Bonjour,

        j'ai bien noté votre sujet. Concernant le travail, vous devez faire quelques recherches sur le thème choisi puis de là, une fois que vous avez un peu de recul, vous pouvez rédiger un petit document (10 pages environ) sur lequel vous amenez tout ce que vous vouliez expliquer au lecteur. Bien sûr, le but est de montrer aussi que vous maîtrisez correctement les différentes notions vues en TP (utilisation des styles, formatage automatique, ajout d'éléments graphiques, ...). L'évaluation sera faite sur un ratio 1/3 fond et 2/3 forme.

Cordialement.

            Jean-Charles BOISSON

Le 20/10/2018 à 21:50, NATHAN TONNELLE a écrit :

Bonjour,   
J'aimerais que vous validez mon sujet de travail,  mon sujet est Le Big Data. Cependant j'aimerais plus de précisions pour faire ce travail notamment sur ce que vous attendez de celui-ci.   
En vous remerciant de votre réponse.   
  
TONNELLE Nathan   
Groupe S1A3A   
Numéro étudiant: 21801754

--

vous recevez ce message pour vous indiquer que j'ai ouvert la zone de dépôt  
du groupe S1A3A dans le cours NUM0101 dans moodle pour le rapport à remettre  
pour la fin des vacances de la Toussaint.

<https://www.lebigdata.fr/definition-big-data>

<https://www.sas.com/fr_fr/insights/big-data/what-is-big-data.html>

<https://www.lemondeinformatique.fr/actualites/lire-les-salaires-des-8-competences-les-plus-recherchees-en-big-data-56610.html>

**<https://www.lebigdata.fr/definition-big-data>**

# Définition : Qu’est-ce que le Big Data ?

## Le phénomène Big Data

**L’explosion quantitative des données numériques** a obligé les chercheurs à trouver de nouvelles manières de voir et d’analyser le monde. Il s’agit de découvrir **de nouveaux ordres de grandeur concernant la capture, la recherche, le partage, le stockage, l’analyse et la présentation des données**. Ainsi est né le « **Big Data** ». Il s’agit d’un concept permettant de stocker un nombre indicible d’informations sur une base numérique. Selon les archives de la bibliothèque numérique de l’Association for Computing Machinery (ou ACM) dans des articles scientifiques concernant les défis technologiques à relever pour visualiser les « grands ensembles de données », cette appellation est apparue en octobre 1997.

## Le Big Data, c’est quoi ?

Littéralement, ces termes signifient **mégadonnées**, grosses données ou encore **données massives**. Ils désignent un ensemble très volumineux de données qu’aucun outil classique de gestion de base de données ou de gestion de l’information ne peut vraiment travailler. En effet, nous procréons **environ 2,5 trillions d’octets de données tous les jours**. Ce sont les **informations provenant de partout** : messages que nous nous envoyons, vidéos que nous publions, informations climatiques, signaux GPS, [enregistrements transactionnels d’achats](https://www.lebigdata.fr/comment-le-big-data-revolutionne-shopping) en ligne et bien d’autres encore. **Ces données sont baptisées Big Data ou volumes massifs de données**. Les géants du Web, au premier rang desquels Yahoo ([mais aussi Facebook](https://www.lebigdata.fr/facebook-pinterest-transcendent-big-data) et Google), ont été les tous premiers à déployer ce type de technologie.

Cependant, **aucune définition précise ou universelle ne peut être donnée au Big Data**. Etant un objet complexe polymorphe, sa définition varie selon les communautés qui s’y intéressent en tant qu’usager ou fournisseur de services. Une approche transdisciplinaire permet d’appréhender le comportement des différents acteurs : les concepteurs et [fournisseurs d’outils](https://www.lebigdata.fr/big-data-revendeurs-3108) (les informaticiens), les catégories d’utilisateurs (gestionnaires, responsables d’entreprises, [décideurs politiques](https://www.lebigdata.fr/gouvernement-colombien-veut-se-servir-big-data), chercheurs), les acteurs de la santé et les usagers.

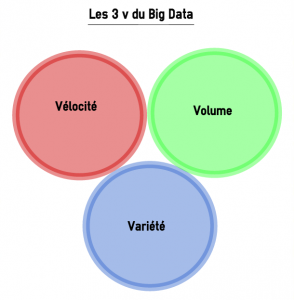
Le big data ne dérive pas des règles de toutes les technologies, il est aussi un **système technique dual**. En effet,**il apporte des bénéfices mais peut également**[**générer des inconvénients**](https://www.lebigdata.fr/5-etapes-gerer-risques-dun-environnement-big-data)**.** Ainsi, il sert aux spéculateurs sur les marchés financiers, de manière autonome avec, à la clé, la constitution des bulles hypothétiques.

