont plus étanches.

2.2 Technologie

La technologie sous-jacente a subi d'énormes évolutions en parallèle et a permis de répondre à ces besoins de puissance de calcul, de capacité de stockage, de capacité d'échanges de données et enfin de miniaturisation.

▶ Puissance de calcul, capacité de stockage et bande passante

Trois lois célèbres donnent une idée de la rapidité d'évolution :
- puissance de calcul : elle double tous les 18 mois (Loi de Moore) ;
- capacité de stockage : elle double tous les ans ;
- bande passante sur fibre optique : elle triple tous les ans (Loi de Guilder).



▶ Miniaturisation

La miniaturisation a très largement facilité cet accroissement considérable en réduisant les volumes et les surfaces, donc l'échauffement, ainsi que le coût de production. Les derniers processeurs sont gravés avec des composants élémentaires (les couches conductrices ou résistantes qui composent les transistors) qui mesurent 20 nano mètres, soit 20 milliardièmes de mètre ou encore 10 fois le diamètre de la double hélice de l'ADN. Ces ordres de grandeur sont à mettre en perspective avec le premier transistor (1947) qui mesurait quelques centimètres. Les processeurs grand public en contiennent aujourd'hui plus d'un milliard.

La miniaturisation a aussi permis de faire évoluer la capacité de stockage d'un seul disque dur d'un facteur d'un million en l'espace de 50 ans. Le coût du Gigaoctet a, dans le même temps, chuté d'un facteur d'un milliard.

Dans le même esprit, la réduction des mémoires vives en taille et surtout en coût permet aujourd'hui de monter des volumes considérables de données en mémoire vive (ou en mémoire cache) et ainsi de réaliser ces calculs en mémoire en réduisant significativement le besoin d'accéder aux données sur disque. Ces techniques dites « in memory » tendent à devenir incontournables chez la plupart des éditeurs de logiciels de traitement de l'information.