**L’arrivée du Big Data** est maintenant présentée par de nombreux articles comme**une nouvelle révolution industrielle semblable à la découverte de la vapeur**(début du 19e siècle), de l’électricité (fin du 19e siècle) et de l’informatique (fin du 20esiècle). D’autres, un peu plus mesurés, qualifient ce phénomène comme étant **la dernière étape de la troisième révolution industrielle**, laquelle  est en fait celle de « l’information ». Dans tous les cas, le Big Data est considéré comme une source de [bouleversement profond de la société](https://www.lebigdata.fr/marche-big-data-atteindrait-67-milliards-de-dollars-2021).

## Big Data : l’analyse de données en masse

Inventé par les géants du web,**le Big Data** se présente comme une solution dessinée pour**permettre à tout le monde d’accéder en temps réel à des bases de données géantes**. Il vise à proposer un choix aux solutions classiques de bases de données et d’analyse (plate-forme de Business Intelligence en serveur SQL…).

Selon le Gartner,**ce concept regroupe une famille d’outils** qui répondent à une triple problématique dite **règle des 3V.** Il s’agit notamment d’un**Volume** de données considérable à traiter, une grande **Variété** d’informations (venant de diverses sources, non-structurées, organisées, Open…), et un certain niveau de **Vélocité** à atteindre, autrement dit de fréquence de création, collecte et partage de ces données.

[](https://www.lebigdata.fr/wp-content/uploads/2015/08/Capture-d%E2%80%99%C3%A9cran-2015-08-20-%C3%A0-13.49.36.png)

## Les évolutions technologiques derrière le Big Data

Les créations technologiques qui ont facilité la venue et **la croissance du Big Data**peuvent globalement être catégorisées en **deux familles** : d’une part, **les technologies de stockage**, portées particulièrement par le déploiement du [**Cloud Computing**](https://www.lebigdata.fr/definition-cloud-computing). D’autre part, l’arrivée de **technologies de traitement ajustées**, spécialement le développement[de nouvelles bases de données adaptées aux données non-structurées (Hadoop) et](https://www.lebigdata.fr/hadoop) la mise au point de [modes de calcul à haute performance (MapReduce)](https://www.lebigdata.fr/mapr).

Il existe plusieurs solutions qui peuvent entrer en jeu pour **optimiser les temps de traitement** sur des bases de données géantes à savoir les bases de données**NoSQL**(comme **MongoDB**, **Cassandra** ou **Redis**), les infrastructures du serveur pour la distribution des traitements sur les nœuds et le stockage des données en mémoire :

La première solution permet d’implémenter les systèmes de stockage considérés comme plus performants que le traditionnel SQL pour l’analyse de données en masse (orienté clé/valeur, document, colonne ou graphe).

La deuxième est aussi appelée le traitement massivement parallèle. Le Framework Hadoop en est un exemple. Celui-ci combine le système de fichiers distribué**HDFS**, la base **NoSQL** HBase et l’algorithme **MapReduce**.

Quant à la dernière solution, elle accélère le temps de traitement des requêtes.

## Evolution du Big Data : le développement de Spark et la fin de MapReduce

[](https://www.lebigdata.fr/wp-content/uploads/2015/08/Spark.jpg)Spark prend la place de MapReduce

Chaque technologie, appartenant au système **mégadonnée**, a son utilité, **ses atouts et ses inconvénients**. Etant un milieu en perpétuelle évolution, **le Big Data** cherche toujours à optimiser les performances des outils. Ainsi, **son paysage technologique bouge très vite, et de nouvelles solutions naissent très fréquemment**, avec pour but d’optimiser encore plus les technologies existantes. Pour illustrer cette évolution, **MapReduce et**[**Spark**](https://www.lebigdata.fr/apache-spark-tout-savoir)[représentent des exemples très concrets.](https://www.lebigdata.fr/apache-spark-tout-savoir)

**Décrit par Google en 2004**, **MapReduce** est un pattern implémenté ultérieurement dans le projet Nutch de Yahoo, qui deviendra le projet Apache Hadoop en 2008. **Cet algorithme dispose d’une grande capacité** en matière de stockage de données. Le seul hic est qu’il est un peu lent. Cette lenteur est notamment visible sur des volumes modestes. Malgré cela, les solutions, souhaitant proposer des traitements quasi-instantanés sur ces volumes, commencent à délaisser MapReduce. **En 2014, Google a donc annoncé qu’il sera succédé par une solution SaaS dénommée Google Cloud Dataflow.**

**Spark est aussi**[**une solution emblématique**](https://www.lebigdata.fr/spark-2-0-approche-big-data) permettant d’écrire simplement des applications distribuées et **proposant des bibliothèques de traitement classique**. Entre-temps, avec une performance remarquable, **il peut travailler sur des données sur disque ou des données chargées en RAM**. Certes, il est plus jeune mais il dispose d’une communauté énorme. C’est aussi un des projets Apache ayant une vitesse de développement rapide. En somme,**c’est une solution qui s’avère être le successeur de MapReduce**, d’autant qu’il a l’avantage de fusionner une grande partie des outils nécessaires dans un cluster Hadoop.

## Les principaux acteurs du marché

La filière Big Data en a attiré plusieurs. Ces derniers se sont positionnés rapidement dans divers secteurs. **Dans le secteur IT**, on retrouve les fournisseurs historiques de solutions IT comme Oracle, HP, SAP [ou encore IBM](https://www.lebigdata.fr/ibm-souhaite-deployer-son-cloud-134022). Il y a aussi **les acteurs du Web** dont Google, Facebook, ou Twitter. Quant aux spécialistes **des solutions Data et Big Data**, on peut citer MapR, Teradata, EMC ou Hortonworks. CapGemini, Sopra, Accenture ou Atos sont des intégrateurs, toujours des acteurs principaux dans les méga données. Dans **le secteur de l’analytique**, comme **éditeurs BI**, on peut citer SAS, Micro-strategy et Qliktech. Cette filière comporte aussi des **fournisseurs spécialisés dans l’analytique** comme Datameer ou Zettaset. En parallèle à ces principaux participants, de nombreuses PME spécialisées dans le Big Data sont apparues, sur toute la chaîne de valeur du secteur. **En France, les pionniers ont été Hurence et Dataiku pour les équipements et logiciels de Big Data** ; Criteo, Squid, Captain Dash et Tiny Clues pour l’analyse de données et Ysance pour le conseil.

## Formation continue en Big Data : ce que proposent les grandes écoles

Désormais, des **grandes écoles proposent**[**des formations dans le Big Data**](https://www.lebigdata.fr/25-meilleures-ecoles-formations-universitaires-big-data). La pédagogie veut accorder une large part à des études de cas et retours d’expérience. Elle met aussi en exergue les “fils rouges”. Il s’agit de projets de mise en situation professionnelle que certaines grandes entreprises telles que EDF ou encore Capgemini proposent.

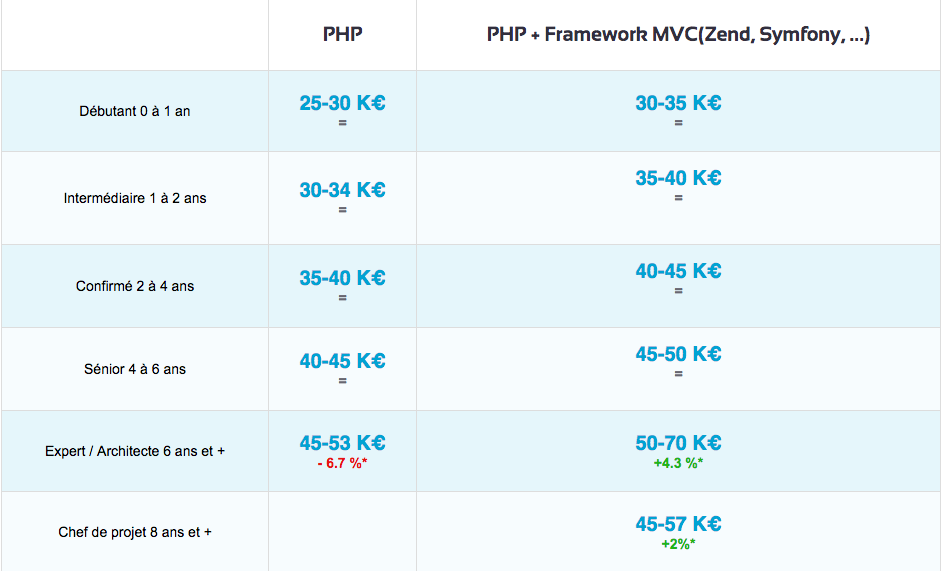
**Ce genre de formation n’est pas limité à un cadre théorique**. Les apprentis sont aussi amenés à faire des pratiques en renforçant leur formation par un stage. **Pour intégrer ces écoles, il faut être un titulaire d’un diplôme d’ingénieur** en informatique ou en télécommunication, **ou d’un master universitaire scientifique ou technique, en informatique ou en mathématiques appliquées.** **Elles acceptent souvent les bac +4** scientifique à condition que la personne dispose d’au moins 3 ans d’expérience professionnelle.

## Les salaires / rémunérations dans le domaine du Big Data

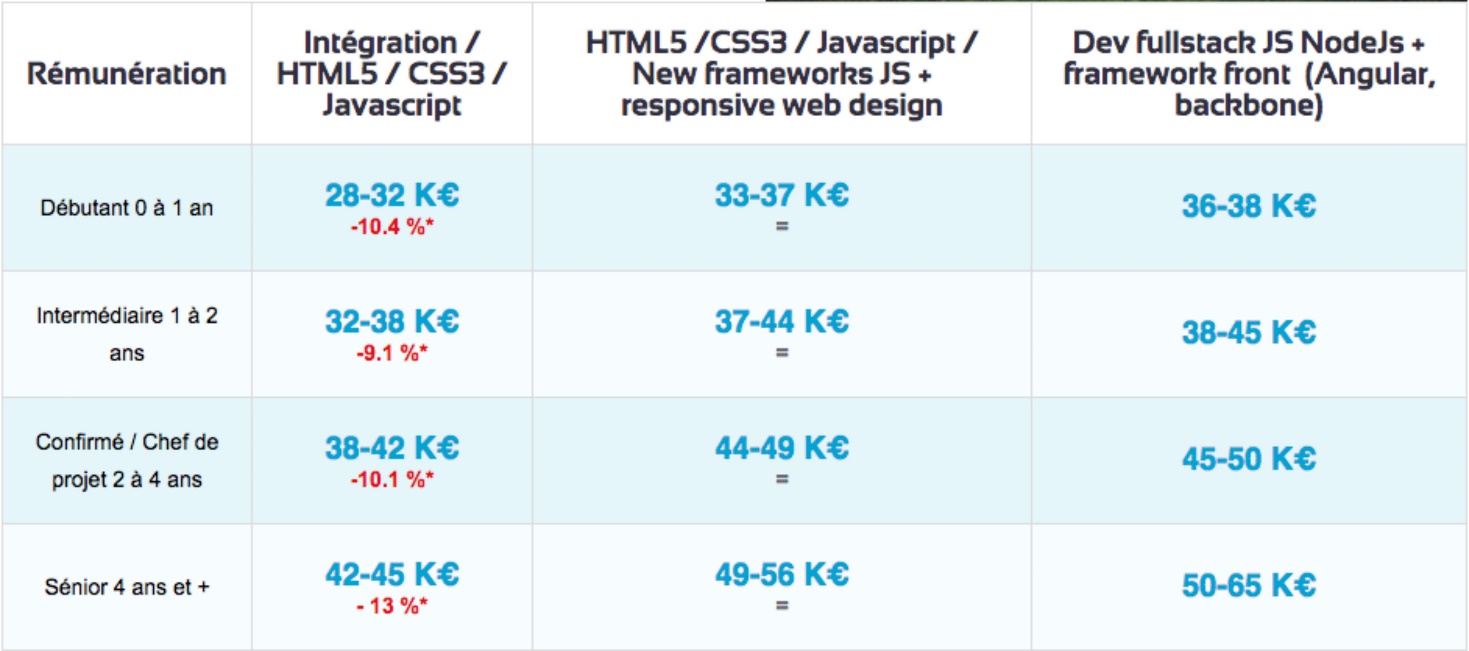
D’après Esilv.fr,**les études de salaire des développeurs** révèlent que le domaine du Big Data en 2015 est **en tête**.

Voici en comparaison les salaires de développeurs PHP et les salaires de développeurs en Big Data d’après Urban Linker.

### Salaires de développeurs PHP :

[](https://www.lebigdata.fr/wp-content/uploads/2015/08/Capture-d%E2%80%99%C3%A9cran-2015-08-20-%C3%A0-14.41.461.png)

### Salaires de développeurs en Big Data :



## Big Data : des innovations disruptives qui changent la donne

**Le Big Data et les analytics** sont utilisés dans presque tous les domaines. Ils se sont même construits une place importante dans la société. **Ils se traduisent sous plusieurs formes** à ne citer que l’usage de statistiques dans le sport de haut niveau, **le programme de surveillance PRISM de la NSA,** la médecine analytique ou encore les algorithmes de recommandation d’Amazon.

En entreprise particulièrement, l’usage d’outils Big Data & Analytics répond généralement à plusieurs objectifs comme l**’amélioration de l’expérience client, l’optimisation des processus et de la performance opérationnelle, le renforcement ou diversification du business model.**

De **nouvelles opportunités** significatives de différenciation concurrentielle sont **générées par l’ère de la gestion d’importants volumes de données**[**et de leur analyse**](https://www.lebigdata.fr/dossier-analyse-predictive-definition-secteurs-dapplication). Pour les organisations, plusieurs raisons peuvent les inciter à se tourner vers cette nouvelle administration de données à savoir **la gestion rentable des données, l’optimisation du stockage d’informations, la possibilité de faire des analyses programmables ou encore la facilité de la manipulation des données.**

## Big Data, exclusivement pour les fonctions Marketing et commerciales ?

Cette technologie **représente aux yeux de tous** un enjeu commercial privilégié compte tenu de sa **capacité à impacter le commerce en profondeur dans l’économie mondiale** intégrée. En effet, **les entreprises,**[**peu importe leur taille**](https://www.lebigdata.fr/dossier-leurope-cloud-pme), font partie des premières  **à bénéficier des avantages** obtenus à partir d’une bonne manipulation des données massives.

Cependant, **les mégadonnées** jouent également **un rôle essentiel dans la transformation des processus**,de la chaîne logistique,  des échanges de type « **Machine-to-Machine** » dans le but de **développer un meilleur « écosystème informationnel ».** Ils permettent aussi de prendre des décisions **plus véloces et plus crédibles**, prenant en considération des informations internes mais également externes à l’organisation. Ils peuvent entre-temps servir d’appui pour la gestion des risques et de la fraude.

## Devant tant d’informations, comment trier le bon grain de l’ivraie ?

Comme le dit le vieil adage « **trop d’informations tuent l’information** ». Il s’agit en fait du principal problème avec les mégadonnées. **La quantité énorme des informations est un des obstacles**. L’autre obstacle provient évidemment du niveau de certitude qu’on peut avoir sur une donnée.

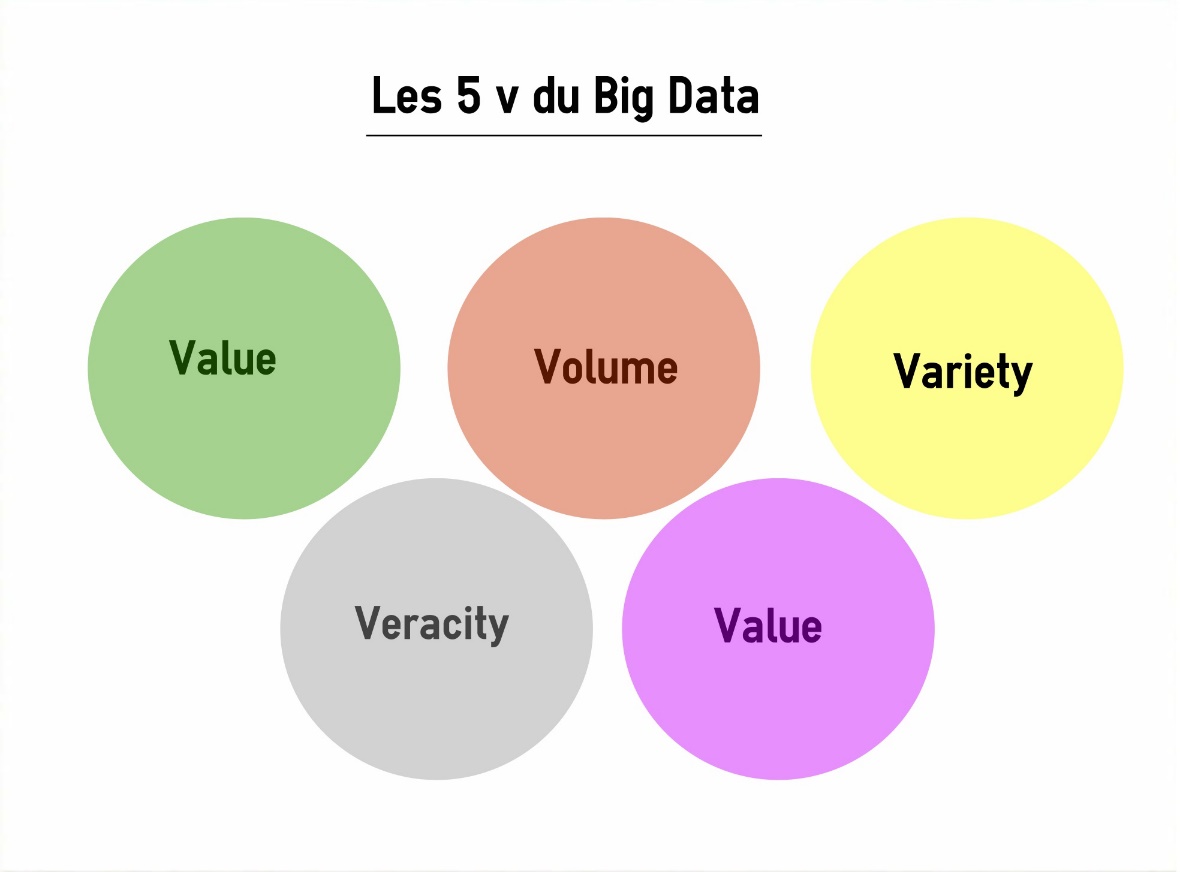
En effet,**les données qui découlent du**[**marketing numérique**](https://www.lebigdata.fr/marketing-big-data-outil-indispensable-campagne-a-succes) peuvent être considérées comme **des informations « incertaines »**, dans la mesure par exemple où on ne peut être sûr de l’identité de qui est en train de cliquer sur une offre incluse dans une URL. **Le volume de données** associé au **manque de crédibilité** de celles-ci rend son exploitation plus alambiquée.

Pour autant, **grâce aux algorithmes statistiques, des solutions existent**. C’est  d’ailleurs, avant même de se demander s’il serait **possible de collecter et stocker le big data**, qu’on devrait toujours commencer par s’interroger de son **aptitude à les analyser et de leur utilité.**

Avec un but convenablement déterminé et des données d’une qualité suffisante, **les algorithmes et méthodes statistiques** permettent désormais de concevoir de la valeur alors que ce n’était pas encore faisable il y a encore quelques années. A ce propos, on peut distinguer d**eux types d’écoles dans le domaine prédictif à savoir**[**l’intelligence artificielle ou « machine learning »**](https://www.lebigdata.fr/intelligence-artificielle-3-secteurs)**et la statistique**. Ces deux secteurs bien qu’ils soient distincts se rejoignent finalement de plus en plus. De plus, ils peuvent être utilisés en simultanéité de manière vertueuse et intelligente pour mener à bien un projet.

Là où l’usage des mégadonnées en gestion devient un enjeu vital pour les entreprises.

Parmi les utilisateurs les plus enthousiastes du Big Data, on retrouve les gestionnaires et les économistes. Ces derniers définissent ce phénomène par [**la règle des 5V**(Volume, Velocity, Variety, Veracity, Value)](https://www.lebigdata.fr/infographie-quatre-v-big-data-expliques-ibm).

[](https://www.lebigdata.fr/wp-content/uploads/2015/08/Les5vdubigdata.jpg)

**Le volume**

**Le volume correspond à la masse d’informations produite chaque seconde**. Selon des études, pour avoir une idée de l’accroissement exponentiel de la masse de données, on considère que **90 % des données ont été engendrées durant les années où l’usage d’internet et des réseaux sociaux a connu une forte croissance**. L’ensemble de toutes les données produites depuis le début des temps jusqu’à la fin de l’année 2008, conviendrait maintenant à la masse de celles qui sont générées chaque minute. Dans le monde des affaires, le volume de données collecté chaque jour est d’une importance vitale.

**La velocité**

**La velocité équivaut à la rapidité de l’élaboration et du déploiement des nouvelles données**. Par exemple, si on diffuse des messages sur les réseaux sociaux, ils peuvent devenir « viraux » et se répandre en un rien de temps. Il s’agit d’analyser les données au décours de leur lignée (appelé parfois analyse en mémoire) sans qu’il soit indispensable que ces informations soient entreposées dans une base de données.

**La variété**

**Seulement 20% des données sont structurées puis stockées** dans des tables de bases de données relationnelles similaire à celles utilisées en gestion comptabilisée. **Les 80% qui restent sont non-structurées**. Cela peut être des images, des vidéos, des textes, des voix, et bien d’autres encore… La technologie Big Data, permet de faire l’analyse, la comparaison, la reconnaissance, le classement des données de différents types comme des conversations ou messages sur les réseaux sociaux, des photos sur différents sites etc. Ce sont les différents éléments qui constituent la variété offerte par le Big Data.

**La véracité**

**La véracité concerne la fiabilité et la crédibilité des informations collectées**. Comme le Big Data permet de collecter un nombre indéfini et plusieurs formes de données,**il est difficile de justifier l’authenticité des contenus**, si l’on considère les post Twitter avec les abréviations, le langage familier, les hashtags, les coquilles etc. Toutefois, les génies de l’informatique sont en train de développer de **nouvelles techniques qui devront permettre de faciliter la gestion de ce type de données notamment par le W3C.**

**La valeur**

**La notion de valeur correspond au profit qu’on puisse tirer de l’usage du Big Data.** Ce sont généralement les entreprises qui commencent à obtenir des avantages incroyables de leurs Big Data.**Selon les gestionnaires et les économistes, les entreprises qui ne s’intéressent pas sérieusement au Big Data risquent d’être pénalisées et écartées**. Puisque l’outil existe, ne pas s’en servir conduirait à perdre un privilège concurrentiel.

## L’essor des mégadonnées en médecine

La médecine est un art qui use des sciences. En effet, un médecin praticien est en simultanéité un scientifique qui a obtenu des connaissances en biophysique, sémiologie médicale et chirurgicale, anatomie, biochimie, physiologie, biologie, … et un artiste qui maîtrise des habiletés pour effectuer des gestes thérapeutiques adaptés. Désormais, **les connaissances traditionnelles ne suffisent plus pour mieux amplifier**[**le pouvoir d’un médecin dans l’investigation et le soin**](https://www.lebigdata.fr/big-data-transforme-soins-de-sante)**.** Il a également appris à maîtriser des technologies de plus en plus sophistiquées dans les différentes spécialités médicales.**On assiste à l’essor du génie biologique médical ou GBM**. Cette alternative offre aux médecins de nouvelles possibilités de diagnostic à savoir, les appareils d’imagerie : scintigraphie, échographes, imagerie par résonance magnétique (IRM) etc. Les automates d’analyse biologique, les appareils d’analyse de signaux comme l’électrocardiogramme (ECG) ou encore l’électroencéphalogramme (EEG), ainsi que les appareils de traitement des pathologies (dialyse, laser, assistance respiratoire, médecine nucléaire,…) figurent aussi parmi les fruits de l’alliance technologie/médecine.

Majoritairement pilotés par des ordinateurs spécialisés qui sont directement ou indirectement connectés à un réseau informatique, ces dispositifs permettent de collecter des informations diverses concernant les patients. **Ils se présentent comme de nouveaux moyens d’investigation, d’acquisition et de stockage de données,** de comparaison de l’information que les médecins traitants peuvent mettre en œuvre afin d’accroître leur réactivité dans les différentes étapes cliniques essentielles à la prise en charge de leurs patients. **Ils peuvent aussi s’en servir pour mener des études épidémiologiques des maladies dans la population.**

## L’avenir du Big Data

Etant une tendance lourde, **le Big Data n’est pas une mode**. Dans le domaine de l’usage,**il satisfait une nécessité de travailler la donnée plus profondément, pour créer de la valeur, conjointement à des aptitudes technologiques qui n’existaient pas dans le passé.** Cependant, compte tenu de l’évolution des technologies qui ne semble pas vouloir s’estomper, on ne peut pas alors parler d’une norme véritable ou de standards dans le domaine du Big data.

Beaucoup d’applications du Big Data n’en sont qu’à leurs préludes et on peut s’attendre à voir apparaître [des utilisations auxquelles on ne s’attend pas encore aujourd’hui](https://www.lebigdata.fr/futur-big-data).**En quelque sorte, le Big Data est un tournant pour les organisations au moins aussi important qu’internet en son temps.** Chaque entreprise [doit donc s’y mettre dès maintenant](https://www.lebigdata.fr/3-questions-a-se-poser-projet-big-data). Dans le cas contraire, il y a un risque qu’elle se rendent comptent d’ici quelques années qu’elles se sont faites dépasser par la concurrence. Les gouvernements et [les organismes publics se penchent également sur la question à travers l’ open data.](https://www.lebigdata.fr/open-data-definition)

## Les données massives : un marché mondial en plein épanouissement

D’ici quelques années,**le marché du big data va se mesurer en centaines de milliards de dollars**. C’est un nouvel eldorado pour le business. Selon des études, il s’agit même d’une vague de fond où l’on retrouve la combinaison de la BI (business intelligence), de l’analytics et de l’internet des objets. **IDC affirme qu’il devrait passer au-delà des 125 milliards de dollars avant la fin 2015.** En effet, plusieurs études affluent sur cette affirmation et toutes confirment que les budgets que les entreprises vont consacrer au Big Data ne vont connaître que des fortes progressions. Ainsi, rien que**le marché des solutions visuelles de découvertes des informations**liées à la gestion des données massives**va grimper de 2,5 fois plus rapidement que celui des solutions de BI d’ici à 2018.**

D’après le calcul effectué par le cabinet Vanson Bourne, dans le monde,**l’ensemble des dépenses consacrées au Big data, dans les budgets IT des grandes entreprises, devrait représenter un quart du budget total IT en 2018**, s’il en est encore à **18% actuellement**. Le Cap Gemini a aussi commandité une étude en mars 2015. Le résultat a montré que **61% des entreprises sont conscientes de l’utilité du Big Data en tant que “moteur de croissance à part entière”**. De ce fait,on lui accorde beaucoup plus d’importance que leurs produits et services existants. Cette même étude a encore indiqué que **43% d’entre elles se sont déjà réorganisées ou se restructurent présentement pour exploiter le potentiel du Big Data.**

Par Loïc Bremme.

[**https://www.sas.com/fr\_fr/insights/big-data/what-is-big-data.html**](https://www.sas.com/fr_fr/insights/big-data/what-is-big-data.html)

Si l'expression « big data » est relativement récente, la démarche consistant à recueillir et conserver de grandes quantités d'informations dans le but de les analyser est vieille comme le monde. Le concept a pris de l'ampleur au début des années 2000 lorsque l'analyste sectoriel Doug Laney a énoncé les trois caractéristiques désormais bien connues des big data, les « trois V » :

**Volume.** Les entreprises collectent des données de diverses sources : dont les transactions commerciales, médias sociaux et informations issues de capteurs ou de transactions de machine à machine. Auparavant, le stockage de ces données aurait posé problème, mais cette tâche est désormais simplifiée par les nouvelles technologies, notamment Hadoop.

**Vitesse.** Les données nous parviennent à un rythme de plus en plus soutenu et doivent être traitées rapidement. Balises RFID, capteurs et compteurs intelligents déversent des torrents d'informations qu'il faut analyser en temps quasi réel.

**Variété.** Les données revêtent tous types de formats : des données numériques structurées dans des bases de données traditionnelles aux documents texte non structurés, en passant par les e-mails, les vidéos, les fichiers audio, les données boursières et les transactions financières.

Chez SAS, nous y associons deux dimensions supplémentaires :

**Variabilité.** De plus en plus rapides et variés, les flux de données peuvent aussi être très irréguliers, avec des périodes de pointe. Une nouvelle tendance émerge sur les réseaux sociaux ? Les pics de données quotidiens, saisonniers ou déclenchés par un événement peuvent s'avérer difficiles à gérer, en particulier lorsqu'il s'agit de données non structurées.

**Complexité.** De nos jours, les données proviennent de multiples sources, ce qui complique l'établissement de liens et de correspondances, le nettoyage et la transformation entre différents systèmes. Cependant, il est nécessaire de connecter et corréler les relations, hiérarchies et nombreux liens entre les données, qui risquent sinon d’échapper à tout contrôle.

**<https://www.lemondeinformatique.fr/actualites/lire-les-salaires-des-8-competences-les-plus-recherchees-en-big-data-56610.html>**

**Spécialistes des outils d'ETL, développeurs Hadoop, data scientists, développeurs OLAP, spécialistes en analyse prédictive ou architectes de données. Un cabinet de recrutement américain a livré une grille de salaires annuels moyens pour ces différents profils pour lesquels la demande grimpe avec le développement des projets big data.**

Dans les projets de big data, les data scientists constituent une nouvelle catégorie de professionnels capables de mettre en oeuvre leurs techniques d'organisation de données dans des applications métiers.

Au fur et à mesure que les entreprises avancent dans leur démarche big data, la demande en compétences spécialisées progresse pour mettre en place la collecte, l'organisation et l'analyse de ces énormes volumes de données provenant de multiples sources. Aux États-Unis, le cabinet de recrutement Kforce a dressé une liste des principaux profils recherchés et indiqué pour chacun une fourchette de salaires annuels moyens, publiés par nos confrères de Computerworld. Mais attention, il s'agit d'une grille valant pour le marché américain, où les charges sociales ne sont absolument pas les mêmes qu'en France. Kforce précise qu'il s'est basé sur une semaine de travail typique de 40 heures.  
  
**1- Spécialistes des outils d'ETL**  
  
Avec l'explosion de données de toutes sortes dont les entreprises cherchent à tirer parti, le besoin en professionnels capables de récupérer et d'intégrer ces big data a augmenté de façon significative. Parmi eux figurent notamment les familiers des solutions d'ETL, ces outils permettant d'extraire les données de leur source initiale, de les mettre au bon format et de les charger dans la base cible (extraction, transformation, loading). Des logiciels tels qu'en proposent le Français Talend, ou encore Informatica et Pentaho pour ne citer qu'eux. Comme les logiciels d'ETL ont acquis de la maturité, les profils spécialisés sur ces outils sont des compétences dont on aura besoin dans la durée, souligne Greg Jones, CTO de Kforce. Dans ce domaine, les entreprises associent généralement des collaborateurs internes et des prestations de services.  
  
***Salaire annuel moyen : 110 à 130 000 $ (80 à 95 000 €)***  
  
**2 - Développeurs Hadoop**  
  
Le cabinet de recrutement a également enregistré une hausse de la demande autour du framework Open Source Hadoop, adapté au traitement des jeux de données très volumineux, et de toutes les technologies associées comme Hive, HBase, MapReduce, Pig, etc. Le traitement de ces teraoctets ou petaoctets serait trop coûteux avec des outils de BI classiques et prendrait beaucoup trop de temps si on ne recourait pas aux capacités massivement distribuées d'Hadoop. Dans l'actuel paysage big data, les personnes possédant une expérience sur ce framework disponible dans de nombreuses distributions (dont Hortonworks et Cloudera) sont les plus recherchées, selon Greg Jones. Pour l'instant, les entreprises ont plutôt tendance à s'appuyer sur des compétences externes.  
  
***Salaire annuel moyen : 150 à 175 000 $ (109 à 127 000 €)***  
  
**3 - Spécialistes des outils de visualisation**  
  
Analyser d'énormes quantités de données s'avère souvent difficile. Les nouveaux outils de visualisation tels qu'en proposent Tibco Spotfire, Qlickview ou Tableau Software permettent une exploration rapide et intuitive de ces données. On peut penser que les compétences requises pour les utiliser se rapprochent de celles d'un profil spécialisé en BI. Mais, précise KForce, avec la montée en force d'Hadoop, cela représente une nouvelle catégorie de profils spécialisés. Pour l'instant, ces ressources sont recherchées à court terme, en externe. Avec le temps et la maturité des outils, la demande pour ces profils se modérera et ces ressources seront plutôt utilisées en interne, estime Greg Jones.  
  
***Salaire annuel moyen : 150 à 175 000 $ (109 à 127 000 €)***  
  
**4 - Data scientists**  
  
Précédemment appelés également data architects, évoluant dans la sphère IT, les data scientists constituent une nouvelle catégorie de professionnels capables de mettre en oeuvre leurs techniques d'organisation de données dans des applications métiers. Ils doivent aussi savoir communiquer pour expliquer les résultats trouvés dans les données à la fois à la DSI et aux directions métiers. Typiquement, ces data scientists ont leur propre sandbox dans laquelle ils explorent et examinent les données de l'entreprise en accompagnant la démarche d'innovation. Anjul Bhambhri, vice président responsable des produits big data chez IBM, les décrit de façon imagée : « En partie analyste et en partie artiste, un data scientist a de la curiosité, il observe les données et repère des tendances. » C'est, selon lui, un profil qui veut vraiment apprendre et transformer une entreprise.  
  
  
***Salaire annuel moyen : 125 à 140 000 $ (91 à 102 000 €)***  
  
**5 - Développeurs OLAP**  
  
L'expertise des développeurs OLAP consiste à optimiser l'organisation des données. Ce sont eux qui récupèrent les données de sources relationnelles ou non structurées pour créer des modèles multidimensionnels -souvent désignés sous le nom de schémas en étoile ou en flocon de neige- et qui construisent ensuite l'interface utilisateur pour accéder aux données à partir de requêtes prédéfinies.  
  
***Salaire annuel moyen : 98 à 115 550 $ (71 500 à à 84 000 €)***